THE UNIVERSITY OF CHICAGO

ESTIMATION AND INFERENCE FOR HIGH DIMENSIONAL DATA WITH STRUCTURED SIGNALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

FAN YANG

CHICAGO, ILLINOIS

JUNE 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Rina Barber and Tracy Ke, for their constant support and encouragement. They are great mentors, advisors and collaborators. Their thoughtful guidance as well as active involvement in my research experience drives me forward through my graduate study. I would also like to thank my committee member Chao Gao for his valuable discussion and support. I would like to thank all faculty and staff in the Department of Statistics where I have spent my wonderful five years of graduate school. I am thankful to the HELIOS groups member for their constructive comments on research and my fellow graduate student for the great academic environment here. I would like to thank my friend Wooseok Ha for his constant encouragement in my early PhD study and Haoyang Liu for his helpful discussion through my research life, and many other friends I met over the past few years for their patience, care and help. Finally, I would like to thank my parents, Bin Yang and Weihua Wang, for their love, encouragement, and patience.

ABSTRACT

In high dimensional statistics, estimation and inference are often done by making use of the underlying signal structures. We consider the cases of sparse, low rank and shape-restricted signal structures for a variety of problems, and propose new approaches for estimating related quantities and make valid inference. The methods we develop can be used in many real world applications such as gene expression data analysis in genetics, portfolio management in finance and experimental A/B testing in industry.

Chapter 2 discusses selective inference for group sparse linear models. We develop tools to construct confidence intervals and p-values for testing selected groups of variables in a linear model with group sparsity. Chapter 3 studies one dimensional isotonic regression which is an example of shape-restricted nonparametric regression. We characterize the contractive property of the isotonic projection with respect to any norm and use this to analyze the convergence properties of isotonic regression. Chapter 4 considers variable ranking in high dimensional sparse linear regression with rare and weak signals. We propose a two step approach to rank variables so that signal variables tend to have higher rank than noise variables. Chapter 5 considers the problem of decomposing a large covariance matrix into a low rank part plus a diagonally dominant part. We propose several algorithms to perform such tasks and demonstrate its usefulness in estimating large covariance matrices for high dimensional data. Chapter 6 discusses estimation and inference for zero-inflated semi-continuous data. We propose several machine learning approaches to estimate related quantities in both one sample setting and two group setting.

CHAPTER 1

INTRODUCTION

For high dimensional data, the signals we are trying to recover are often associated with certain structures: either those signals do have certain characteristics by the nature of the problem, or we are willing to assume such structures so that we could have sufficient information from data for accurate estimation and inference. Examples of those structures are sparse structure for coefficients in linear regression, (approximate) low rank structures for signal matrix in covariance matrix estimation, and shape-constrained structure for signal sequence in nonparametric regression. In this thesis, we will cover a handful of scenarios where signals exhibit certain structures and we propose new methods designed for specific problems.

In Chapter 2, we consider a high dimensional linear model with group sparsity and discuss how to do inference for groups selected by certain algorithms. The groups of covariates are predetermined and the sparsity of coefficient is on group level. We consider three algorithms for selecting signal groups: forward stepwise regression [1, 2], iterative hard thresholding [3] and group lasso [4]. For those selected groups, we are willing to do inference on them and we have to take a smart approach to correct for the fact that we have been using the same data twice. In this regard, we develop a selective inference method following the methodology proposed in [5] to produce valid p-values and construct conservative one-sided confidence intervals. These are finite-sample results and the technical tools we develop can be applied to other potential selection methods.

In Chapter 3, we look at isotonic regression which is an example of shape restricted nonparametric regression. In this case, we observe noisy data where the underlying signal sequence x satisfies a monotonicity constraint, that is, x lies in the isotonic cone $\{x \in \mathbb{R}^n : x_1 \leq \cdots \leq x_n)\}$. We study the least squares estimator where the estimated signal is given by the isotonic projection operator (projection to the isotonic cone) applied on observed data. We find a necessary and sufficient condition characterizing all norms with respect to which this projection is contractive. We then define a new norm called sliding window norm which we use to establish the convergence

properties of isotonic regression and to construct a data adaptive confidence bands for signal sequence. In particular, we establish convergence rates in l_2 norm for signals with bounded variation and a uniformly bound for locally Lipschitz signals, where both rates match known results in the literature. The tools we develop can be applied to study the Grenander estimator in shape restricted density estimation problem as well.

In Chapter 4, we turn to variable ranking problem for high dimensional linear model. The goal here is to rank the variables to maximize the area under ROC curve with respect to classifying variables. We assume the Gram matrix of design is the sum of a low rank matrix plus a approximate sparse matrix, and the coefficients are sparse and individually small, i.e. rare and weak. We propose a two-step method called factor-adjusted covariate assisted ranking (FA-CAR) to rank variables. In the FA-step, we use principal component analysis to reduce the linear model into a new one where the Gram matrix is approximately sparse; in the CAR-step, we exploit the local covariate structure to rank variables. Compared to marginal ranking, our approach is proved to overcome the signal cancellation problem. FA-CAR can be extended to the generalized linear model case as well, where the signal cancellation problem still exists for logistic regression in the random design case.

In Chapter 5, we consider decomposing a large covariance matrix into the sum of a low rank matrix and a diagonally dominant matrix, and we call this problem diagonally-dominant principal component analysis (DD-PCA). We propose two ADMM algorithms and an iterative projection algorithm for solving different versions of DD-PCA, and demonstrate the usefulness of this new decomposition through two applications: large covariance matrix estimation and global null testing problem. In the covariance matrix estimation problem, we combine DD-PCA with the idea of POET [6] to create a new approach DD-POET, and demonstrate its strength in applications of portfolio management and high dimensional linear discriminant analysis. In the global null testing problem, we combine DD-PCA with the idea of Higher Criticism (HC) [7] to obtain DD-HC that shows better performance than its competitors in numerical studies.

In Chapter 6, we focus on regression problem with zero-inflated semi-continuous responses.

There is few literature in this area and we try to lay out a general framework that enables the use of advanced machine learning algorithms. We build a generative model for the observed data and defines some new metrics of interest. We then propose several machine learning algorithms for estimating related quantities and make some fair comparison. In particular, we discuss how to estimate the conditional mean of the positive part of response in the one sample setting, and the heterogeneous treatment effect on the positive part in the two sample setting where we have a control group and a treatment group. Empirical studies show some interesting results and suggests research directions for future investigation.

1.1 Summary

This thesis is intended to investigate statistical models with structured signals in different scenarios. In Chapter 2, we focus on selective inference problem where the coefficients of the linear model is group sparse. In Chapter 3, we work on isotonic regression where the signal sequence has certain shape restrictions. Chapter 4 describes variable ranking problem where signals are rare and weak and the Gram matrix is approximately low rank. In Chapter 5, we suppose a large covariance matrix can be decomposed into a low rank part plus a diagonally dominant part. Chapter 6 discusses regression with zero-inflated semi-continuous responses.

1.2 Notation

Throughout we will use the following notation. We will write $\mathscr{P}_{\mathscr{L}}$ for the projection to any closed and convex set $\mathscr{L} \subseteq \mathbb{R}^n$, and $\mathscr{P}_{\mathscr{L}}^{\perp}$ for the projection to its orthogonal complement if $\mathscr{L}$ is a linear subspace. For $y \in \mathbb{R}^n$, $\mathrm{dir}_{\mathscr{L}}(y) = \frac{\mathscr{P}_{\mathscr{L}} y}{\|\mathscr{P}_{\mathscr{L}} y\|_2} \in \mathscr{L} \cap \mathbb{S}^{n-1}$ is the unit vector in the direction of $\mathscr{P}_{\mathscr{L}} y$. This direction is not defined if $\mathscr{P}_{\mathscr{L}} y = 0$.

For positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = o(b_n)$, $a_n = O(b_n)$ and $a_n \lesssim b_n$, if $\lim_{n\to\infty}(a_n/b_n) = 0$, $\limsup_{n\to\infty}(a_n/b_n) < \infty$, $\max\{a_n - b_n, 0\} = o(1)$, respectively. Given $0 \le q \le \infty$, for any vector x, $\|x\|_q$ denotes the L_q-norm of x; when $q = 2$ it's the Euclidean norm

and we will sometimes omit the subscript to write it as $\|x\|$. For any $m \times n$ matrix A, $\|A\|_q$ denotes the matrix L_q-norm of A, i.e. $\|A\|_q = \max_{x \neq 0} \frac{\|Ax\|_q}{\|x\|_q}$; when $q = 2$, it coincides with the the spectral norm, and we will generally omit the subscript to simply write it as $\|A\|$. $\|A\|_F$ denotes the Frobenius norm and $\|A\|_{\max}$ denotes the entrywise max norm. When A is symmetric, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues, respectively. For two sets $\mathscr{I} \subset \{1, 2, \ldots, m\}$ and $\mathscr{J} \subset \{1, 2, \ldots, n\}$, $A^{\mathscr{I},\mathscr{J}}$ denotes the submatrix of A formed by restricting the rows and columns of A to sets $\mathscr{I}$ and $\mathscr{J}$. For a vector $x \in \mathbb{R}^p$ and set $\mathscr{I} \subset \{1, 2, \ldots, p\}$, $x^{\mathscr{I}}$ denotes the sub-vector of x formed by restricting coordinates to set $\mathscr{I}$.

CHAPTER 2

SELECTIVE INFERENCE FOR GROUP SPARSE LINEAR MODEL

Significant progress has been recently made on developing inference tools to complement the feature selection methods that have been intensively studied in the past decade [8, 9, 5, 10]. The goal of selective inference is to make accurate uncertainty assessments for the parameters estimated using a feature selection algorithm, such as the lasso [11]. The fundamental challenge is that after the data have been used to select a set of coefficients to be studied, this selection event must then be accounted for when performing inference, using the same data. A specific goal of selective inference is to provide p-values and confidence intervals for the fitted coefficients. As the sparsity pattern is chosen using nonlinear estimators, the distribution of the estimated coefficients is typically non-Gaussian and can have multiple modes, even under a standard Gaussian noise model, making classical techniques unusable for accurate inference. It is of particular interest to develop finite-sample, non-asymptotic results.

In this chapter, [1] we present new results for selective inference in the setting of group sparsity [4, 13, 14]. We consider the linear model $Y = X\beta + \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ where $X \in \mathbb{R}^{n\times p}$ is a fixed design matrix. In many applications, the p columns or features of X are naturally grouped into blocks $\mathscr{C}_1, \ldots, \mathscr{C}_G \subseteq \{1, \ldots, p\}$. In the high dimensional setting, the working assumption is that only a few of the corresponding blocks of the coefficients β contain nonzero elements; that is, $\beta_{\mathscr{C}_g} = 0$ for most groups g. This group-sparse model can be viewed as an extension of the standard sparse regression model. Algorithms for fitting this model, such as the group lasso [4], extend well-studied methods for sparse linear regression to this grouped setting. We provide a tool for constructing confidence intervals as well as p-values for testing selected groups. In contrast to the (non-grouped) sparse regression setting, the confidence interval construction does not follow immediately from the p-value calculation, and requires a careful analysis of non-centered multivariate normal distributions.

1. The work presented in this chapter is published in Yang et al. [12].

2.1 Problem formulation

We focus on the linear model $Y = X\beta + \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, where $X \in \mathbb{R}^{n\times p}$ is fixed and $\sigma^2 > 0$ is assumed to be known. More generally, our model is $Y \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I}_n)$ with $\mu \in \mathbb{R}^n$ unknown and σ^2 known. For a given block of variables $\mathscr{C}_g \subseteq [p]$, we write X_g to denote the $n \times |\mathscr{C}_g|$ submatrix of X consisting of all features of this block. For a set $\mathscr{S} \subseteq [G]$ of blocks, $X_{\mathscr{S}}$ consists of all features that lie in any of the blocks in $\mathscr{S}$.

When we refer to "selective inference," we are generally interested in the distribution of subsets of parameters that have been chosen by some model selection procedure. After choosing a set of groups $\mathscr{S} \subseteq [G]$, we would like to test whether the true mean μ is correlated with a group X_g for each $g \in \mathscr{S}$ after controlling for the remaining selected groups, i.e. after regressing out all the other groups, indexed by $\mathscr{S}\backslash g$. Thus, the following question is central to selective inference:

$$\text{Question}_{g,\mathscr{S}}: \text{ What is the magnitude of the projection of } \mu \text{ onto the span of } \mathscr{P}^{\perp}_{X_{\mathscr{S}\backslash g}} X_g? \quad (2.1)$$

In particular, we are interested in a hypothesis test to determine if μ is orthogonal to this span, that is, whether block g should be removed from the model with group-sparse support determined by $\mathscr{S}$; this is the question studied by Loftus and Taylor [5] for which they compute p-values. Alternatively, we may be interested in a confidence interval on $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$, where $\mathscr{L} = \text{span}(\mathscr{P}^{\perp}_{X_{\mathscr{S}\backslash g}} X_g)$. Since $\mathscr{S}$ and g are themselves determined by the data Y, any inference on these questions must be performed "post-selection," by conditioning on the event that $\mathscr{S}$ is the selected set of groups.

2.1.1 Background: the polyhedral lemma

In the more standard sparse regression setting without grouped variables, after selecting a set $\mathscr{S} \subseteq [p]$ of features corresponding to columns of X, we might be interested in testing whether the column X_j should be included in the model obtained by regressing Y onto $X_{\mathscr{S}\backslash j}$. We may want to test the null hypothesis that $X_j^{\top} \mathscr{P}^{\perp}_{X_{\mathscr{S}\backslash j}} \mu$ is zero, or to construct a confidence interval for this inner product.

In the setting where $\mathscr{S}$ is the output of the lasso, Lee et al. [9] characterize the selection

event as a polyhedron in $\mathbb{R}^n$: for any set $\mathscr{S} \subseteq [p]$ and any signs $s \in \{\pm 1\}^{\mathscr{S}}$, the event that the lasso (with a fixed regularization parameter λ) selects the given support with the given signs is equivalent to the event $Y \in \mathscr{A} = \{y : Ay < b\}$, where A is a fixed matrix and b is a fixed vector, which are functions of X, $\mathscr{S}$, s, λ. The inequalities are interpreted elementwise, yielding a convex polyhedron $\mathscr{A}$. To test the regression question described above, one then tests $\eta^\top \mu$ for a fixed unit vector $\eta \propto \mathscr{P}^\perp_{X_{\mathscr{S}\setminus j}} X_j$. The "polyhedral lemma" [9, Theorem 5.2] proves that the distribution of $\eta^\top Y$, after conditioning on $\{Y \in \mathscr{A}\}$ and on $\mathscr{P}^\perp_\eta Y$, is given by a truncated normal distribution, with density

$$f(r) \propto \exp\left\{-(r - \eta^\top \mu)^2 / 2\sigma^2\right\} \cdot \mathbb{1}\{a_1(Y) \le r \le a_2(Y)\}. \tag{2.2}$$

The interval endpoints $a_1(Y), a_2(Y)$ depend on Y only through $\mathscr{P}^\perp_\eta Y$ and are defined to include exactly those values of r that are feasible given the event $Y \in \mathscr{A}$. That is, the interval contains all values r such that $r \cdot \eta + \mathscr{P}^\perp_\eta Y \in \mathscr{A}$.

Examining (2.2), we see that under the null hypothesis $\eta^\top \mu = 0$, this is a truncated *zero-mean* normal density, which can be used to construct a p-value testing $\eta^\top \mu = 0$. To construct a confidence interval for $\eta^\top \mu$, we can instead use (2.2) with nonzero $\eta^\top \mu$, which is a truncated *noncentral* normal density.

2.1.2 The group-sparse case

In the group-sparse regression setting, Loftus and Taylor [5] extend the work of Lee et al. [9] to questions where we would like to test $\mathscr{P}_\mathscr{L} \mu$, the projection of the mean μ to some potentially multi-dimensional subspace, rather than simply testing $\eta^\top \mu$, which can be interpreted as a projection to a one-dimensional subspace, $\mathscr{L} = \text{span}(\eta)$. For a fixed set $\mathscr{A} \subseteq \mathbb{R}^n$ and a fixed subspace $\mathscr{L}$ of dimension k, Loftus and Taylor [5, Theorem 3.1] prove that, after conditioning on $\{Y \in \mathscr{A}\}$, on $\text{dir}_\mathscr{L}(Y)$, and on $\mathscr{P}^\perp_\mathscr{L} Y$, under the null hypothesis $\mathscr{P}_\mathscr{L} \mu = 0$, the distribution of $\|\mathscr{P}_\mathscr{L} Y\|_2$ is given by a truncated χ_k distribution,

$$\|\mathscr{P}_\mathscr{L} Y\|_2 \sim (\sigma \cdot \chi_k \text{ truncated to } \mathscr{R}_Y) \text{ where } \mathscr{R}_Y = \left\{r : r \cdot \text{dir}_\mathscr{L}(Y) + \mathscr{P}^\perp_\mathscr{L} Y \in \mathscr{A}\right\}. \tag{2.3}$$

In particular, this means that, if we would like to test the null hypothesis $\mathcal{P}_{\mathcal{L}}\mu = 0$, we can compute a p-value using the truncated χ_k distribution as our null distribution. To better understand this null hypothesis, suppose that we run a group-sparse model selection algorithm that chooses a set of blocks $\mathcal{S} \subseteq [G]$. We might then want to test whether some particular block $g \in \mathcal{S}$ should be retained in this model or removed. In that case, we would set $\mathcal{L} = \text{span}(\mathcal{P}_{X_{\mathcal{S}\setminus g}}^{\perp} X_g)$ and test whether $\mathcal{P}_{\mathcal{L}}\mu = 0$.

Examining the parallels between this result and the work of Lee et al. [9], where (2.2) gives either a truncated zero-mean normal or truncated noncentral normal distribution depending on whether the null hypothesis $\eta^\top \mu = 0$ is true or false, we might expect that the result (2.3) of Loftus and Taylor [5] can extend in a straightforward way to the case where $\mathcal{P}_{\mathcal{L}}\mu \neq 0$. More specifically, we might expect that (2.3) might then be replaced by a truncated *noncentral* χ_k distribution, with its noncentrality parameter determined by $\|\mathcal{P}_{\mathcal{L}}\mu\|_2$. However, this turns out not to be the case. To understand why, observe that $\|\mathcal{P}_{\mathcal{L}}Y\|_2$ and $\text{dir}_{\mathcal{L}}(Y)$ are the length and the direction of the vector $\mathcal{P}_{\mathcal{L}}Y$; in the inference procedure of Loftus and Taylor [5], they need to condition on the direction $\text{dir}_{\mathcal{L}}(Y)$ in order to compute the truncation interval $\mathcal{R}_Y$, and then they perform inference on $\|\mathcal{P}_{\mathcal{L}}Y\|_2$, the length. These two quantities are independent for a centered multivariate normal, and therefore if $\mathcal{P}_{\mathcal{L}}\mu = 0$ then $\|\mathcal{P}_{\mathcal{L}}Y\|_2$ follows a χ_k distribution even if we have conditioned on $\text{dir}_{\mathcal{L}}(Y)$. However, in the general case where $\mathcal{P}_{\mathcal{L}}\mu \neq 0$, we do not have independence between the length and the direction of $\mathcal{P}_{\mathcal{L}}Y$, and so while $\|\mathcal{P}_{\mathcal{L}}Y\|_2$ is marginally distributed as a noncentral χ_k, this is no longer true after conditioning on $\text{dir}_{\mathcal{L}}(Y)$.

In the following section, we consider the problem of computing the distribution of $\|\mathcal{P}_{\mathcal{L}}Y\|_2$ after conditioning on $\text{dir}_{\mathcal{L}}(Y)$, which is the setting that we require for inference. This leads to the main contribution of this work, where we are able to perform inference on $\mathcal{P}_{\mathcal{L}}\mu$ beyond simply testing the null hypothesis that $\mathcal{P}_{\mathcal{L}}\mu = 0$.

2.2 Theoretical results

Now we present our key lemma and its application to group sparse regression methods.

2.2.1 Key lemma: truncated projections of Gaussians

Before presenting our key lemma, we introduce some further notation. Let $\mathscr{A} \subseteq \mathbb{R}^n$ be any fixed open set and let $\mathscr{L} \subseteq \mathbb{R}^n$ be a fixed subspace of dimension k. For any $y \in \mathscr{A}$, consider the set

$$\mathscr{R}_y = \{r > 0 : r \cdot \mathrm{dir}_{\mathscr{L}}(y) + \mathscr{P}_{\mathscr{L}}^{\perp} y \in \mathscr{A}\} \subseteq \mathbb{R}_+.$$

Note that $\mathscr{R}_y$ is an open subset of $\mathbb{R}_+$, and its construction does not depend on $\|\mathscr{P}_{\mathscr{L}} y\|_2$, but we see that $\|\mathscr{P}_{\mathscr{L}} y\|_2 \in \mathscr{R}_y$ by definition.

Lemma 2.2.1 (Truncated projection)**.** *Let $\mathscr{A} \subseteq \mathbb{R}^n$ be a fixed open set and let $\mathscr{L} \subseteq \mathbb{R}^n$ be a fixed subspace of dimension k. Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I}_n)$. Then, conditioning on the values of $\mathrm{dir}_{\mathscr{L}}(Y)$ and $\mathscr{P}_{\mathscr{L}}^{\perp} Y$ and on the event $Y \in \mathscr{A}$, the conditional distribution of $\|\mathscr{P}_{\mathscr{L}} Y\|_2$ has density*[2]

$$f(r) \propto r^{k-1} \exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2r \cdot \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu\rangle\right)\right\} \cdot \mathbb{1}\{r \in \mathscr{R}_Y\}.$$

We pause to point out two special cases that are treated in the existing literature.

Special case 1: $k = 1$ and $\mathscr{A}$ is a convex polytope. Suppose $\mathscr{A}$ is the convex polytope $\{y : Ay < b\}$ for fixed $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. In this case, this almost exactly yields the "polyhedral lemma" of Lee et al. [9, Theorem 5.2]. Specifically, in their work they perform inference on $\eta^\top \mu$ for a fixed vector η; this corresponds to taking $\mathscr{L} = \mathrm{span}(\eta)$ in our notation. Then since $k = 1$, Lemma 2.2.1 yields a truncated Gaussian distribution, coinciding with Lee et al. [9]'s result (2.2). The only difference relative to [9] is that our lemma implicitly conditions on $\mathrm{sign}(\eta^\top Y)$, which is not required in [9].

Special case 2: the mean μ is orthogonal to the subspace $\mathscr{L}$. In this case, without conditioning on $\{Y \in \mathscr{A}\}$, we have $\mathscr{P}_{\mathscr{L}} Y = \mathscr{P}_{\mathscr{L}}(\mu + \mathcal{N}(0, \sigma^2 \mathbf{I})) = \mathscr{P}_{\mathscr{L}}(\mathcal{N}(0, \sigma^2 \mathbf{I}))$, and so $\|\mathscr{P}_{\mathscr{L}} Y\|_2 \sim \sigma \cdot \chi_k$. Without conditioning on $\{Y \in \mathscr{A}\}$ (or equivalently, taking $\mathscr{A} = \mathbb{R}^n$), the resulting density

2. Here and throughout the paper, we ignore the possibility that $Y \perp \mathscr{L}$ since this has probability zero.

is then

$$f(r) \propto r^{k-1} e^{-r^2/2\sigma^2} \cdot \mathbb{1}\{r > 0\}$$

which is the density of the χ_k distribution (rescaled by σ), as expected. If we also condition on $\{Y \in \mathscr{A}\}$ then this is a truncated χ_k distribution, as proved in Loftus and Taylor [5, Theorem 3.1].

2.2.2 Selective inference on truncated projections

We now show how the key result in Lemma 2.2.1 can be used for group-sparse inference. In particular, we show how to compute a p-value for the null hypothesis $H_0 : \mu \perp \mathscr{L}$, or equivalently, $H_0 : \|\mathscr{P}_{\mathscr{L}}\mu\|_2 = 0$. In addition, we show how to compute a one-sided confidence interval for $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$, specifically, how to give a lower bound on the size of this projection.

Theorem 2.2.1 (Selective inference for projections)**.** *Under the setting and notation of Lemma 2.2.1, define*

$$P = \frac{\int_{r \in \mathscr{R}_Y, r > \|\mathscr{P}_{\mathscr{L}} Y\|_2} r^{k-1} e^{-r^2/2\sigma^2} \, \mathrm{d}r}{\int_{r \in \mathscr{R}_Y} r^{k-1} e^{-r^2/2\sigma^2} \, \mathrm{d}r}. \tag{2.4}$$

If $\mu \perp \mathscr{L}$ (or, more generally, if $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle = 0$), then $P \sim$ Uniform$[0,1]$. Furthermore, for any desired error level $\alpha \in (0,1)$, there is a unique value $L_\alpha \in \mathbb{R}$ satisfying

$$\frac{\int_{r \in \mathscr{R}_Y, r > \|\mathscr{P}_{\mathscr{L}} Y\|_2} r^{k-1} e^{-(r^2 - 2rL_\alpha)/2\sigma^2} \, \mathrm{d}r}{\int_{r \in \mathscr{R}_Y} r^{k-1} e^{-(r^2 - 2rL_\alpha)/2\sigma^2} \, \mathrm{d}r} = \alpha, \tag{2.5}$$

and we have

$$\mathbb{P}\{\|\mathscr{P}_{\mathscr{L}}\mu\|_2 \geq L_\alpha\} \geq \mathbb{P}\{\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle \geq L_\alpha\} = 1 - \alpha.$$

Finally, the p-value and the confidence interval agree in the sense that $P < \alpha$ if and only if $L_\alpha > 0$.

From the form of Lemma 2.2.1, we see that we are actually performing inference on $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$. Since $\|\mathscr{P}_{\mathscr{L}}\mu\|_2 \geq \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$, this means that any lower bound on $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$ also gives a lower bound on $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$. For the p-value, the statement $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle = 0$ is implied by the stronger null hypothesis $\mu \perp \mathscr{L}$. We can also use Lemma 2.2.1 to give a two-sided confidence

interval for $\langle \text{dir}_{\mathscr{L}}(Y), \mu \rangle$; specifically, $\langle \text{dir}_{\mathscr{L}}(Y), \mu \rangle$ lies in the interval $[L_{\alpha/2}, L_{1-\alpha/2}]$ with probability $1-\alpha$. However, in general this cannot be extended to a two-sided interval for $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$.

2.2.3 Application to group sparse regression methods: General recipe

With a fixed design matrix, the outcome of any group-sparse selection method is a function of Y. For example, a forward stepwise procedure determines a particular sequence of groups of variables. We call such an outcome a *selection event*, and assume that the set of all selection events forms a countable partition of $\mathbb{R}^n$ into disjoint open sets: $\mathbb{R}^n = \cup_e \mathscr{A}_e$.[3] Each data vector $y \in \mathbb{R}^n$ determines a selection event, denoted $e(y)$, and thus $y \in \mathscr{A}_{e(y)}$.

Let $\mathscr{S}(y) \subseteq [G]$ be the set of feature groups that are selected for testing. This is assumed to be a function of $e(y)$, i.e. $\mathscr{S}(y) = \mathscr{S}_e$ for all $y \in \mathscr{A}_e$. For any $g \in \mathscr{S}_e$, define $\mathscr{L}_{e,g} = \text{span}(\mathscr{P}^{\perp}_{X_{\mathscr{S}_e \setminus g}} X_g)$; this is the subspace of $\mathbb{R}^n$ indicating correlation with group X_g beyond what can be explained by the other selected groups, $X_{\mathscr{S}_e \setminus g}$.

Write $\mathscr{R}_Y = \left\{ r > 0 : r \cdot U + Y_\perp \in \mathscr{A}_{e(Y)} \right\}$, where $U = \text{dir}_{\mathscr{L}_{e(Y),g}}(Y)$ and $Y_\perp = \mathscr{P}^{\perp}_{\mathscr{L}_{e(Y),g}} Y$. If we condition on the event $\{Y \in \mathscr{A}_e\}$ for some e, then as soon as we have calculated the region $\mathscr{R}_Y \subseteq \mathbb{R}_+$, Theorem 2.2.1 will allow us to perform inference on the quantity of interest $\|\mathscr{P}_{\mathscr{L}_{e,g}}\mu\|_2$ by evaluating the expressions (2.4) and (2.5). In other words, we are testing whether μ is significantly correlated with the group X_g, after controlling for all the other selected groups, $\mathscr{S}(Y) \setminus g = \mathscr{S}_e \setminus g$.

To evaluate these expressions accurately, ideally we would like an explicit characterization of the region $\mathscr{R}_Y \subseteq \mathbb{R}_+$. To gain a better intuition for this set, define $z_Y(r) = r \cdot U + Y_\perp \in \mathbb{R}^n$ for $r > 0$, and note that $z_Y(r) = Y$ when we plug in $r = \|\mathscr{P}_{\mathscr{L}_{e(Y),g}} Y\|_2$. Then we see that

$$\mathscr{R}_Y = \left\{ r > 0 : e(z_Y(r)) = e(Y) \right\}. \tag{2.6}$$

In other words, we need to find the range of values of r such that, if we replace Y with $z_Y(r)$, then this does not change the output of the model selection algorithm, i.e. $e(z_Y(r)) = e(Y)$. For the

3. Since the distribution of Y is continuous on $\mathbb{R}^n$, we ignore sets of measure zero without further comment.

forward stepwise and IHT methods, we find that we can calculate $\mathscr{R}_Y$ explicitly. For the group lasso, we cannot calculate $\mathscr{R}_Y$ explicitly, but we can nonetheless compute the integrals required by Theorem 2.2.1 through numerical approximations. We present the details for each of these methods in the following three sections.

2.2.4 Application to Forward stepwise regression

Forward stepwise regression [2, 1] is a simple and widely used method. We will use the following version:[4] for design matrix X and response $Y = y$,

1. Initialize the residual $\widehat{\varepsilon}_0 = y$ and the model $\mathscr{S}_0 = \varnothing$.
2. For $t = 1, 2, \ldots, T$,
   (a) Let $g_t = \arg\max_{g \in [G] \setminus \mathscr{S}_{t-1}} \{\|X_g^\top \widehat{\varepsilon}_{t-1}\|_2\}$.
   (b) Update the model, $\mathscr{S}_t = \{g_1, \ldots, g_t\}$, and update the residual, $\widehat{\varepsilon}_t = \mathscr{P}^\perp_{X_{\mathscr{S}_t}} y$.

Testing all groups at time T. First we consider the inference procedure where, at time T, we would like to test each selected group g_t for $t = 1, \ldots, T$. Our selection event $e(Y)$ is the ordered sequence $g_1, \ldots, g_T$ of selected groups. For a response vector $Y = y$, this selection event is equivalent to

$$\|X_{g_k}^\top \mathscr{P}^\perp_{X_{\mathscr{S}_{k-1}}} y\|_2 > \|X_g^\top \mathscr{P}^\perp_{X_{\mathscr{S}_{k-1}}} y\|_2 \text{ for all } k = 1, \ldots, T, \text{ for all } g \notin \mathscr{S}_k. \tag{2.7}$$

Now we would like to perform inference on the group $g = g_t$, while controlling for the other groups in $\mathscr{S}(Y) = \mathscr{S}_T$. Define U, $Y_\perp$, and $z_Y(r)$ as before. Then, to determine $\mathscr{R}_Y = \{r > 0 : z_Y(r) \in \mathscr{A}_{e(Y)}\}$, we check whether all of the inequalities in (2.7) are satisfied with $y = z_Y(r)$: for

4. In practice, we would add some correction for the scale of the columns of X_g or for the number of features in group g; this can be accomplished with simple modifications of the forward stepwise procedure.

each $k = 1, \ldots, T$ and each $g \notin \mathcal{S}_k$, the corresponding inequality of (2.7) can be expressed as

$$r^2 \cdot \|X_{g_k}^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp U\|_2^2 + 2r \cdot \langle X_{g_k}^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp U, X_{g_k}^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp Y_\perp \rangle + \|X_{g_k}^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp Y_\perp\|_2^2$$
$$> r^2 \cdot \|X_g^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp U\|_2^2 + 2r \cdot \langle X_g^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp U, X_g^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp Y_\perp \rangle + \|X_g^\top \mathcal{P}_{X_{\mathcal{S}_{k-1}}}^\perp Y_\perp\|_2^2.$$

Solving this quadratic inequality over $r \in \mathbb{R}_+$, we obtain a region $\mathcal{I}_{k,g} \subseteq \mathbb{R}_+$ which is either a single interval or a union of two disjoint intervals, whose endpoints we can calculate explicitly with the quadratic formula. The set $\mathcal{R}_Y$ is then given by all values r that satisfy the full set of inequalities:

$$\mathcal{R}_Y = \bigcap_{k=1,\ldots,T} \bigcap_{g \in [G] \setminus \mathcal{S}_k} \mathcal{I}_{k,g}.$$

This is a union of finitely many disjoint intervals, whose endpoints are calculated explicitly as above.

Sequential testing. Now suppose we carry out a sequential inference procedure, testing group g_t at its time of selection, controlling only for the previously selected groups $\mathcal{S}_{t-1}$. In fact, this is a special case of the non-sequential procedure above, which shows how to test g_T while controlling for $\mathcal{S}_T \setminus g_T = \mathcal{S}_{T-1}$. Applying this method at each stage of the algorithm yields a sequential testing procedure. (The method developed in [5] computes p-values for this problem, testing whether $\mu \perp \mathcal{P}_{X_{\mathcal{S}_{t-1}}}^\perp X_{g_t}$ at each time t.) Detailed pseudo-code of our inference algorithm for forward selection is presented in Algorithm 1 . Here, we compute the P value as well as confidence interval for each selected group conditioned on our previous selections. The algorithm is efficient and only overhead above the Forward Selection method is computation of the integral of a one-dimensional density over different intervals (see Step 15).

2.2.5 Application to Iterative hard thresholding (IHT)

The iterative hard thresholding algorithm finds a k-group-sparse solution to the linear regression problem, iterating gradient descent steps with hard thresholding to update the model choice as

Algorithm 1 Post-selection Inference for Forward Selection

Input : Response Y, design matrix X, groups $\mathscr{C}_1,\dots,\mathscr{C}_G \subseteq \{1,\dots,p\}$, maximum number of selected groups T, desired accuracy α

Initialize : $\mathscr{S}_0 = \varnothing$, residual $\widehat{\varepsilon}_0 = Y$, $\mathscr{R}_Y = \mathbb{R}_+$

for $t = 1,2,\dots,T$ **do**

  $g_t = \arg\max_{g\in[G]\setminus\mathscr{S}_{t-1}} \{\|X_g^\top \widehat{\varepsilon}_{t-1}\|_2\}$

  Update the model, $\mathscr{S}_t = \{g_1,\dots,g_t\}$, and the residual, $\widehat{\varepsilon}_t = \mathscr{P}^{\perp}_{X_{\mathscr{S}_t}} Y$

  $\mathscr{L}_t \leftarrow \mathrm{span}(\mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} X_{g_t}),\ U_t \leftarrow \frac{\mathscr{P}_{\mathscr{L}_t} Y}{\|\mathscr{P}_{\mathscr{L}_t} Y\|_2},\ Y^t_\perp \leftarrow \mathscr{P}^{\perp}_{\mathscr{L}_t} Y$

  for $g \notin \mathscr{S}_t$ **do**

    $a_{t,g} \leftarrow \|X_{g_t}^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} U_t\|_2^2 - \|X_g^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} U_t\|_2^2$

    $b_{t,g} \leftarrow \langle X_{g_t}^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} U_t, X_{g_t}^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} Y^t_\perp\rangle - \langle X_g^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} U_t, X_g^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} Y^t_\perp\rangle$

    $c_{t,g} \leftarrow \|X_{g_t}^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} Y^t_\perp\|_2^2 - \|X_g^\top \mathscr{P}^{\perp}_{X_{\mathscr{S}_{t-1}}} Y^t_\perp\|_2^2$

    $\mathscr{I}_{t,g} \leftarrow \{r \in \mathbb{R}_+ : a_{t,g} r^2 + 2 b_{t,g} r + c_{t,g} \geq 0\}$

    $\mathscr{R}_Y \leftarrow \mathscr{R}_Y \cap \mathscr{I}_{t,g}$

  end for

  $P_t = \dfrac{\int_{r\in\mathscr{R}_Y, r>\|\mathscr{P}_{\mathscr{L}_t} Y\|_2} r^{k-1} e^{-r^2/2\sigma^2}\,\mathrm{d}r}{\int_{r\in\mathscr{R}_Y} r^{k-1} e^{-r^2/2\sigma^2}\,\mathrm{d}r}$

  $L^t_\alpha = \beta$ s.t. $\dfrac{\int_{r\in\mathscr{R}_Y, r>\|\mathscr{P}_{\mathscr{L}_t} Y\|_2} r^{k-1} e^{-(r^2-2r\beta)/2\sigma^2}\,\mathrm{d}r}{\int_{r\in\mathscr{R}_Y} r^{k-1} e^{-(r^2-2r\beta)/2\sigma^2}\,\mathrm{d}r} = \alpha$

end for

Output : Selected groups $\{g_1,\dots,g_T\}$, p-values $\{P_1,\dots,P_T\}$, confidence interval lower bounds $\{L^1_\alpha,\dots,L^T_\alpha\}$

needed [15, 3]. Given $k \geq 1$, number of iterations T, step sizes η_t, design matrix X and response $Y = y$,

1. Initialize the coefficient vector, $\mathbf{b}_0 = 0 \in \mathbb{R}^p$ (or any other desired initial point).

2. For $t = 1,2,\dots,T$,

   (a) Take a gradient step, $\widetilde{\mathbf{b}}_t = \mathbf{b}_{t-1} - \eta_t X^\top (X\mathbf{b}_{t-1} - y)$.

   (b) Compute $\|(\widetilde{\mathbf{b}}_t)_{\mathscr{C}_g}\|_2$ for each $g \in [G]$ and let $\mathscr{S}_t \subseteq [G]$ index the k largest norms.

   (c) Update the fitted coefficients $\mathbf{b}_t$ via $(\mathbf{b}_t)_j = (\widetilde{\mathbf{b}}_t)_j \cdot \mathbb{1}\left\{j \in \cup_{g\in\mathscr{S}_t} \mathscr{C}_g\right\}$.

Here we are typically interested in testing $\mathrm{Question}_{g,\mathscr{S}_T}$ for each $g \in \mathscr{S}_T$. We condition on the selection event, $e(Y)$, given by the sequence of k-group-sparse models $\mathscr{S}_1,\dots,\mathscr{S}_T$ selected at each

stage of the algorithm, which is characterized by the inequalities

$$\|(\widetilde{\mathbf{b}}_t)_{\mathscr{C}_g}\|_2 > \|(\widetilde{\mathbf{b}}_t)_{\mathscr{C}_h}\|_2 \quad \text{for all } t = 1,\dots,T, \text{ and all } g \in \mathscr{S}_t, h \notin \mathscr{S}_t. \tag{2.8}$$

Fixing a group $g \in \mathscr{S}_T$ to test, determining $\mathscr{R}_Y = \{r > 0 : z_Y(r) \in \mathscr{A}_{e(Y)}\}$ involves checking whether all of the inequalities in (2.8) are satisfied with $y = z_Y(r)$. First, with the response Y replaced by $y = z_Y(r)$, we show that we can write $\widetilde{\mathbf{b}}_t = r \cdot c_t + d_t$ for each $t = 1,\dots,T$, where $c_t, d_t \in \mathbb{R}^p$ are independent of r; we derive c_t, d_t inductively as

$$\begin{cases} c_1 = \frac{\eta_1}{n} X^\top U, \\ d_1 = (\mathbf{I} - \frac{\eta_1}{n} X^\top X)\mathbf{b}_0 + \frac{\eta_1}{n} X^\top Y_\perp, \end{cases} \quad \begin{cases} c_t = (\mathbf{I}_p - \frac{\eta_t}{n} X^\top X)\mathscr{P}_{\mathscr{S}_{t-1}} c_{t-1} + \frac{\eta_t}{n} X^\top U, \\ d_t = (\mathbf{I}_p - \frac{\eta_t}{n} X^\top X)\mathscr{P}_{\mathscr{S}_{t-1}} d_{t-1} + \frac{\eta_t}{n} X^\top Y_\perp \end{cases} \text{for } t \geq 2.$$

In fact, at time $t = 1$,

$$\begin{aligned} \widetilde{\mathbf{b}}_1 = b_0 - \eta_1 \nabla f(b_0) &= b_0 - \eta_1 \left(\frac{1}{n} X^\top (X b_0 - z_Y(r)) \right) \\ &= r \cdot \left[\frac{\eta_1}{n} X^\top U \right] + \left[(\mathbf{I} - \frac{\eta_1}{n} X^\top X) b_0 + \frac{\eta_1}{n} X^\top Y_\perp \right] =: r \cdot c_1 + d_1. \end{aligned}$$

Next, at each time $t = 2,\dots,T$, assume that $\widetilde{\mathbf{b}}_{t-1} = c_{t-1} r + d_{t-1}$. Then, writing $\mathscr{P}_{\mathscr{S}_{t-1}}$ as the matrix in $\mathbb{R}^{p \times p}$ which acts as the identity on groups in $\mathscr{S}_{t-1}$ and sets all other groups to zero, we have $b_{t-1} = \mathscr{P}_{\mathscr{S}_{t-1}} \widetilde{\mathbf{b}}_{t-1} = \mathscr{P}_{\mathscr{S}_{t-1}} c_{t-1} r + \mathscr{P}_{\mathscr{S}_{t-1}} d_{t-1}$, and so

$$\begin{aligned} &\widetilde{\mathbf{b}}_t = b_{t-1} - \eta_t \nabla f(b_{t-1}) = b_{t-1} - \eta_t \left(\frac{1}{n} X^\top (X b_{t-1} - z_Y(r)) \right) \\ &= r \cdot \left[(\mathbf{I}_p - \frac{\eta_t}{n} X^\top X)\mathscr{P}_{\mathscr{S}_{t-1}} c_{t-1} + \frac{\eta_t}{n} X^\top U \right] + \left[(\mathbf{I}_p - \frac{\eta_t}{n} X^\top X)\mathscr{P}_{\mathscr{S}_{t-1}} d_{t-1} + \frac{\eta_t}{n} X^\top Y_\perp \right] =: r \cdot c_t + d_t. \end{aligned}$$

Now we compute the region $\mathscr{R}_Y$. For each $t = 1,\dots,T$ and each $g \in \mathscr{S}_t, h \notin \mathscr{S}_t$, the corre-

sponding inequality in (2.8), after writing $\widetilde{\mathbf{b}}_t = r \cdot c_t + d_t$, can be expressed as

$$r^2 \cdot \|(c_t)_{\mathscr{C}_g}\|_2^2 + 2r \cdot \langle (c_t)_{\mathscr{C}_g}, (d_t)_{\mathscr{C}_g} \rangle + \|(d_t)_{\mathscr{C}_g}\|_2^2 > r^2 \cdot \|(c_t)_{\mathscr{C}_h}\|_2^2 + 2r \cdot \langle (c_t)_{\mathscr{C}_h}, (d_t)_{\mathscr{C}_h} \rangle + \|(d_t)_{\mathscr{C}_h}\|_2^2.$$

As for the forward stepwise procedure, solving this quadratic inequality over $r \in \mathbb{R}_+$, we obtain a region $\mathscr{I}_{t,g,h} \subseteq \mathbb{R}_+$ that is either a single interval or a union of two disjoint intervals whose endpoints we can calculate explicitly. Finally, we obtain $\mathscr{R}_Y = \bigcap_{t=1,\ldots,T} \bigcap_{g \in \mathscr{S}_t} \bigcap_{h \in [G] \setminus \mathscr{S}_t} \mathscr{I}_{t,g,h}$.

2.2.6 *Application to Group Lasso*

The group lasso, first introduced by Yuan and Lin [4], is a convex optimization method for linear regression where the form of the penalty is designed to encourage group-wise sparsity of the solution. It is an extension of the lasso method [11] for linear regression. The method is given by

$$\widehat{\beta} = \arg\min_{\mathbf{b}} \Big\{ \frac{1}{2} \|y - X\mathbf{b}\|_2^2 + \lambda \sum_g \|\mathbf{b}_{\mathscr{C}_g}\|_2 \Big\},$$

where $\lambda > 0$ is a penalty parameter. The penalty $\sum_g \|\mathbf{b}_{\mathscr{C}_g}\|_2$ promotes sparsity at the group level.[5]

For this method, we perform inference on the group support $\mathscr{S}$ of the fitted model $\widehat{\beta}$. We would like to test Question$_{g,\mathscr{S}}$ for each $g \in \mathscr{S}$. In this setting, for groups of size ≥ 2, we believe that it is not possible to analytically calculate $\mathscr{R}_Y$, and furthermore, that there is no additional information that we can condition on to make this computation possible, without losing all power to do inference.

We thus propose a numerical approximation that circumvents the need for an explicit calculation of $\mathscr{R}_Y$. Examining the calculation of the p-value P and the lower bound L_α in Theorem 2.2.1,

5. Our method can also be applied to a modification of group lasso designed for overlapping groups [13] with a nearly identical procedure but we do not give details here.

we see that we can write $P = f_Y(0)$ and can find L_α as the unique solution to $f_Y(L_\alpha) = \alpha$, where

$$f_Y(t) = \frac{\mathbb{E}_{r\sim\sigma\cdot\chi_k}\left[e^{rt/\sigma^2}\cdot\mathbb{1}\left\{r\in\mathscr{R}_Y, r > \|\mathscr{P}_{\mathscr{L}}Y\|_2\right\}\right]}{\mathbb{E}_{r\sim\sigma\cdot\chi_k}\left[e^{rt/\sigma^2}\cdot\mathbb{1}\left\{r\in\mathscr{R}_Y\right\}\right]},$$

where we treat Y as fixed in this calculation and set $k = \dim(\mathscr{L}) = \operatorname{rank}(X_{\mathscr{S}\setminus g})$. Both the numerator and denominator can be approximated by taking a large number B of samples $r \sim \sigma\cdot\chi_k$ and taking the empirical expectations. Checking $r\in\mathscr{R}_Y$ is equivalent to running the group lasso with the response replaced by $y = z_Y(r)$, and checking if the resulting selected model remains unchanged.

This may be problematic, however, if $\mathscr{R}_Y$ is in the tails of the $\sigma\cdot\chi_k$ distribution. We implement an importance sampling approach by repeatedly drawing $r\sim\psi$ for some density ψ; we find that $\psi = \|\mathscr{P}_{\mathscr{L}}Y\|_2 + \mathscr{N}(0,\sigma^2)$ works well in practice. Given samples $r_1,\ldots,r_B\sim\psi$ we then estimate

$$f_Y(t) \approx \widehat{f}_Y(t) := \frac{\sum_b \frac{\psi_{\sigma\cdot\chi_k}(r_b)}{\psi(r_b)}\cdot e^{r_b t/\sigma^2}\cdot\mathbb{1}\left\{r_b\in\mathscr{R}_Y, r_b > \|\mathscr{P}_{\mathscr{L}}Y\|_2\right\}}{\sum_b \frac{\psi_{\sigma\cdot\chi_k}(r_b)}{\psi(r_b)}\cdot e^{r_b t/\sigma^2}\cdot\mathbb{1}\left\{r_b\in\mathscr{R}_Y\right\}}$$

where $\psi_{\sigma\cdot\chi_k}$ is the density of the $\sigma\cdot\chi_k$ distribution. We then estimate $P\approx\widehat{P} = \widehat{f}_Y(0)$. Finally, since $\widehat{f}_Y(t)$ is continuous and strictly increasing in t, we estimate L_α by numerically solving $\widehat{f}_Y(t) = \alpha$.

2.3 Empirical results

We present results from experiments on simulated and real data, performed in R [16].

2.3.1 Simulated data

We fix sample size $n = 500$ and $G = 50$ groups each of size 10. For each trial, we generate a design matrix X with i.i.d. $\mathscr{N}(0,1/n)$ entries, set β with its first 50 entries (corresponding to first $s = 5$ groups) equal to τ and all other entries equal to 0, and set $Y = X\beta + \mathscr{N}(0,\mathbf{I}_n)$.

IHT We run IHT to select $k = 10$ groups over $T = 5$ iterations, with step sizes $\eta_t = 2$ and initial point $\mathbf{b}_0 = 0$. For a moderate signal strength $\tau = 1.5$, we plot the p-values for each selected group across 200 trials in Figure 2.1; each group displays p-values only for those trials in which it was selected. The histogram of p-values for the s true signals and for the $G - s$ nulls are also shown. We see that, at this moderate signal strength, the the distribution of p-values for the true signals concentrates near zero while the null p-values are roughly uniformly distributed.

Next we look at the confidence intervals given by our method, examining their empirical coverage across different signal strengths τ in Figure 2.2. We fix confidence level 0.9 (i.e. $\alpha = 0.1$) and run 2,000 trials to obtain empirical coverage with respect to both $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ and $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$, with results shown separately for true signals and for nulls. For true signals, we see that the confidence interval for $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ is somewhat conservative while the coverage for $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$ is right at the target level, as expected from our theory. As signal strength τ increases, the gap is reduced for the true signals; this is because $\mathrm{dir}_{\mathscr{L}}(Y)$ becomes an increasingly more accurate estimate of $\mathrm{dir}_{\mathscr{L}}(\mu)$, and so the gap in the inequality $\|\mathscr{P}_{\mathscr{L}}\mu\|_2 \geq \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$ is reduced. For the nulls, if the set of selected groups contains the support of the true model, which is nearly always true for higher signal levels τ in this experiment, then the two are equivalent (as $\|\mathscr{P}_{\mathscr{L}}\mu\|_2 = \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle = 0$), and coverage is at the target level. At low signal levels τ, however, one or more of the s true groups is occasionally missed, in which case we again have a gap in the inequality $\|\mathscr{P}_{\mathscr{L}}\mu\|_2 \geq \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$.

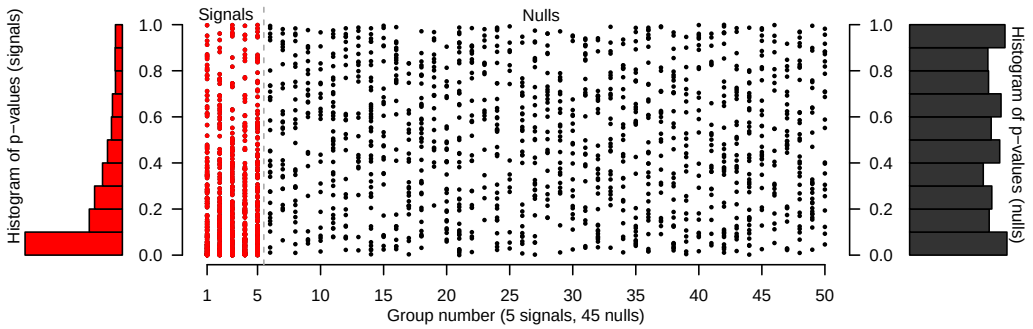


Figure 2.1: Iterative hard thresholding (IHT). For each group, we plot its p-value for each trial in which that group was selected, for 200 trials. Histograms of the p-values for true signals (left, red) and for nulls (right, gray) are attached.

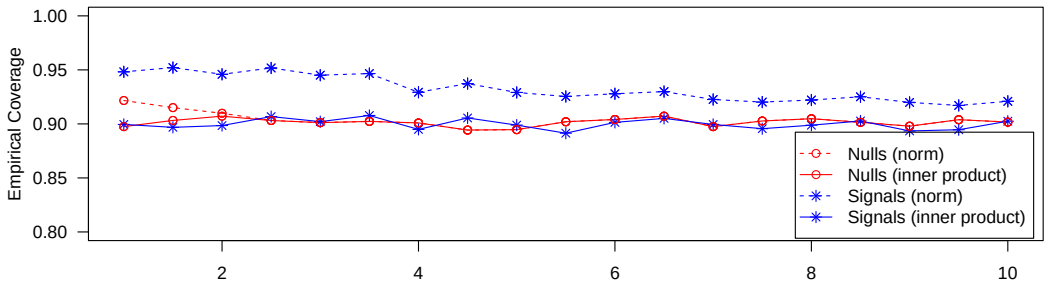


Figure 2.2: Iterative hard thresholding (IHT). Empirical coverage over 200 trials with signal strength τ. "Norm" and "inner product" refer to coverage of $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ and $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$, respectively.

Group lasso The group lasso is run with penalty parameter $\lambda = 4$. The group lasso algorithm is run via the R package `gglasso` [17]. Figure 2.3 shows the p-values obtained with the group lasso, while Figure 2.4 displays the coverage for the norms $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ and the inner products $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$; these plots are produced exactly as Figures 2.1 and 2.2 for IHT, except that only 200 trials are shown for the coverage plot due to the slower run time of this method. We observe very similar trends for this method as for IHT.

Forward stepwise The forward stepwise method is implemented with $T = 10$ many steps, and p-values and confidence intervals are computed by considering all 10 selected groups simultaneously at the end of the procedure (rather than sequentially) so that the results are more comparable to the other two methods. Figure 2.5 shows the p-values obtained with the forward stepwise method, while Figure 2.6 displays the coverage for the norms $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ and the inner products $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$; these plots are produced exactly as Figures 2.1 and 2.2 for IHT. We again observe similar trends in the results.

2.3.2 California health data

We examine the 2015 California county health data[6] which was also studied by Loftus and Taylor [5]. We fit a linear model where the response is the log-years of potential life lost and the covariates

6. Available at `http://www.countyhealthrankings.org`

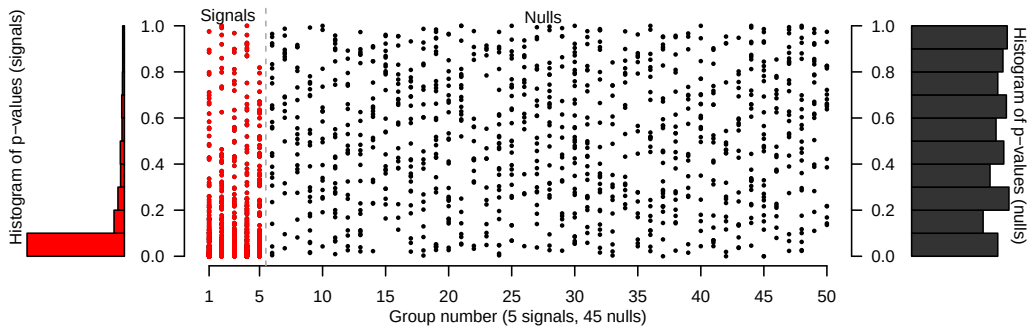


Figure 2.3: Group lasso. For each group, we plot its p-value for each trial in which that group was selected, for 200 trials. Histograms of the p-values for true signals (left, red) and for nulls (right, gray) are attached.

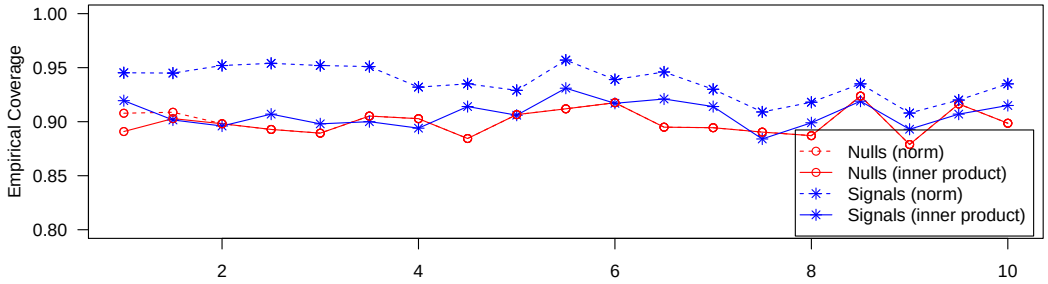


Figure 2.4: Group lasso. Empirical coverage over 200 trials with signal strength τ. "Norm" and "inner product" refer to coverage of $\|\mathcal{P}_{\mathcal{L}}\mu\|_2$ and $\langle \mathrm{dir}_{\mathcal{L}}(Y), \mu \rangle$, respectively.

are the 34 predictors in this data set. We first let each predictor be its own group (i.e., group size 1) and obtain p-values by running each of the three algorithms considered in Section 2.2. Next, we form a grouped model by expanding each predictor into a group of size three using the first three non-constant Legendre polynomials, thus expanding predictor X_j to the group $(X_j, \frac{1}{2}(3X_j^2 - 1), \frac{1}{2}(5X_j^3 - 3X_j))$. In each case we set parameters so that 8 groups are selected. The selected groups and their corresponding p-values are given in Table 2.1; interestingly, even when the same predictor is selected by multiple methods, its p-value can differ substantially across the different methods.

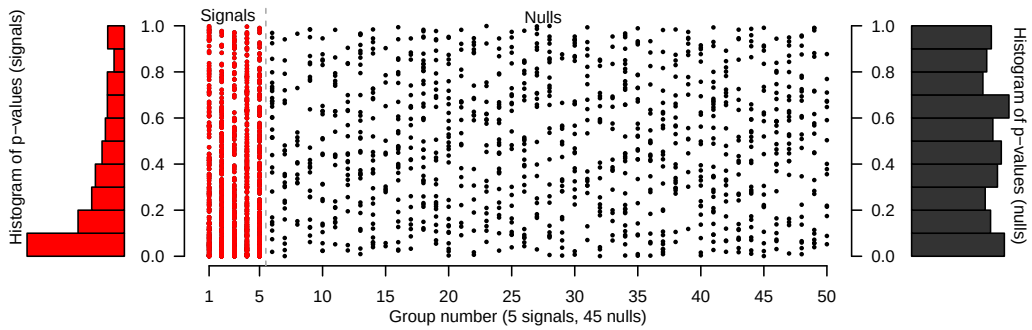


Figure 2.5: Forward stepwise regression. For each group, we plot its p-value for each trial in which that group was selected, for 200 trials. Histograms of the p-values for true signals (left, red) and for nulls (right, gray) are attached.

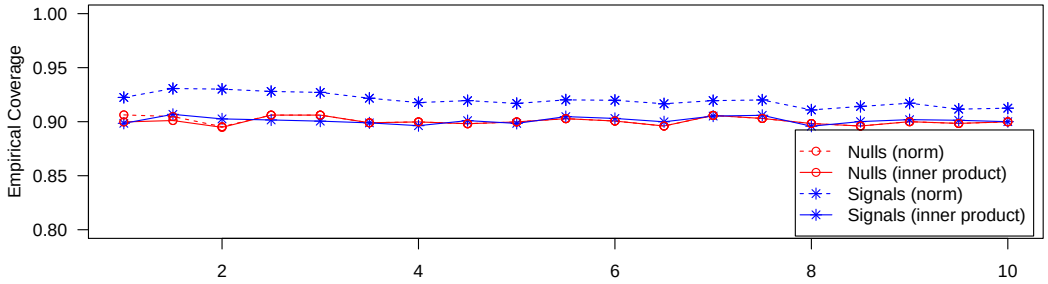


Figure 2.6: Forward stepwise regression. Empirical coverage over 2000 trials with signal strength τ. "Norm" and "inner product" refer to coverage of $\|\mathscr{P}_{\mathscr{L}}\mu\|_2$ and $\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu \rangle$, respectively.

2.4 Proofs

2.4.1 Proof of Theorem 2.2.1

For any $y \in \mathscr{A}$, define a function $f_y : \mathbb{R} \to [0,1]$ as

$$f_y(t) = \frac{\int_{r \in \mathscr{R}_y, r > \|\mathscr{P}_{\mathscr{L}} y\|_2} r^{k-1} e^{-(r^2-2rt)/2\sigma^2} \, \mathrm{d}r}{\int_{r \in \mathscr{R}_y} r^{k-1} e^{-(r^2-2rt)/2\sigma^2} \, \mathrm{d}r}.$$

(As always we ignore the case $\mathscr{P}_{\mathscr{L}} y = 0$ to avoid degeneracy.) By examining the integrals, we can immediately see that, for any fixed y, $f_y(t)$ is strictly increasing as a function of t, with $\lim_{t \to -\infty} f_y(t) = 0$ and $\lim_{t \to \infty} f_y(t) = 1$. These properties guarantee that, for any fixed y and any fixed $\alpha \in (0,1)$, there is a unique $t \in \mathbb{R}$ with $f_y(t) = \alpha$, i.e. this proves the existence and uniqueness

Group size	Forward stepwise p-value / seq. p-value		Iterative hard thresholding p-value		Group lasso p-value	
1	80th percentile income	0.116 / 0.000	80th percentile income	0.000	80th percentile income	0.000
	Injury death rate	0.000 / 0.000	Injury death rate	0.000	% Obese	0.007
	Violent crime rate	0.016 / 0.000	% Smokers	0.004	% Physically inactive	0.040
	% Receiving HbA1c	0.591 / 0.839	% Single-parent household	0.009	Violent crime rate	0.055
	% Obese	0.481 / 0.464	% Children in poverty	0.332	% Single-parent household	0.075
	Chlamydia rate	0.944 / 0.975	Physically unhealthy days	0.716	Injury death rate	0.235
	% Physically inactive	0.654 / 0.812	Food environment index	0.807	% Smokers	0.701
	% Alcohol-impaired	0.104 / 0.104	Mentally unhealthy days	0.957	Preventable hospital stays rate	0.932
3	80th percentile income	0.001 / 0.000	Injury death rate	0.000	80th percentile income	0.000
	Injury death rate	0.044 / 0.000	80th percentile income	0.000	Injury death rate	0.000
	Violent crime rate	0.793 / 0.617	% Smokers	0.000	% Single-parent household	0.038
	% Physically inactive	0.507 / 0.249	% Single-parent household	0.005	% Physically inactive	0.043
	% Alcohol-impaired	0.892 / 0.933	Food environment index	0.057	% Obese	0.339
	% Severe housing problems	0.119 / 0.496	% Children in poverty	0.388	% Alcohol-impaired	0.366
	Chlamydia rate	0.188 / 0.099	Physically unhealthy days	0.713	% Smokers	0.372
	Preventable hospital stays rate	0.421 / 0.421	Mentally unhealthy days	0.977	Violent crime rate	0.629

Table 2.1: Selective p-values for selected groups in the California county health data experiment. The predictors obtained with forward stepwise are tested both simultaneously at the end of the procedure (first p-value shown), and also tested sequentially (second p-value shown), and are displayed in the selected order. For IHT and group lasso the predictors are shown in order of increasing selective p-value.

of L_α as required.

Furthermore, Lemma 2.2.1 immediately implies that, after conditioning on the event $Y \in \mathcal{A}$, and on the values of $\mathrm{dir}_{\mathcal{L}}(Y)$ and $\mathcal{P}_{\mathcal{L}}^{\perp} Y$, the conditional density of $\|\mathcal{P}_{\mathcal{L}} Y\|_2$ is

$$\propto r^{k-1} e^{-(r^2-2rt_Y)/2\sigma^2} \cdot \mathbb{1}\{r \in \mathcal{R}_Y\}$$

for $t_Y := \langle \mathrm{dir}_{\mathcal{L}}(Y), \mu \rangle$, and therefore, $f_Y(t_Y) \sim \mathrm{Uniform}[0,1]$. In the case that $\mu \perp \mathcal{L}$, we have $t_Y = 0$ always and therefore $P = f_Y(0) \sim \mathrm{Uniform}[0,1]$, as desired. In the general case, by definition of L_α, we have $f_Y(L_\alpha) = \alpha$ and so, again using the fact that $f_Y(\cdot)$ is strictly increasing,

$$f_Y(t_Y) \leq \alpha = f_Y(L_\alpha) \;\Leftrightarrow\; t_Y \leq L_\alpha,$$

and so by definition of t_Y,

$$\mathbb{P}\{\langle \mathrm{dir}_{\mathcal{L}}(Y), \mu \rangle < L_\alpha\} = \mathbb{P}\{t_Y < L_\alpha\} = \mathbb{P}\{f_Y(t_Y) < \alpha\} = \alpha.$$

Furthermore, we know that $\|\mathscr{P}_{\mathscr{L}}\mu\|_2 \geq \langle \mathrm{dir}_{\mathscr{L}}(Y), \mu\rangle = t_Y$, and so

$$\mathbb{P}\{\|\mathscr{P}_{\mathscr{L}}\mu\|_2 < L_\alpha\} \leq \mathbb{P}\{\langle \mathrm{dir}_{\mathscr{L}}(Y), \mu\rangle < L_\alpha\} = \alpha.$$

Finally, we see that since $P = f_Y(0)$ while $\alpha = f_Y(L_\alpha)$, $P < \alpha$ if and only if $0 < L_\alpha$.

2.4.2 Proof of Lemma 2.2.1

We begin with the following elementary calculation (for completeness the proof is given below):

Lemma 2.4.1. *Suppose that $\widetilde{Y} \sim \mathscr{N}(\widetilde{\mu}, \sigma^2\mathbf{I}_k)$. Let $R = \|\widetilde{Y}\|_2 \in \mathbb{R}_+$ and $U = \mathrm{dir}(\widetilde{Y}) \in \mathbb{S}^{k-1}$ be the radius and direction of the random vector $\widetilde{Y}$. Then the joint distribution of (R,U) has density*

$$f(r,u) \propto r^{k-1}\exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2r\cdot\langle u, \widetilde{\mu}\rangle\right)\right\} \text{ for } (r,u) \in \mathbb{R}_+ \times \mathbb{S}^{k-1}.$$

Next, let $\mathbf{V} \in \mathbb{R}^{n\times k}$ be an orthonormal basis for $\mathscr{L}$ and let

$$\widetilde{Y} = \mathbf{V}^\top Y \sim \mathscr{N}(\widetilde{\mu}, \sigma^2\mathbf{I}_k) \text{ where } \widetilde{\mu} = \mathbf{V}^\top\mu.$$

Now let $R = \|\widetilde{Y}\|_2 = \|\mathscr{P}_{\mathscr{L}}Y\|_2$ and let $U = \mathrm{dir}(\widetilde{Y}) = \mathbf{V}^\top \mathrm{dir}_{\mathscr{L}}(Y)$; note that $\mathrm{dir}_{\mathscr{L}}(Y) = \mathbf{V}U$.

Defining $W = \mathscr{P}^{\perp}_{\mathscr{L}}Y$, we see that $Y = r\cdot\mathbf{V}u + w$, and that $\widetilde{Y} \perp\!\!\!\perp W$ by properties of the normal distribution. Combining this with the result of Lemma 2.4.1, we see that the joint density of (R,U,W) is given by

$$f_{R,U,W}(r,u,w) \propto r^{k-1}\exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2r\cdot\langle u, \widetilde{\mu}\rangle\right)\right\}\cdot\exp\left\{-\frac{1}{2\sigma^2}\|w - \mathscr{P}^{\perp}_{\mathscr{L}}\mu\|_2^2\right\}$$

for $(r,u,w) \in \mathbb{R}_+ \times \mathbb{S}^{k-1} \times \mathscr{L}_\perp$. After conditioning on the event $\{Y \in \mathscr{A}\}$, this density becomes

$$\propto r^{k-1}\exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2r\cdot\langle u, \widetilde{\mu}\rangle\right)\right\}\cdot\exp\left\{-\frac{1}{2\sigma^2}\|w - \mathscr{P}^{\perp}_{\mathscr{L}}\mu\|_2^2\right\}\cdot\mathbb{1}\{r\cdot\mathbf{V}u + w \in \mathscr{A}\}.$$

Next note that the event $\{Y \in \mathscr{A}\}$ is equivalent to $\{R \in \mathscr{R}_Y\}$ where $\mathscr{R}_Y = \{r > 0 : r \cdot \mathbf{V}U + W \in \mathscr{A}\}$, and so the conditional density of R, after conditioning on U, W, and on the event $Y \in \mathscr{A}$, is

$$\propto r^{k-1} \exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2r \cdot \langle U, \widetilde{\mu}\rangle\right)\right\} \cdot \mathbb{1}\{R \in \mathscr{R}_Y\},$$

as desired. Now we prove our supporting result, Lemma 2.4.1.

Proof of Lemma 2.4.1. It's easier to work with the parametrization (Z,U) where $Z = \log(R)$. By a simple change of variables calculation, the claim in the lemma is equivalent to showing that

$$f_{Z,U}(z,u) \propto e^{kz} \exp\left\{-\frac{1}{2\sigma^2}\left(e^{2z} - 2e^z \cdot \langle u, \widetilde{\mu}\rangle\right)\right\} \text{ for } (z,u) \in \mathbb{R} \times \mathbb{S}^{k-1}.$$

Fix any $\varepsilon \in (0,1)$ and, for each $(z,u) \in \mathbb{R} \times \mathbb{S}^{k-1}$, consider the region $(z-\varepsilon, z+\varepsilon) \times \mathscr{C}_u^\varepsilon \subseteq \mathbb{R} \times \mathbb{S}^{k-1}$, where $\mathscr{C}_u^\varepsilon$ is a spherical cap, $\mathscr{C}_u^\varepsilon := \{v \in \mathbb{S}^{k-1} : \|v-u\|_2 < \varepsilon\}$. Let s_ε be the surface area of $\mathscr{C}_u^\varepsilon \subseteq \mathbb{S}^{k-1}$ (note that this surface area does not depend on u since it's rotation invariant).

To check that our density is correct, it is sufficient to check that

$$\mathbb{P}\left\{(Z,U) \in (z-\varepsilon, z+\varepsilon) \times \mathscr{C}_u^\varepsilon\right\} \propto$$
$$\text{Volume}\left((z-\varepsilon, z+\varepsilon) \times \mathscr{C}_u^\varepsilon\right) \cdot e^{kz} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(e^{2z} - 2e^z \cdot \langle u, \widetilde{\mu}\rangle\right)\right\} \cdot (1 + o(1)),$$

where the $o(1)$ term is with respect to the limit $\varepsilon \to 0$ while (z,u) is held fixed, and where the constant of proportionality is independent of ε and of z,u. We can also calculate $\text{Volume}\left((z-\varepsilon, z+\varepsilon) \times \mathscr{C}_u^\varepsilon\right) = 2\varepsilon \cdot s_\varepsilon$.

Now consider

$$\mathscr{Y}_{z,u}^\varepsilon = \left\{y \in \mathbb{R}^n : \frac{y}{\|y\|_2} \in \mathscr{C}_u^\varepsilon, \log(\|y\|_2) \in (z-\varepsilon, z+\varepsilon)\right\} \subseteq \mathbb{R}^n.$$

We have

$$\mathbb{P}\left\{(Z,U) \in (z-\varepsilon, z+\varepsilon) \times \mathscr{C}_u^\varepsilon\right\} = \mathbb{P}\left\{\widetilde{Y} \in \mathscr{Y}_{z,u}^\varepsilon\right\}.$$

Since $\mathscr{Y}^{\varepsilon}_{z,u} = \cup_{t\in(e^{z-\varepsilon},e^{z+\varepsilon})}(t\cdot\mathscr{C}^{\varepsilon}_u)$, and the surface area of $t\cdot C^{\varepsilon}_u \subseteq t\cdot\mathbb{S}^{k-1}$ is equal to $s_\varepsilon t^{k-1}$, we can also calculate

$$\mathsf{Volume}(\mathscr{Y}^{\varepsilon}_{z,u}) = \int_{t=e^{z-\varepsilon}}^{e^{z+\varepsilon}} s_\varepsilon t^{k-1}\,\mathrm{d}t = \frac{1}{k}s_\varepsilon t^k\Big|_{t=e^{z-\varepsilon}}^{e^{z+\varepsilon}} = \frac{1}{k}s_\varepsilon\cdot(e^{k(z+\varepsilon)}-e^{k(z-\varepsilon)}) = 2\varepsilon\cdot s_\varepsilon\cdot e^{kz}\cdot(1+o(1)),$$

since $e^{k(z+\varepsilon)}-e^{k(z-\varepsilon)} = e^{kz}\cdot 2k\varepsilon\cdot(1+o(1))$. And, since $\max_{y\in\mathscr{Y}^{\varepsilon}_{z,u}}\|y-e^z\cdot u\|_2\to 0$ as $\varepsilon\to 0$, then for any $y\in\mathscr{Y}^{\varepsilon}_{z,u}$, the density of $\widetilde{Y}$ at this point is given by

$$f_{\widetilde{Y}}(y) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{1}{2\sigma^2}\|y-\widetilde{\mu}\|_2^2} = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{1}{2\sigma^2}\|e^z\cdot u-\widetilde{\mu}\|_2^2}\cdot(1+o(1)),$$

where again the $o(1)$ term is with respect to the limit $\varepsilon\to 0$ while (z,u) is held fixed. So, we have

$$\begin{aligned}
&\mathbb{P}\left\{(Z,U)\in(z-\varepsilon,z+\varepsilon)\times\mathscr{C}^{\varepsilon}_u\right\} = \mathbb{P}\left\{\widetilde{Y}\in\mathscr{Y}^{\varepsilon}_{z,u}\right\} = \int_{y\in\mathscr{Y}^{\varepsilon}_{z,u}} f_{\widetilde{Y}}(y)\,\mathrm{d}y\\
&= \mathsf{Volume}(\mathscr{Y}^{\varepsilon}_{z,u})\cdot\frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{1}{2\sigma^2}\|e^z\cdot u-\widetilde{\mu}\|_2^2}\cdot(1+o(1))\\
&= 2\varepsilon\cdot s_\varepsilon\cdot e^{kz}\cdot(1+o(1))\cdot\frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{1}{2\sigma^2}\|e^z\cdot u-\widetilde{\mu}\|_2^2}\cdot(1+o(1))\\
&= 2\varepsilon\cdot s_\varepsilon\cdot e^{kz}\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(e^{2z}-2e^z\cdot\langle u,\widetilde{\mu}\rangle\right)\right\}\cdot\left[\frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{1}{2\sigma^2}\|\widetilde{\mu}\|_2^2}\right]\cdot(1+o(1)),
\end{aligned}$$

which gives the desired result since the term in square brackets is constant with respect to z,u,ε.

□

CHAPTER 3

CONTRACTIONS AND UNIFORM CONVERGENCE OF ISOTONIC REGRESSION

Isotonic regression is a powerful nonparametric tool used for estimating a monotone signal from noisy data. Specifically, our data consists of observations $y_1, \ldots, y_n \in \mathbb{R}$, which are assumed to be noisy observations of some monotone increasing signal—for instance, we might assume that $\mathbb{E}[y_1] \leq \cdots \leq \mathbb{E}[y_n]$. Isotonic (least-squares) regression solves the optimization problem

$$\text{Minimize } \|y - x\|_2^2 \text{ subject to } x_1 \leq \cdots \leq x_n$$

in order to estimate the underlying signal.

This regression problem can be viewed as a convex projection, since $\mathscr{K}_{\mathsf{iso}} = \{x \in \mathbb{R}^n : x_1 \leq \cdots \leq x_n\}$ is a convex cone. We will write

$$\mathsf{iso}(y) := \mathscr{P}_{\mathscr{K}_{\mathsf{iso}}}(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \|y - x\|_2^2 : x_1 \leq \cdots \leq x_n \right\}$$

to denote the projection to this cone, which solves the least-squares isotonic regression problem. This projection can be computed in finite time with the Pool Adjacent Violators Algorithm (PAVA), developed by Barlow et al. [18].

In fact, mapping y to $\mathsf{iso}(y)$ is known to also solve the isotonic binary regression problem. This arises when the data is binary, that is, $y \in \{0,1\}^n$. If we assume that $y_i \sim \text{Bernoulli}(x_i)$, then the constrained maximum likelihood estimator is exactly equal to the projection $\mathsf{iso}(y)$ (Robertson et al. [19]).

In this Chapter, [1] we examine the properties of the isotonic projection operator $x \mapsto \mathsf{iso}(x)$, with respect to different norms $\|\cdot\|$ on $\mathbb{R}^n$. We examine the conditions on $\|\cdot\|$ needed in order to ensure that $x \mapsto \mathsf{iso}(x)$ is contractive with respect to this norm, and in particular, we define the

1. The work presented in this chapter is published in Yang et al. [20].

new "sliding window norm" which measures weighted averages over "windows" of the vector x, i.e. contiguous stretches of the form $(x_i, x_{i+1}, \ldots, x_{j-1}, x_j)$ for some indices $1 \leq i \leq j \leq n$. This sliding window norm then provides a tool for analyzing the convergence behavior of isotonic regression in a setting where our data is given by $y_i = x_i + \text{noise}$. If the underlying signal x is believed to be (approximately) monotone increasing, then $\mathsf{iso}(y)$ will provide a substantially better estimate of x than the observed vector y itself. By using our results on contractions with respect to the isotonic projection operator, we obtain clean, finite-sample bounds on the pointwise errors, $\left|x_i - \mathsf{iso}(y)_i\right|$, which are locally adaptive to the behavior of the signal x in the region around the index i and hold uniformly over the entire sequence.

3.1 Related work

There is extensive literature studying convergence rates of univariate isotonic regression, in both finite-sample and asymptotic settings. For an asymptotic formulation of the problem, since the signal $x \in \mathbb{R}^n$ must necessarily change as $n \to \infty$, a standard method for framing this as a sequence of problems indexed by n is to consider a fixed function $f : [0,1] \to \infty$, and then for each n, define $x_i = f(i/n)$ (or more generally, $x_i = f(t_i)$ for points t_i that are roughly uniformly spaced). Most models in the literature assume that $y_i = x_i + \sigma \cdot \varepsilon_i$, where the noise terms ε_i are i.i.d. standard normal variables (or, more generally, are zero-mean variables that satisfy some moment assumptions or are subgaussian).

One class of existing results treats *global* convergence rates, where the goal is to bound the error $\|x - \mathsf{iso}(y)\|_2$, or more generally to bound $\|x - \mathsf{iso}(y)\|_p$ for some ℓ_p norm. The estimation error under the ℓ_2 norm was studied by Van de Geer [21], Wang and Chen [22], Meyer and Woodroofe [23], among others. Van de Geer [24] obtains the asymptotic risk bounds for certain 'bounded' isotonic regression under Hellinger distance, whereas Zhang [25] establishes the non-asymptotic risk bounds for general ℓ_p norm—in particular, for $p = 2$, they show that the least-squares estimator $\mathsf{iso}(y)$ of the signal x has error scaling as $\|x - \widehat{x}\|_2/\sqrt{n} \sim n^{-1/3}$. Recent work by Chatterjee et al. [26] considers non-asymptotic minimax rates for the estimation error, focusing specifically on

$\|x-\widehat{x}\|_2$ for any estimator $\widehat{x}$ to obtain a minimax rate. Under a Gaussian noise model, they prove that the minimax rate scales as $\|x-\widehat{x}\|_2/\sqrt{n} \gtrsim n^{-1/3}$ over the class of monotone and Lipschitz signals x, which matches the error rate of the constrained maximum likelihood estimator (i.e. the isotonic least-squares projection, $\mathsf{iso}(y)$) established earlier. They also study minimax rates in a range of settings, including piecewise constant signals, which we will discuss later on.[2] The piecewise constant case is further studied by Gao et al. [27], where they prove that the sharp adaptation rate in the mean squared risk in this case can be achieved by a penalized lease square estimator.

A separate class of results considers *local* convergence rates, where the error at a particular index, i.e. $\left|x_i - \mathsf{iso}(y)_i\right|$ for some particular i, may scale differently in different regions of the vector. In an asymptotic setting, where $n \to \infty$ and the underlying signal comes from a function $f : [0,1] \to \mathbb{R}$, we may consider an estimator $\widehat{f} : [0,1] \to \mathbb{R}$, where $\widehat{f}(t)$ is estimated via $\mathsf{iso}(y)_i$ for $t \approx i/n$. Results in the literature for this setting study the asymptotic rate of convergence of $\left|f(t) - \widehat{f}(t)\right|$, which depends on the local properties of f near t. Brunk [28] establishes the convergence rate as well as the limiting distribution when $f'(t)$ is positive, whereas Wright [29] generalizes the result to the case of t lying in a flat region, i.e. $f'(t) = 0$. Cator [30] shows that the isotonic estimator adapts to the unknown function locally and is asymptotically minimax optimal for local behavior. Relatedly, Dümbgen [31] gives confidence bands in the related Gaussian white noise model, by taking averages over windows of the data curve, i.e. ranges of the form $[t_0, t_1]$ near the point t of interest.

In addition, many researchers have considered the related problem of monotone density estimation, where we aim to estimate a monotone decreasing density from n samples drawn from that distribution. This problem was first studied by Grenander [32], and has attracted much attention since then, see Rao [33], Groeneboom [34], Birgé [35], Birgé and Massart [36], Carolan and Dykstra [37], Balabdaoui et al. [38], Jankowski [39], among others. Birgé [35] proves a $n^{-1/3}$

2. Chatterjee et al. [26]'s results, which they describe as "local minimax" bounds, are "local" in the sense that the risk bound they provide is specific to an individual signal $x \in \mathbb{R}^n$, but the error is nonetheless measured with respect to the ℓ_2 norm, i.e. "globally" over the entire length of the signal.

minimax rate for the ℓ_2 error in estimating the true monotone density $f(t)$—the same rate as for the isotonic regression problem. The pointwise i.e. ℓ_∞ error has also been studied—Durot et al. [40] shows that, for Lipschitz and bounded densities on $[0,1]$, asymptotically the error rate for estimating $f(t)$ scales as $(n/\log(n))^{-1/3}$, uniformly over all t bounded away from the endpoints. Adaptive convergence rates are studied by Cator [30]. Later we will show that our results yield a non-asymptotic error bound for this problem as well, which matches this known rate.

Several related problems for isotonic regression have also been studied. First, assuming the model $y_i = x_i + \sigma \cdot \varepsilon_i$ for standard normal error terms ε_i, estimating σ has been studied by Meyer and Woodroofe [23], among others. Estimators of σ for general distribution of ε_i are also available, see Rice [41], Gasser et al. [42]. We discuss the relevance of these tools for constructing our confidence bands in Section 3.3. Second, we can hope that our estimator $\mathsf{iso}(y)$ can recover x accurately only if x itself is monotone (or approximately monotone); thus, testing this hypothesis is important for knowing whether our confidence band can be expected to cover x itself or only its best monotone approximation, $\mathsf{iso}(x)$. Drton and Klivans [43] study the problem of testing the null hypothesis $x \in \mathscr{K}_{\mathsf{iso}}$ (or more generally, whether the signal x belongs to some arbitrary pre-specified cone $\mathscr{K}$), based on the volumes of lower-dimensional faces of the cone (see Drton and Klivans [43, Theorem 2 and Section 3]).

Recently, there is some work that considers isotonic regression beyond the univariate case. Chatterjee et al. [44] studies isotonic regression in dimension $d = 2$ and they show that the least square estimator is nearly rate minimax up to a logarithmic factor for a wide range of scenarios under Gaussian noise. Han et al. [45] extends the results of Chatterjee et al. [44] to the $d > 2$ case and also proves parallel results for random designs. Deng and Zhang [46] considers block estimators for isotonic regression on a general graph and develops ℓ_q risk bounds for such estimators.

3.2 Contraction results

We will establish the contraction results and define a new norm called "sliding window norm".

3.2.1 Contractions under isotonic projection

In this section, we examine the contractive behavior of the isotonic projection,

$$\mathsf{iso}(x) = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ \|x-y\|_2 : y_1 \leq \cdots \leq y_n \right\},$$

with respect to various norms on $\mathbb{R}^n$. Since this operator projects x onto a convex set (the cone $\mathscr{K}_{\mathsf{iso}}$ of all ordered vectors), it is trivially true that

$$\|\mathsf{iso}(x) - \mathsf{iso}(y)\|_2 \leq \|x-y\|_2,$$

but we may ask whether the same property holds when we consider norms other than the ℓ_2 norm.

Formally, we defined our question as follows:

Definition 3.2.1. For a seminorm $\|\cdot\|$ on $\mathbb{R}^n$, we say that isotonic projection is contractive with respect to $\|\cdot\|$ if

$$\|\mathsf{iso}(x) - \mathsf{iso}(y)\| \leq \|x-y\| \text{ for all } x, y \in \mathbb{R}^n.$$

We recall that a seminorm must satisfy a scaling law, $\|c \cdot x\| = |c| \cdot \|x\|$, and the triangle inequality, $\|x+y\| \leq \|x\| + \|y\|$, but may have $\|x\| = 0$ even if $x \neq 0$. From this point on, for simplicity, we will simply say "norm" to refer to any seminorm.

For which types of norms can we expect this contraction property to hold? To answer this question, we first define a simple property to help our analysis:

Definition 3.2.2. For a norm $\|\cdot\|$ on $\mathbb{R}^n$, we say that $\|\cdot\|$ is nonincreasing under neighbor averaging (NUNA) if

$$\left\| \left(x_1, \ldots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, \ldots, x_n \right) \right\| \leq \|x\|$$

for all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n-1$.

Our first main result proves that the NUNA property exactly characterizes the contractive behavior of isotonic projection—NUNA is both necessary and sufficient for isotonic projection to be

contractive.

Theorem 3.2.1. *For any norm $\|\cdot\|$ on $\mathbb{R}^n$, isotonic projection is contractive with respect to $\|\cdot\|$ if and only if $\|\cdot\|$ is nonincreasing under neighbor averaging (NUNA).*

In particular, this theorem allows us to easily prove that isotonic projection is contractive with respect to the ℓ_p norm for any $p \in [1,\infty]$, and more generally as well, via the following lemma:

Lemma 3.2.1. *Suppose that $\|\cdot\|$ is a norm that is invariant to permutations of the entries of the vector, that is, for any $x \in \mathbb{R}^n$ and any permutation π on $\{1,\dots,n\}$,*

$$\|x\| = \|x_\pi\| \text{ where } x_\pi := (x_{\pi(1)},\dots,x_{\pi(n)}).$$

(In particular, the ℓ_p norm, for any $p \in [1,\infty]$, satisfies this property.) Then $\|\cdot\|$ satisfies the NUNA property, and therefore isotonic projection is a contraction with respect to $\|\cdot\|$.

Proof of Lemma 3.2.1. Let π swap indices i and $i+1$, so that

$$x_\pi = (x_1,\dots,x_{i-1},x_{i+1},x_i,x_{i+2},\dots,x_n).$$

Then

$$\left\|\left(x_1,\dots,x_{i-1},\frac{x_i+x_{i+1}}{2},\frac{x_i+x_{i+1}}{2},x_{i+2},\dots,x_n\right)\right\| = \left\|\frac{x+x_\pi}{2}\right\| \le \frac{1}{2}\left(\|x\|+\|x_\pi\|\right) = \|x\|,$$

where we apply the triangle inequality, and the assumption that $\|x_\pi\| = \|x\|$. This proves that $\|\cdot\|$ satisfies NUNA. By Theorem 3.2.1, this implies that isotonic projection is contractive with respect to $\|\cdot\|$. □

3.2.2 The sliding window norm

We now introduce a *sliding window* norm, which will later be a useful tool for obtaining uniform convergence guarantees for isotonic regression. For any pair of indices $1 \le i \le j \le n$, we write $i:j$

to denote the stretch of $j-i+1$ many coordinates indexed by $\{i,\dots,j\}$,

$$x = (x_1,\dots,x_{i-1},\underbrace{x_i,\dots,x_j}_{\text{window } i:j},x_{j+1},\dots,x_n).$$

Fix any function

$$\psi : \{1,\dots,n\} \to \mathbb{R}_+ \text{ such that } \psi \text{ is nondecreasing and } i \mapsto i/\psi(i) \text{ is concave.} \tag{3.1}$$

The sliding window norm is defined as

$$\|x\|_\psi^{\mathsf{SW}} = \max_{1\le i\le j\le n} \Big\{ \big|\bar{x}_{i:j}\big| \cdot \psi(j-i+1) \Big\},$$

where $\bar{x}_{i:j} = \frac{x_i+\cdots+x_j}{j-i+1}$ denotes the average over the window $i:j$.

The following key lemma proves that our contraction theorem, Theorem 3.2.1, can be applied to this sliding window norm.

Lemma 3.2.2. *For any function ψ satisfying the conditions* (3.1)*, the sliding window norm $\|\cdot\|_\psi^{\mathsf{SW}}$ satisfies the NUNA property, and therefore, isotonic projection is contractive with respect to this norm.*

This lemma is a key ingredient to our convergence analyses for isotonic regression. It will allow us to use the sliding window norm to understand the behavior of $\mathsf{iso}(y)$ as an estimator of $\mathsf{iso}(x)$, where y is a vector of noisy observations of some target signal x. In particular, we will consider the special case of subgaussian noise [3] . The following lemma can be proved with a very basic union bound argument:

Lemma 3.2.3. *Let $x \in \mathbb{R}^n$ be a fixed vector, and let $y_i = x_i + \sigma\varepsilon_i$, where the ε_i's are independent,*

3. We call a random variable X subgaussian if $\mathbb{P}\{|X-\mathbb{E}[X]| \ge t\} \le 2\exp(-t^2/2)$ for any $t > 0$.

zero-mean, and subgaussian. Then taking $\psi(i) = \sqrt{i}$*, we have*

$$\mathbb{E}\left[\|x-y\|_\psi^{\mathsf{SW}}\right] \le \sqrt{2\sigma^2\log(n^2+n)} \text{ and } \mathbb{E}\left[\left(\|x-y\|_\psi^{\mathsf{SW}}\right)^2\right] \le 8\sigma^2\log(n^2+n),$$

and for any $\delta > 0$*,*

$$\mathbb{P}\left\{\|x-y\|_\psi^{\mathsf{SW}} \le \sqrt{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}\right\} \ge 1-\delta.$$

As a specific example, in a Bernoulli model, if the signal is given by $x \in [0,1]^n$ and our observations are given by $y_i \sim \text{Bernoulli}(x_i)$ (each drawn independently), then this model satisfies the subgaussian noise model with $\sigma = 1$.

3.3 Convergence rates and estimation bands

In this section, we will develop a range of results bounding our estimation error when we observe a (nearly) monotone signal plus noise. These results will all use the sliding window contraction result in Lemma 3.2.2 as the main ingredient in our analysis.

3.3.1 A deterministic result

We begin with a deterministic statement that is a straightforward consequence of the sliding window contraction result:

Theorem 3.3.1. *For any* $x, y \in \mathbb{R}^n$*, for all indices* $k = 1, \dots, n$*,*

$$\max_{1\le m\le k}\left\{\overline{\mathsf{iso}(y)}_{(k-m+1):k} - \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\} \le \mathsf{iso}(x)_k \le \min_{1\le m\le n-k+1}\left\{\overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\} \tag{3.2}$$

and

$$\max_{1\le m\le k}\left\{\overline{\mathsf{iso}(x)}_{(k-m+1):k} - \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\} \le \mathsf{iso}(y)_k \le \min_{1\le m\le n-k+1}\left\{\overline{\mathsf{iso}(x)}_{k:(k+m-1)} + \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\}. \tag{3.3}$$

Note that these two statements are symmetric; they are identical up to reversing the roles of x and y.

Proof of Theorem 3.3.1. We have $\mathsf{iso}(x)_k \geq \overline{\mathsf{iso}(x)}_{(k-m+1):k} \geq \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}$, where the first inequality uses the monotonicity of $\mathsf{iso}(x)$ while the second uses the definition of the sliding window norm along with the fact that $\|\mathsf{iso}(x) - \mathsf{iso}(y)\|_\psi^{\mathsf{SW}} \leq \|x-y\|_\psi^{\mathsf{SW}}$ by Lemma 3.2.2. This proves the lower bound for (3.2); the upper bound, and the symmetric result (3.3), are proved analogously. □

This simple reformulation of our contraction result, in fact forms the backbone of all our estimation band guarantees.

These bounds bound the difference between $\mathsf{iso}(x)$ and $\mathsf{iso}(y)$, computed using either y (as in (3.2)) or x (as in (3.3)). Thus far, the two results are entirely symmetrical—they are the same if we swap the vectors x and y.

We will next study the statistical setting where we aim to estimate a signal x based on noisy observations y, in which case the vectors x and y play distinct roles, and so the two versions of the bands will carry entirely different meanings. Before proceeding, we note that the above bounds cannot give results on x itself, but only on its projection $\mathsf{iso}(x)$. If x is far from monotonic, we cannot hope that the monotonic vector $\mathsf{iso}(y)$ would give a good estimate of x. We will consider a relaxed monotonicity constraint: we say that $x \in \mathbb{R}^n$ is $\varepsilon_{\mathsf{iso}}$-monotone if

$$x_i \leq x_j + \varepsilon_{\mathsf{iso}} \text{ for all } 1 \leq i \leq j \leq n.$$

(If x is monotonic then we can simply set $\varepsilon_{\mathsf{iso}} = 0$.) We find that $\varepsilon_{\mathsf{iso}}$ corresponds roughly to the ℓ_∞ distance between x and its isotonic projection $\mathsf{iso}(x)$:

Lemma 3.3.1. *For any $x \in \mathbb{R}^n$ that is $\varepsilon_{\mathsf{iso}}$-monotone,*

$$\|x - \mathsf{iso}(x)\|_\infty \leq \varepsilon_{\mathsf{iso}}.$$

Conversely, any $x \in \mathbb{R}^n$ with $\|x - \mathsf{iso}(x)\|_\infty \le \varepsilon$ must be (2ε)-monotone.

With this in place, we turn to our results for the statistical setting.

3.3.2 Statistical setting

We will consider a subgaussian noise model, where $x \in \mathbb{R}^n$ is a fixed signal, and the observation vector y is generated as

$$y_i = x_i + \sigma\varepsilon_i, \text{ where the } \varepsilon_i\text{'s are independent, zero-mean, and subgaussian.} \tag{3.4}$$

Lemma 3.2.3 proves that, in this case, setting $\psi(m) = \sqrt{m}$ would yield $\|x - y\|_\psi^{\mathsf{SW}} \le \sqrt{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}$ with probability at least $1 - \delta$. Of course, we could consider other models as well, e.g. involving correlated noise or heavy-tailed noise, but restrict our attention to this simple model for the sake of giving an intuitive illustration of our results.

In order for this bound on the sliding window to be useful in practice, we need to obtain a bound or an estimate for the noise level σ. Under the Bernoulli model $y_i \sim \text{Bernoulli}(x_i)$, we can simply set $\sigma = 1$. More generally, it may be possible to estimate σ from the data itself, for instance if the noise terms ε_i are i.i.d. standard normal, Meyer and Woodroofe [23] propose estimating the noise level σ with the maximum likelihood estimator (MLE), $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - \mathsf{iso}(y)_i)^2$, or the bias-corrected MLE given by

$$\widehat{\sigma}^2 = \frac{\sum_i (y_i - \mathsf{iso}(y)_i)^2}{n - c_1 \cdot \mathsf{df}\big(\mathsf{iso}(y)\big)},$$

where c_1 is a known constant while $\mathsf{df}\big(\mathsf{iso}(y)\big)$ is the number of "degrees of freedom" in the monotone vector $\mathsf{iso}(y)$, i.e. the number of distinct values in this vector.

We next consider the two different types of statistical guarantees that can be obtained, using the two symmetric formulations in Theorem 3.3.1 above.

3.3.3 Data-adaptive bands

We first consider the problem of providing a confidence band for the signal x in a practical setting, where we can only observe the noisy data y and do not have access to other information. In this setting, the bound (3.2) in Theorem 3.3.1, combined with Lemma 3.2.3's bound on $\|x-y\|_\psi^{\mathsf{SW}}$ for the subgaussian model, yields the following result:

Theorem 3.3.2. *For any signal $x \in \mathbb{R}^n$ and any $\delta > 0$, under the subgaussian noise model* (3.4)*, then with probability at least $1-\delta$, for all $k = 1, \dots, n$,*

$$\begin{aligned} &\max_{1 \le m \le k} \left\{ \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} \\ &\le \mathsf{iso}(x)_k \le \min_{1 \le m \le n-k+1} \left\{ \overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\}. \end{aligned} \tag{3.5}$$

If additionally x is $\varepsilon_{\mathsf{iso}}$-monotone, then we also have

$$\begin{aligned} &\max_{1 \le m \le k} \left\{ \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} - \varepsilon_{\mathsf{iso}} \\ &\le x_k \le \min_{1 \le m \le n-k+1} \left\{ \overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} + \varepsilon_{\mathsf{iso}}. \end{aligned} \tag{3.6}$$

We emphasize that these bounds give us a uniform confidence band for $\mathsf{iso}(x)$ (or for x itself, if it is monotone) that can be computed without assuming anything about the properties of the signal; for instance, we do not assume that the signal is Lipschitz with some known constant, or anything of this sort. We only need to know the noise level σ, which can be estimated as discussed in Section 3.3.2. In this sense, the bounds are data-adaptive—they are computed using the observed projection $\mathsf{iso}(y)$, and adapt to the properties of the signal (for instance, if x is locally constant near

k, then the upper and lower confidence bounds will be closer together).

Comparison to existing work The flavor of our data-adaptive band is close to that given in Dümbgen [31], where the author gives confidence bands for signals in a continuous Gaussian white noise model. Although in Section 5 of Dümbgen [31] the result is applied to the discrete case, the confidence band there is only valid asymptotically as pointed our by the author, whereas our band is valid for finite samples. Moreover, the computation of the band in Dümbgen [31] involves Monte Carlo simulation to estimate several key quantiles, and hence is much heavier than the computation of our band. Another difference is that Dümbgen [31] employes kernel estimators in their bands while we use the isotonic least squares estimator in our construction.

3.3.4 Convergence rates

While the results of Theorem 3.3.2 give data-adaptive bounds that do not depend on properties of x, from a theoretical point of view we would also like to understand how the estimation error depends on these properties. For the data-adaptive bands, we used the result (3.2) relating $\mathsf{iso}(x)$ and $\mathsf{iso}(y)$, but for this question, we will use the symmetric result (3.3) instead, which immediately yields the following theorem.

Theorem 3.3.3. *For any signal $x \in \mathbb{R}^n$ and any $\delta > 0$, under the subgaussian noise model* (3.4)*, then with probability at least $1-\delta$, for all $k = 1, \ldots, n$,*

$$-\min_{1\le m\le k}\left\{\left(\mathsf{iso}(x)_k - \overline{\mathsf{iso}(x)}_{(k-m+1):k}\right) + \sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}\right\}$$

$$\le \mathsf{iso}(y)_k - \mathsf{iso}(x)_k \le \min_{1\le m\le n-k+1}\left\{\left(\overline{\mathsf{iso}(x)}_{k:(k+m-1)} - \mathsf{iso}(x)_k\right) + \sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}\right\}. \quad (3.7)$$

If additionally x is $\varepsilon_{\mathsf{iso}}$-monotone, then we also have

$$-\min_{1\le m\le k}\left\{\left(x_k-\overline{x}_{(k-m+1):k}\right)+\sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}\right\}-\varepsilon_{\mathsf{iso}}$$
$$\le \mathsf{iso}(y)_k-x_k\le \min_{1\le m\le n-k+1}\left\{\left(\overline{x}_{k:(k+m-1)}-x_k\right)+\sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}\right\}+\varepsilon_{\mathsf{iso}}. \quad (3.8)$$

Proof of Theorem 3.3.3. For the first bound (3.7), we simply subtract $\mathsf{iso}(x)_k$ from the inequalities (3.3). For the second bound (3.8) in the case that x is approximately monotone, we instead subtract x_k from (3.3), and also use the fact that $\|x-\mathsf{iso}(x)\|_\infty\le\varepsilon_{\mathsf{iso}}$ by Lemma 3.3.1, which implies that $\left|\overline{x}_{k:(k+m-1)}-\overline{\mathsf{iso}(x)}_{k:(k+m-1)}\right|\le\varepsilon_{\mathsf{iso}}$, and similarly $\left|\overline{x}_{(k-m+1):k}-\overline{\mathsf{iso}(x)}_{(k-m+1):k}\right|\le\varepsilon_{\mathsf{iso}}$. □

Comparison to existing work In the monotone setting (i.e. $x=\mathsf{iso}(x)$), Chatterjee et al. [26] derive related results bounding the pointwise error $\left|x_k-\mathsf{iso}(y)_k\right|$. Specifically, they use the "minmax" formulation of the isotonic projection, $\mathsf{iso}(y)_k=\min_{j\ge k}\max_{i\le k}\overline{y}_{i:j}$, and give the following argument:

$$\begin{aligned}\mathsf{iso}(y)_k-x_k&=\min_{1\le m\le n-k+1}\max_{i\le k}\overline{y}_{i:(k+m-1)}-x_k\\&\le\min_{1\le m\le n-k+1}\left\{\left(\max_{i\le k}\overline{x}_{i:(k+m-1)}-x_k\right)+\max_{i\le k}\left|\overline{x}_{i:(k+m-1)}-\overline{y}_{i:(k+m-1)}\right|\right\}\\&\le\min_{1\le m\le n-k+1}\left\{\left(\overline{x}_{k:(k+m-1)}-x_k\right)+\underbrace{\max_{i\le k}\left|\overline{x}_{i:(k+m-1)}-\overline{y}_{i:(k+m-1)}\right|}_{\text{(Err)}}\right\},\end{aligned}$$

where the first step defines $m=j-k+1$ and uses the "minmax" formulation, while the third uses the assumption that x is monotone. They then bound the error term (Err) in expectation. We can instead bound it as (Err) $\le\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\sqrt{m}}$, which is exactly the same as the upper bound in our result (3.8). Their "minmax" strategy can analogously be used to obtain the corresponding lower

bound as well.

3.3.5 *Locally constant and locally Lipschitz signals*

If the signal x is monotone, Chatterjee et al. [26]'s results, which are analogous to our bounds in (3.8), yield implications for many different classes of signals: for instance, they show that for a piecewise constant signal x taking only s many unique values, the ℓ_2 error scales as

$$\frac{1}{n}\|x - \mathsf{iso}(y)\|_2^2 \leq \frac{16s\sigma^2}{n}\log\left(\frac{en}{s}\right).$$

We therefore see that

$$\left|x_k - \mathsf{iso}(y)_k\right| \lesssim \sqrt{\frac{log(n)}{n}} \tag{3.9}$$

for "most" indices k when the signal is piecewise constant.

We can instead consider a Lipschitz signal: we say that x is L-Lipschitz if $|x_i - x_{i+1}| \leq L/n$ for all i. (Rescaling by n is natural as we often think of $x_i = f(i/n)$ for some underlying function f). In this setting, our results in Theorem 3.3.3 immediately yield the bound

$$\left|x_k - \mathsf{iso}(y)_k\right| \leq \min_{1\leq m\leq k\wedge(n-k+1)} \left\{ \frac{L(m-1)}{2n} + \sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\}, \tag{3.10}$$

where the term $\frac{L(m-1)}{2n}$ is a bound on $\left(\bar{x}_{k:(k+m-1)} - x_k\right)$ and $\left(x_k - \bar{x}_{(k-m+1):k}\right)$ when x is L-Lipschitz. It's easy to see that the optimal scaling is achieved by taking $m = \left\lceil \left(\frac{n\sqrt{\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}}{L} \right)^{2/3} \right\rceil$, in which case we obtain the bound

$$\left|x_k - \mathsf{iso}(y)_k\right| \leq 2\sqrt[3]{\frac{L\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{n}} \tag{3.11}$$

for all $m \leq k \leq n-m+1$. (For indices k nearer to the endpoints, we are forced to choose a smaller

m, and the scaling will be worse.)

We can also compute convergence rates in a more general setting, where the signal x is locally Lipschitz—its behavior may vary across different regions of the signal. As discussed in Section 3.1, many papers in the literature consider asymptotic local convergence rates—local in the sense of giving *pointwise* error bounds, which for a single signal $x = (x_1, \dots, x_n)$, may be larger for indices i falling within a region of the signal that is strictly increasing, and smaller for indices i falling into a locally flat region. We would hope to see some interpolation between the $n^{-1/3}$ rate expected for a strictly increasing stretch of the signal, as in (3.11), versus the improved parametric rate of $n^{-1/2}$ in a locally constant region as in (3.9).

Our confidence bands can also be viewed as providing error bounds that are local in this sense, i.e. that adapt to the local behavior of the signal x as we move from index $i = 1$ to $i = n$. To make this more precise, we will show how our bounds scale locally with the sample size n to obtain the $n^{-1/3}$ and $n^{-1/2}$ rates described above. Consider any monotone signal x. Suppose the signal x is locally constant near k, with $x_{k-cn+1} = \cdots = x_k = \cdots = x_{k+cn-1}$ for some positive constant $c > 0$. Then our bound (3.8) applied with $m = cn$ yields

$$\left|x_k - \mathsf{iso}(y)_k\right| \lesssim \sqrt{\frac{\sigma^2 \log(n)}{n}}. \tag{3.12}$$

For other indices, however, where the signal is locally strictly increasing with a Lipschitz constant L, then taking $m \sim \left(\frac{\sigma^2 n \log(n)}{L}\right)^{2/3}$ yields the $n^{-1/3}$ scaling obtained above in (3.11). It is of course also possible to achieve an interpolation between the $n^{-1/2}$ and $n^{-1/3}$ rates via our results, as well.

Many works in the literature consider the local adaptivity problem in an asymptotic setting; here we will describe the results of Cator [30]. Consider an asymptotic setting where the signal $x = (x_1, \dots, x_n)$ comes from measuring (at n many points) a monotone function $f : [0,1] \to \mathbb{R}$, and we are interested in the local convergence rate at some fixed $t \in (0,1)$. Cator [30] show that if the first α derivatives of f at t satisfy $f^{(1)}(t) = \cdots = f^{(\alpha-1)}(t) = 0$ and $f^{(\alpha)}(t) > 0$, the convergence

rate for estimating $f(t)$ scales as $n^{-\alpha/(2\alpha+1)}$. In particular, if $\alpha = 1$ (f is strictly increasing at t) then they obtain the $n^{-1/3}$ rate seen before, while if $\alpha = \infty$ (f is locally constant near t) then they obtain the faster parametric $n^{-1/2}$ rate. Of course, any α in between 1 and ∞ will produce some power of n between these two. Our work can be viewed as a finite-sample version of these types of results.

3.3.6 Convergence rates in the ℓ_2 norm

We next show that the tools developed in this paper can be used to yield a bound on the ℓ_2 error, achieving the same $n^{-1/3}$ scaling as in Chatterjee et al. [26]. While achieving an $n^{-1/3}$ elementwise requires a Lipschitz condition on the signal (as in our result (3.11) above), here we do not assume any Lipschitz conditions and require only a bound on the total variation,

$$V := \mathsf{iso}(x)_n - \mathsf{iso}(x)_1.$$

Our proof uses similar techniques as Chatterjee et al. [26]'s result.

Theorem 3.3.4. *For any signal $x \in \mathbb{R}^n$, under the subgaussian noise model* (3.4)*, we have*

$$\frac{1}{n}\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 \leq 48\left(\frac{V\sigma^2\log(2n)}{n}\right)^{2/3} + \frac{96\sigma^2\log^2(2n)}{n}.$$

As long as $n \gg \frac{\sigma^2\log^4(2n)}{V^2}$, the first term is the dominant one, matching the result of Chatterjee et al. [26, Theorem 4.1] with a slight improvement in the log term. (The constants in this result are of course far from optimal.)

3.4 Density Estimation

As a second application of the tools developed for the sliding window norm, we consider the problem of estimating a monotone nonincreasing density g on the interval $[0,1]$, using n samples

drawn from this density.

Let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} g$ be n samples drawn from the target density f, sorted into an ordered list $Z_{(1)} \leq \cdots \leq Z_{(n)}$. The Grenander estimator for the monotone density g is defined as follows. Let $\widehat{G}_n$ be the empirical cumulative distribution function for this sample,

$$\widehat{G}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\},$$

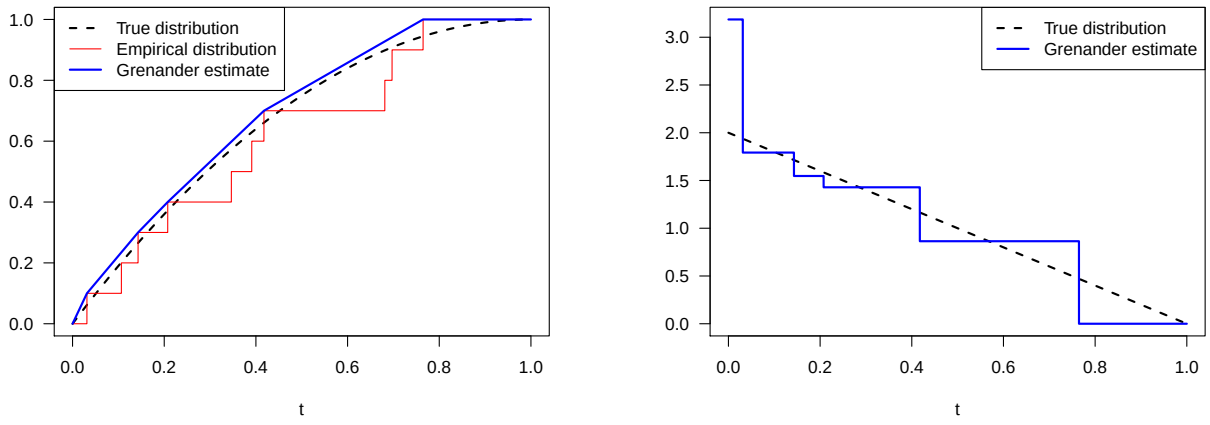
and let $\widehat{G}_{\mathsf{Gren}}$ be the minimal concave upper bound on $\widehat{G}_n$. Finally, define the Grenander estimator of the density, denoted by $\widehat{g}_{\mathsf{Gren}}$, as the left-continuous piecewise constant first derivative of $\widehat{G}_{\mathsf{Gren}}$. This process is illustrated in Figure 3.1. It is known (Robertson et al. [19]) that $\widehat{g}_{\mathsf{Gren}}$ can be computed with a simple isotonic projection of a sequence. Namely, for $i = 1, \ldots, n$, let $y_i = n(Z_{(i)} - Z_{(i-1)})$ where we set $Z_{(0)} \coloneqq 0$, and calculate the isotonic projection $\mathsf{iso}(y)$. Then the Grenander estimator is given by

$$\widehat{g}_{\mathsf{Gren}} = \begin{cases} 1/\mathsf{iso}(y)_1, & 0 \leq t \leq Z_{(1)}, \\ 1/\mathsf{iso}(y)_2, & Z_{(1)} < t \leq Z_{(2)}, \\ \ldots & \\ 1/\mathsf{iso}(y)_n, & Z_{(n-1)} < t \leq Z_{(n)}, \\ 0, & Z_{(n)} < t \leq 1. \end{cases} \tag{3.13}$$

If we assume that f is Lipschitz and lower-bounded, then our error bounds for isotonic regression transfer easily into this setting, yielding the following theorem:

Theorem 3.4.1. *Let $g : [0,1] \to [c, \infty)$ be a nonincreasing L-Lipschitz density, let $Z_1, \ldots, Z_n$ be i.i.d. draws from g, and define the Grenander estimator $\widehat{g}_{\mathsf{Gren}}$ as in* (3.13)*. Then for any $\delta > 0$, if*

$$\Delta \coloneqq 9 \left(\frac{1}{c} + \frac{L}{2c^3} \right) \sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \leq \frac{1}{c+L},$$



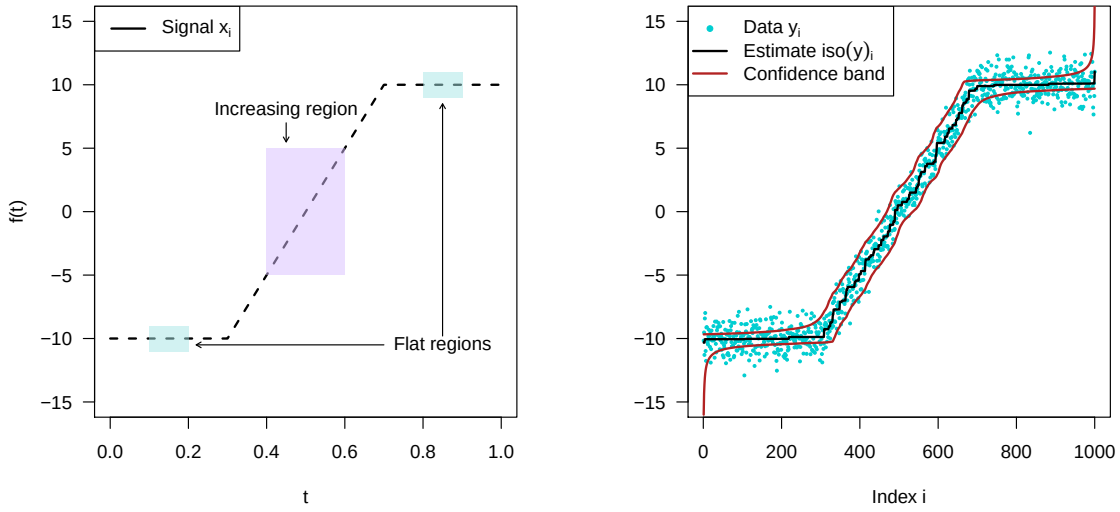
(a) Cumulative distribution function (b) Density

Figure 3.1: Illustration of the Grenander estimator for a monotone decreasing density.

then

$$\mathbb{P}\left\{\sup_{\Delta \le t \le 1-\Delta} \left|g(t) - \widehat{g}_{\mathsf{Gren}}(t)\right| \le \frac{\Delta}{\frac{1}{c+L} \cdot \left(\frac{1}{c+L} - \Delta\right)}\right\} \ge 1 - \delta.$$

This result is similar to that of Durot et al. [40], which also obtains a $(n/\log(n))^{-1/3}$ convergence rate uniformly over t (although in their work, t is allowed to be slightly closer to the endpoints, by a log factor). Their results are asymptotic, while our work gives a finite-sample guarantee. As mentioned in Section 3.1, Cator [30] also derives locally adaptive error bounds whose scaling depends on the local Lipschitz behavior or local derivatives of f. Our locally adaptive results for sequences may also be applied here to obtain a locally adaptive confidence band on the density g, but we do not give details here.



(a) Signal (b) Results for $n = 1000$

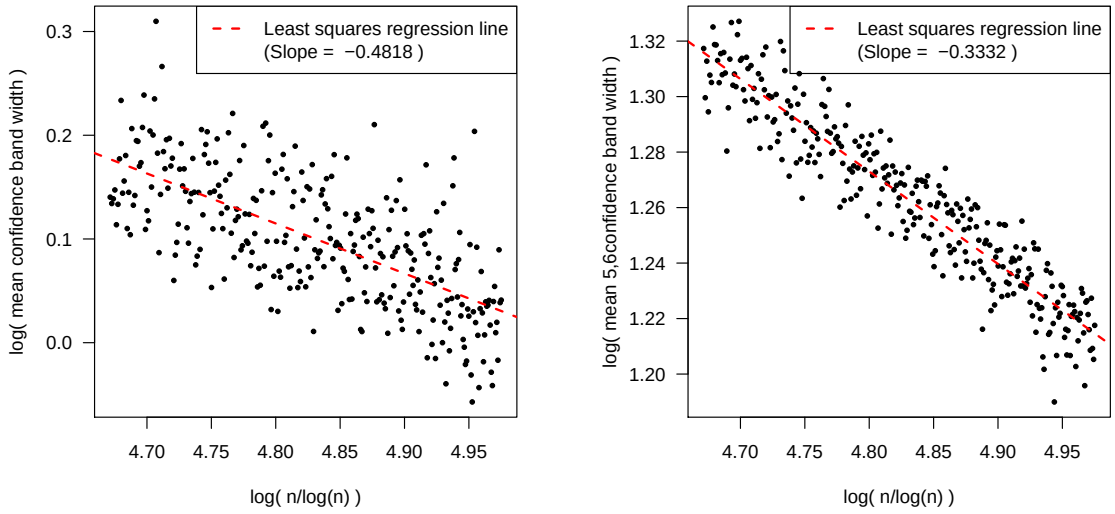
Figure 3.2: (a) The function $f(t)$ used to generate signals $x \in \mathbb{R}^n$ for various n, with flat and increasing regions highlighted. (b) At sample size $n = 1000$, the observed data y, estimated signal $\mathsf{iso}(y)$, and data-adaptive confidence band computed as in (3.5).

3.5 Numerical Study

In this section, we run a simple simulation to demonstrate the local adaptivity of our confidence bands. The signal is generated from an underlying function $f(t)$ defined over $t \in [0,1]$, with

$$f(t) = \begin{cases} -10, & 0 \le t \le 0.3, \\ \text{linearly increasing from } -10 \text{ to } 10, & 0.3 \le t \le 0.7, \\ 10, & 0.7 \le t \le 1, \end{cases}$$

as illustrated in Figure 3.2(a). For a fixed sample size n, we set $x_i = f\left(\frac{i}{n+1}\right)$ and $y_i = x_i + N(0,1)$. We then compute a data-adaptive confidence band as given in (3.6), with known noise level $\sigma = 1$, with target coverage level $1-\delta = 0.9$, and with $\varepsilon_{\mathsf{iso}} = 0$ as the signal x is known to be monotone. For sample size $n = 1000$, the resulting estimate $\mathsf{iso}(y)$ and confidence band are illustrated in Figure 3.2(b).



(a) Results for flat regions

(b) Results for increasing region

Figure 3.3: For each sample size $700 \leq n \leq 1000$, log mean width of the confidence band over a region. (a) Flat region: $t \in [0.1, 0.2] \cup [0.8, 0.9]$, where slope $\approx -1/2$, i.e. pointwise error scales as $(n/\log(n))^{-1/2}$, as predicted in (3.12). (b) Increasing region: $t \in [0.4, 0.6]$, where slope $\approx -1/3$, as predicted in (3.11).

We then repeat this experiment at sample sizes $n = 700, 701, 702, \ldots, 1000$. For each sample size n, we take the mean width of the confidence band averaged over (a) the locally constant ("flat") regions of the signal, defined by all indices i corresponding to values $t \in [0.1, 0.2] \cup [0.8, 0.9]$, and (b) a strictly increasing region, defined by indices i corresponding to $t \in [0.4, 0.6]$. (These regions are illustrated in Figure 3.2(a).)

Our theory predicts that the mean confidence band width scales as $\sim \sqrt{\frac{\log(n)}{n}}$ in the flat regions, and $\sim \sqrt[3]{\frac{\log(n)}{n}}$ in the increasing region. To test this, we take a linear regression of the log of the mean confidence band width against $\log\left(\frac{n}{\log(n)}\right)$, and find a slope $\approx -1/2$ in the flat regions and $\approx -1/3$ in the increasing region, confirming our theory. These results are illustrated in Figure 3.3.

Note that our data-adaptive estimation bands given by Theorem 3.3.2 are calculated without using prior knowledge of the signal's local behavior (locally constant / locally Lipschitz)—the confidence bands computed in Theorem 3.3.2 are able to adapt to this unknown structure automatically.

We next check the empirical coverage level of these confidence bands. Ideally we would want to see that, over repeated simulations, the true monotone sequence $x = (x_1, \ldots, x_n)$ lies entirely in the band roughly $1 - \delta = 90\%$ of the time. While our theory guarantees that coverage will hold with probability *at least* 90%, our bounds are of course somewhat conservative. We observe empirically that the coverage is in fact too high—it is essentially 100%—but nonetheless, the width of the confidence band is not too conservative. In particular, shrinking the width of the confidence band by a factor of ≈ 0.855 empirically leads to achieving the target 90% coverage level; in other words, our confidence bands are around 17% too wide. (Of course, this ratio is specific to our choice of data distribution, and is likely to vary in different settings.)

3.6 Proofs of Theorems

3.6.1 Proof for contractive isotonic projection

In this section, we prove our main result Theorem 3.2.1 showing that, for any semi-norm, the nonincreasing-under-neighbor-averaging (NUNA) property is necessary and sufficient for isotonic projection to be contractive under this semi-norm.

Before proving the theorem, we introduce a few definitions. First, for any index $i = 1, \ldots, n-1$, we define the matrix

$$A_i = \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & \mathbf{I}_{n-i-1} \end{pmatrix} \in \mathbb{R}^{n \times n}, \tag{3.14}$$

which averages entries i and $i+1$. That is,

$$A_i x = \left(x_1, \ldots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, \ldots, x_n \right).$$

We also define an algorithm for isotonic projection that differs from PAVA, and in fact does not

converge in finite time, but is useful for the purpose of theoretical analysis. For any $x \in \mathbb{R}^n$ and any index $i = 1, \ldots, n-1$, define

$$\mathsf{iso}_i(x) = \begin{cases} x, & \text{if } x_i \leq x_{i+1}, \\ A_i x, & \text{if } x_i > x_{i+1}. \end{cases}$$

In other words, if neighboring entries i and $i+1$ violate the monotonicity constraint, then we average them. The following lemma shows that, by iterating this step infinitely many times (while cycling through the indices $i = 1, \ldots, n-1$), we converge to the isotonic projection of x.

Lemma 3.6.1. *Fix any $x = x^{(0)} \in \mathbb{R}^n$, and define*

$$x^{(t)} = \mathsf{iso}_{i_t}(x^{(t-1)}) \text{ where } i_t = 1 + \mathrm{mod}(t-1, n-1) \text{ for } t = 1, 2, 3, \ldots. \tag{3.15}$$

Then

$$\lim_{t \to \infty} x^{(t)} = \mathsf{iso}(x).$$

With this slow projection algorithm in place, we turn to the proof of our theorem.

Proof of Theorem 3.2.1. First suppose that $\|\cdot\|$ satisfies the NUNA property. We will prove that, for any $x, y \in \mathbb{R}^n$ and any index $i = 1, \ldots, n-1$,

$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| \leq \|x - y\|. \tag{3.16}$$

If this is true, then by Lemma 3.6.1, this is sufficient to see that isotonic projection is contractive with respect to $\|\cdot\|$, since the map $x \mapsto \mathsf{iso}(x)$ is just a composition of (infinitely many) steps of the form $x \mapsto \mathsf{iso}_i(x)$. More concretely, defining $x^{(t)}$ and $y^{(t)}$ as in Lemma 3.6.1, (3.16) proves that $\|x^{(t)} - y^{(t)}\| \leq \|x^{(t-1)} - y^{(t-1)}\|$ for each $t \geq 1$. Applying this inductively proves that $\|x^{(t)} - y^{(t)}\| \leq \|x - y\|$ for all $t \geq 1$, and then taking the limit as $t \to \infty$, we obtain $\|\mathsf{iso}(x) - \mathsf{iso}(y)\| \leq \|x - y\|$.

Now we turn to proving (3.16). We will split into four cases.

- Case 1: $x_i \le x_{i+1}$ and $y_i \le y_{i+1}$. In this case, $\mathsf{iso}_i(x) = x$ and $\mathsf{iso}_i(y) = y$, and so trivially,

$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| = \|x - y\|.$$

- Case 2: $x_i > x_{i+1}$ and $y_i > y_{i+1}$. In this case, we have

$$\big[\mathsf{iso}_i(x)\big]_i = \big[\mathsf{iso}_i(x)\big]_{i+1} = \frac{x_i + x_{i+1}}{2}$$

  and

$$\big[\mathsf{iso}_i(y)\big]_i = \big[\mathsf{iso}_i(y)\big]_{i+1} = \frac{y_i + y_{i+1}}{2},$$

  while all entries $j \notin \{i, i+1\}$ are unchanged. Therefore, we can write

$$\mathsf{iso}_i(x) - \mathsf{iso}_i(y) = A_i \cdot (x - y).$$

  Since $\|\cdot\|$ satisfies the NUNA property, therefore,

$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| = \|A_i \cdot (x - y)\| \le \|x - y\|.$$

- Case 3: $x_i \le x_{i+1}$ and $y_i > y_{i+1}$. Let

$$t = \frac{y_i - y_{i+1}}{x_{i+1} - x_i + y_i - y_{i+1}}.$$

  Note that $t \in [0, 1]$ by the definition of this case. A trivial calculation shows that

$$\big[\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\big]_i = x_i - \frac{y_i + y_{i+1}}{2} = (1 - t/2) \cdot (x_i - y_i) + t/2 \cdot (x_{i+1} - y_{i+1})$$

and

$$\left[\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\right]_{i+1} = x_{i+1} - \frac{y_i + y_{i+1}}{2} = t/2 \cdot (x_i - y_i) + (1 - t/2) \cdot (x_{i+1} - y_{i+1})$$

This means that we have

$$\mathsf{iso}_i(x) - \mathsf{iso}_i(y) = (1-t) \cdot (x-y) + t \cdot A_i \cdot (x-y),$$

and so

$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| \leq (1-t) \cdot \|x-y\| + t \cdot \|A_i \cdot (x-y)\| \leq \|x-y\|,$$

since $\|\cdot\|$ satisfies NUNA.

- Case 4: $x_i > x_{i+1}$ and $y_i \leq y_{i+1}$. By symmetry, this is equivalent to Case 3.

This proves (3.16), and therefore, is sufficient to show that isotonic projection is a contraction with respect to $\|\cdot\|$.

Now we prove the converse. Suppose that $\|\cdot\|$ does not satisfy NUNA. Then we can find some x and some i such that

$$\|A_i x\| > \|x\|.$$

Without loss of generality we can assume $x_i \leq x_{i+1}$ (otherwise simply replace x with $-x$—since $\|\cdot\|$ is a norm, we will have $\|-A_i x\| = \|A_i x\| > \|x\| = \|-x\|$).

Let $B = \max_{1 \leq j \leq n-1} |x_j - x_{j+1}|$, and let $\Delta = x_{i+1} - x_i \in [0, B]$.

Now define

$$y = (\Delta - (i-1)B, \Delta - (i-2)B, \ldots, \Delta - B, \underbrace{\Delta}_{\text{entry } i}, \underbrace{0}_{\text{entry } i+1}, B, 2B, \ldots, (n-i-1)B),$$

and $z = y + x$. We can check that $\mathsf{iso}(z) = z$, since

$$z_{j+1} - z_j = \begin{cases} x_{j+1} - x_j + B \geq 0, & \text{if } j \neq i, \\ x_{i+1} - x_i - \Delta = 0, & \text{if } j = i. \end{cases}$$

On the other hand, using the fact that $0 \leq \Delta \leq B$, we have

$$\mathsf{iso}(y) = \left(\Delta - (i-1)B, \Delta - (i-2)B, \ldots, \Delta - B, \frac{\Delta}{2}, \frac{\Delta}{2}, B, 2B, \ldots, (n-i-1)B \right),$$

and so

$$\begin{aligned} \mathsf{iso}(z) - \mathsf{iso}(y) = (z-y) + (y - \mathsf{iso}(y)) &= x + \left(0, \ldots, 0, \frac{\Delta}{2}, -\frac{\Delta}{2}, 0, \ldots, 0 \right) \\ &= \left(x_1, x_2, \ldots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, x_{i+3}, \ldots, x_n \right) = A_i x. \end{aligned}$$

Therefore, $\|\mathsf{iso}(z) - \mathsf{iso}(y)\| > \|z - y\|$, proving that isotonic projection is not contractive with respect to $\|\cdot\|$.

□

3.6.2 Proof of ℓ_2 error rate (Theorem 3.3.4)

First we define the cube $A = [\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]^n \subset \mathbb{R}^n$, and let $z = \mathscr{P}_A(\mathsf{iso}(y))$ be the projection of $\mathsf{iso}(y)$ to this cube, which is computed by truncating each entry $\mathsf{iso}(y)_i$ to the range $[\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]$. Note that $\mathsf{iso}(x) + z$ is now a monotone vector with range given by $(\mathsf{iso}(x) + z)_n - (\mathsf{iso}(x) + z)_1 \leq 2V$. Now, fix any integer $M \geq 1$ and find integers

$$0 = k_0 < \cdots < k_M = n$$

such that

$$\left| (\mathsf{iso}(x) + z)_{k_m} - (\mathsf{iso}(x) + z)_{k_{m-1}+1} \right| \leq \frac{2V}{M} \text{ for all } m = 1, \ldots, M,$$

which we can find since the total variation of the vector $\mathsf{iso}(x)+z$ is bounded by $2V$ (Lemma 11.1 in Chatterjee et al. [26]).

For each $m=1,\dots,M$, let $I_m=\{k_{m-1}+1,\dots,k_m\}$ be the set of indices in the mth bin. We can calculate

$$\begin{aligned}
&\max_{i\in I_m}(z-\mathsf{iso}(x))_i-\min_{i\in I_m}(z-\mathsf{iso}(x))_i \\
&\qquad\le \max_{i\in I_m} z_i+\max_{i\in I_m}\mathsf{iso}(x)_i-\min_{i\in I_m} z_i-\min_{i\in I_m}\mathsf{iso}(x)_i \\
&\qquad\qquad = (\mathsf{iso}(x)+z)_{k_m}-(\mathsf{iso}(x)+z)_{k_{m-1}+1}\le \frac{2V}{M}.
\end{aligned}$$

This implies that

$$\|(z-\mathsf{iso}(x))_{I_m}-\overline{(z-\mathsf{iso}(x))}_{I_m}\cdot\mathbf{1}_{I_m}\|_2\le\frac{2V}{M}\cdot\sqrt{k_m-k_{m-1}}.$$

We also have

$$\left|\bar{z}_{I_m}-\overline{\mathsf{iso}(x)}_{I_m}\right|\le\frac{\|z-\mathsf{iso}(x)\|_\psi^{\mathsf{SW}}}{\sqrt{k_m-k_{m-1}}}$$

by our choice of the sliding window norm. Next, by the triangle inequality we have

$$\begin{aligned}
&\|z_{I_m}-\mathsf{iso}(x)_{I_m}\|_2^2 \\
&\qquad\le\left(\|\bar{z}_{I_m}\cdot\mathbf{1}_{I_m}-\overline{\mathsf{iso}(x)}_{I_m}\cdot\mathbf{1}_{I_m}\|_2+\|(z-\mathsf{iso}(x))_{I_m}-\overline{(z-\mathsf{iso}(x))}_{I_m}\cdot\mathbf{1}_{I_m}\|_2\right)^2 \\
&\qquad=\left(\left|\bar{z}_{I_m}-\overline{\mathsf{iso}(x)}_{I_m}\right|\cdot\sqrt{k_m-k_{m-1}}+\|(z-\mathsf{iso}(x))_{I_m}-\overline{(z-\mathsf{iso}(x))}_{I_m}\cdot\mathbf{1}_{I_m}\|_2\right)^2 \\
&\qquad\le\left(\|z-\mathsf{iso}(x)\|_\psi^{\mathsf{SW}}+\frac{2V}{M}\cdot\sqrt{k_m-k_{m-1}}\right)^2\le 2\big(\|z-\mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2+\frac{8V^2}{M^2}(k_m-k_{m-1}).
\end{aligned}$$

Therefore,

$$\|z-\mathsf{iso}(x)\|_2^2=\sum_{m=1}^M\|z_{I_m}-\mathsf{iso}(x)_{I_m}\|_2^2\le 2M\big(\|z-\mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2+\frac{8V^2}{M^2}\sum_{m=1}^M(k_m-k_{m-1}).$$

Since $\sum_{m=1}^{M}(k_m - k_{m-1}) = n$ trivially, we can simplify this to

$$\|z - \mathsf{iso}(x)\|_2^2 \leq 2M\big(\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2 + \frac{8V^2 n}{M^2}.$$

Now, since z is the projection of $\mathsf{iso}(y)$ to the range $[\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]$, it follows trivially that $\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}} \leq \|\mathsf{iso}(x) - \mathsf{iso}(y)\|_\psi^{\mathsf{SW}}$ and therefore is bounded by $\|x - y\|_\psi^{\mathsf{SW}}$ by our contraction result. Therefore,

$$\|z - \mathsf{iso}(x)\|_2^2 \leq 2M\big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{8V^2 n}{M^2}.$$

Next, we have

$$\begin{aligned}
\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 &\leq 2\|z - \mathsf{iso}(x)\|_2^2 + 2\|\mathsf{iso}(y) - z\|_2^2 \\
&= 2\|z - \mathsf{iso}(x)\|_2^2 + 2\sum_{i=1}^{n}(z_i - \mathsf{iso}(y)_i)_+^2 + 2\sum_{i=1}^{n}(\mathsf{iso}(y)_i - z_i)_+^2 \\
&\leq 4M\big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{16V^2 n}{M^2} + 2\sum_{i=1}^{n}(\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2 + 2\sum_{i=1}^{n}(\mathsf{iso}(y)_i - \mathsf{iso}(x)_n)_+^2.
\end{aligned}$$

It remains to bound these last two terms. First we bound $\sum_{i=1}^{n}(\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2$. If $\mathsf{iso}(y)_1 \geq \mathsf{iso}(x)_1$ then this term is zero, so now we focus on the case that $\mathsf{iso}(y)_1 < \mathsf{iso}(x)_1$. For any $1 \leq j \leq i$, we have

$$\mathsf{iso}(x)_j - \mathsf{iso}(y)_j \geq \mathsf{iso}(x)_1 - \mathsf{iso}(y)_j$$

and so

$$\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i \leq \overline{\mathsf{iso}(x)}_{1:i} - \overline{\mathsf{iso}(y)}_{1:i} \leq \frac{\|x - y\|_\psi^{\mathsf{SW}}}{\sqrt{i}}.$$

Therefore,

$$\sum_{i=1}^{n}(\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2 \leq \sum_{i=1}^{n}\left(\frac{\|x - y\|_\psi^{\mathsf{SW}}}{\sqrt{i}}\right)^2 = \big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2 \cdot 2\log(2n),$$

by bounding the harmonic series $1 + \frac{1}{2} + \cdots + \frac{1}{n}$. We also have $\sum_{i=1}^{n}(\mathsf{iso}(y)_i - \mathsf{iso}(x)_n)_+^2 \leq \big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2$.

$2\log(2n)$ by an identical argument. Combining everything, then,

$$\|\mathsf{iso}(y)-\mathsf{iso}(x)\|_2^2 \le 4M\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2+\frac{16V^2n}{M^2}+8\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2\log(2n).$$

Setting

$$M=\lceil M_0\rceil \text{ where } M_0=\frac{2V^{2/3}n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}},$$

we obtain

$$\|\mathsf{iso}(y)-\mathsf{iso}(x)\|_2^2 \le 4\left(\frac{2V^{2/3}n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}}+1\right)\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{16V^2n}{\left(\frac{2V^{2/3}n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}}\right)^2}+8\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2\log(2n).$$

After simplifying (and assuming $n\ge 2$ to avoid triviality), this bound becomes

$$\frac{1}{n}\|\mathsf{iso}(y)-\mathsf{iso}(x)\|_2^2 \le \frac{12V^{2/3}\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{4/3}}{n^{2/3}}+\frac{12\log(2n)}{n}\cdot\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2.$$

Applying Lemma 3.2.3, we have $\mathbb{E}\left[\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2\right]\le 8\sigma^2\log(2n)$ which implies that

$$\mathbb{E}\left[\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{4/3}\right]\le(8\sigma^2\log(2n))^{2/3}=4\sigma^{4/3}\log^{2/3}(2n)$$

Plugging this in, we then have

$$\mathbb{E}\left[\frac{1}{n}\|\mathsf{iso}(y)-\mathsf{iso}(x)\|_2^2\right]\le 48\left(\frac{V\sigma^2\log(2n)}{n}\right)^{2/3}+\frac{96\sigma^2\log^2(2n)}{n}.$$

3.6.3 Proof of density estimation result (Theorem 3.4.1)

Let $G(t) = \int_{s=0}^{t} g(s)\,\mathrm{d}s$ be the cumulative distribution function for the density g. Since $g(t) \geq c$ everywhere, this means that $G(t)$ is strictly increasing, and is therefore invertible. Using this lower bound on g, and the assumption that g is L-Lipschitz, we can furthermore calculate

$$0 \leq (G^{-1})'(t) = \frac{1}{g(G^{-1}(t))} \leq \frac{1}{c} \text{ and } \left|(G^{-1})''(t)\right| = \left|\frac{-g'(G^{-1}(t))}{(g(G^{-1}(t)))^3}\right| \leq \frac{L}{c^3}. \tag{3.17}$$

Let

$$x_i = n\left(G^{-1}(i/n) - G^{-1}\big((i-1)/n\big)\right) \text{ and } y_i = n\big(Z_{(i)} - Z_{(i-1)}\big)$$

for $i = 1, \ldots, n$, where $Z_{(0)} \coloneqq 0$. Note that x gives the difference in quantiles of the distribution, while y estimates these gaps empirically.

The following lemma gives a concentration result on the $Z_{(i)}$'s:

Lemma 3.6.2. *Let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the order statistics of $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} g$, where the density $g : [0,1] \to [c, \infty)$ is L-Lipschitz. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\left|Z_{(i)} - G^{-1}(i/n)\right| \leq \frac{4}{c} \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}} \tag{3.18}$$

for all $1 \leq i \leq n$, and

$$\begin{aligned}&\left|\big(Z_{(i)} - Z_{(j)}\big) - \Big(G^{-1}(i/n) - G^{-1}(j/n)\Big)\right| \\ &\qquad \leq \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{cn} + \frac{4L|i-j|\sqrt{\log((n^2+n)/\delta)}}{c^3 n^{3/2}}\end{aligned} \tag{3.19}$$

for all $1 \leq i < j \leq n$.

From now on, we assume that these bounds (3.18) and (3.19) both hold. Plugging our defini-

tions of x and y into these two bounds, this proves that

$$\left|\overline{x}_{i:j} - \overline{y}_{i:j}\right| \le \frac{\sqrt{3(j-i+1)\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{c\cdot(j-i+1)} + \frac{4L}{c^3}\cdot\sqrt{\frac{\log((n^2+n)/\delta)}{n}}$$

for all $1 \le i \le j \le n$. (If $i = 1$ then we use the bound (3.18) while if $i > 1$ then we use the bound (3.19).)

Now, defining

$$\psi(i) = \frac{i}{\frac{1}{c}\cdot\left(\sqrt{3i\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)\right) + \frac{4L}{c^3}\cdot i\cdot\sqrt{\frac{\log((n^2+n)/\delta)}{n}}},$$

we see that

$$\|x-y\|_\psi^{\mathsf{SW}} = \max_{1\le i\le j\le n} \left|\overline{x}_{i:j} - \overline{y}_{i:j}\right|\cdot\psi(j-i+1) \le 1.$$

(Note that ψ is nondecreasing and $i \mapsto i/\psi(i)$ is concave, as required by (3.1).)

Next we check that x is a Lipschitz sequence. We have

$$\begin{aligned} n^{-1}\left(x_{i+1} - x_i\right) &= \left(G^{-1}\left(\frac{i+1}{n}\right) - G^{-1}(i/n)\right) + \left(G^{-1}\left((i-1)/n\right) - G^{-1}(i/n)\right) \\ &= (G^{-1})'(i/n)\cdot\frac{1}{n} + \frac{1}{2}(G^{-1})''\left(\frac{i+s}{n}\right)\cdot\frac{1}{n^2} \\ &\qquad + (G^{-1})'(i/n)\cdot -\frac{1}{n} + \frac{1}{2}(G^{-1})''\left(\frac{i-1+t}{n}\right)\cdot\frac{1}{n^2} \end{aligned}$$

for some $s,t \in [0,1]$, by Taylor's theorem. The first-order terms cancel, and we know by (3.17) that $(G^{-1})''$ is bounded by $\frac{L}{c^3}$. Therefore, x is $\frac{L}{c^3}$-Lipschitz. Finally, x is monotone nondecreasing since g is a monotone nonincreasing density.

We then apply the calculations (3.10) for the Lipschitz signal setting (with $\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}$ taking the place of $\sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}$, which was specific to the subgaussian noise model setting). We see that

for any index $m \geq 1$,

$$\max_{m \leq k \leq n-m+1} \left| x_k - \mathsf{iso}(y)_k \right| \leq \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)} + \frac{L(m-1)}{2nc^3}$$
$$\leq \frac{\frac{1}{c} \cdot \left(\sqrt{3m \log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta) \right) + \frac{4L}{c^3} \cdot m \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}}}{m} + \frac{L(m-1)}{2nc^3}.$$

Set $m = \left\lceil \left(n\sqrt{\log((n^2+n)/\delta)} \right)^{2/3} \right\rceil$. Since $\Delta < \frac{1}{c+L} \leq 1$ we know $\log((n^2+n)/\delta) \leq n$, so we can simplify the above bound to

$$\max_{m \leq k \leq n-m+1} \left| x_k - \mathsf{iso}(y)_k \right| \leq \left(\frac{4}{c} + \frac{4.5L}{c^3} \right) \sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}.$$

Now we show how this uniform bound on the difference $x-y$, translates to an error bound on the Grenander density estimator $\widehat{g}_{\mathsf{Gren}}$. First, we check that $Z_{(m)} \leq \Delta$ and $Z_{(n-m+1)} \geq 1-\Delta$. We have

$$1 \geq \|x-y\|_\psi^{\mathsf{SW}} \geq \psi(m) \cdot \left| \bar{x}_{1:m} - \bar{y}_{1:m} \right| = \frac{\psi(m)}{m} \cdot n \cdot \left| G^{-1}(m/n) - Z_{(m)} \right|.$$

We also know that $G^{-1}(m/n) \leq \frac{m}{cn}$ since G^{-1} is $(1/c)$-Lipschitz as calculated in (3.17), and so

$$Z_{(m)} \leq G^{-1}(m/n) + \frac{m}{n\psi(m)} \leq$$
$$\frac{m}{cn} + \frac{\sqrt{3m\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{cn} + \frac{4Lm\sqrt{\log((n^2+n)/\delta)}}{c^3 n^{3/2}}$$
$$\leq \left(\frac{5}{c} + \frac{4L}{c^3} \right) \sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \leq \Delta,$$

using the fact that $\log((n^2+n)/\delta) \leq n$ as before. Similarly $Z_{(n-m+1)} \geq 1-\Delta$.

Next, for any t with $\Delta \leq t \leq 1-\Delta$, find index k such that

$$Z_{(k-1)} < t \leq Z_{(k)}.$$

By the work above we will have $m \le k \le n-m+1$. Then $\widehat{g}_{\mathsf{Gren}}(t) = \frac{1}{\mathsf{iso}(y)_k}$ by definition of the Grenander estimator. Therefore, we have

$$\left|\widehat{g}_{\mathsf{Gren}}(t) - g(t)\right| = \left|\frac{1}{\mathsf{iso}(y)_k} - g(t)\right| \le \left|\frac{1}{\mathsf{iso}(y)_k} - \frac{1}{x_k}\right| + \left|\frac{1}{x_k} - g(t)\right|.$$

By Lemma 3.6.2, we know

$$Z_{(k)} \le G^{-1}\left(\frac{k}{n}\right) + \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

and

$$Z_{(k-1)} \ge G^{-1}\left(\frac{k-1}{n}\right) - \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

so we have

$$G^{-1}\left(\frac{k-1}{n}\right) - \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} < t \le G^{-1}\left(\frac{k}{n}\right) + \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

We calculate

$$x_k = n\left(G^{-1}\left(\frac{k}{n}\right) - G^{-1}\left(\frac{k-1}{n}\right)\right) = n(G^{-1})'\left(\frac{k-1+s}{n}\right)\cdot\frac{1}{n} = \frac{1}{g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right)}$$

by Taylor's theorem for some $s \in [0,1]$, and so

$$\begin{aligned}\left|\frac{1}{x_k} - g(t)\right| = \left|g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right) - g(t)\right| &\le L\cdot\left|G^{-1}\left(\frac{k-1+s}{n}\right) - t\right| \\ &\le L\left|G^{-1}\left(\frac{k}{n}\right) - G^{-1}\left(\frac{k-1}{n}\right)\right| + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \\ &\le \frac{L}{cn} + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}},\end{aligned}$$

since g is L-Lipschitz and G^{-1} is $(1/c)$-Lipschitz, as proved before. Finally,

$$\left|\frac{1}{\mathsf{iso}(y)_k}-\frac{1}{x_k}\right| = \frac{|x_k-\mathsf{iso}(y)_k|}{x_k\cdot\mathsf{iso}(y)_k}$$
$$\leq \frac{\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{x_k\cdot\mathsf{iso}(y)_k} \leq \frac{\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{x_k\cdot\left(x_k-\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)},$$

from the bound on $\left|x_k-\mathsf{iso}(y)_k\right|$ above. And, we know that $x_k = \frac{1}{g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right)}$ for some $s\in[0,1]$ as above, so $x_k \geq \frac{1}{\max_{s\in[0,1]} g(s)}$. Now, since g is lower-bounded by c and is L-Lipschitz, we see that $g(s)\leq c+L$, and so $x_k \geq \frac{1}{c+L}$. Combining everything,

$$\left|\widehat{g}_{\mathsf{Gren}}(t)-g(t)\right| \leq$$
$$\frac{\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{\frac{1}{c+L}\cdot\left(\frac{1}{c+L}-\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)} + \frac{L}{cn} + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$
$$\leq \frac{\left(\frac{1}{c}\left(4+\frac{5L}{c+L}\right)+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{\frac{1}{c+L}\cdot\left(\frac{1}{c+L}-\left(\frac{4}{c}+\frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)} \leq \frac{\Delta}{\frac{1}{c+L}(\frac{1}{c+L}-\Delta)}.$$

3.7 Proofs of Lemmas

Proof of Lemma 3.2.2. By Theorem 3.2.1, we only need to prove that $\|\cdot\|_\psi^{\mathsf{SW}}$ satisfies NUNA. Fix any $x\in\mathbb{R}^n$ and any index $k=1,\ldots,n-1$. Let $y=A_kx$. Note that we have

$$y_i = \begin{cases} x_i, & \text{if } i<k \text{ or } i>k+1,\\ \frac{x_k+x_{k+1}}{2}, & \text{if } i=k \text{ or } i=k+1.\end{cases}$$

Take any indices $1\leq i\leq j\leq n$. We need to prove that $|\bar{y}_{i:j}|\cdot\psi(j-i+1)\leq\|x\|_\psi^{\mathsf{SW}}$.

- Case 1: if $j<k$ or if $i>k+1$, then neither of the indices $k,k+1$ are included in the window

$i:j$, and therefore $x_{i:j} = y_{i:j}$ (i.e. all entries in the stretch of indices $i:j$ are equal). So,

$$|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{x}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_{\psi}^{\mathsf{SW}}.$$

- Case 2: If $i \leq k$ and $j \geq k+1$, then both indices $k, k+1$ are included in the window $i:j$. Since $y_k + y_{k+1} = x_k + x_{k+1}$ and all other entries of x and y coincide, we can trivially see that

$$|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{x}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_{\psi}^{\mathsf{SW}}.$$

- Case 3: if $i < k$ and $j = k$, then

$$\begin{aligned}
&|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{y}_{i:k}| \cdot \psi(k-i+1) \\
&= \frac{\left|\sum_{\ell=i}^{k-1} x_\ell + \frac{x_k + x_{k+1}}{2}\right| \cdot \psi(k-i+1)}{k-i+1} \\
&= \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k-1} x_\ell + \frac{1}{2}\sum_{\ell=i}^{k+1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1} \\
&\leq \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k-1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1} + \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k+1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1} \\
&= \frac{\psi(k-i+1)}{k-i+1} \cdot \left(\frac{1}{2} |\overline{x}_{i:(k-1)}| \cdot \psi(k-i) \cdot \frac{k-i}{\psi(k-i)} \right. \\
&\left. + \frac{1}{2} |\overline{x}_{i:(k+1)}| \cdot \psi(k-i+2) \cdot \frac{k-i+2}{\psi(k-i+2)} \right) \\
&\leq \|x\|_{\psi}^{\mathsf{SW}} \cdot \frac{1}{2} \left[\frac{k-i}{\psi(k-i)} + \frac{k-i+2}{\psi(k-i+2)} \right] \cdot \frac{\psi(k-i+1)}{k-i+1} \\
&\leq \|x\|_{\psi}^{\mathsf{SW}} \cdot \frac{k-i+1}{\psi(k-i+1)} \cdot \frac{\psi(k-i+1)}{k-i+1} = \|x\|_{\psi}^{\mathsf{SW}},
\end{aligned}$$

  where the last inequality holds since $i \mapsto i/\psi(i)$ is concave by assumption on ψ.

- Case 4: if $i = k+1$ and $j > k+1$, by symmetry this case is analogous to Case 3.

- Case 5: if $i = j = k$, then

$$|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |y_k| \cdot \psi(1) = |x_k + x_{k+1}| \cdot \psi(1)/2$$
$$= |\overline{x}_{k:(k+1)}| \cdot \psi(1) \le |\overline{x}_{k:(k+1)}| \cdot \psi(2) \le \|x\|_\psi^{\mathsf{SW}},$$

  since $\psi(1) \le \psi(2)$ due to the assumption that ψ is nondecreasing.

- Case 6: if $i = j = k+1$, then by symmetry this case is analogous to Case 5.

Therefore, $|\overline{y}_{i:j}| \cdot \psi(j-i+1) \le \|x\|_\psi^{\mathsf{SW}}$ for all indices $1 \le i \le j \le n$, and so $\|y\|_\psi^{\mathsf{SW}} \le \|x\|_\psi^{\mathsf{SW}}$, as desired. □

Proof of Lemma 3.2.3. For any indices $1 \le i \le j \le n$,

$$\overline{y}_{i:j} - \overline{x}_{i:j} = \sigma \overline{\varepsilon}_{i:j},$$

and we know that $\sqrt{j-i+1} \cdot \overline{\varepsilon}_{i:j}$ is subgaussian, that is,

$$\mathbb{P}\left\{ \sqrt{j-i+1} \cdot |\overline{\varepsilon}_{i:j}| > t \right\} \le 2e^{-t^2/2}$$

for any $t \ge 0$. Now we set $t = \sqrt{2\log\left(\frac{n^2+n}{\delta}\right)}$, and take a union bound over all $n + \binom{n}{2} = \frac{n^2+n}{2}$ possible pairs of indices $i \le j$. We then have

$$\mathbb{P}\left\{ \max_{1 \le i \le j \le n} \sqrt{j-i+1} \cdot |\overline{\varepsilon}_{i:j}| \le \sqrt{2\log\left(\frac{n^2+n}{\delta}\right)} \right\} \ge 1 - \delta.$$

Setting $\psi(t) = \sqrt{t}$ proves that, on this event, $\|x-y\|_\psi^{\mathsf{SW}} \le \sigma\sqrt{2\log\left(\frac{n^2+n}{\delta}\right)}$, as desired. For the bound in expectation, we have a similar calculation: it is known that $\mathbb{E}\left[\max_{k=1,\dots,N} |Z_k|\right] \le \sqrt{2\log(2N)}$ and $\mathbb{E}\left[\max_{k=1,\dots,N} |Z_k|^2\right] \le 8\log(2N)$ for any (not necessarily independent) subgaussian random variables Z_k. Setting $Z_k = \sqrt{j-i+1} \cdot \overline{\varepsilon}_{i:j}$ for each of the $N = \frac{n^2+n}{2}$ possible pairs i, j,

we obtain

$$\mathbb{E}\left[\max_{1\le i\le j\le n}\sqrt{j-i+1}\cdot\left|\overline{\varepsilon}_{i:j}\right|\right]\le\sqrt{2\log(n^2+n)}$$

and

$$\mathbb{E}\left[\left(\max_{1\le i\le j\le n}\sqrt{j-i+1}\cdot\left|\overline{\varepsilon}_{i:j}\right|\right)^2\right]\le 8\log(n^2+n).$$

□

Proof of Lemma 3.3.1. Assume that x is $\varepsilon_{\mathsf{iso}}$-monotone, and fix any index $1\le i\le n$. Let $j=\max\{k\le n:\mathsf{iso}(x)_k=\mathsf{iso}(x)_i\}$. Then $i\le j\le n$, $\mathsf{iso}(x)_i=\mathsf{iso}(x)_j$, and either $j=n$ or $\mathsf{iso}(x)_j<\mathsf{iso}(x)_{j+1}$. Therefore, we must have $x_j\le\mathsf{iso}(x)_j$ by properties of the isotonic projection. (This is because, if $x_j>\mathsf{iso}(x)_j$, then writing $\mathbf{e}_j$ for the jth basis vector and taking some sufficiently small $\varepsilon>0$, the vector $\mathsf{iso}(x)+\varepsilon\cdot\mathbf{e}_j$ is an isotonic vector that is strictly closer to x than $\mathsf{iso}(x)$, which is a contradiction.) Therefore, $x_i\le x_j+\varepsilon_{\mathsf{iso}}\le\mathsf{iso}(x)_j+\varepsilon_{\mathsf{iso}}=\mathsf{iso}(x)_i+\varepsilon_{\mathsf{iso}}$. The reverse bound is proved similarly.

Now we turn to the converse. For any $1\le i\le j\le n$, we have $x_i\le\mathsf{iso}(x)_i+\varepsilon\le\mathsf{iso}(x)_j+\varepsilon\le x_j+2\varepsilon$, where the first and third inequalities use the bound $\|x-\mathsf{iso}(x)\|_\infty\le\varepsilon$, while the second uses the fact that $\mathsf{iso}(x)$ is monotone. □

Proof of Lemma 3.6.1. For $i=1,\ldots,n-1$, let $\mathscr{K}_i=\{x\in\mathbb{R}^n:x_i\le x_{i+1}\}$, which is a closed convex cone in $\mathbb{R}^n$. We have $\mathscr{K}_{\mathsf{iso}}=\cap_{i=1}^{n-1}\mathscr{K}_i$ and it's easy to see that $\mathsf{iso}_i(x)=\mathscr{P}_{\mathscr{K}_i}(x)$. Hence the slow projection algorithm defined in (3.15) is actually a cyclic projection algorithm, that is, the iterates are given by

$$x^{(1)}=\mathscr{P}_{\mathscr{K}_1}(x^{(0)}),\ \ x^{(2)}=\mathscr{P}_{\mathscr{K}_2}(x^{(1)}),\ \ \ldots,\ \ x^{(n)}=\mathscr{P}_{\mathscr{K}_1}(x^{(n-1)}),\ \ \ldots.$$

In general, it is known that a cyclic projection algorithm starting at some point $x=x^{(0)}$ is guaranteed to converge to some point in the intersection of the respective convex sets, i.e. $\lim_{t\to\infty}x^{(t)}=x^*\in\cap_{i=1}^{n-1}\mathscr{K}_i=\mathscr{K}_{\mathsf{iso}}$, but without any assumptions on the nature of the convex sets $\mathscr{K}_i$, this point may not necessarily be the projection of x onto the intersection of the sets (Bregman [47], Han

[48]). Therefore, we need to check that for our specific choice of the sets $\mathscr{K}_i$, the cyclic projection algorithm (3.15) in fact converges to $\mathsf{iso}(x)$ as claimed in the lemma.

We first claim that

$$\mathsf{iso}\big(\mathsf{iso}_i(x)\big) = \mathsf{iso}(x) \tag{3.20}$$

for all $x \in \mathbb{R}^n$ and all $i = 1, \dots, n-1$. Assume for now that this is true. Since $\mathsf{iso}(\cdot)$ is contractive with respect to the ℓ_2 norm, the convergence $x^{(t)} \to x^*$ implies that $\mathsf{iso}(x^{(t)}) \to \mathsf{iso}(x^*)$. Applying (3.20) inductively, we know that $\mathsf{iso}(x^{(t)}) = \mathsf{iso}(x^{(0)}) = \mathsf{iso}(x)$ for all $t \geq 1$. On the other hand, since $x^* \in \mathscr{K}_{\mathsf{iso}}$, this means that $x^* = \mathsf{iso}(x^*)$. Combining everything, then, we obtain

$$\lim_{t\to\infty} x^{(t)} = x^* = \mathsf{iso}(x^*) = \lim_{t\to\infty} \mathsf{iso}(x^{(t)}) = \mathsf{iso}(x).$$

Finally, we need to prove (3.20). Fix any index i and any $x \in \mathbb{R}^n$. If $x_i \leq x_{i+1}$, then $\mathsf{iso}_i(x) = x$ and the statement holds trivially. If not, then $x_i > x_{i+1}$ and we have $\mathsf{iso}_i(x) = A_i x$ (recalling the definition of A_i in (3.14) earlier). Now let $y = \mathsf{iso}(x)$ and $z = \mathsf{iso}(A_i x)$. It is trivially true that, since $x_i > x_{i+1}$, we must have $y_i = y_{i+1}$. Also, $\langle x - y, z - y\rangle \leq 0$ by properties of projection to the convex set $\mathscr{K}_{\mathsf{iso}}$, so we can calculate

$$\begin{aligned} \langle A_i x - y, z - y\rangle &= \langle x - y, z - y\rangle + \frac{x_{i+1} - x_i}{2} \cdot (z_i - y_i - z_{i+1} + y_{i+1}) \\ &\leq \frac{x_{i+1} - x_i}{2} \cdot (z_i - z_{i+1}) \leq 0, \end{aligned} \tag{3.21}$$

where the last step holds since $z_i \leq z_{i+1}$ due to $z \in \mathscr{K}_{\mathsf{iso}}$ and $x_i \geq x_{i+1}$ by assumption. We also have $\|A_i x - z\|_2^2 \leq \|A_i x - y\|_2^2$ since $z = \mathsf{iso}(A_i x)$, which combined with (3.21) proves that $y = z$. Thus (3.20) holds, as desired. □

Proof of Lemma 3.6.2. Let $U_1, \dots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, and let $G(t) = \int_{s=0}^{t} g(s)\, \mathrm{d}s$ be the cumulative distribution function for the density g. Since $g \geq c > 0$, $G : [0,1] \to [0,1]$ is strictly increasing, and is therefore invertible. It is known that setting $Z_{(i)} = G^{-1}(U_{(i)})$ recovers the desired distribution for the ordered sample points $Z_{(1)} \leq \cdots \leq Z_{(n)}$.

Next, by Lemma 3.7.2 below, with probability at least $1-\delta$, for all indices $0 \le i < j \le n$,

$$\left| U_{(i)} - U_{(j)} - \frac{i-j}{n} \right| \le \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n}. \tag{3.22}$$

From this point on, assume that this bound holds. In particular, by taking $i = 0$, this implies that

$$\left| U_{(j)} - \frac{j}{n} \right| \le \frac{\sqrt{3j\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n} \le 4\sqrt{\frac{\log((n^2+n)/\delta)}{n}}, \tag{3.23}$$

for all $j = 1, \ldots, n$, by assuming that $\log((n^2+n)/\delta) \le n$ (if not, then this bound holds trivially since $U_{(j)}$ and j/n both lie in $[0,1]$).

Then, since g is L-Lipschitz, for $1 \le i < j \le n$ we compute

$$\begin{aligned}
&\left| \left(Z_{(i)} - Z_{(j)}\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right) \right| \\
&\qquad = \left| \left(G^{-1}(U_{(i)}) - G^{-1}(U_{(j)})\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right) \right| \\
&\qquad = \left| (U_{(i)} - U_{(j)}) \cdot (G^{-1})' \left(sU_{(i)} + (1-s)U_{(j)}\right) - \frac{i-j}{n} \cdot (G^{-1})' \left(\frac{si + (1-s)j}{n}\right) \right|,
\end{aligned}$$

where the last step holds by Taylor's theorem applied to the function

$$s \mapsto \left(G^{-1}\left(sU_{(i)} + (1-s)U_{(j)}\right) - G^{-1}\left(\frac{si + (1-s)j}{n}\right) \right).$$

We can rewrite this as

$$\begin{aligned}
&\left| \left(Z_{(i)} - Z_{(j)}\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right) \right| \\
&\qquad \le \left| (U_{(i)} - U_{(j)}) - \frac{i-j}{n} \right| \cdot (G^{-1})' \left(sU_{(i)} + (1-s)U_{(j)}\right) + \\
&\qquad\qquad \frac{j-i}{n} \cdot \left| (G^{-1})' \left(sU_{(i)} + (1-s)U_{(j)}\right) - (G^{-1})' \left(\frac{si + (1-s)j}{n}\right) \right|,
\end{aligned}$$

Since G^{-1} has bounded first and second derivatives as in (3.17), we then have

$$\begin{aligned}
&\left|\left(Z_{(i)} - Z_{(j)}\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right)\right| \\
&\quad \le \left|(U_{(i)} - U_{(j)}) - \frac{i-j}{n}\right| \cdot \frac{1}{c} + \frac{j-i}{n} \cdot \frac{L}{c^3} \cdot \left|\left(sU_{(i)} + (1-s)U_{(j)}\right) - \frac{si + (1-s)j}{n}\right| \\
&\quad \le \frac{\sqrt{3(j-i)\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n} \cdot \frac{1}{c} + \frac{j-i}{n} \cdot \frac{4L}{c^3} \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}},
\end{aligned}$$

applying the bounds obtained above in (3.22) and (3.23). This proves the bound (3.19) in the lemma. To prove the simpler bound (3.18), we calculate

$$\left|Z_{(i)} - G^{-1}(i/n)\right| = \left|G^{-1}(U_{(i)}) - G^{-1}(i/n)\right| \le \frac{1}{c}\left|U_{(i)} - i/n\right| \le \frac{4}{c}\sqrt{\frac{\log((n^2+n)/\delta)}{n}},$$

since G^{-1} is $(1/c)$-Lipschitz and we can apply (3.23). □

Lemma 3.7.1. *Let $U_{(1)} \le \cdots \le U_{(n)}$ be the order statistics of $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$. For any $\delta > 0$,*

$$\mathbb{P}\left\{\left|U_{(i)} - \frac{i}{n}\right| \le \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n} \text{ for all } i = 1, \ldots, n\right\} \ge 1 - \delta.$$

Proof of Lemma 3.7.1. Fix any index i. If $i < 3\log(2n/\delta)$, then

$$\frac{i}{n} - \frac{\sqrt{3i\log(2n/\delta)}}{n} \le 0$$

and so trivially we have $U_{(i)} \ge \frac{i}{n} - \frac{\sqrt{3i\log(2n/\delta)}}{n}$. If instead $i \ge 3\log(2n/\delta)$, then suppose that $U_{(i)} \le \frac{i}{n} - \frac{\sqrt{3i\log(2n/\delta)}}{n} =: p$. This means that at least i many of the U_k's lie in the interval $[0, p]$.

Then

$$\begin{aligned}
\mathbb{P}\left\{U_{(i)} \le \frac{i}{n} - \frac{\sqrt{3i\log(2n/\delta)}}{n}\right\} &= \mathbb{P}\left\{U_{(i)} \le p\right\} = \mathbb{P}\left\{\text{Binomial}(n,p) \ge i\right\} \\
&= \mathbb{P}\left\{\text{Binomial}(n,p) \ge np \cdot \left(1 + \frac{\sqrt{3i\log(2n/\delta)}}{i - \sqrt{3i\log(2n/\delta)}}\right)\right\} \\
&\le \exp\left\{-\frac{1}{3}np\left(\frac{\sqrt{3i\log(2n/\delta)}}{i-\sqrt{3i\log(2n/\delta)}}\right)^2\right\} = \exp\left\{-\frac{1}{3}\frac{\left(\sqrt{3i\log(2n/\delta)}\right)^2}{i-\sqrt{3i\log(2n/\delta)}}\right\} \le \frac{\delta}{2n},
\end{aligned}$$

where the inequality uses the multiplicative Chernoff bound. Next, suppose that instead we have

$$U_{(i)} \ge \frac{i}{n} + \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n} =: p'.$$

This means that at most $i-1$ of the U_k's lie in the interval $[0,p']$. Then

$$\begin{aligned}
&\mathbb{P}\left\{U_{(i)} \ge \frac{i}{n} + \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n}\right\} = \mathbb{P}\left\{U_{(i)} \ge p'\right\} \\
\le\ & \mathbb{P}\left\{\text{Binomial}(n,p') \le i\right\} \\
=\ & \mathbb{P}\left\{\text{Binomial}(n,p') \le np' \cdot \left(1 - \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{i + \sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right)\right\} \\
\le\ & \exp\left\{-\frac{1}{2}np'\left(\frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{i+\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right)^2\right\} \\
=\ & \exp\left\{-\frac{1}{2}\frac{\left(\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)\right)^2}{i+\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right\} \le \frac{\delta}{2n},
\end{aligned}$$

where again the first inequality uses the multiplicative Chernoff bound. Combining these two calculations,

$$\mathbb{P}\left\{\left|U_{(i)} - \frac{i}{n}\right| \le \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n}\right\} \ge 1 - \delta/n.$$

Finally, taking a union bound over all i, we have proved the lemma. □

Lemma 3.7.2. *Let $U_{(1)} \le \cdots \le U_{(n)}$ be the order statistics of $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, and let*

$U_{(0)} = 0$. *For any* $\delta > 0$,

$$\mathbb{P}\left\{ \left| U_{(i)} - U_{(j)} - \frac{i-j}{n} \right| \le \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n} \right.$$
$$\left. \text{for all } 0 \le i < j \le n \right\} \ge 1 - \delta.$$

Proof of Lemma 3.7.2. First, it is known that $U_{(j)} - U_{(i)} \sim \text{Beta}\big((j-i), (n+1)-(j-i)\big)$ for all $0 \le i < j \le n$. In particular, $U_{(j)} - U_{(i)}$ has the same distribution as $U_{(j-i)}$, and so by Lemma 3.7.1 applied with $k = j - i$ and with $2\delta/(n^2+n)$ in place of δ/n,

$$\mathbb{P}\left\{ \left| U_{(j)} - U_{(i)} - \frac{j-i}{n} \right| > \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n} \right\} \le \frac{2\delta}{n^2+n}.$$

Taking a union bound over all $\binom{n+1}{2} = \frac{n^2+n}{2}$ pairs of indices i, j, then, we obtain the desired bound. □

CHAPTER 4

COVARIATE ASSISTED VARIABLE RANKING

Consider a high dimensional linear model where the number of features far exceeds the number of samples. We are interested in variable ranking, a problem related to variable selection but not the same. Scientific experiments are constrained by budget and manpower, and it is often impossible to completely separate the signals from the noise. An alternative is then to identify a few most promising variables for follow-up lab experiments. This is where variable ranking comes in.

In this chapter, [1] we mainly consider the variable ranking problem for high dimensional linear regression. We are interested in the Rare/Weak signal setting where all but a small fraction of the entries of β are nonzero, and the nonzero entries are relatively small individually. We propose *Factor-adjusted Covariate Assisted Ranking* (FA-CAR) as a two-step approach to variable ranking. FA-CAR is easy to use and computationally fast, and it is effective in resolving signal cancellation, a challenge we face in regression models. We then extend our methods to the generalized linear model case.

4.1 Problem formulation

Consider the linear model in the $p \gg n$ setting:

$$y = X\beta + z, \qquad X = [x_1, \cdots, x_p] \in \mathbb{R}^{n,p}, \qquad z \sim N(0, \sigma^2 I_n). \tag{4.1}$$

We call a nonzero entry of β a "signal" and a zero entry a "noise." We are interested in the Rare/Weak signal regime:

- (*Rare*). All but a small fraction of the entries of β are nonzero.
- (*Weak*). Individually, the nonzero entries are relatively small.

1. The work in this chapter except section 4.4.1 can be found in [49]. The results in section 4.4.1 are unpublished.

We assume the Gram matrix $\Theta = (1/n)X'X$ follows an *approximate factor model* [50, 51]:

$$\Theta = \sum_{k=1}^{K} \lambda_k v_k v_k' + G_0, \qquad K \ll \min\{n, p\}, \tag{4.2}$$

where G_0 is positive definite and approximately sparse (in the sense that each row has relatively few large entries, with all other entries relatively small), $\lambda_1, \cdots, \lambda_K$ are positive, and $v_1, \cdots, v_K$ are mutually orthogonal unit-norm vectors.

Model (4.2) is popular in finance [52], where the low-rank part represents a few risk factors and G_0 is the covariance matrix of the (weakly-correlated) idiosyncratic noise. It is also useful in microarray analysis, where the low-rank part represents technical, environmental, demographic, or genetic factors [53].

For variable ranking, Marginal Ranking (MR) is an approach that is especially popular in genomics and genetics [54, 55]. Recall that $X = [x_1, x_2, \ldots, x_p]$. MR ranks variables according to the marginal regression coefficients $|(x_j, y)|/(x_j, x_j)$, $1 \leq j \leq p$, where $(\cdot, \cdot)$ denotes the inner product. Marginal ranking has advantages: (a) It directly provides an explicit ranking that is reasonable in many cases; (b) It is easy-to-use and computationally fast; (c) It does not need tuning parameters; (d) It has a relatively less stringent requirement on noise distributions and provides reasonable results even when the noise distribution is unknown or when the features/samples are correlated.

Our goal is to improve MR. Despite many good aspects of MR, we recognize that it faces two challenges:

- The K factors in Θ may have a dominant effect, and the ranking by MR is only reasonable when these factors are removed.
- MR faces the so-called challenge of “signal cancellation" [56].

The “signal cancellation” problem arises when the conditional effect of a covariate on response Y

is large but due to confounding the overall marginal correlation is close to zero. Note that

$$E[(x_j,y)/(x_j,x_j)] = (x_j,x_j)^{-1} \sum_{k:\beta_k \neq 0} (x_j,x_k)\beta_k.$$

"Signal cancellation" means that due to correlations among x_j's, signals may have a mutual canceling effects, and variable j may receive a relatively low rank even when β_j is top-ranked among $\beta_1, \beta_2, \ldots, \beta_p$.

To overcome the challenges, we propose FA-CAR as a two-stage ranking method. FA-CAR contains a Factor-Adjusting (FA) step, where we use PCA for factor removal and reduce the linear model to a new one where the Gram matrix is sparse. In the Covariate-Assisted Ranking (CAR) step, we rank variables using covariate structures. We recognize that "signal cancellation" is only severe when the predictors are heavily correlated, and so by exploiting the covariate structures, we can significantly alleviate the canceling effects.

4.1.1 Two illustrating examples

It is instructive to use two simple examples to illustrate why FA and CAR are useful.

Example 1 (*One-factor design*). Consider a case where the Gram matrix is given by

$$\Theta = I_p + \omega_p \xi \xi', \qquad \omega_p > 0 \text{ is a parameter},$$

where $\xi = \eta/\|\eta\|$ with $\eta \sim N(0, I_n)$ and $\sigma^2 = 0$ so there is no noise. We assume β has s nonzeros and each nonzero equals to τ ($\tau > 0$). Even in this simple setting, many methods do not perform well. Take MR for example. As long as $w_p \ll p$, we have $n^{-1}(x_j,x_j) \approx 1$ and

$$|(x_j,y)|/(x_j,x_j) \sim |\omega_p \cdot (\xi,\beta)\xi_j + \beta_j|.$$

Since $|(\xi,\beta)\xi_j| = O_p(\tau\sqrt{s}/p)$, whenever $w_p\sqrt{s}/p \gg 1$, the factor has a non-negligible effect: the ranking depends more on ξ instead of β, and many signal variables may receive lower rankings

than the noise variables.

Seemingly, the problem can be fixed if we use a factor removal step. Consider the Singular Value Decomposition (SVD) of the design matrix X:

$$X = \sum_{k=1}^{n} \lambda_k u_k v_k' \equiv \lambda_1 u_1 v_1' + \tilde{X}, \qquad \text{where } v_1 = \xi \text{ and } \tilde{X} = \sum_{k=2}^{n} \lambda_k u_k v_k'.$$

We have two ways to remove the factor ξ: one is to project the columns of X using the projection matrix $H_u = I_n - u_1 u_1'$, and the other one is to project the rows of X using the projection matrix $H_v = I_p - v_1 v_1'$. However, while both projections produce the same matrix:

$$\tilde{X} = H_u X = X H_v,$$

only the first one reduces Model (4.1) to a new linear model with the same vector β. In particular, letting $\tilde{y} = H_u y$ and $\tilde{\varepsilon} = H_u \varepsilon$, we have

$$\tilde{y} = \tilde{X}\beta + \tilde{z}, \qquad \text{where } \tilde{z} \sim N(0, \sigma^2 H_u) \text{ and } n^{-1}\tilde{X}'\tilde{X} = H_v = I_p - \xi\xi'.$$

Similarly, if we write $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p]$ and apply MR, then $n^{-1}(\tilde{x}_j, \tilde{x}_j) \approx 1$ and

$$|(\tilde{y}, \tilde{x}_j)|/(\tilde{x}_j, \tilde{x}_j) \sim |(\xi, \beta)\xi_j + \beta_j|,$$

where $|(\xi, \beta)\xi_j| = O_p(\tau\sqrt{s}/p)$ and has a negligible effect on the ranking. We therefore have a successful ranking scheme if we first remove the factor and then apply MR (in this simple setting, the sparse component G_0 in Θ is diagonal). This is the basic idea of the FA step, which can be conveniently extended to cases where we have more than 1 factors.

Example 2. (*Block-wise diagonal design*). Suppose p is even and Θ is block-wise diagonal,

where each diagonal block takes the form of

$$\begin{pmatrix} 1 & h \\ h & 1 \end{pmatrix}, \qquad h \in (-1,1) \text{ is a parameter.}$$

The parameter σ^2 is equal to zero so there is no noise. The vector β only has three nonzeros (but we don't know either the number of signals, or the locations or strengths of them):

$$\beta_1 = \tau, \qquad \beta_2 = \beta_3 = a\tau, \qquad \text{where } \tau > 0 \text{ and } a \in \mathbb{R}.$$

In this simple setting, even there is no factors in Θ, MR still does not perform well. For example, by direct calculations,

$$\beta_2 = a\tau, \quad \beta_4 = 0, \quad |(x_2, y)|/(x_2, x_2) = |a-h|\tau, \quad |(x_4, y)|/(x_4, x_4) = |ah|\tau.$$

Therefore, we may face severe signal cancellation at location 2, and variable 2 (a signal variable) is ranked under variable 4 (a noise variable) when

$$|ah| > |a-h|.$$

We recognize that this problem can be resolved by exploiting local covariate structures. For each variable j, let

$$\mathscr{A}_j = \{\mathscr{I}_1, \mathscr{I}_2\}, \quad \mathscr{I}_1 = \{j\}, \ \ \mathscr{I}_2 = \{j, j+1\} \text{ for } j \text{ odd and } \mathscr{I}_2 = \{j-1, j\} \text{ for } j \text{ even.}$$

Each element $\mathscr{I} \in \mathscr{A}_j$ is called a "neighborhood" of j. For each $\mathscr{I} \in \mathscr{A}_j$, we measure the "significance" of variable j in $\mathscr{I}$ by

$$T_{j|\mathscr{I}} = \|P_{\mathscr{I}} y\|^2 - \|P_{\mathscr{I}\setminus\{j\}} y\|^2,$$

where $P_{\mathscr{I}}$ is the projection from $\mathbb{R}^n$ to the space spanned by $\{x_j, j \in \mathscr{I}\}$. Neglecting the influence of all variables outside the set $\mathscr{I}$, $T_{j|\mathscr{I}}$ is the likelihood ratio for testing whether $\mathrm{supp}(\beta) = \mathscr{I}$ or $\mathrm{supp}(\beta) = \mathscr{I} \setminus \{j\}$. Take an odd j for example, where $\mathscr{I}_1 = \{j\}$ and $\mathscr{I}_2 = \{j, j+1\}$. By direct calculations,

$$T_{j|\mathscr{I}_1} = n(\beta_j + h\beta_{j+1})^2, \qquad T_{j|\mathscr{I}_2} = n(1-h^2)\beta_j^2.$$

When both variables j and $(j+1)$ are signals, signal cancellation only affects $T_{j|\mathscr{I}_1}$ but not $T_{j|\mathscr{I}_2}$, so the latter is preferred. When variable j is a signal and variable $(j+1)$ is a noise, signal cancellation affects neither of them; since $T_{j|\mathscr{I}_1} = n\beta_j^2 \geq T_{j|\mathscr{I}_2}$ in this case, $T_{j|\mathscr{I}_1}$ is preferred. This motivates us to assess the significance of variable j by combining these scores:

$$T_j^* = \max\big\{T_{j|\mathscr{I}} : \mathscr{I} \in \mathscr{A}_j\big\}.$$

In the above example,

$$T_2^* = n\max\big\{a^2(1-h^2), (a-h)^2\big\}\tau^2, \qquad T_4^* = n(ah)^2\tau^2,$$

and variables 2 and 4 are ranked correctly as long as $|h| < 1/\sqrt{2} \approx 0.7$.

In more general cases, we use Θ to construct a graph and let a "neighborhood" of j be a connected subgraph that contains j, and the above idea can thus be conveniently extended. This is the main idea of the CAR step.

4.1.2 Our methods: FA-CAR

We extend the intuition gained in illustrating examples and develop a ranking method that works for a general design from Model (4.2). The method consists of a Factor-Adjusting (FA) step and a Covariate Assisted Ranking (CAR) step.

In the FA step, let $X = \sum_{k=1}^n \hat{\sigma}_k \hat{u}_k \hat{v}_k'$ be the SVD of X, where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_n > 0$ are the singular values, and $\hat{u}_k \in \mathbb{R}^n$ and $\hat{v}_k \in \mathbb{R}^p$ are the k-th (unit-norm) left and right singular vectors,

respectively. Introduce

$$\tilde{y} = y - \sum_{k=1}^{K} (\hat{u}_k' y)\hat{u}_k, \qquad \tilde{X} = X - \sum_{k=1}^{K} \hat{\sigma}_k \hat{u}_k \hat{v}_k'. \tag{4.3}$$

If we consider the two projection matrices $H_u = I_n - \sum_{k=1}^{K} \hat{u}_k \hat{u}_k'$ and $H_v = I_p - \sum_{k=1}^{K} \hat{v}_k \hat{v}_k'$, then it follows from elementary linear algebra that $\tilde{y} = H_u y$, $\tilde{X} = H_u X = X H_v$, and $(1/n)\tilde{X}'\tilde{X} = H_v \Theta H_v$. As a result,

$$\tilde{y} = \tilde{X}\beta + \tilde{z}, \quad \text{where } \tilde{z} \sim N(0, \sigma^2 H_u) \text{ and } n^{-1}\tilde{X}'\tilde{X} = \Theta - \sum_{k=1}^{K} (\hat{\sigma}_k^2/n)\hat{v}_k \hat{v}_k'. \tag{4.4}$$

This gives a new linear model with the same β but a different Gram matrix.

Note that $(\hat{\sigma}_k^2/n)$ and $\hat{v}_k$ are the k-th eigenvalue and eigenvector of Θ, respectively. In Model (4.2), the component G_0 is a sparse matrix. Therefore, the leading eigenvalues (eigenvectors) of Θ are approximately equal to the leading eigenvalues (eigenvectors) of $(\Theta - G_0)$, i.e., $(\hat{\sigma}_k^2/n) \approx \lambda_k$ and $\hat{v}_k \approx v_k$ for $1 \le k \le K$. We thus have

$$(1/n)\tilde{X}'\tilde{X} \approx \Theta - \sum_{k=1}^{K} \lambda_k v_k v_k' = G_0.$$

So the Gram matrix for Model (4.4) is sparse.

In the CAR step, we focus on Model (4.4). Write $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_p]$ and $G = (1/n)\tilde{X}'\tilde{X}$. Given a threshold $\delta \in (0,1)$, let $\mathscr{G}^\delta$ be the graph with nodes $\{1, 2, \ldots, n\}$ such that nodes i and j are connected by an undirected edge if

$$|G(i,j)|/\sqrt{G(i,i)G(j,j)} > \delta, \qquad 1 \le i \ne j \le p. \tag{4.5}$$

For each variable j, any connected subgraph $\mathscr{I}$ of $\mathscr{G}^\delta$ that contains j is called a "neighborhood" of j. Consider a collection of such local "neighborhoods"

$$\mathscr{A}_{\delta,j}(m) = \big\{ \mathscr{I} \text{ is a connected subgraph of } \mathscr{G}^\delta : j \in \mathscr{I},\ |\mathscr{I}| \le m \big\}, \tag{4.6}$$

where $m \geq 1$ is an integer that controls the maximum size of selected neighborhoods. For $\mathscr{I} \in \mathscr{A}_{\delta,j}(m)$, we measure the "significance" of variable j in $\mathscr{I}$ by

$$T_{j|\mathscr{I}} = \|P_{\mathscr{I}}\tilde{y}\|^2 - \|P_{\mathscr{I}\setminus\{j\}}\tilde{y}\|^2, \tag{4.7}$$

where $P_{\mathscr{I}}\tilde{y}$ is the projection of $\tilde{y}$ onto the space spanned by $\{\tilde{x}_j : j \in \mathscr{I}\}$. We then measure the "significance" of variable j by combining these scores:

$$T_j^* = \max\left\{T_{j|\mathscr{I}} : \mathscr{I} \in \mathscr{A}_{\delta,j}(m)\right\}. \tag{4.8}$$

The scores $T_1^*, T_2^*, \ldots, T_p^*$ are used to rank variables.

FA-CAR has tuning parameters (m, K, δ), but the ideal choice of tuning parameters is insensitive to the unknown β (it mainly depends on the design X). Therefore, tuning here is not as critical as it is for variable selection. In practice, we recommend using $m = 2$ and $\delta = 0.5$ and choosing K as the elbow point in the scree plot of the Gram matrix Θ (see Section 4.3.1).

The computational cost of our method comes from two parts: the SVD on X and the CAR step. SVD is a scalable algorithm even for large matrices [57]. The computational cost of the CAR step is determined by the total number of subsets in the collection $\mathscr{A}_\delta(m) \equiv \cup_{j=1}^p \mathscr{A}_{\delta,j}(m)$. By graph theory [58],

$$|\mathscr{A}_\delta(m)| \leq pm(2.72d_p)^m, \qquad d_p\text{: maximum node degree.}$$

Since $G \approx G_0$ is sparse, d_p grows slowly with p. So the computational cost of the CAR step is only moderately larger than that of MR.

4.1.3 Comparison of the sure-screening model size

We use the blockwise-diagonal example in Section 4.1.1 to demonstrate the advantage of exploiting local covariate structures for ranking. We use the sure-screening model size as the loss function, which is the minimum number of top-ranked variables one needs to select such that all signals are

retained (then, all signal variables will be included in the follow-up lab experiments, say).

We adopt a Rare/Weak (RW) signal model, which has been used a lot in the literature [7, 59].

Fixing $\vartheta \in (0,1)$ and $r > 0$, we assume the vector β is generated from (ν_a: a point mass at a)

$$\beta_j \overset{iid}{\sim} (1-\varepsilon_p)\nu_0 + \frac{\varepsilon_p}{2}\nu_{\tau_p} + \frac{\varepsilon_p}{2}\nu_{-\tau_p}, \quad \varepsilon_p = p^{-\vartheta}, \tau_p = \sigma\sqrt{2r\log(p)/n}. \tag{4.9}$$

Under (4.9), the total number of signals is approximately $s_p \equiv p^{1-\vartheta}$; as p grows, the signals become increasingly sparser. The two parameters (ϑ, r) characterize the signal rareness and signal weakness, respectively. For any threshold $t > 0$, let $FN_p(t) = \sum_{j=1}^{p} \mathbb{P}(\beta_j \neq 0, T_j^* \leq t)$ and $FP_p(t) = \sum_{j=1}^{p} \mathbb{P}(\beta_j = 0, T_j^* > t)$ be the expected number of false negative and false positives, respectively. Define the sure-screening model size as

$$SS_p^*(\vartheta, r, h) = s_p + \min_{t: FN_p(t) < 1} FP_p(t).$$

For the blockwise diagonal design, the Gram matrix is already sparse, so the FA step is not needed. We compare CAR with two other ideas, MR and LSR, where LSR simultaneously runs least-squares on each pair of variables $\{2j-1, 2j\}$ for $j = 1, 2, \ldots, p/2$ and uses these least-squares coefficients to rank variables. We note that the least-squares estimator coincides with the recent de-biased lasso estimator [60, 61] in this design. The following lemma shows that the convergence rate of $SS_p^*(\vartheta, r, h)$ for CAR is always no slower than those of the other two methods.

Lemma 4.1.1 (Sure-screening model size)**.** *Consider Model* (4.1) *with the blockwise-diagonal design as in Section 4.1.1, where the RW model* (4.9) *holds. Let L_p denote a generic multi-*$\log(p)$ *term such that $L_p p^{-\delta} \to 0$ and $L_p p^{\delta} \to \infty$ for all $\delta > 0$. Given any $(\vartheta, r, h) \in (0,1) \times (0,\infty) \times (-1,1)$, for each of the three methods, there is a constant $\eta^*(\vartheta, r, h) \in [0,1]$ such that $SS_p^*(\vartheta, r, h) = L_p p^{\eta^*(\vartheta, r, h)}$. Furthermore, for all (ϑ, r, h),*

$$\eta_{CAR}^*(\vartheta, r, h) \leq \min\big\{\eta_{MR}^*(\vartheta, r, h),\ \ \eta_{LSR}^*(\vartheta, r, h)\big\}.$$

The explicit expression of $\eta^*(\vartheta, r, h)$ for all three methods can be found in Lemma 4.5.1. Using the results there, we can find settings where the convergence rate of CAR is strictly faster; see Table 4.1.

Table 4.1: The exponent $\eta^*(\vartheta, r, h)$ for the blockwise-diagonal design.

$(\vartheta, r, \lvert h \rvert)$	$(.8, 1.5, .4)$	$(.5, 2, .8)$	$(.3, 2, .2)$
CAR	.395	.500	.700
MR	.395	.920	.751
LSR	.543	.980	.700

4.1.4 Connection to the literature

Our method is related to the recent ideas of Graphlet Screening (GS) [62] and Covariate Assisted Screening and Estimation (CASE) [63]. These methods also use Θ to construct a graph and use local graphical structures to improve inference. However, our settings and goals are very different, and our method/theory can not be deduced from previous works: (a) GS and CASE are for variable selection and it is unclear how to use them for variable ranking. (b) GS and CASE have more stringent assumptions on the Gram matrix and do not work for the general designs considered in this paper.

Our FA step is related to the idea of using PCA to remove factor structures in multiple testing [64] and covariance estimation [65], but our FA step is designed for linear models and is thus very different. [66] used PCA to improve marginal screening, which is similar to our FA step; however, their PCA approach is only justified for a random design that comes from an exact factor model, and their theory is insufficient for justifying our FA step.

Our work is related to the literatures on ranking differently expressed genes [67]. Common gene-ranking approaches (e.g., p-value, fold-change) are connected to the idea of Marginal Ranking. The key idea of our method is to exploit correlation structures among variables to improve MR, and an extension of our method (Section 4.4) can be potentially used for gene-ranking. On a high level, our work is also related to feature ranking problem in machine learning [68], but

most methods in these literatures (e.g., wrappers, filters) are algorithm-based and are not designed specifically for linear models.

4.2 Theoretical analysis

We describe the asymptotic settings in Section 4.2.1 and present the main results in Section 4.2.2; our main results contain the rate of convergence of the sure-screening model size. Section 4.2.3 contains some new perturbation bounds for PCA; they are the key for studying Factor Adjusting and are also useful technical tools for other problems. Section 4.2.4 contains the proof of the main result.

4.2.1 Assumptions

We assume Θ has unit diagonals without loss of generality. Let S be the support of β and let $s_p = |S|$. We assume

$$\log(p)/n \to 0, \qquad s_p \leq p^{1-\vartheta} \quad \text{for some } \vartheta \in (0,1). \tag{4.10}$$

Under (4.10), it is known that $n^{-1/2}\sqrt{\log(p)}$ is the minimax order of signal strength for successful variable selection [59]. We focus on the most subtle region that nonzero β_j's are constant multiples of $n^{-1/2}\sqrt{\log(p)}$. Fixing a constant $r > 0$ that calibrates the signal strength and a constant $a > 0$, we assume for any $j \in S$,

$$\tau_p \leq |\beta_j| \leq a\tau_p, \qquad \text{where} \quad \tau_p = n^{-1/2}\sigma\sqrt{2r\log(p)}. \tag{4.11}$$

Model (4.10)-(4.11) is a non-stochastic version of the Rare/Weak signal model in the literatures [69].

The Gram matrix Θ satisfies model (4.2). For any integer $1 \leq m \leq p$ and matrix $\Omega \in \mathbb{R}^{p,p}$, define $\nu_m^*(\Omega)$ as the minimum possible eigenvalue of any $m \times m$ principal submatrix of Ω. Fixing

$\gamma \in (0,1)$, $c_0, C_0 > 0$ and an integer $g \geq 1$, we introduce a class of sparse covariance matrices:

$$\mathscr{M}_p(g,\gamma,c_0,C_0) = \Big\{\Omega \in \mathbb{R}^{p,p} \text{ is p.s.d.} : v_g^*(\Omega) \geq c_0,\ \max_{1\leq i\leq p}\sum_{j=1}^{p}|\Omega(i,j)|^\gamma \leq C_0\Big\}.$$

Recall that G_0 is the sparse component in Model (4.2). We assume

$$G_0 \in \mathscr{M}_p(g,\gamma,c_0,C_0), \quad \lambda_1 \leq c_1\lambda_K, \quad \lambda_K/\max\{s_p,\log(p)\} \to \infty, \tag{4.12}$$

where $c_1 > 0$ is a constant. Fixing a constant $b > 0$, let $\mathscr{G}_0^\delta$ be the undirected graph whose nodes are $\{1,\cdots,p\}$ and there is an edge between nodes i and j if and only if

$$|G_0(i,j)|/\sqrt{G_0(i,i)G_0(j,j)} > \delta_p, \qquad \text{where} \quad \delta_p = b/\log(p).$$

This graph can be viewed as the "oracle" graph, and the graph $\mathscr{G}^\delta$ used in our method is an approximation to $\mathscr{G}_0^\delta$. Let $\mathscr{G}_{0,S}^\delta$ be the induced graph by restricting nodes to S. We assume that for a positive integer $\ell_0 \leq g$,

$$\text{each component of } \mathscr{G}_{0,S}^\delta \text{ consists of } \leq \ell_0 \text{ nodes.}^2 \tag{4.13}$$

This is an assumption on the correlation structures among signal variables. It implies that the signal variables divide into many groups, each consisting of $\leq \ell_0$ variables, such that signals in distinct groups are only weakly correlated after the factors are removed.

FA-CAR has tuning parameters (K,m,δ). We choose K adaptively by

$$K = \hat{K}_p = \max\big\{1 \leq k \leq n : \hat{\sigma}_k^2 > n\log(p)\big\}, \tag{4.14}$$

where $\hat{\sigma}_k$ is the k-th leading singular value of X. We choose (m,δ) such that, for some constant

2. A component is a subgraph in which any two nodes are connected to each other by a path, and which is connected to no additional nodes in the graph.

$C > 0$,

$$\ell_0 \le m \le g, \qquad 1.01\delta_p \le \delta \le C\delta_p, \tag{4.15}$$

where $\delta_p = b/\log(p)$, g and ℓ_0 are defined in (4.12) and (4.13) respectively.

4.2.2 Main result: Sure-screening model size

Given the scores $T_1^*, \ldots, T_p^*$, if we threshold them by

$$t_p(q) = 2q\sigma^2\log(p), \qquad q > 0 \text{ is a constant}, \tag{4.16}$$

the set of retained variables is $\hat{S}(q) = \hat{S}_p(q; X, y) = \{1 \le j \le p : T_j^* > t_p(q)\}$. Recall that S is the support of β and $s_p = |S|$. The (asymptotic) sure-screening model size is defined as

$$SS_p^*(\vartheta, r; X, \beta) = s_p + \min\Big\{\mathbb{E}(|\hat{S}(q) \setminus S|) : q \text{ satisfies } \lim_{p\to\infty} \mathbb{E}(|S \setminus \hat{S}(q)|) = 0\Big\}. \tag{4.17}$$

To describe the asymptotic behavior of SS_p^*, we introduce the quantities $\omega_j(r,m)$, $1 \le j \le p$, where r calibrates the signal strength and $m \ge 1$ is a parameter in FA-CAR. By assumption (4.13), the set of signal variables S has the decomposition $S = \mathscr{I}_1 \cup \mathscr{I}_2 \cup \ldots \cup \mathscr{I}_M$, where nodes in each $\mathscr{I}_k$ form a component of $\mathscr{G}_{0,S}^{\delta}$ and $\max_{1\le k\le M} |\mathscr{I}_k| \le \ell_0$. Fix j. There exists a unique $\mathscr{I}_k$ which contains j. For any $\mathscr{I} \subset \mathscr{I}_k$, let $N = \mathscr{I} \setminus \{j\}$, $F = \mathscr{I}_k \setminus \mathscr{I}$, and $A_{j|\mathscr{I}}^0 = G_0^{j,j} - G_0^{j,N}(G_0^{N,N})^{-1}G_0^{N,j}$. Define

$$\omega_{j|\mathscr{I}}(r; G_0, \beta, \delta_p) = \frac{nA_{j|\mathscr{I}}^0}{2\sigma^2\log(p)}\Big\{\beta_j + (A_{j|\mathscr{I}}^0)^{-1}[G_0^{j,F} - G_0^{j,N}(G_0^{N,N})^{-1}G_0^{N,F}]\beta^F\Big\}^2. \tag{4.18}$$

For each $m \ge 1$, define

$$\omega_j(r, m; G_0, \beta, \delta_p) = \max\Big\{\omega_{j|\mathscr{I}}(r; G_0, \beta, \delta_p) : \mathscr{I} \in \mathscr{A}_{\delta,j}(m), \mathscr{I} \subset \mathscr{I}_k\Big\}. \tag{4.19}$$

We notice that $\omega_j(r,m)$ is a monotone increasing function of m. The following definition is useful:

Definition 4.2.1. L_p, as a positive sequence indexed by p, is called a multi-$\log(p)$ term if for any fixed $c > 0$, $L_p p^c \to \infty$ and $L_p p^{-c} \to 0$ as $p \to \infty$.

The following theorem gives an upper bound for rate of convergence

Theorem 4.2.1 (Sure-screening model size)**.** *Under Model* (4.1)-(4.2)*, suppose* (4.10)-(4.13) *hold for fixed* $(\vartheta, r, g, \ell_0, \gamma, c_0, C_0, c_1, a, b)$ *such that* $g \geq \ell_0$*, and suppose the tuning parameters* (K, m, δ) *satisfy* (4.14)-(4.15)*. Define the constant*

$$q^*(\vartheta, r, m; G_0, \beta, \delta_p) = \inf \left\{ q \geq 0 : \varliminf_{p\to\infty} \frac{\log\left(\sum_{j\in S} p^{-[(\sqrt{\omega_j(r,m)}-\sqrt{q})_+]^2}\right)}{\log(p)} > 0 \right\}.$$

Then, as $p \to \infty$*,*

$$SS_p^*(\vartheta, r; X, \beta) \leq L_p p^{1-\min\{\vartheta,\, q^*(\vartheta,r,m)\}}.$$

We use Theorem 4.2.1 to draw some conclusions. First, we introduce a lower bound for the quantities $\omega_j(r,m)$. Fix j and let $\mathscr{I}_k$ be the component of $\mathscr{G}_{0,S}^{\delta}$ that contains j. Write $N = \mathscr{I}_k \setminus \{j\}$ and $A^0_{j|\mathscr{I}_k} = G_0^{j,j} - G_0^{j,N}(G_0^{N,N})^{-1}G_0^{N,j}$. Define

$$\omega_j^*(r; G_0, \beta, \delta_p) = \frac{nA^0_{j|\mathscr{I}_k}}{2\sigma^2 \log(p)}\beta_j^2. \tag{4.20}$$

This quantity depends on β only through β_j, so there should be no "signal cancellation" involved in our method, as justified in the following corollary.

Corollary 4.2.1 (No signal cancellation)**.** *Suppose the conditions of Theorem 4.2.1 hold. Let* c_0 *be the same as that in* (4.12)*. Then,*

$$SS_p^*(\vartheta, r; X, \beta) \leq L_p p^{1-\min\{\vartheta,\, [(\sqrt{c_0 r}-\sqrt{1-\vartheta})_+]^2\}}.$$

As a result, as long as r is properly large, $SS_p^* \leq L_p s_p$*.*

Due to signal cancellation, no matter how large r is, there still exist choices of the signs and

locations of nonzero β_j's such that MR ranks some signal variables strictly lower than many noise variables and that $SS_p^* \gg L_p s_p$. In contrast, Corollary 4.2.1 demonstrates that FA-CAR successfully overcomes the "signal cancellation" issue.

Next, we compare FA-CAR with an alternative approach which applies MR after the FA step.[3]

Corollary 4.2.2 (Advantage over FA-MR)**.** *Suppose the conditions of Theorem 4.2.1 hold. Let* $\widetilde{SS}_p^*(\vartheta, r; X, \beta)$ *be the sure-screening model size for FA-MR. Then,*

$$SS_p^*(\vartheta, r; X, \beta) \leq L_p \cdot \widetilde{SS}_p^*(\vartheta, r; X, \beta).$$

Corollary 4.2.2 demonstrates that FA-CAR is always no worse than FA-MR. Additionally, we have seen examples in Section 4.1.3 where FA-CAR is strictly better. This justifies the need of exploiting local covariate structures.

4.2.3 Perturbation bounds for PCA

The success of the FA step relies on a tight bound for $\|G - G_0\|_{\max}$. To bound this quantity, we need develop to new perturbation results for PCA. We can rewrite

$$G - G_0 = \sum_{k=1}^{K} (\hat{\sigma}_k^2/n)\hat{v}_k\hat{v}_k' - \sum_{k=1}^{K} \lambda_k v_k v_k',$$

where v_k and $\hat{v}_k$ are the k-th eigenvector of $(\Theta - G_0)$ and G_0, respectively. In the simplest case of $K = 1$, the problem reduces to deriving a sharp bound for $\|\hat{v}_1 - v_1\|_\infty$. Unfortunately, the standard tool of sine-theta theorem [70] only yields a bound for $\|\hat{v}_1 - v_1\|$, which is often too loose if used as a bound for $\|\hat{v}_1 - v_1\|_\infty$. We need the following lemma:

Lemma 4.2.1 (Perturbation of leading eigenvector)**.** *Consider* $\Theta = \lambda_1 v_1 v_1' + G_0$*, where* $\lambda_1 > 0$*,* $\|v_1\| = 1$*, and* $G_0 \in \mathbb{R}^{p,p}$ *is symmetric. Let* $\hat{v}_1$ *be the leading eigenvector of* Θ*. If* $3\|G_0\|_\infty \leq \lambda_1$*,*

3. Since the Gram matrix for Model (4.4) has unequal diagonals, we first normalize the columns of $\tilde{X}$ to have the same ℓ^2-norm and then apply MR.

then

$$\min\{\|\hat{v}_1 - v_1\|_\infty, \|\hat{v}_1 + v_1\|_\infty\} \le 12\lambda_1^{-1}\|G_0\|_\infty\|v_1\|_\infty.$$

We compare it with the sine-theta theorem, which gives that $\min\{\|\hat{v}_1 - v_1\|_\infty, \|\hat{v}_1 + v_1\|_\infty\} \le \min\{\|\hat{v}_1 - v_1\|, \|\hat{v}_1 + v_1\|\} \le C\lambda_1^{-1}\|G_0\|$. Consider a case where each row of G_0 has at most d_p nonzero entries. Since $\|G_0\|_\infty \le \sqrt{d_p}\|G_0\|$, our bound is sharper if $\sqrt{d_p}\|v_1\|_\infty = o(1)$.

For the case $K > 1$, the eigenvectors are generally not unique (unless all the eigenvalues are distinct from each other). It makes more sense to bound $\|\sum_{k=1}^K \hat{v}_k\hat{v}_k' - \sum_{k=1}^K v_k v_k'\|_{\max}$. We have the following theorem:

Theorem 4.2.2. *Let $\Theta = \sum_{k=1}^K \lambda_k v_k v_k' + G_0$, where $\lambda_1 \ge \lambda_2 \ge \cdots \lambda_K > 0$, $v_1, \cdots, v_K \in \mathbb{R}^p$ are unit-norm, mutually orthogonal vectors, and $G_0 \in \mathbb{R}^{p,p}$ is symmetric. For $1 \le k \le K$, let $(\hat{\lambda}_k, \hat{v}_k)$ be the k-th leading eigenvalue and associated eigenvector of Θ. Write $V = [v_1, \cdots, v_K]$, $\hat{V} = [\hat{v}_1, \cdots, \hat{v}_K]$ and $G = \Theta - \sum_{k=1}^K \hat{\lambda}_k\hat{v}_k\hat{v}_k'$. If $\lambda_K > C_1\|G_0\|_\infty$ for some constant $C_1 > 2$, then*

$$\|VV' - \hat{V}\hat{V}'\|_{\max} \le C_2(\lambda_1/\lambda_K)^2 \cdot \lambda_K^{-1}\|G_0\|_\infty \cdot \max_{1\le k\le K}\|v_k\|_\infty^2,$$

and

$$\|G - G_0\|_{\max} \le C_2'(\lambda_1/\lambda_K)^2 \cdot \|G_0\|_\infty \cdot \max_{1\le k\le K}\|v_k\|_\infty^2,$$

where $C_2, C_2' > 0$ are constants that only depend on (C_1, K).

The proof of Lemma 4.2.1 uses a similar approach as the proof of Lemma 3.1 in [71]. The proof of Theorem 4.2.2 is new and highly non-trivial since we do not assume any gap between $\lambda_1, \ldots, \lambda_K$. That it requires no eigen-gaps makes this result very different from other recent perturbation results (e.g., [72]).

4.2.4 Proof of Theorem 4.2.1

Recall that $G = (1/n)\tilde{X}'\tilde{X}$ is the Gram matrix of Model (4.4). Using the results in Section 4.2.3, we can show that G is entry-wise close to G_0:

Lemma 4.2.2. *Suppose the conditions of Theorem 4.2.1 hold. Then,* $\|G - G_0\|_{\max} = o(\delta_p)$ *and* $s_p\|G - G_0\|_{\max} = o(1)$.

The key of the proof is to study the distribution of $T_{j|\mathscr{I}}$, for each $j \in S$ and $\mathscr{I} \in \mathscr{A}_{\delta,j}(m)$. The following lemma is proved in Section 4.5.

Lemma 4.2.3. *Suppose the conditions of Theorem 4.2.1 hold. Fix* $j \in S$ *and let* $\mathscr{I}^{(j)}$ *be the unique component of* $\mathscr{G}^{\delta}_{0,S}$ *that contains* j*. For any* $\mathscr{I} \subset \mathscr{I}^{(j)} \cap \mathscr{G}^{\delta}_{S}$ *that contains* j*,*

$$T_{j|\mathscr{I}} = (W + \Delta)^2, \quad W \sim \mathscr{N}\left(\sqrt{2\omega_{j|\mathscr{I}}(r)\sigma^2 \log(p)},\, \sigma^2\right),\ |\Delta| = o_P\left(\sqrt{\log(p)}\right),$$

where $\omega_{j|\mathscr{I}}(r)$ *is the same as that in Section 4.2.2.*

The proof of Lemma 4.2.3 is lengthy, and we provide some illustration. For simplicity, we only consider a special case where $\mathscr{I}$ is exactly the component of $\mathscr{G}^{\delta}_{S}$ that contains j. By definition and elementary calculations,

$$T_{j|\mathscr{I}} = n^{-1}(\eta^{\mathscr{I}})'\left((G^{\mathscr{I},\mathscr{I}})^{-1} - \begin{bmatrix} (G^{N,N})^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right)\eta^{\mathscr{I}}, \quad \eta = \tilde{X}'\tilde{y},\ N = \mathscr{I} \setminus \{j\}, \tag{4.21}$$

Since $\eta \sim \mathscr{N}(nG\beta, \sigma^2 nG)$, we have

$$n^{-1}\mathbb{E}[\eta^{\mathscr{I}}] = (G\beta)^{\mathscr{I}} = G^{\mathscr{I},\mathscr{I}}\beta^{\mathscr{I}} + G_0^{\mathscr{I},\mathscr{I}^c}\beta^{\mathscr{I}^c} + (G - G_0)^{\mathscr{I},\mathscr{I}^c}\beta^{\mathscr{I}^c}.$$

It can be proved that the third term is negligible as a result of Lemma 4.2.2 and the second term is negligible due to the sparsity of G_0 and the definition of the graph $\mathscr{G}^{\delta}$. It follows that $E[\eta^{\mathscr{I}}] \approx nG^{\mathscr{I},\mathscr{I}}\beta^{\mathscr{I}}$. We plug it into (4.21) and find that

$$T_{j|\mathscr{I}} \approx n(\beta^{\mathscr{I}})'\left[(G^{\mathscr{I},\mathscr{I}})^{-1} - G^{\mathscr{I},N}(G^{N,N})^{-1}G^{N,\mathscr{I}}\right]\beta^{\mathscr{I}} = nA_{j|\mathscr{I}}\beta_j^2,$$

where $A_{j|\mathscr{I}} = G^{j,j} - G^{j,N}(G^{N,N})^{-1}G^{N,j}$ is a counterpart of $A^0_{j|\mathscr{I}_k}$ in (4.20) and the last equality

is a result of the matrix inverse formula in linear algebra. It remains to characterize the difference between $A_{j|\mathscr{I}}$ and $A^0_{j|\mathscr{I}_k}$; recall that $\mathscr{I}_k$ is the unique component of $\mathscr{G}^{\delta}_{0,S}$ that contains j. Using Lemma 4.2.2, we can prove that, if we restrict $\mathscr{G}^{\delta}_{0,S}$ to $\mathscr{I}_k$, it splits into a few components and one component is exactly $\mathscr{I}$. Such an observation allows us to show that

$$A_{j|\mathscr{I}} = A^0_{j|\mathscr{I}_k}[1+o(1)].$$

The proof of Lemma 4.2.3 follows a similar idea as the above derivation but is much more complicated.

Once we have the distribution of $T_{j|\mathscr{I}}$, we can quantify the type I and type II errors associated with any threshold $t_p(q)$.

Lemma 4.2.4 (Type I and Type II errors)**.** *Suppose the conditions of Theorem 4.2.1 hold. Consider $\hat{S}(q)$, the set of selected variables associated with the threshold $t_p(q)$ as in* (4.16)*. Then,*

$$\mathbb{E}(|S\backslash\hat{S}(q)|) \le L_p \sum_{j\in S} p^{-[(\sqrt{\omega_j(r,m)}-\sqrt{q})_+]^2},$$

and

$$\mathbb{E}(|\hat{S}(q)\backslash S|) \le Cs_p[\log(p)]^{\gamma m} + L_p p^{1-q}.$$

where $\omega_j(r,m)$ is as in (4.19) *and γ is the same as that in $\mathscr{M}_p(g,\gamma,c_0,C_0)$.*

We now derive the upper bound for $SS^*_p(\vartheta,r;X,\beta)$. By Lemma 4.2.4 and the definition of $q^*(\vartheta,r,m)$, for any $q < q^*(\vartheta,r,m)$, there is an $\varepsilon > 0$ such that

$$\mathbb{E}(|S\backslash\hat{S}(q)|) \le L_p p^{-\varepsilon} \to 0, \qquad \text{for all sufficiently large } p.$$

As a result, for any $q < q^*(\vartheta,r,m)$,

$$SS^*_p(\vartheta,r;X,\beta) \le s_p + \mathbb{E}(|\hat{S}(q)\backslash S|) \le L_p p^{1-\min\{\vartheta,q\}}.$$

Taking the limit of $q \to q^*(\vartheta, r, m)$ gives the claim of Theorem 4.2.1.

4.3 Empirical analysis

4.3.1 Simulation study

We investigate the performance of FA-CAR in simulations. In all experiments below, given a covariance matrix $\Sigma \in \mathbb{R}^{p,p}$, the rows of X are independently sampled from $\mathcal{N}(0,\Sigma)$. We consider four different types of designs where the corresponding Σ is:

- *Tridiagonal.* $\Sigma(j,j) = 1$ for all $1 \le j \le p$, and $\Sigma(i,j) = \rho \cdot 1\{|i-j| = 1\}$ for any $1 \le i \ne j \le p$. We set $\rho = 0.5$.
- *Autoregressive.* $\Sigma(i,j) = \rho^{|i-j|}$ for all $1 \le i, j \le p$. We set $\rho = 0.6$.
- *Equal correlation.* $\Sigma(j,j) = 1$ for $1 \le j \le p$, and $\Sigma(i,j) = \rho$ for $1 \le i \ne j \le p$. We set $\rho = 0.6$.
- *Two factors.* $\Sigma = \frac{\rho}{2} a_1 a_1' + \frac{\rho}{2} a_2 a_2' + (1-\rho)\Sigma_1$, where $a_1 = (1,1,\ldots,1)'$, $a_2 = (1,-1,1,-1,\ldots,1,-1)'$ and Σ_1 is an autoregressive covariance matrix, i.e., $\Sigma_1(i,j) = \rho_1^{|i-j|}$. We set $\rho = 0.5$ and $\rho_1 = 0.6$.

Fixing (n, p, η, s), we generate β as follows: The first s coordinates of β are independently sampled from $\mathcal{N}(0, \eta^2)$, the other coordinates all equal to 0. We then generate y using Model (4.1) with $\sigma^2 = 1$.

Our method has three tuning parameters (K, δ, m). In Experiments 1-2, we set $m = 2$ and $\delta = 0.5$, and use the ideal choice of K, that is, $K = 0, 0, 1, 2$ for the above four types of designs. In Experiment 3, we investigate the sensitivity of our method to tuning parameters.

Experiment 1: Comparison of ROC curves We compare the ROC curves of our method and three other methods: (1) Marginal Ranking (MR) [73], (2) HOLP [74], which uses the coordinates

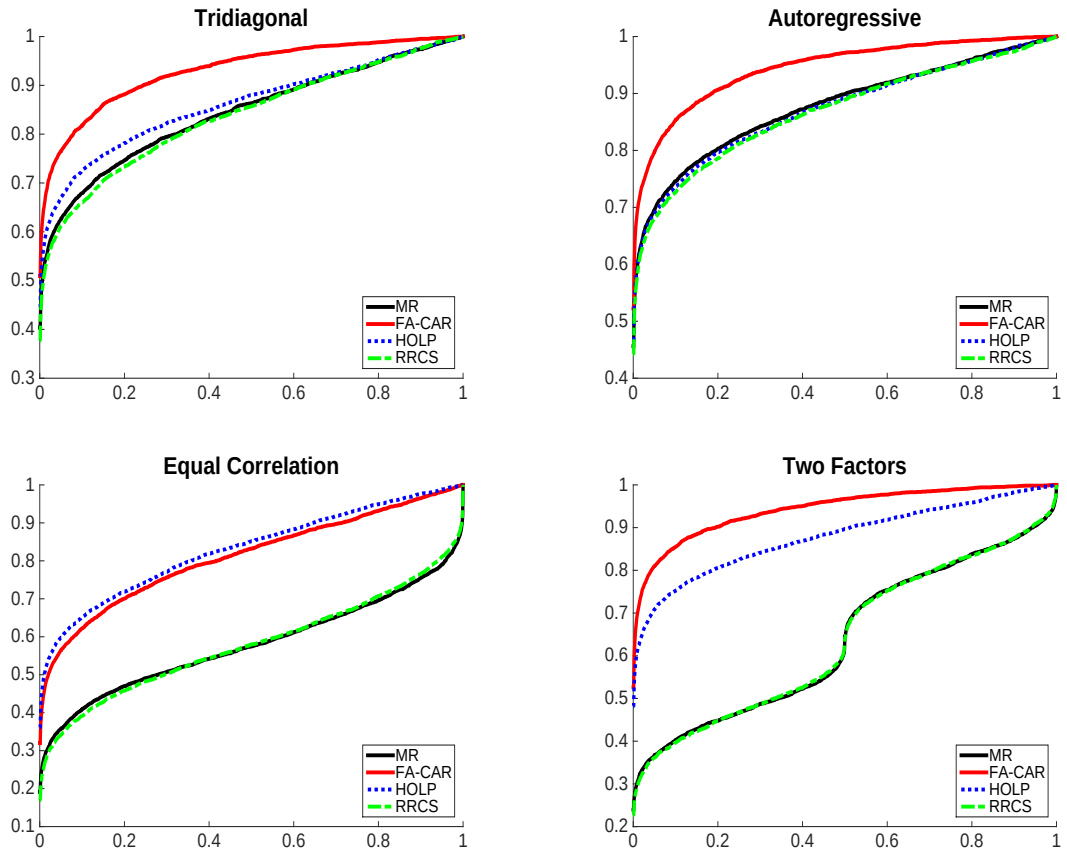


Figure 4.1: ROC curves in Experiment 1. $(n, p, \eta, s) = (200, 1000, 3, 20)$.

of $X'(XX')^{-1}y$ for ranking, and (3) RRCS [75] which uses the marginal Kendall's τ correlation coefficients for ranking. Fix $(n, p, \eta, s) = (200, 1000, 3, 20)$. For each of the four design types, we generate 200 datasets and output the average ROC curves of these 200 repetitions. The results are displayed in Figure 4.1. For the tridiagonal, autoregressive, and two-factor designs, our method significantly outperforms the other methods. For the equal correlation design, our method significantly outperforms MR and RRCS and is similar to HOLP.

Experiment 2: Various (n, p, η, s) We consider four choices of (n, p, η, s), for each of the four types of designs. There are 16 different settings in total. We measure the performance of different methods using several criteria: (a) Sure screening probability (SP): the probability that all the signal variables are selected when retaining n variables in total. (b) Type II: the number of type II errors when retaining n variables in total. (c) Sure screening model size (Size): the minimum

number L such that all signal variables are selected when retaining L variables in total. The results are shown in Table 4.2.

Table 4.2: Results of Experiment 2. For Type II, we report the mean over 200 repetitions, and for Size, we report the median over 200 repetitions.

Designs	Setting: (n,p,η,s)	Measure	Method			
			FA-CAR	MR	HOLP	RRCS
Tridiag.	(200,1000,3,5)	SP/Type II	**0.91/0.11**	0.45/0.64	0.51/0.58	0.45/0.65
		Size	**6**	246	195	247
	(200,1000,3,20)	SP/Type II	**0.11/2.45**	0.01/5.19	0.01/4.49	0.01/5.47
		Size	**518.5**	865.5	861	874.5
	(200,1000,0.5,5)	SP/Type II	**0.73/0.35**	0.38/0.86	0.36/0.94	0.35/0.95
		Size	**39.5**	336.5	384.5	382
	(200,5000,0.5,5)	SP/Type II	**0.57/0.66**	0.20/1.39	0.18/1.42	0.17/1.47
		Size	**132.5**	1789.5	1942	1964
Autoreg.	(200,1000,3,5)	SP/Type II	**0.95/0.06**	0.67/0.43	0.62/0.46	0.67/0.44
		Size	**6**	65	85	58.5
	(200,1000,3,20)	SP/Type II	**0.17/2.00**	0.02/4.05	0.00/4.19	0.01/4.37
		Size	**422**	840	848	850.5
	(200,1000,0.5,5)	SP/Type II	**0.79/0.28**	0.53/0.67	0.42/0.83	0.51/0.77
		Size	**22**	179.5	293	184
	(200,5000,0.5,5)	SP/Type II	**0.6/0.61**	0.35/1.18	0.35/1.20	0.315/1.29
		Size	**57**	937	980	1077
Equal corr.	(200,1000,3,5)	SP/Type II	0.46/0.72	0.06/1.85	**0.46/0.68**	0.07/1.84
		Size	247	998	**230**	997
	(200,1000,3,20)	SP/Type II	0.00/6.09	0.00/10.66	**0.00/5.73**	0.00/10.86
		Size	909.5	1000	**863.5**	1000
	(200,1000,0.5,5)	SP/Type II	0.16/**1.38**	0.05/2.06	**0.17**/1.47	0.03/2.08
		Size	**577**	969	589	957
	(200,5000,0.5,5)	SP/Type II	**0.07/2.00**	0.00/2.65	0.06/2.03	0.00/2.72
		Size	**2600.5**	4856	2689.5	4844.5
Two factors	(200,1000,3,5)	SP/Type II	**0.93/0.09**	0.16/1.83	0.62/0.47	0.17/1.82
		Size	**6**	690.5	88.5	687.5
	(200,1000,3,20)	SP/Type II	**0.21/2.03**	0.01/11.06	0.02/3.99	0.01/11.08
		Size	**454**	988.5	832.5	983.5
	(200,1000,0.5,5)	SP/Type II	**0.73/0.43**	0.17/1.94	0.38/1.06	0.16/1.95
		Size	**43.5**	674.5	387	678
	(200,5000,0.5,5)	SP/Type II	**0.47/0.89**	0.08/2.46	0.28/1.45	0.07/2.49
		Size	**274**	3592.5	1382	3633

Experiment 3: Sensitivity to tuning parameters. We study how the performance of FA-CAR changes as the tuning parameters (K,δ,m) vary. We fix $(n,p,\eta,s)=(200,1000,0.5,5)$, and focus on the autoregressive designs and two-factor designs. We implement FA-CAR for $K\in\{0,1,2,3\}$, $m\in\{2,3\}$ and $\delta\in\{.2,.25,.3,.35,\cdots,.9\}$. The results are shown in Figures 4.2; to save space,

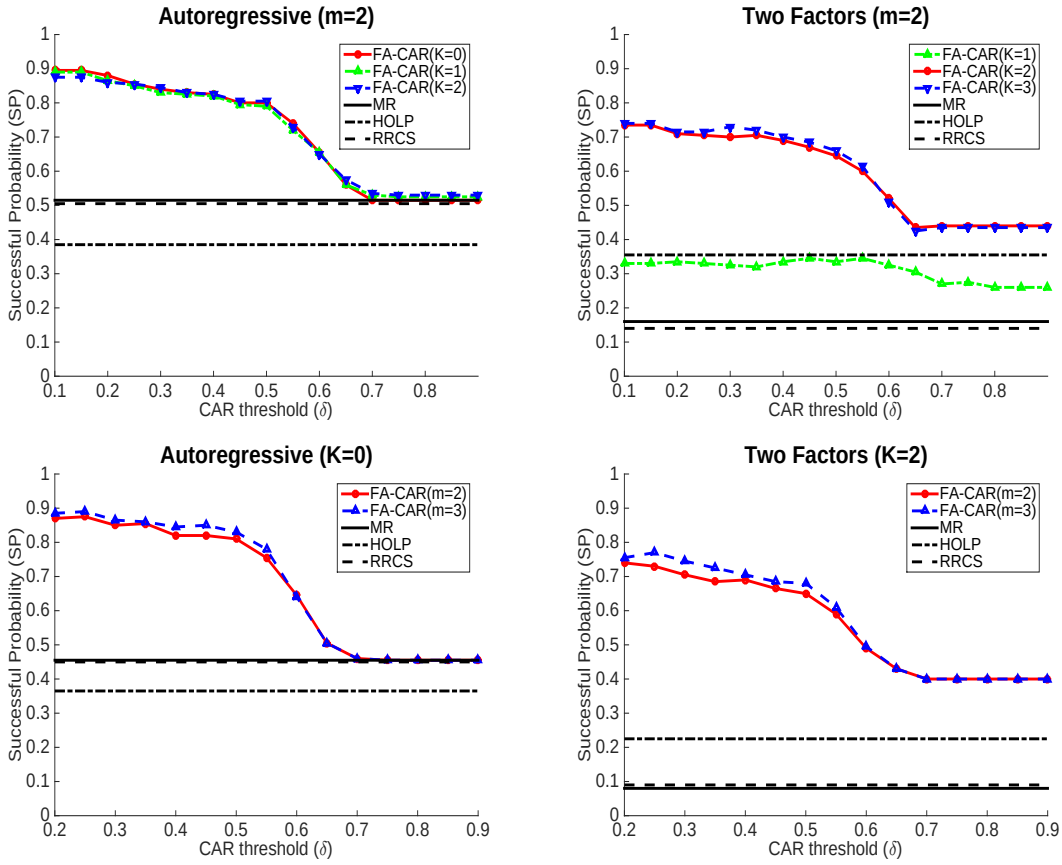


Figure 4.2: Experiment 3: sensitivity to tuning parameters. The ideal choice of K is $K = 0$ for the autoregressive design and $K = 2$ for the two-factor design.

we only report the sure screening probability (SP). We also report the computing time for different values of δ in Figure 4.3.

Choice of K. The top two panels of Figure 4.2 suggest that overshooting of K makes almost no difference in the performance, but undershooting of K could render the performance worse (e.g., $K = 1$ for the two factors design). Even with an undershooting K, FA-CAR still significantly outperforms MR and RRCS, and is comparable with HOLP for a wide range of δ.

Choice of m. The bottom two panels of Figure 4.2 suggest that increasing m from 2 to 3 slightly improves the performance especially when δ is small, but we pay a price in computational cost. In general, $m = 2$ is a good choice.

Choice of δ. From Figure 4.3, smaller δ tends to yield better performance of FA-CAR; but as long as $\delta < 0.5$, the performance is more or less similar (and is much better than the other

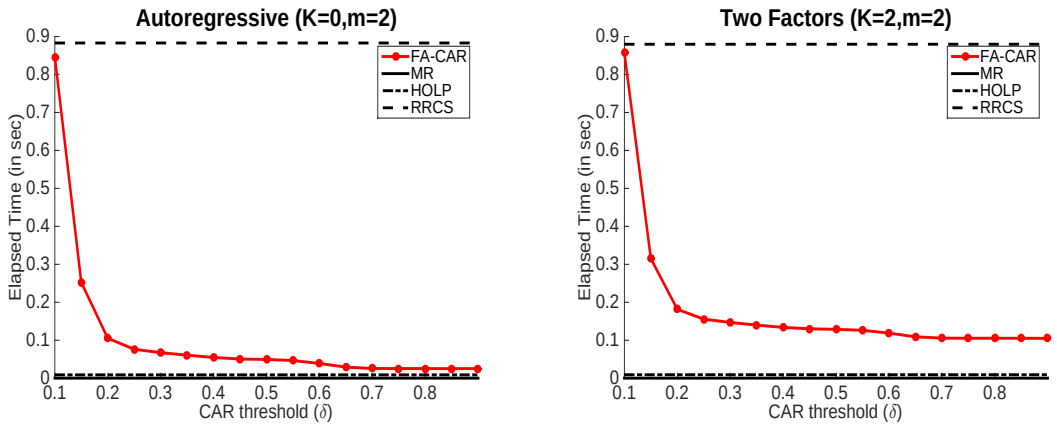


Figure 4.3: Computing time in Experiment 3.

methods). From Figure 4.3, the computing time decreases as δ increases. Combining Figures 4.2-4.3, we find that $\delta = 0.5$ achieves a good balance between statistical accuracy and computational cost.

4.3.2 Application to a microarray dataset

We investigate the performance of our method using a gene microarray dataset [76]. It contains the gene expressions of human immortalized B cells for $p = 4238$ genes and $n = 148$ subjects (CEPH-Utah subpopulation). We use this $n \times p$ data matrix as the design. Figure 4.4 compares the two respective Gram matrices for Model (4.1) and Model (4.4), and it shows that the Gram matrix for Model (4.4) is much sparser. This suggests that our assumption (4.2) fits the data well and that the FA step is effective in removing the factors.

In our experiment, fixing parameters (η, s), we first generate β by drawing its first s coordinates $i.i.d$ from $N(0, \eta^2)$ and setting the other coordinates to be 0 and then generate y using Model (4.1) with $\sigma = 1$. Here, (η, s) control the signal strength and signal sparsity, respectively. For each method, we report the average ROC curves over 200 repetitions; the results are displayed in Figure 4.5. When $s = 50$, FA-CAR always yields the best performance, and it is especially advantageous when η is small (i.e., the signals are “weak"). When $s = 10$, FA-CAR performs

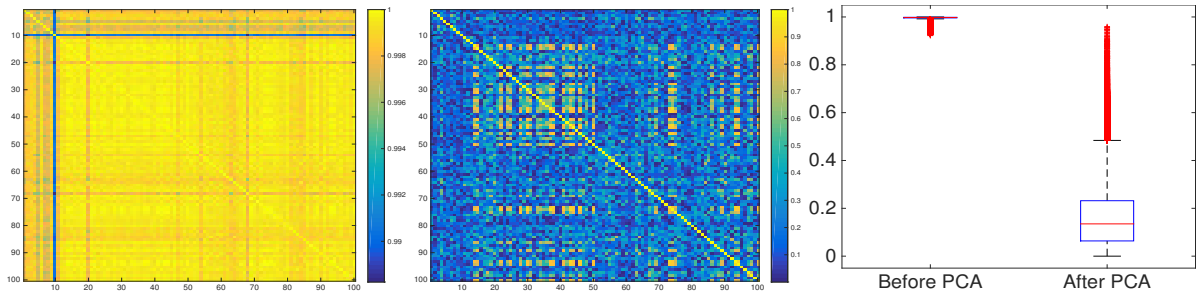


Figure 4.4: Left two panels: the Gram matrix before and after Factor Adjusting (for presentation purpose, both matrices have been normalized so that the diagonals are 1; only the upper left 100×100 block is displayed). Right panel: Boxplots of the off-diagonal entries (in absolute value) of two Gram matrices.

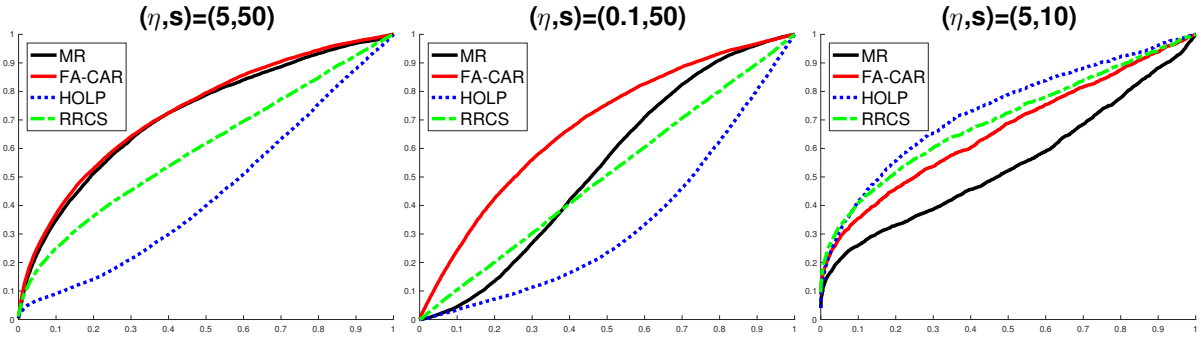


Figure 4.5: The ROC. Design: gene-microarray. The curves are averaged over 200 repetitions.

reasonably well, and it is better than MR. It is a little worse than HOLP and RRSC, but these two methods are unsatisfactory in the other settings. In terms of the overall performance, we conclude that FA-CAR is the best among the four methods.

4.4 Extension to generalized linear models

In bioinformatics and machine learning, it is often the case that the responses are not continuous, and the generalized linear models (GLM) is more appropriate for modeling the data. Consider a GLM with the canonical link: The responses $y_1, \ldots, y_n$ are independent of each other, and each y_i has a probability density from the exponential family:

$$f(y_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}, \qquad 1 \le i \le n,$$

where $b(\cdot)$ and $c(\cdot)$ are known functions satisfying that $E[y_i] = b'(\theta_i)$. The parameter θ_i is called the canonical or natural parameter. GLM models that

$$\theta_i = \theta_i(X_i) = \beta_0 + X_i'\beta, \qquad 1 \le i \le n.$$

The parameters $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are unknown.[4] Same as before, we call a nonzero entry of β a "signal". We are interested in ranking the variables such that the top ranked variables contain as many signals as possible.

Marginal Ranking (MR) can be conveniently extended to GLM, where the marginal correlation coefficients are replaced by the maximum marginal likelihoods or maximum marginal likelihood estimators [77, 78]. However, signal cancellation problem exists in GLM as well and it may hurt the performance of MR. In the following sections, we consider a random design and use logistic regression as an example to illustrate this point.

4.4.1 Signal cancellation in logistic regression

We consider a logistic regression with random design. Suppose Y follows Bernoulli distribution, with

$$\mathbf{E}(Y|X_1,\cdots,X_p) = g^{-1}(\eta) \qquad \eta = \beta_0 + \sum_{i=1}^{p} \beta_i X_i = \beta_0 + \beta^T X$$

where

$$g(x) = \log(\frac{x}{1-x}) \quad \text{with} \quad g^{-1}(x) = \frac{e^x}{1+e^x}$$

We assume $X = (X_1,\ldots,X_p)$ are jointly normal distributed with mean $\mathbf{0}$ and covariance matrix Σ.

Here we consider the omitted variable case where we model

$$\eta = \beta_0^* + \sum_{i=1}^{d} \beta_i^* X_i, \quad d < p \tag{4.22}$$

In marginal ranking, $d = 1$ and $|\beta_1^*|$ is the rank score for the first variable.

4. We can also add a dispersion parameter σ^2 and all the results continue to hold.

We write

$$\begin{pmatrix} X_{1:d} \\ X_{(d+1):p} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

The following theorem characterize the population quantity $\beta^* = (\beta_0^*, \beta_{1:d}^*)^T$, where β^* is defined as the limit which MLE for misspecified model converges to as sample size goes to infinity [5].

Theorem 4.4.1. *Under misspecified model* (4.22)*, the population estimate* β^* *satisfies*

$$\beta_{1:d}^* = \left(\beta_{1:d} + \Sigma_{11}^{-1}\Sigma_{12}\beta_{(d+1):p}\right) \frac{\mathbf{E}\left[(g^{-1})'\left(\beta_0 + \sum_{i=1}^p \beta_i X_i\right)\right]}{\mathbf{E}\left[(g^{-1})'\left(\beta_0^* + \sum_{i=1}^d \beta_i^* X_i\right)\right]} \tag{4.23}$$

As a corollary, we know β_j^* always have the same sign as the jth coordinate of $\beta_{1:d} + \Sigma_{11}^{-1}\Sigma_{12}\beta_{(d+1):p}$. Due to the presence of term $\Sigma_{11}^{-1}\Sigma_{12}\beta_{(d+1):p}$ it is possible that $\beta_j = 0$ while $\beta_j^* \neq 0$, and vice versa. Therefore, Theorem 4.4.1 indicates that signal cancellation does exist for logistic regression at least in the population level.

4.4.2 FA-CAR in GLM

We now describe the GLM version of FA-CAR. In the FA step, let $X = \sum_{k=1}^n \hat{\sigma}_k \hat{u}_k \hat{v}_k'$ be the SVD of X and define

$$\tilde{X} = X - \sum_{k=1}^K \hat{\sigma}_k \hat{u}_k \hat{v}_k'. \tag{4.24}$$

Write $\hat{U} = [\hat{u}_1, \cdots, \hat{u}_K]$. Let $\hat{U}_i'$ and $\tilde{X}_i'$ be the i-th row of $\hat{U}$ and $\tilde{X}$, respectively, $1 \leq i \leq n$. We consider a new GLM where $y_1, \ldots, y_n$ are independent and each y_i has the probability density

$$f(y_i) = \exp\{y_i\tilde{\theta}_i - b(\tilde{\theta}_i) + c(y_i)\}, \qquad \tilde{\theta}_i = \beta_0 + \hat{U}_i'\alpha + \tilde{X}_i'\beta. \tag{4.25}$$

5. Here we consider the classical setting where p and d are fixed.

The log-likelihood of the new GLM is

$$\ell(\beta_0,\beta;y,\tilde{X},\hat{U})=\sum_{i=1}^{n}\left[\left(\beta_0+\hat{U}_i'\alpha+\tilde{X}_i'\beta\right)y_i-b\left(\beta_0+\hat{U}_i'\alpha+\tilde{X}_i'\beta\right)+c(y_i)\right].$$

In the special case of linear models (we assume $\beta_0=0$), Model (4.25) becomes $y=\hat{U}\alpha+\tilde{X}\beta+\mathcal{N}(0,I_n)$. Since $\alpha\in\mathbb{R}^K$ is low-dimensional and the columns of $\hat{U}$ are orthogonal to the columns of $\tilde{X}$, we can regress y on $\hat{U}$ only to get the least-squares estimator $\hat{\alpha}^{ols}$ and subtract $\hat{U}\hat{\alpha}^{ols}$ from y. This gives $\tilde{y}$. So we have recovered Model (4.4).

In the CAR step, we introduce a "local log-likelihood" for each subset $V\subset\{1,\ldots,p\}$:

$$\ell(\alpha,\beta_0,\beta_V;y,\tilde{X},\hat{U})=\sum_{i=1}^{n}\left[\left(\beta_0+\hat{U}_i'\alpha+\tilde{X}_{i,V}'\beta_V\right)y_i-b\left(\beta_0+\hat{U}_i'\alpha+\tilde{X}_{i,V}'\beta_V\right)+c(y_i)\right],$$

where $\tilde{X}_{i,V}$ is obtained from restricting $\tilde{X}_i$ to the coordinates in V. Define the maximum partial log-likelihood as

$$\hat{\ell}_V(y,\tilde{X},\hat{U})=\max_{\alpha,\beta_0,\beta_V}\ell(\alpha,\beta_0,\beta_V;y,\tilde{X},\hat{U}).$$

This quantity $\hat{\ell}_V(y,\tilde{X},\hat{U})$ serves as a counterpart of $\|P_V\tilde{y}\|^2$ in the case of linear models. We then introduce a counterpart of $T_{j|\mathscr{I}}$ for GLM:

$$T_{j|\mathscr{I}}^{glm}=\hat{\ell}_{\mathscr{I}}(y;\tilde{X},\hat{U})-\hat{\ell}_{\mathscr{I}\setminus\{j\}}(y,\tilde{X},\hat{U}). \tag{4.26}$$

Let $\mathscr{G}^{\delta}$ and $A_{\delta,j}(m)$ be the same as in (4.5) and (4.6). The final scores are

$$T_j^*=\max\left\{T_{j|\mathscr{I}}^{glm}:\mathscr{I}\in\mathscr{A}_{\delta,j}(m)\right\}. \tag{4.27}$$

We use a numerical example to compare FA-CAR with two GLM versions of MR: MR-1 [77] uses the maximum marginal likelihood estimator to rank variables, and MR-2 [78] uses the maximum marginal log-likelihood to rank variables. We are not aware of any direct extensions of HOLP and RRCS for GLM, so we omit the comparison with them. Fixing $(n,p,\eta,s)=(200,1000,3,5)$,

we generate the designs similarly as in Section 4.3.1 and generate binary y_i's using the logistic regression setting. In Table 4.3, we report the performance of three methods, where the measures are the same as those in Experiment 2 in Section 4.3.1. It suggests a significant advantage of FA-CAR over the other two methods.

Table 4.3: Comparison of ranking methods for the logistic regression. The measures, SP, Type II, and Size, are defined the same as those in Table 4.2.

Designs	Measure	Method		
		FA-CAR	MR-1	MR-2
Tridiagonal	SP/Type II	**0.87/0.15**	0.49/0.62	0.49/0.62
	Size	**13**	225	226.5
Autoregressive	SP/Type II	**0.84/0.20**	0.55/0.59	0.55/0.59
	Size	**11.5**	160	159.5
Equal corr.	SP/Type II	**0.26/1.10**	0.05/2.03	0.05/2.02
	Size	**510**	977	980
Two factors	SP/Type II	**0.86/0.22**	0.13/1.95	0.13/1.94
	Size	**22**	709.5	707.5

4.5 Proofs

4.5.1 Proofs of Theorems

Proof of Theorem 4.2.2 As preparation, we introduce $\tilde{v}_k$, defined in (4.29), as a counterpart of $\hat{v}_k$ for $1 \le k \le K$. By Weyl's inequality, for any $1 \le k \le K$, $|\hat{\lambda}_k - \lambda_k| \le \|G_0\| \le \|G_0\|_\infty \le C_1^{-1}\lambda_K \le C_1^{-1}\lambda_k$. It follows that

$$\frac{C_1 - 1}{C_1}\lambda_k \le \hat{\lambda}_k \le \frac{C_1 + 1}{C_1}\lambda_k \tag{4.28}$$

Write $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_K)$ and $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_K)$. Recall that $V = [v_1, \cdots, v_K]$ and $\hat{V} = [\hat{v}_1, \cdots, \hat{v}_K]$. By definition,

$$\hat{\lambda}_k \hat{v}_k = \Theta \hat{v}_k = (V\Lambda V' + G_0)\hat{v}_k,$$

which implies $(\hat{\lambda}_k I_p - G_0)\hat{v}_k = V\Lambda V'\hat{v}_k$. By (4.28), $(\hat{\lambda}_k I_p - G_0)$ is positive definite. Hence,

$$\hat{v}_k = (I_p - \hat{\lambda}_k^{-1} G_0)^{-1}\tilde{v}_k, \qquad \text{where} \quad \tilde{v}_k \equiv \hat{\lambda}_k^{-1}(V\Lambda V')\hat{v}_k. \tag{4.29}$$

Write $\tilde{V} = [\tilde{v}_1, \cdots, \tilde{v}_K]$.

We now show the first claim about $\|\hat{V}\hat{V}' - VV'\|_{\max}$. It is seen that

$$\begin{aligned}\|\hat{V}\hat{V}' - VV'\|_{\max} &\le \|\hat{V}\hat{V}' - \tilde{V}\tilde{V}'\|_{\max} + \|\tilde{V}\tilde{V}' - VV'\|_{\max} \\ &\le \sum_{k=1}^{K} \|\hat{v}_k\hat{v}_k' - \tilde{v}_k\tilde{v}_k'\|_{\max} + \|\tilde{V}\tilde{V}' - VV'\|_{\max} \equiv I + II. \end{aligned} \tag{4.30}$$

First, we bound I. For $1 \le k \le K$, letting $\Delta_k = (I_p - \hat{\lambda}_k^{-1} G_0)^{-1} - I_p$, we have $\|\hat{v}_k - \tilde{v}_k\|_\infty \le \|\Delta_k\|_\infty \|\tilde{v}_k\|_\infty$ and

$$\begin{aligned}\|\hat{v}_k\hat{v}_k' - \tilde{v}_k\tilde{v}_k'\|_{\max} &\le \|\hat{v}_k - \tilde{v}_k\|_\infty^2 + 2\|\hat{v}_k - \tilde{v}_k\|_\infty \|\tilde{v}_k\|_\infty \\ &\le \|\tilde{v}_k\|_\infty^2 (\|\Delta_k\|_\infty^2 + 2\|\Delta_k\|_\infty) \end{aligned} \tag{4.31}$$

We consider $\|\Delta_k\|_\infty$ and $\|\tilde{v}_k\|_\infty$ separately. Observing that $\Delta_k = \hat{\lambda}_k^{-1} G_0 + \Delta_k \hat{\lambda}_1^{-1} G_0$, we apply the triangle inequality to get $\|\Delta_k\|_\infty \le \hat{\lambda}_k^{-1}\|G_0\|_\infty + \|\Delta_k\|_\infty \hat{\lambda}_k^{-1}\|G_0\|_\infty$. By (4.28) and the assumption, $\hat{\lambda}_k^{-1}\|G_0\|_\infty \le \frac{C_1}{C_1 - 1}\lambda_k^{-1}\|G_0\|_\infty \le \frac{1}{C_1 - 1}$. It follows that

$$\|\Delta\|_\infty \le \frac{\hat{\lambda}_k^{-1}\|G_0\|_\infty}{1 - \hat{\lambda}_k^{-1}\|G_0\|_\infty} \le \frac{C_1 - 1}{C_1 - 2}\frac{\|G_0\|_\infty}{\hat{\lambda}_k} \le \frac{C_1}{C_1 - 2}\frac{\|G_0\|_\infty}{\lambda_k}.$$

Recalling that $\tilde{v}_k = \hat{\lambda}_k^{-1} V \Lambda V' \hat{v}_k$, we have $\|\tilde{v}_k\|_\infty \le \hat{\lambda}_k^{-1}\|V\|_{\max}\|\Lambda V'\hat{v}_k\|_1 \le \hat{\lambda}_k^{-1}\|V\|_{\max} \cdot \sqrt{K}\|\Lambda V'\hat{v}_k\|$. Since $\|V\| = 1$ and $\|\hat{v}_k\| = 1$, $\|\Lambda V'\hat{v}_k\| \le \lambda_1$. It follows that

$$\|\tilde{v}_k\|_\infty \le \sqrt{K}\Big(\frac{\lambda_1}{\hat{\lambda}_k}\Big)\|V\|_{\max} \le \sqrt{K}\frac{C_1}{C_1 - 1}\Big(\frac{\lambda_1}{\lambda_k}\Big)\|V\|_{\max}.$$

Combining the above with (4.31) and noting that $\|\Delta_k\|_\infty^2 \le C\|\Delta_k\|_\infty$, we find that

$$\|\hat{v}_k\hat{v}_k' - \tilde{v}_k\tilde{v}_k'\|_{\max} \le C\lambda_k^{-1}\Big(\frac{\lambda_1}{\lambda_k}\Big)^2 \|V\|_{\max}^2 \|G_0\|_\infty. \tag{4.32}$$

We then bound II. It is seen that

$$\begin{aligned}\tilde{V}\tilde{V}' - VV' &= \sum_{k=1}^{K} \hat{\lambda}_k^{-2} V\Lambda V'\hat{v}_k\hat{v}_k' V\Lambda V' - VV' \\ &= V\Lambda V'\hat{V}\hat{\Lambda}^{-2}\hat{V}'V\Lambda V' - VV' \\ &= V(M'M - I_K)V', \qquad \text{where} \quad M \equiv \hat{\Lambda}^{-1}\hat{V}'V\Lambda. \end{aligned} \tag{4.33}$$

We now derive a bound for $\|M'M - I_K\|$. By definition,

$$(V\Lambda V' + G_0)\hat{V} = \Theta\hat{V} = \hat{V}\hat{\Lambda}.$$

Multiplying both sides by V' from the left and noting that $V'V = \hat{V}\hat{V}' = I_K$, we find that $\Lambda(V'\hat{V}) + V'G_0\hat{V} = (V'\hat{V})\hat{\Lambda}$. This yields an equation for $\hat{V}'V$:

$$(\hat{V}'V)\Lambda = \hat{\Lambda}(\hat{V}'V) - \hat{V}'G_0V.$$

As a result, we can write

$$M = \hat{\Lambda}^{-1}\left[\hat{\Lambda}(\hat{V}'V) - \hat{V}'G_0V\right] = \hat{V}'V - \hat{\Lambda}^{-1}(\hat{V}'G_0V). \tag{4.34}$$

Write $B = -\hat{\Lambda}^{-1}(\hat{V}'G_0V)$. It follows from (4.34) that

$$\begin{aligned}\|M'M - I_K\| &= \|(\hat{V}'V + B)'(\hat{V}'V + B) - I_K\| \\ &\leq \|V'\hat{V}\hat{V}'V - I_K\| + 2\|B\|\|V'\hat{V}\| + \|B\|^2 \\ &\leq \|V'(\hat{V}\hat{V}' - VV')V\| + (2\|B\| + \|B\|^2) \\ &\leq \|\hat{V}\hat{V}' - VV'\| + (2\|B\| + \|B\|^2),\end{aligned}$$

where the third inequality is because $\|V'\hat{V}\| \leq 1$ and $V'V = I_K$. Applying the sine-theta theorem [70], we obtain $\|\hat{V}\hat{V}' - VV'\| \leq \frac{\|G_0\|}{\lambda_K - \|G_0\|}$. Combining it with $\|G_0\| \leq C_1^{-1}\lambda_K$ gives $\|\hat{V}\hat{V}' - VV'\| \leq$

$\frac{C_1}{C_1-1}\lambda_K^{-1}\|G_0\|$. Moreover, $\|B\| \le \hat{\lambda}_K^{-1}\|G_0\| \le \frac{C_1}{C_1-1}\lambda_K^{-1}\|G_0\|$ by (4.28). We plug these results into the above inequality and find that

$$\|M'M - I_K\| \le C\lambda_K^{-1}\|G_0\|. \tag{4.35}$$

Combining (4.35) with (4.33) gives

$$\begin{aligned}\|\tilde{V}\tilde{V}' - VV'\|_{\max} &\le \|V(M'M - I_K)\|_\infty \|V'\|_{\max} \\ &\le K\sqrt{K}\|M'M - I_K\|\|V\|_{\max}^2 \\ &\le C\lambda_K^{-1}\|G_0\|\|V\|_{\max}^2. \end{aligned} \tag{4.36}$$

We plug (4.32) and (4.36) into (4.30), and note that $\|G_0\| \le \|G_0\|_\infty$ and $\lambda_k \ge \lambda_K$ for all $1 \le k \le K$. It follows that

$$\|\hat{V}\hat{V}' - VV'\|_{\max} \le C_2\lambda_K^{-1}\Big(\frac{\lambda_1}{\lambda_K}\Big)^2\|V\|_{\max}^2\|G_0\|_\infty$$

This proves the first claim.

We then show the second claim about $\|G - G_0\|_{\max}$. Note that

$$\begin{aligned}\|G - G_0\|_{\max} &= \|\hat{V}\hat{\Lambda}\hat{V}' - V\Lambda V'\|_{\max} \\ &\le \|\hat{V}\hat{\Lambda}\hat{V}' - \tilde{V}\hat{\Lambda}\tilde{V}'\|_{\max} + \|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max} \\ &\le \sum_{k=1}^K \hat{\lambda}_k\|\hat{v}_k\hat{v}_k' - \tilde{v}_k\tilde{v}_k'\|_{\max} + \|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max} \\ &\le C\sum_{k=1}^K \Big(\frac{\lambda_1}{\lambda_k}\Big)\|V\|_{\max}^2\|G_0\|_\infty + \|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max}, \end{aligned} \tag{4.37}$$

where we have used (4.32) and (4.28) in the last inequality. It remains to bound $\|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max}$.

We recall the definition of $\tilde{v}_k$ in (4.29) and M in (4.33). By direct calculations,

$$\begin{aligned}\|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max} &= \Big\| \sum_{k=1}^{K} \hat{\lambda}_k^{-1}(V\Lambda V')\hat{v}_k\hat{v}_k'(V\Lambda V') - V\Lambda V' \Big\|_{\max} \\ &= \|V\Lambda V'(\hat{V}\hat{\Lambda}^{-1}\hat{V}')V\Lambda V' - V\Lambda V'\|_{\max} \\ &= \|V\Lambda(V'\hat{V}M)V' - V\Lambda V'\|_{\max} \\ &\le K\sqrt{K}\|V\|_{\max}^2\|\Lambda\|\|V'\hat{V}M - I_K\|.\end{aligned}$$

By (4.34), $M = \hat{V}'V + B$, where $B = -\hat{\Lambda}^{-1}(\hat{V}'G_0V)$. In the proof of (4.35), we have seen that $\|\hat{V}\hat{V}' - VV'\| \le \frac{C_1}{C_1-1}\lambda_K^{-1}\|G_0\|$ and $\|B\| \le \frac{C_1}{C_1-1}\lambda_K^{-1}\|G_0\|$. It follows that

$$\begin{aligned}\|V'\hat{V}M - I_K\| &= \|(V'\hat{V}\hat{V}'V - I_K) + V'\hat{V}B\| \\ &\le \|\hat{V}\hat{V}' - VV'\| + \|B\| \le C\lambda_K^{-1}\|G_0\|.\end{aligned}$$

Combining the above gives

$$\|\tilde{V}\hat{\Lambda}\tilde{V}' - V\Lambda V'\|_{\max} \le C\Big(\frac{\lambda_1}{\lambda_k}\Big)\|V\|_{\max}^2\|G_0\| \tag{4.38}$$

We plug (4.38) into (4.37), and note that $\|G_0\| \le \|G_0\|_\infty$ and $\lambda_k \ge \lambda_K$ for all $1 \le k \le K$. It yields that

$$\|G - G_0\|_{\max} \le C_2'\Big(\frac{\lambda_1}{\lambda_k}\Big)^2\|V\|_{\max}^2\|G_0\|_\infty.$$

This proves the second claim.

Proof of Theorem 4.2.4 For any $j \in S$ and any $\mathscr{I} \in \mathscr{A}_{\delta,j}(m)$ with $\mathscr{I} \subset \mathscr{I}^{(j)}$, we know by Lemma 4.2.3 and Mill's ratio,

$$\begin{aligned} P(T_{j|\mathscr{I}} \le t_p(q)) &\lesssim P\left(|\mathscr{N}(\sqrt{2\omega_{j|\mathscr{I}}(r)\sigma^2\log(p)}, \sigma^2)| \le \sqrt{2q\sigma^2\log(p)}\right) \\ &\le P\left(\mathscr{N}(0,1) \ge \sqrt{2\omega_{j|\mathscr{I}}(r)\log(p)} - \sqrt{2q\log(p)}\right) \\ &\le L_p p^{-[(\sqrt{\omega_{j|\mathscr{I}}(r)}-\sqrt{q})_+]^2} \end{aligned}$$

which implies that

$$P(T_j^* \le t_p(q)) \le \min_{\mathscr{I} \in \mathscr{A}_{\delta,j}(m), \mathscr{I} \subset \mathscr{I}^{(j)}} P(T_{j|\mathscr{I}} \le t_p(q)) \le L_p p^{-[(\sqrt{\omega_j(r,m)}-\sqrt{q})_+]^2}$$

Therefore, we get

$$E(|S \setminus \hat{S}(q)|) = \sum_{j \in S} P(T_j^* \le t_p(q)) \le L_p \sum_{j \in S} p^{-[(\sqrt{\omega_j(r,m)}-\sqrt{q})_+]^2}$$

Now we look at the second term. Recall G^δ defined in Lemma 4.5.4. We show that each row of G^δ has at most $C(\log(p))^\gamma$ nonzeros for some constant C. For any $1 \le i \le p$, suppose there are K_i nonzeros at ith row of G^δ. By Lemma 4.2.2, when p is sufficiently large $\|G - G_0\|_{\max} \le \delta/2$. Hence we have

$$C_0 \ge \sum_{j=1}^p |G_0(i,j)|^\gamma = \sum_{j=1}^p |G(i,j) - (G(i,j) - G_0(i,j))|^\gamma \ge K_i(\delta - \delta/2)^\gamma$$

which implies that

$$K_i \le C_0(\delta/2)^{-\gamma} \le C(\log(p))^\gamma$$

By a classical result in graph theory [58], we have for any $1 \le j \le p$,

$$|\mathscr{A}_{\delta,j}(m+1)| \le (m+1)(e \max_i K_i)^m \le C(\log(p))^{\gamma m} \tag{4.39}$$

Define

$$S_\delta^*(m) = \{1 \le j \le p : j \text{ is connected to } S \text{ through a path of length} \le m \text{ in } \mathscr{G}^\delta\}$$

For $j \in S_\delta^*(m)$, we know there exists a node $i \in S \cap \mathscr{I}$ for some $\mathscr{I} \in \mathscr{A}_{\delta,j}(m+1)$. This implies that $j \in \mathscr{I} \in \mathscr{A}_{\delta,j}(m+1)$. Hence we have

$$|S_\delta^*(m)| \le \sum_{i \in S} m|\mathscr{A}_{\delta,j}(m)| \le Cs_p(\log(p))^{\gamma m}$$

For $j \notin S_\delta^*(m)$ and any $\mathscr{A}_{\delta,j}(m)$, by Lemma 4.5.2 we can write $T_{j|\mathscr{I}} = W^2$ where $W \sim \mathscr{N}(w, \sigma^2)$ and

$$w = n^{1/2} A_{j|\mathscr{I}}^{-1/2} G^{j,\mathscr{I}^c} \beta^{\mathscr{I}^c} - n^{1/2} A_{j|\mathscr{I}}^{-1/2} G^{j,N} (G^{N,N})^{-1} G^{N,\mathscr{I}^c} \beta^{\mathscr{I}^c}$$

By definition of $S_\delta^*(m)$, we know $(G^\delta)^{\mathscr{I},\mathscr{I}^c} \beta^{\mathscr{I}^c} = (G^\delta)^{\mathscr{I},S} \beta^S = 0$. By Lemma 4.5.4 we have

$$\|G^{\mathscr{I},\mathscr{I}^c} \beta^{\mathscr{I}^c}\|_\infty = \| \left(G^{\mathscr{I},\mathscr{I}^c} - (G^\delta)^{\mathscr{I},\mathscr{I}^c} \right) \beta^{\mathscr{I}^c} \|_\infty = o(\tau_p)$$

which implies that $w = o(n^{1/2}\tau_p) = o(\sqrt{\log(p)})$. Hence we have

$$T_{j|\mathscr{I}} \sim \sigma^2 \chi_1^2 \left(o(\log(p))\right)$$

which suggests

$$P(T_{j|\mathscr{I}} > t_p(q)) \le P\left(\mathscr{N}(0,1) > \sqrt{2q\log(p)} - o(\sqrt{\log(p)})\right) \lesssim L_p p^{-q}$$

Hence by union bound, we have

$$P(T_j^* > t_p(q)) \le \sum_{\mathscr{I} \in \mathscr{A}_{\delta,j}(m)} P(T_{j|\mathscr{I}} > t_p(q)) \le L_p p^{-q} |\mathscr{A}_{\delta,j}(m)| \le L_p p^{-q}$$

Therefore, we have

$$\begin{aligned} E(|\hat{S}(q)|) &= \sum_{j\in S^*_\delta(m)} P(T^*_j > t_p(q)) + \sum_{j\notin S^*_\delta(m)} P(T^*_j > t_p(q)) \\ &\leq |S^*_\delta(m)| + p\cdot L_p p^{-q} \\ &\leq Cs_p(\log(p))^{\gamma m} + L_p p^{1-q} \end{aligned}$$

which proves the theorem.

Proof of Theorem 4.4.1 Write $Z = \sum_{i=d+1}^{p} \beta_i X_i$. We first consider the case where Z and $X_{1:d} = (X_1, \cdots, X_d)^T$ are independent. Denote σ_Z^2 as the variance of Z. Define

$$\begin{aligned} &\mu(x_1,\ldots,x_d) = \mu(x_{1:d}) = \mathbf{E}\,(Y|X_{1:d} = x_{1:d}) \\ = &\int \mathbf{E}(Y|X_{1:d} = x_{1:d}, Z = z) f_Z(z) dz = \frac{1}{\sigma_Z}\int g^{-1}(\beta_0 + \beta_{1:d}^T x_{1:d} + z)\phi(z/\sigma_Z)dz \end{aligned}$$

where we made use of the independence between $X_{1:d}$ and Z. According to [79], the MLE for misspecified model will converge to

$$(\beta_0^*, \beta_{1:d}^*) = \text{argmin}\ \mathbf{E}_{X_{1:d},Z}\mathbf{E}_{Y|X_{1:d},Z} \log\left(\frac{p^Y(1-p)^{1-Y}}{q^Y(1-q)^{1-Y}}\right)$$

where

$$p = g^{-1}(\beta_0 + \beta_{1:d}^T X_{1:d} + Z)$$

and

$$q = g^{-1}(\beta_0^* + \beta_{1:d}^{*,T} X_{1:d})$$

By some simple algebra, we know

$$(\beta_0^*, \beta_{1:d}^*) = \text{argmax}\ \mathbf{E}_{X_{1:d},Z}\,(p\log(q) + (1-p)\log(1-q))$$

and the first order condition is

$$\mathbf{E}_{X_{1:d},Z}\left(\frac{p-q}{q(1-q)}\frac{\partial q}{\partial \beta^*}\right)=0$$

Since

$$\frac{\partial q}{\partial \beta^*}=(g^{-1})'(\beta_0^*+\beta_{1:d}^{*,T}X_{1:d})\begin{pmatrix}1\\X_{1:d}\end{pmatrix}=q(1-q)\begin{pmatrix}1\\X_{1:d}\end{pmatrix}$$

we know

$$\mathbf{E}_{X_{1:d}}\left[\mu(X_{1:d})-g^{-1}(\beta_0^*+\beta_{1:d}^{*,T}X_{1:d})\right]=0 \tag{4.40}$$

and

$$\mathbf{E}_{X_{1:d}}\left[\left(\mu(X_{1:d})-g^{-1}(\beta_0^*+\beta_{1:d}^{*,T}X_{1:d})\right)X_{1:d}\right]=0 \tag{4.41}$$

Recall Σ_{11} is the covariance matrix of $X_{1:d}$ and we write its eigenvalue decomposition as $\Sigma_{11}=Q^T\Lambda Q$ where $\Lambda=\text{diag}\{\lambda_1,\ldots,\lambda_d\}$. Let $Y_{1:d}=(Y_1,\cdots,Y_d)^T=QX_{1:d}$, we know $Y_{1:d}\sim\mathcal{N}(0,\Lambda)$ so components of $Y_{1:d}$ are mutually independent. Moreover, write $\widetilde{\beta}_{1:d}^*=(\widetilde{\beta}_1^*,\ldots,\widetilde{\beta}_d^*)^T=Q\beta_{1:d}^*$, and $\widetilde{\beta}_{1:d}=(\widetilde{\beta}_1,\ldots,\widetilde{\beta}_d)^T=Q\beta_{1:d}$. We know that $\sum_{i=1}^d\widetilde{\beta}_i^*Y_i=\sum_{i=1}^d\beta_i^*X_i$

For any $1\le j\le d$, by Stein's lemma we have

$$\begin{aligned}
&\mathbf{E}\left(g^{-1}\left(\beta_0^*+\sum_{i=1}^d\widetilde{\beta}_i^*Y_i\right)Y_j\right)=\mathbf{E}_{Y_{1:d/j}}\mathbf{E}_{Y_j}\left(g^{-1}\left(\beta_0^*+\sum_{i=1}^d\widetilde{\beta}_i^*Y_i\right)Y_j\right)\\
=\;&\mathbf{E}_{Y_{1:d/j}}\left(\widetilde{\beta}_j^*\lambda_j\mathbf{E}_{Y_j}\left[(g^{-1})'\left(\beta_0^*+\sum_{i=1}^d\widetilde{\beta}_i^*Y_i\right)\right]\right)\\
=\;&\widetilde{\beta}_j^*\lambda_j\mathbf{E}\left[(g^{-1})'\left(\beta_0^*+\sum_{i=1}^d\widetilde{\beta}_i^*Y_i\right)\right]
\end{aligned}$$

where $Y_{1:d/j}$ denote the Y vector without jth component and we made use of the fact that $Y_{1:d/j}$ is independent of Y_j.

On the other hand, we can write

$$\mu(X_{1:d}) = \int g^{-1}(\beta_0 + \sum_{i=1}^{d} \beta_i X_i + z) f_Z(z) dz = \int g^{-1}(\beta_0 + \sum_{i=1}^{d} \widetilde{\beta}_i Y_i + z) f_Z(z) dz$$

and similarly by Stein's lemma we have for any $1 \leq j \leq d$,

$$\mathbf{E}\left[\mu(X_{1:d}) Y_j\right] = \widetilde{\beta}_j \lambda_j \mathbf{E}\left[(g^{-1})'\left(\beta_0 + \sum_{i=1}^{d} \widetilde{\beta}_i Y_i + Z\right)\right]$$

Therefore, by (4.41) we have for any $1 \leq j \leq d$

$$\widetilde{\beta}_j^* = \widetilde{\beta}_j \frac{\mathbf{E}\left[(g^{-1})'\left(\beta_0 + \sum_{i=1}^{d} \widetilde{\beta}_i Y_i + Z\right)\right]}{\mathbf{E}\left[(g^{-1})'\left(\beta_0^* + \sum_{i=1}^{d} \widetilde{\beta}_i^* Y_i\right)\right]}$$

Since the multiplicative factor does not depend on j, it's easy to see that for all $1 \leq j \leq d$

$$\beta_j^* = \beta_j \frac{\mathbf{E}\left[(g^{-1})'\left(\beta_0 + \sum_{i=1}^{d} \beta_i X_i + Z\right)\right]}{\mathbf{E}\left[(g^{-1})'\left(\beta_0^* + \sum_{i=1}^{d} \beta_i^* X_i\right)\right]} \tag{4.42}$$

Now we turn to the general case where independence of Z and $X_{1:d}$ is not assumed. It's easy to see that

$$\begin{pmatrix} X_{1:d} \\ Z \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12}\beta_{(d+1):p} \\ \beta_{(d+1):p}^T \Sigma_{21} & \beta_{(d+1):p}^T \Sigma_{22} \beta_{(d+1):p} \end{pmatrix}\right)$$

We can write

$$Z = \beta_{(d+1):p}^T \Sigma_{21} \Sigma_{11}^{-1} X_{1:d} + W$$

where W is independent of $X_{1:d}$. Hence we can write

$$\eta = \beta_0 + \beta_{1:d}^T X_{1:d} + Z = \beta_0 + \left(\beta_{1:d} + \Sigma_{11}^{-1}\Sigma_{12}\beta_{(d+1):p}\right)^T X_{1:d} + W$$

and using the result in the independence case, i.e. (4.42), we have

$$\beta^*_{1:d} = \left(\beta_{1:d} + \Sigma_{11}^{-1}\Sigma_{12}\beta_{(d+1):p}\right) \frac{\mathbf{E}\left[(g^{-1})'\left(\beta_0 + \Sigma_{i=1}^{p}\beta_i X_i\right)\right]}{\mathbf{E}\left[(g^{-1})'\left(\beta_0^* + \Sigma_{i=1}^{d}\beta_i^* X_i\right)\right]}$$

which implies the theorem.

4.5.2 Proof of Corollaries 4.2.1-4.2.2

Consider Corollary 4.2.1. It suffices to prove

$$\omega_j(r,m) \geq c_0 r, \qquad \text{for all } j \in S. \tag{4.43}$$

Once (4.43) is true, the $q^*(\vartheta, r, m)$ defined in Theorem 4.2.1 satisfies $q^*(\vartheta, r, m) \geq (\sqrt{c_0 r} - \sqrt{1-\vartheta})_+^2$. Then, Corollary 4.2.1 follows.

We show (4.43). Fix $j \in S$ and let $\mathscr{I}_k$ be the unique component of $\mathscr{G}_S^\delta$ that contains j. By (4.13) and (4.15), $\mathscr{I}_k \in \mathscr{A}_{\delta,j}(m)$. It then follows from (4.19) that $\omega_j(r,m) \geq \omega_{j|\mathscr{I}_k}(r)$. Furthermore, by arguments in Lemma 4.5.3, $\omega_{j|\mathscr{I}_k}(r) = \omega_j^*(r) + o(\omega_j^*(r))$. Combining the above gives

$$\omega_j(r,m) \gtrsim \omega_j^*(r) = \frac{nA^0_{j|\mathscr{I}_k}}{2\sigma^2\log(p)}\beta_j^2.$$

Note that $A^0_{j|\mathscr{I}_k} = G_0^{j,j} - G_0^{j,N}(G_0^{N,N})^{-1}G_0^{N,j}$, where $N = \mathscr{I}_k \setminus \{j\}$. We arrange indices in $\mathscr{I}_k$ such that j is the first index. By the matrix inverse formula, $A^0_{j|\mathscr{I}_k}$ is the inverse of the first diagonal of $(G_0^{\mathscr{I}_k,\mathscr{I}_k})^{-1}$. As a result,

$$A^0_{j|\mathscr{I}_k} \geq \lambda_{\min}(G_0^{\mathscr{I}_k,\mathscr{I}_k}) \geq c_0,$$

where the last inequality comes from $G_0 \in \mathscr{M}_p(g,\gamma,c_0,C_0)$ and $|\mathscr{I}_k| \leq \ell_0 \leq g$. Combining it with $|\beta_j| \geq \tau_p$ gives (4.43).

Consider Corollary 4.2.2. In FA-MR, since the columns of $\tilde{X}$ have unequal norms, we first nor-

malize them: $\tilde{x}_j^* = (\sqrt{n}/\|\tilde{x}_j\|)\tilde{x}_j$. We then rank variables by the marginal correlation coefficients

$$|(\tilde{x}_j^*, \tilde{y})|/(\tilde{x}_j^*, \tilde{x}_j^*) = n^{1/2}|(\tilde{x}_j/\|\tilde{x}_j\|, \tilde{y})| = n^{1/2}\|P_{\{j\}}\tilde{y}\|,$$

where we recall that $P_{\{j\}}\tilde{y}$ is the projection of $\tilde{y}$ onto $\tilde{x}_j$. So FA-MR is a special case of FA-CAR with $m = 1$. The claim then follows from the fact that $\omega_j(r,m)$ is a monotone increasing function of m.

4.5.3 Proofs of Lemmas

Proof of Lemma 4.1.1 We shall prove the following lemma, and Lemma 4.1.1 follows immediately.

Lemma 4.5.1. *Suppose the conditions of Lemma 4.1.1 hold. For all methods,*

$$\eta^*(\vartheta, r, h) = 1 - \min\{\vartheta,\, q^*(\vartheta, r, h)\},$$

where (notation: $a_+^2 = \max\{a, 0\}^2$ for any $a \in \mathbb{R}$)

$$q^*_{LSR}(\vartheta, r, h) = (\sqrt{(1-h^2)r} - \sqrt{1-\vartheta})_+^2,$$

$$q^*_{MR}(\vartheta, r, h) = \begin{cases} (\sqrt{r} - \sqrt{1-\vartheta})_+^2, & \vartheta \geq 1/2, \\ \min\big\{(\sqrt{r} - \sqrt{1-\vartheta})_+^2,\, \big((1-|h|)\sqrt{r} - \sqrt{1-2\vartheta}\big)_+^2\big\}, & \vartheta < 1/2, \end{cases}$$

$$q^*_{CAR}(\vartheta, r, h) = \begin{cases} (\sqrt{r} - \sqrt{1-\vartheta})_+^2, & \vartheta \geq 1/2, \\ \min\big\{(\sqrt{r} - \sqrt{1-\vartheta})_+^2,\, \big(\sqrt{(1-h^2)r} - \sqrt{1-2\vartheta}\big)_+^2\big\}, & \vartheta < 1/2. \end{cases}$$

We now prove Lemma 4.5.1. Consider using $t_p(q) = 2q\sigma^2\log(p)$ to threshold

$$n^{-1}|(x_j, y)|^2 \text{ in MR}, \quad T_j^* \text{ in CAR}, \quad \text{and } n(1-h^2)|\hat{\beta}_j^{ols}|^2 \text{ in LSR}. \tag{4.44}$$

We claim that, for all three methods,

$$FP_p(t_p(q)) = L_p p^{1-\rho_1(q;\vartheta,r,h)}, \qquad FN_p(t_p(q)) = L_p p^{1-\rho_2(q;\vartheta,r,h)}, \tag{4.45}$$

where the exponents are

$$\begin{aligned}
&\rho_1^{MR}(q) = \rho_1^{CAR}(q) = \min\{q,\ \vartheta + (\sqrt{q} - |h|\sqrt{r})_+^2\},\\
&\rho_1^{LSR}(q) = q,\\
&\rho_2^{MR}(q) = \min\{\vartheta + (\sqrt{r} - \sqrt{q})_+^2,\ 2\vartheta + [(1-|h|)\sqrt{r} - \sqrt{q}]_+^2\},\\
&\rho_2^{CAR}(q) = \min\{\vartheta + (\sqrt{r} - \sqrt{q})_+^2,\ 2\vartheta + [\sqrt{(1-h^2)r} - \sqrt{q}]_+^2\},\\
&\rho_2^{LSR}(q) = \vartheta + (\sqrt{(1-h^2)r} - \sqrt{q})_+^2.
\end{aligned}$$

Given (4.45), for each of the three methods, the quantity $q^* = q^*(\vartheta, r, h)$ in Lemma 4.5.1 is the solution of $\rho_2(q) = 1$. As a result, $FN_p(t_p(q)) \to 0$ for any $q < q^*$, and $FN_p(t_p(q)) \to \infty$ for any $q > q^*$. It follows that

$$SS_p^* = s_p + FP_p(t_p(q^*)) = p^{1-\vartheta} + L_p p^{1-\rho_1(q^*)} = L_p p^{1-\min\{\vartheta, q^*\}}.$$

Here the last equality comes from the expressions of $\rho_1(q)$ for all three methods. This proves Lemma 4.5.1.

It remains to prove (4.45). Let M_j be a symbol that represents the scores in (4.44) for each method. For an even j, define the events:

$$\begin{aligned}
&B_{j1} = \{\beta_{j-1} = 0, \beta_j \neq 0, M_j < t_p(q)\},\ \ B_{j2} = \{\beta_{j-1} \neq 0, \beta_j \neq 0, M_j < t_p(q)\},\\
&D_{j1} = \{\beta_{j-1} = 0, \beta_j = 0, M_j > t_p(q)\},\ \ D_{j2} = \{\beta_{j-1} \neq 0, \beta_j = 0, M_j > t_p(q)\}.
\end{aligned}$$

There is a false negative at location j over the events B_{j1} and B_{j2}, and there is a false positive over the events D_{j1} and D_{j2}. We can similarly define these four events for an odd j, by replacing

$(j-1)$ by $(j+1)$. It is seen that

$$FN_p(t_p(q)) = \sum_{j=1}^{p} \big[\mathbb{P}(B_{j1}) + \mathbb{P}(B_{j2})\big], \qquad FP_p(t_p(q)) = \sum_{j=1}^{p} \big[\mathbb{P}(D_{j1}) + \mathbb{P}(D_{j2})\big],$$

Therefore, to show (4.45), it suffices to calculate the probabilities of the above events. We only consider an even j, and the case for an odd j is similar.

First, consider MR, where the score for variable j is $M_j = n^{-1}|(x_j, y)|^2$. Note that

$$n^{-1/2}(x_j'y) = \mathcal{N}\left(\sqrt{n}(h\beta_{j-1} + \beta_j),\ \sigma^2\right).$$

So M_j/σ^2 has a non-central chi-square distribution with the non-centrality parameter equal to $n\sigma^{-2}|h\beta_{j-1} + \beta_j|^2$. On the event B_{j1}, $n\sigma^{-2}|h\beta_{j-1} + \beta_j|^2 = n\sigma^{-2}\beta_j^2 = 2r\log(p)$. It follows that

$$\begin{aligned}\mathbb{P}(B_{j1}) &= \varepsilon_p(1-\varepsilon_p) \cdot \mathbb{P}\left(\chi_1^2(2r\log(p)) < 2q\log(p)\right) \\ &= \varepsilon_p(1-\varepsilon_p) \cdot L_p p^{-(\sqrt{r}-\sqrt{q})_+^2} = L_p p^{-\vartheta-(\sqrt{r}-\sqrt{q})_+^2}.\end{aligned}$$

Here, the second equality is due to Mills' ratio and elementary properties of non-central chi-square distributions. On the event B_{j2}, if $h \geq 0$,

$$n\sigma^{-2}|h\beta_{j-1} + \beta_j|^2 = \begin{cases} (1+|h|)^2 \cdot 2r\log(p), & \text{if } \mathrm{sign}(\beta_{j-1}) = \mathrm{sign}(\beta_j), \\ (1-|h|)^2 \cdot 2r\log(p), & \text{if } \mathrm{sign}(\beta_{j-1}) \neq \mathrm{sign}(\beta_j). \end{cases}$$

If $h < 0$, we have similar results except that the two cases swap. As a result,

$$\begin{aligned}\mathbb{P}(B_{j2}) &= (\varepsilon_p^2/2) \cdot \mathbb{P}\left(\chi_1^2\big(2r(1+|h|)^2\log(p)\big) < 2q\log(p)\right) \\ &\quad + (\varepsilon_p^2/2) \cdot \mathbb{P}\left(\chi_1^2\big(2r(1-|h|)^2\log(p)\big) < 2q\log(p)\right) \\ &= L_p p^{-2\vartheta-[(1+|h|)\sqrt{r}-\sqrt{q}]_+^2} + L_p p^{-2\vartheta-[(1-|h|)\sqrt{r}-\sqrt{q}]_+^2} \\ &= L_p p^{-2\vartheta-[(1-|h|)\sqrt{r}-\sqrt{q}]_+^2}.\end{aligned}$$

Combining the above results, we have found that

$$FN_p(t_p(q)) = \sum_{j=1}^{p} L_p p^{-\min\{\vartheta+(\sqrt{r}-\sqrt{q})_+^2,\, 2\vartheta+[(1-|h|)\sqrt{r}-\sqrt{q}]_+^2\}} = L_p p^{1-\rho_2^{MR}(q)}.$$

Similarly, on the event D_{j1}, $n\sigma^{-2}|h\beta_{j-1}+\beta_j|^2 = 0$, and on the event D_{j2}, $n\sigma^{-2}|h\beta_{j-1}+\beta_j|^2 = h^2 \cdot 2r\log(p)$. We then have

$$\begin{aligned}
&\mathbb{P}(D_{j1}) = (1-\varepsilon_p)^2 \cdot \mathbb{P}\left(\chi_1^2(0) > 2q\log(p)\right) = L_p p^{-q}, \\
&\mathbb{P}(D_{j2}) = \varepsilon_p(1-\varepsilon_p) \cdot \mathbb{P}\left(\chi_1^2(2rh^2\log(p)) > 2q\log(p)\right) = L_p p^{-\vartheta-(\sqrt{q}-|h|\sqrt{r})_+^2}.
\end{aligned}$$

As a result,

$$FP_p(t_p(q)) = \sum_{j=1}^{p} L_p p^{-\min\{q,\, \vartheta+(\sqrt{q}-|h|\sqrt{r})_+^2\}} = L_p p^{1-\rho_1^{MR}(q)}.$$

Next, consider LSR. The score $M_j = n(1-h^2)|\hat{\beta}_j^{ols}|^2$. The least squares estimator satisfies that $\hat{\beta} = (X'X)^{-1}X'y \sim \mathcal{N}(\beta, n^{-1}\sigma^2\Theta^{-1})$. Here, Θ is blockwise diagonal with two-by-two blocks. We immediately have

$$\hat{\beta}_j^{ols} \sim \mathcal{N}\left(\beta_j,\, \frac{\sigma^2}{n(1-h^2)}\right).$$

So M_j/σ^2 has a non-central chi-square distribution with the non-centrality parameter $n(1-h^2)\sigma^{-2}\beta_j^2$. On both of the events B_{j1} and B_{j2}, the non-centrality parameter is equal to $(1-h^2)\cdot 2r\log(p)$; it is easy to see that the probability of B_{j1} dominates. It follows that

$$\begin{aligned}
FN_p(t_p(q)) &= L_p \sum_{j=1}^{p} \mathbb{P}(B_{j1}) = L_p p\varepsilon_p \cdot \mathbb{P}\left(\chi_1^2(2r(1-h^2)\log(p)) < 2q\log(p)\right) \\
&= L_p p^{1-\vartheta-(\sqrt{(1-h^2)r}-\sqrt{q})_+^2} = L_p p^{1-\rho_2^{LSR}(q)}.
\end{aligned}$$

On both of the events D_{j1} and D_{j2}, the non-centrality parameter is equal to 0, and the probability

of D_{j1} dominates. It follows that

$$FP_p(t_p(q)) = L_p \sum_{j=1}^{p} \mathbb{P}(D_{j1}) = L_p p \cdot \mathbb{P}\left(\chi_1^2(0) > 2q\log(p)\right)$$
$$= L_p p^{1-q} = L_p p^{1-\rho_1^{LSR}(q)}.$$

Last, consider CAR. Note that $T_j^* = \max\{T_{j|\{j\}}, T_{j|\{j-1,j\}}\}$. It is easy to see that $T_{j|\{j\}}$ coincides with the score in MR. To obtain the distribution of $T_{j|\{j-1,j\}}$, we apply (4.21). Let $\eta = (x'_{j-1}y, x'_j y)'$ and H be the two-by-two matrix with unit diagonals and off-diagonals h. It follows from (4.21) that

$$T_{j|\{j-1,j\}} = n^{-1}(\eta' H^{-1}\eta - \eta_1^2) = \frac{1}{n(1-h^2)}(\eta_2 - h\eta_1)^2.$$

Write $W = \frac{1}{\sqrt{n(1-h^2)}}(\eta_2 - h\eta_1)$. Then, $T_{j|\{j-1,j\}} = W^2$. Since $\eta \sim \mathcal{N}(nH\beta, nH)$,

$$W \sim \mathcal{N}\left(\sqrt{n(1-h^2)}\beta_j,\ \sigma^2\right).$$

To summarize, we have found that

$$T_{j|\{j\}}/\sigma^2 \sim \chi_1^2\left(n\sigma^{-2}|\beta_j + h\beta_{j-1}|^2\right),$$
$$T_{j|\{j-1,j\}}/\sigma^2 \sim \chi_1^2\left(n\sigma^{-2}(1-h^2)\beta_j^2\right). \tag{4.46}$$

Consider the type II errors. We use a simply fact that $\max\{T_{j|\{j\}}, T_{j|\{j-1,j\}}\} < t_p(q))$ has a probability that is upper bounded by either the probability of $T_{j|\{j\}} < t_p(q)$ or the probability of $T_{j|\{j-1,j\}} < t_p(q)$, so we can take the minimum of these two probabilities as an upper bound. On the event B_{j1}, the non-centrality parameters for the two statistics are $n\beta_j^2$ and $n(1-h^2)\beta_j^2$.

Therefore, the type II error is determined by the behavior of $T_{j|\{j\}}$. It follows that

$$\begin{aligned}\mathbb{P}(B_{j1}) &\leq \varepsilon_p(1-\varepsilon_p)\cdot\mathbb{P}(T_{j|\{j\}} < t_p(q)) \\ &= \varepsilon_p(1-\varepsilon_p)\cdot\mathbb{P}\big(\chi_1^2(2r\log(p)) < t_p(q)\big) \\ &= L_p p^{-\vartheta-(\sqrt{r}-\sqrt{q})_+^2}.\end{aligned}$$

On the event B_{j2}, the non-centrality parameter for $T_{j|\{j-1,j\}}/\sigma^2$ is the same as before, which is $n(1-h^2)\sigma^{-2}\beta_j^2 = (1-h^2)\cdot 2r\log(p)$. The non-centrality parameter for $T_{j|\{j\}}/\sigma^2$ has been studied in the MR case, which is equal to $(1\pm|h|)^2\cdot 2r\log(p)$. In the case of $(1+|h|)^2\cdot 2r\log(p)$, since $(1+|h|)^2 \geq 1-h^2$, the type II error is determined by the behavior of $T_{j|\{j\}}$. In the case of $(1-|h|)^2\cdot 2r\log(p)$, since $1-h^2 \geq (1-|h|)^2$, the type II error is determined by the behavior of $T_{j|\{j-1,j\}}$. As a result,

$$\begin{aligned}\mathbb{P}(B_{j2}) &\leq (\varepsilon_p^2/2)\cdot\mathbb{P}\big(T_{j|\{j\}} < t_p(q)\big) + (\varepsilon_p^2/2)\cdot\mathbb{P}\big(T_{j|\{j-1,j\}} < t_p(q)\big) \\ &= (\varepsilon_p^2/2)\cdot\mathbb{P}\big(\chi_1^2(2r(1+|h|)^2\log(p)) < t_p(q)\big) \\ &\quad + (\varepsilon_p^2/2)\cdot\mathbb{P}\big(\chi_1^2(2r(1-h^2)\log(p)) < t_p(q)\big) \\ &= L_p p^{-2\vartheta-[(1+|h|)\sqrt{r}-\sqrt{q}]_+^2} + L_p p^{-2\vartheta-(\sqrt{(1-h^2)r}-\sqrt{q})_+^2} \\ &= L_p p^{-2\vartheta-(\sqrt{(1-h^2)r}-\sqrt{q})_+^2}.\end{aligned}$$

Combining the above results, we have

$$FN_p(t_p(q)) = \sum_{j=1}^{p} L_p p^{-\min\left\{\vartheta+(\sqrt{r}-\sqrt{q})_+^2,\, 2\vartheta+(\sqrt{(1-h^2)r}-\sqrt{q})^2\right\}} = L_p p^{1-\rho_2^{CAR}(q)}.$$

Consider the type I errors. On the event D_{j1}, both non-centrality parameters in (4.46) become 0.

We then use the probability union bound to get

$$\begin{aligned}\mathbb{P}(D_{j1}) &\leq (1-\varepsilon_p)^2 \cdot \left[\mathbb{P}(T_{j|\{j\}} > t_p(q)) + \mathbb{P}(T_{j|\{j-1,j\}} > t_p(q))\right] \\ &= (1-\varepsilon_p)^2 \cdot 2\mathbb{P}\big(\chi_1^2(0) > t_p(q)\big) \\ &= L_p p^{-q}.\end{aligned}$$

Similarly, on the event D_{j2},

$$\begin{aligned}\mathbb{P}(D_{j2}) &\leq \varepsilon_p(1-\varepsilon_p) \cdot \left[\mathbb{P}(T_{j|\{j\}} > t_p(q)) + \mathbb{P}(T_{j|\{j-1,j\}} > t_p(q))\right] \\ &= \varepsilon_p(1-\varepsilon_p) \cdot \Big[\mathbb{P}\big(\chi_1^2(2h^2 r\log(p)) > t_p(q)\big) + \mathbb{P}\big(\chi_1^2(0) > t_p(q)\big)\Big] \\ &= L_p p^{-\vartheta-(\sqrt{q}-|h|\sqrt{r})_+^2} + L_p p^{-\vartheta-q}.\end{aligned}$$

It follows that

$$FP_p(t_p(q)) = \sum_{j=1}^{p} L_p p^{-\min\{q,\, \vartheta+(\sqrt{q}-|h|\sqrt{r})_+^2\}} = L_p p^{1-\rho_1^{CAR}(q)}.$$

The proof is now complete.

Proof of Lemma 4.2.1 Without loss of generality, we assume $v_1'\hat{v}_1 \geq 0$. By definition, $\Theta\hat{v}_1 = \hat{\lambda}_1\hat{v}_1$, where $\Theta = \lambda_1 v_1 v_1' + G_0$. It follows that

$$\lambda_1(v_1'\hat{v}_1)v_1 + G_0\hat{v}_1 = \hat{\lambda}_1\hat{v}_1. \tag{4.47}$$

By Weyl's inequality, $|\hat{\lambda}_1 - \lambda_1| \leq \|G_0\| \leq \|G_0\|_\infty \leq \lambda_1/3$. As a result,

$$(2/3)\lambda_1 \leq \hat{\lambda}_1 \leq (4/3)\lambda_1. \tag{4.48}$$

In particular, the minimum eigenvalue of $\hat{\lambda}_1 I_p - G_0$ is lower bounded by $(2/3)\lambda_1 - \|G_0\| \geq \lambda_1/3$. So $(\hat{\lambda}_1 I_p - G_0)$ is always positive definite. So we can solve from (4.47) to get

$$\hat{v}_1 = (I_p - \hat{\lambda}_1^{-1} G_0)^{-1} \cdot \frac{\lambda_1 (v_1' \hat{v}_1)}{\hat{\lambda}_1} v_1. \tag{4.49}$$

We now show the claim. Write $\Delta = (I_p - \hat{\lambda}_1^{-1} G_0)^{-1} - I_p$ and $\varepsilon = \frac{\lambda_1 (v_1' \hat{v}_1)}{\hat{\lambda}_1} - 1$. We have

$$\begin{aligned} \|\hat{v}_1 - v_1\|_\infty &= \big\|(I_p + \Delta)(1+\varepsilon) v_1 - v_1\big\|_\infty \\ &\leq \|\Delta v_1\|_\infty + \|\varepsilon v_1 + \varepsilon \Delta v_1\|_\infty \\ &\leq \|\Delta\|_\infty \|v_1\|_\infty + |\varepsilon| \cdot \big(\|v_1\|_\infty + \|\Delta\|_\infty \|v_1\|_\infty\big). \end{aligned} \tag{4.50}$$

First, we bound $\|\Delta\|_\infty$. Since $(\Delta + I_p)(I_p - \hat{\lambda}_1^{-1} G_0) = I_p$, we have

$$\Delta = \hat{\lambda}_1^{-1} G_0 + \Delta \hat{\lambda}_1^{-1} G_0.$$

Using the triangular inequality, $\|\Delta\|_\infty \leq \hat{\lambda}_1^{-1} \|G_0\|_\infty + \|\Delta\|_\infty \hat{\lambda}_1^{-1} \|G_0\|_\infty$. It follows that

$$\|\Delta\|_\infty \leq \frac{\hat{\lambda}_1^{-1} \|G_0\|_\infty}{1 - \hat{\lambda}_1^{-1} \|G_0\|_\infty}.$$

By assumption, $\|G_0\|_\infty \leq \lambda_1/3$; by (4.48), $\hat{\lambda}^{-1} \leq \frac{3}{2} \lambda_1^{-1}$. So the denominator $1 - \hat{\lambda}_1^{-1} \|G_0\|_\infty \geq 1/2$. It follows that

$$\|\Delta\|_\infty \leq 3 \lambda_1^{-1} \|G_0\|_\infty. \tag{4.51}$$

Next, we bound $|\varepsilon|$. Note that

$$|\varepsilon| = \Big|\frac{\lambda_1 (v_1' \hat{v}_1)}{\hat{\lambda}_1} - 1\Big| \leq |1 - \frac{\lambda_1}{\hat{\lambda}_1}| + \frac{\lambda_1}{\hat{\lambda}_1} \cdot |1 - v_1' \hat{v}_1| \leq \frac{3}{2} \frac{\|G_0\|}{\lambda_1} + \frac{3}{2} |1 - v_1' \hat{v}_1|,$$

where we use $|\hat{\lambda}_1 - \lambda_1| \leq \|G_0\|$ and (4.48) in the last inequality. We now consider $|1 - v_1' \hat{v}_1|$.

Multiplying both sides of (4.47) by $\hat{v}_1'$ from the left, we get $\lambda_1(v_1'\hat{v}_1)^2 + \hat{v}_1' G_0 \hat{v}_1 = \hat{\lambda}_1$. So

$$(v_1'\hat{v}_1)^2 = \frac{\hat{\lambda}_1}{\lambda_1} - \frac{\hat{v}_1' G_0 \hat{v}_1}{\lambda_1}.$$

As a result,

$$|1 - v_1'\hat{v}_1| \le 1 - (v_1'\hat{v}_1)^2 \le |1 - \frac{\hat{\lambda}_1}{\lambda_1}| + \frac{|\hat{v}_1' G_0 \hat{v}_1|}{\lambda_1} \le \frac{\|G_0\|}{\lambda_1} + \frac{\|G_0\|}{\lambda_1} = 2\frac{\|G_0\|}{\lambda_1}.$$

Combining the above gives

$$|\varepsilon| \le \frac{9}{2}\lambda_1^{-1}\|G_0\| \le \frac{9}{2}\lambda_1^{-1}\|G_0\|_\infty. \tag{4.52}$$

We plug (4.51)-(4.52) into (4.50), and use $\|G_0\|_\infty \le \lambda_1/3$. It yields

$$\|\hat{v}_1 - v_1\|_\infty \le \|v_1\|_\infty \left(\frac{3\|G_0\|_\infty}{\lambda_1} + \frac{9\|G_0\|_\infty}{2\lambda_1}(1 + \frac{3\|G_0\|_\infty}{\lambda_1}) \right) \le \frac{12\|v_1\|_\infty \|G_0\|_\infty}{\lambda_1}.$$

This proves the claim.

Proof of Lemma 4.2.2 By (4.12) and Weyl's inequality, we know

$$\hat{\sigma}_K^2/n \ge \lambda_K - \|G_0\| \gg \log(p) \quad \hat{\sigma}_{K+1}^2/n \le \|G_0\| \ll \log(p)$$

and hence $\hat{K}_p = K$ where $\hat{K}_p$ is defined in (4.14).

For any $1 \le i \le K$ and $1 \le j \le p$, since $\Theta(j,j) = G_0(j,j) + \sum_{k=1}^K \lambda_k v_k(j)^2 = 1$, we have

$$\lambda_K v_i(j)^2 \le \lambda_i v_i(j)^2 \le \sum_{k=1}^K \lambda_k v_k(j)^2 = 1 - G_0(j,j) \le 1$$

and hence by (4.12), we have

$$\max_{1\le k\le K}\|v_k\|_\infty^2 \le \lambda_K^{-1} = o(1/\max\{s_p,\log(p)\})$$

Then we can get the desired result by directly applying Theorem 4.2.2.

Proof of Lemma 4.2.3 We give several technical lemmas that are used frequently in the main proofs. Lemma 4.5.2 gives the exact distribution of the statistic $T_{j|\mathscr{I}}$. Lemma 4.5.3 implies that, if we replace $\mathscr{G}_{0,S}^{\delta}$ with $\mathscr{G}_S^{\delta}$ in the definition of $\omega_{j|\mathscr{I}}(r)$, the resulting change is negligible. Lemma 4.5.4 shows that those small entries of G (recall that G is the Gram matrix of model (4.4)) has negligible effects on screening. In this section, we write $G(j,j) = G^{j,j}$ and $G^{\{j\},N} = G^{j,N}$ for notation convenience, similarly for $G_0^{j,j}$ and $G_0^{j,N}$.

Lemma 4.5.2. *Under the conditions of Theorem 4.2.1, for $\mathscr{I} \subset \{1,\cdots,p\}$ such that $|\mathscr{I}| \le g$ and any $j \in \mathscr{I}$, $T_{j|\mathscr{I}}$ has the same distribution as W^2, where $W \sim \mathscr{N}(w,\sigma^2)$,*

$$w = n^{1/2}A_{j|\mathscr{I}}^{1/2}\left[\beta_j + A_{j|\mathscr{I}}^{-1}(G^{j,\mathscr{I}^c}\beta^{\mathscr{I}^c} - G^{j,N}(G^{N,N})^{-1}G^{N,\mathscr{I}^c}\beta^{\mathscr{I}^c})\right],$$

and $A_{j|\mathscr{I}} = G^{j,j} - G^{j,N}(G^{N,N})^{-1}G^{N,j}$ with $N = \mathscr{I}\setminus\{j\}$.

If there is an edge between i and j in $\mathscr{G}^{\delta}$, then $|G_0(i,j)| \ge |G(i,j)| - \|G-G_0\|_{\max} \ge \delta - o(\delta_p) \gtrsim 1.01\delta_p$, where we have used Lemma 4.2.2 and the assumption (4.15). So there must be an edge between i and j in $\mathscr{G}_0^{\delta}$. In other words, $\mathscr{G}_S^{\delta}$ is a subgraph of $\mathscr{G}_{0,S}^{\delta}$ by removing some edges. Fix $j \in S$ and $\mathscr{I} \subset \mathscr{I}^{(j)}$ where $\mathscr{I}^{(j)}$ is the unique component of $\mathscr{G}_{0,S}^{\delta}$ that contains j. We introduce a counterpart of $\omega_{j|\mathscr{I}}(r)$ in (4.18) when $\mathscr{I} \subset \mathscr{G}_S^{\delta}$:

$$\tilde{\omega}_{j|\mathscr{I}}(r) = \frac{nA_{j|\mathscr{I}}}{2\sigma^2\log(p)}\left\{\beta_j + A_{j|\mathscr{I}}^{-1}[G^{j,F} - G^{j,N}(G^{N,N})^{-1}G^{N,F}]\beta^F\right\}^2.$$

where $A_{j|\mathscr{I}} = G^{j,j} - G^{j,N}(G^{N,N})^{-1}G^{N,j}$ with $N = \mathscr{I}\setminus\{j\}$ and $F = \mathscr{I}^{(j)}\setminus\mathscr{I}$.

Lemma 4.5.3. *Under conditions of Theorem 4.2.1. For any $j \in S$, if $\mathscr{I}^{(j)}$, the unique component of $\mathscr{G}^{\delta}_{0,S}$ that contains j, has a size $\leq g$, then for any $\mathscr{I} \subset \mathscr{I}^{(j)}$ we have $A^0_{j|\mathscr{I}} \geq c_0$ and $|A^0_{j|\mathscr{I}} - A_{j|\mathscr{I}\cap\mathscr{G}^{\delta}_S}| = o(\delta_p)$. Moreover, if $\mathscr{I} \subset \mathscr{G}^{\delta}_S$ then $|\tilde{\omega}_{j|\mathscr{I}}(r) - \omega_{j|\mathscr{I}}(r)|/\omega_{j|\mathscr{I}}(r) = o(1)$.*

Lemma 4.5.4. *Define the matrix $G^{\delta} \in \mathbb{R}^{p,p}$ by $G^{\delta}(i,j) = G(i,j)\mathbf{1}\{|G(i,j)| > \delta\}$ for $1 \leq i,j \leq p$. Under conditions of Theorem 4.2.1, for any $\mathscr{I} \subset \{1,2,\ldots,p\}$ and $\mathscr{J} \subset \{1,2,\ldots,p\}$,*

$$\left\| \left(G^{\mathscr{I},\mathscr{J}} - (G^{\delta})^{\mathscr{I},\mathscr{J}} \right) \beta^{\mathscr{J}} \right\|_{\infty} \leq C \left(\log(p)^{-(1-\gamma)} + s_p \|G - G_0\|_{\max} \right) \tau_p = o(\tau_p).$$

Now we prove Lemma 4.2.3. We denote $\mathscr{I}^{(j)}$ as $\mathscr{I}_k$ for some $1 \leq k \leq M$. We know by Lemma 4.5.2 that $T_{j|\mathscr{I}} \sim \mathscr{N}^2(w_1, \sigma^2)$ where

$$w_1 = n^{1/2} A^{1/2}_{j|\mathscr{I}} \left(\beta_j + A^{-1}_{j|\mathscr{I}_k} [G^{j,F} - G^{j,N}(G^{N,N})^{-1} G^{N,F}] \beta^F \right) + \mathrm{I} + \mathrm{II}$$

and

$$\mathrm{I} = n^{1/2} A^{-1/2}_{j|\mathscr{I}} G^{j,\mathscr{I}_k^c} \beta^{\mathscr{I}_k^c}, \mathrm{II} = -n^{1/2} A^{-1/2}_{j|\mathscr{I}_k} G^{j,N}(G^{N,N})^{-1} G^{N,\mathscr{I}_k^c} \beta^{\mathscr{I}_k^c}$$

It's easy to see that there exists a constant C such that $\|G^{j,N}(G^{N,N})^{-1}\|_{\infty} \leq C$. By definition of $\mathscr{I}_k$ and the fact that $\mathscr{G}^{\delta}_S \subset \mathscr{G}^{\delta}_{0,S}$, we know $(G^{\delta})^{\mathscr{I}_k,\mathscr{I}_k^c} \beta^{\mathscr{I}_k^c} = 0$ where G^{δ} is defined in Lemma 4.5.4. Hence we have

$$G^{\mathscr{I}_k,\mathscr{I}_k^c} \beta^{\mathscr{I}_k^c} = (G\beta)^{\mathscr{I}_k} - G^{\mathscr{I}_k,\mathscr{I}_k} \beta^{\mathscr{I}_k} = \left(G^{\mathscr{I}_k,\mathscr{I}_k^c} - (G^{\delta})^{\mathscr{I}_k,\mathscr{I}_k^c} \right) \beta^{\mathscr{I}_k^c}$$

By Lemma 4.5.4, we know

$$\|G^{\mathscr{I}_k,\mathscr{I}_k^c} \beta^{\mathscr{I}_k^c}\|_{\infty} = \|(G\beta)^{\mathscr{I}_k} - G^{\mathscr{I}_k,\mathscr{I}_k} \beta^{\mathscr{I}_k}\|_{\infty} = o(\tau_p) = o(n^{-1/2}\sqrt{\log(p)})$$

which suggests that

$$\max\{|\mathrm{I}|, |\mathrm{II}|\} = o(\sqrt{\log(p)})$$

By Lemma 4.5.3, we know

$$w_1 = \sigma\sqrt{2\log(p)\tilde{\omega}_{j|\mathscr{I}}(r)} + o(\sqrt{\log(p)}) = \sigma\sqrt{2\log(p)\omega_{j|\mathscr{I}}(r)} + o(\sqrt{\log(p)})$$

which implies Lemma 4.2.3.

Proof of Lemma 4.5.2 We need some preparations. First, we show that

$$A_{j|\mathscr{I}} \geq \nu_{\min}(G^{\mathscr{I},\mathscr{I}}) \gtrsim c_0, \tag{4.53}$$

so that $A_{j|\mathscr{I}}$ is always positive. A helpful result is the matrix blockwise inverse fomular

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BMCA^{-1} & -A^{-1}BM \\ -MCA^{-1} & M \end{bmatrix} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B \\ -I \end{bmatrix} M \begin{bmatrix} C & -I \end{bmatrix},$$

with $M = (D - CA^{-1}B)^{-1}$. Without loss of generality, we assume j is the first index in $\mathscr{I}$. Applying the above formula, we see that $(A_{j|\mathscr{I}})^{-1}$ is the $(1,1)$-th entry of $(G^{\mathscr{I},\mathscr{I}})^{-1}$. It follows that $A_{j|\mathscr{I}} \geq \nu_{\min}(G^{\mathscr{I},\mathscr{I}})$. Since $|\mathscr{I}| \leq g$, it suffices to show that $\nu_g^*(G) \gtrsim c_0$ where $\nu_g^*(G)$ is the same as in Section 4.2.1. For any $g \times g$ matrix $\tilde{E}$ which is a principal submatrix of G, let E_0 be the corresponding principal submatrix of G_0. We know $\nu_{\min}(E_0) \geq c_0$. By Weyl's inequality and Lemma 4.2.2,

$$|\nu_{\min}(\tilde{E}) - \nu_{\min}(E_0)| \leq \|\tilde{E} - E_0\|_2 \leq g\|\tilde{E} - E_0\|_{\max} = o(1/\log(p))$$

which implies that $\nu_{\min}(\tilde{E}) \gtrsim c_0$ and hence $\nu_g^*(G) \gtrsim c_0$ as p goes to infinity.

Second, we introduce $y_1 = \tilde{X}'\tilde{y}$ and show that

$$y_1 \sim \mathscr{N}(nG\beta, \sigma^2 nG). \tag{4.54}$$

Since $\tilde{y} \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 H)$ where $H = I_n - \sum_{k=1}^{K} \hat{u}_k \hat{u}_k'$, we have $y_1 = \tilde{X}'\tilde{y} \sim \mathcal{N}(\tilde{X}'\tilde{X}\beta, \sigma^2 \tilde{X}' H \tilde{X})$. Noting that $\tilde{X} = HX$ and $G = (1/n)\tilde{X}'\tilde{X}$, we obtain $\tilde{X}' H \tilde{X} = (HX)' H (HX) = X' H^2 X = (HX)'(HX) = \tilde{X}'\tilde{X} = nG$. So (4.54) follows.

We now show the claim. By definition,

$$\begin{aligned} T_{j|\mathscr{I}} &= \|P_{\mathscr{I}}\tilde{y}\|^2 - \|P_N \tilde{y}\|^2 \\ &= \tilde{y}'\tilde{X}^{\mathscr{I}} \left((\tilde{X}^{\mathscr{I}})'\tilde{X}^{\mathscr{I}} \right)^{-1} (\tilde{X}^{\mathscr{I}})'\tilde{y} - \tilde{y}'\tilde{X}^N \left((\tilde{X}^N)'\tilde{X}^N \right)^{-1} (\tilde{X}^N)'\tilde{y} \\ &= n^{-1} (y_1^{\mathscr{I}})' (G^{\mathscr{I},\mathscr{I}})^{-1} y_1^{\mathscr{I}} - n^{-1} (y_1^N)' (G^{N,N})^{-1} y_1^N \\ &= n^{-1} (y_1^{\mathscr{I}})' \left((G^{\mathscr{I},\mathscr{I}})^{-1} - \begin{bmatrix} (G^{N,N})^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) y_1^{\mathscr{I}} \end{aligned}$$

where we assume j is the last index in $\mathscr{I}$ for the presentation purpose. Applying the matrix inverse formula, we obtain

$$T_{j|\mathscr{I}} = n^{-1} (y_1^{\mathscr{I}})' B' A_{j|\mathscr{I}}^{-1} B y_1^{\mathscr{I}}, \qquad B = \left[-G^{j,N} (G^{N,N})^{-1}, 1 \right]. \tag{4.55}$$

Therefore, $T_{j|\mathscr{I}} = W^2$ for $W = n^{-1/2} A_{j|\mathscr{I}}^{-1/2} B(y_1^{\mathscr{I}})$.

It remains to calculate the mean and variance of W. First, by (4.54), the variance of W is $\sigma^2 A_{j|\mathscr{I}}^{-1} (B G^{\mathscr{I},\mathscr{I}} B')$, where by definition of B and elementary calculations, $B G^{\mathscr{I},\mathscr{I}} B' = A_{j|\mathscr{I}}$. So $\mathrm{var}(W) = \sigma^2$. Second, it is seen that $W = n^{-1/2} A_{j|\mathscr{I}}^{-1/2} (y_1^j - G^{j,N} (G^{N,N})^{-1} y_1^N)$. It follows from (4.54) that

$$\begin{aligned} E[W] &= n^{-1/2} A_{j|\mathscr{I}}^{-1/2} \left[(G\beta)^j - G^{j,N} (G^{N,N})^{-1} (G\beta)^N \right] \\ &= n^{-1/2} A_{j|\mathscr{I}}^{-1/2} \left[G^{j,\mathscr{I}} \beta^{\mathscr{I}} - G^{j,N} (G^{N,N})^{-1} G^{N,\mathscr{I}} \beta^{\mathscr{I}} + rem \right] \\ &= n^{-1/2} A_{j|\mathscr{I}}^{-1/2} \left[G^{j,j}\beta_j + G^{j,N}\beta^N - G^{j,N} (G^{N,N})^{-1} G^{N,j} \beta_j - G^{j,N}\beta^N + rem \right] \\ &= n^{-1/2} A_{j|\mathscr{I}}^{-1/2} \left[A_{j|\mathscr{I}} \beta_j + rem \right], \end{aligned}$$

where $rem = G^{j,\mathscr{I}^c}\beta^{\mathscr{I}^c} - G^{j,N}(G^{N,N})^{-1}G^{N,\mathscr{I}^c}\beta^{\mathscr{I}^c}$. So $E[W] = w$.

Proof of Lemma 4.5.3 Fix j and for any $\mathscr{I} \subset \mathscr{I}^{(j)}$ denote by $\tilde{\mathscr{I}} = \mathscr{I} \cap \mathscr{G}_S^{\delta}$. Without loss of generality, we assume j is the first index of both sets $\mathscr{I}$ and $\tilde{\mathscr{I}}$. By Lemma 4.5.2, we have seen that $A^{-1}_{j|\tilde{\mathscr{I}}}$ equals to the $(1,1)$-th entry of $(G^{\tilde{\mathscr{I}},\tilde{\mathscr{I}}})^{-1}$; similarly, $(A^0_{j|\mathscr{I}})^{-1}$ equals to the $(1,1)$-th entry of $(G_0^{\mathscr{I},\mathscr{I}})^{-1}$. Since $|\mathscr{I}| \le g$

$$A^0_{j|\mathscr{I}} \ge \lambda_{\min}(G_0^{\mathscr{I},\mathscr{I}}) \ge \nu_g^*(G_0) \ge c_0.$$

This proves the first claim.

We now show the second claim. Since both $A^0_{j|\mathscr{I}}$ and $A_{j|\tilde{\mathscr{I}}}$ are upper bounded by some constant, it suffices to show that

$$|(A^0_{j|\mathscr{I}})^{-1} - A^{-1}_{j|\tilde{\mathscr{I}}}| = O(\delta_p). \tag{4.56}$$

By triangular inequality,

$$\begin{aligned}
|(A^0_{j|\mathscr{I}})^{-1} - A^{-1}_{j|\tilde{\mathscr{I}}}| &= |(G_0^{\mathscr{I},\mathscr{I}})^{-1}(1,1) - (G^{\tilde{\mathscr{I}},\tilde{\mathscr{I}}})^{-1}(1,1)| \\
&\le |(G_0^{\mathscr{I},\mathscr{I}})^{-1}(1,1) - (G^{\mathscr{I},\mathscr{I}})^{-1}(1,1)| + |(G^{\mathscr{I},\mathscr{I}})^{-1}(1,1) - (G^{\tilde{\mathscr{I}},\tilde{\mathscr{I}}})^{-1}(1,1)| \\
&\equiv I + II.
\end{aligned}$$

Consider I. First, since $|\mathscr{I}| \le g$, $\|G^{\mathscr{I},\mathscr{I}} - G_0^{\mathscr{I},\mathscr{I}}\| \le g\|G - G_0\|_{\max} = o(\delta_p)$ by Lemma 4.2.2. Second, $\lambda_{\min}(G_0^{\mathscr{I},\mathscr{I}}) \ge \nu_g^*(G_0) \ge c_0$. It follows that

$$\begin{aligned}
I &\le \|(G_0^{\mathscr{I},\mathscr{I}})^{-1} - (G^{\mathscr{I},\mathscr{I}})^{-1}\| \le \|(G_0^{\mathscr{I},\mathscr{I}})^{-1}\|\|G^{\mathscr{I},\mathscr{I}} - G_0^{\mathscr{I},\mathscr{I}}\|\|(G^{\mathscr{I},\mathscr{I}})^{-1}\| \\
&\lesssim c_0^{-2}\|G^{\mathscr{I},\mathscr{I}} - G_0^{\mathscr{I},\mathscr{I}}\| = o(\delta_p).
\end{aligned} \tag{4.57}$$

Consider II. By definition, we have $\tilde{\mathscr{I}} \subset \mathscr{I}$. If $\tilde{\mathscr{I}} = \mathscr{I}$ then $II = 0$. Otherwise, write $N = \mathscr{I} \setminus \tilde{\mathscr{I}}$ and assume w.l.o.g. that the first $|\tilde{\mathscr{I}}|$ indices in $\mathscr{I}$ are from $\tilde{\mathscr{I}}$. Since $\tilde{\mathscr{I}} = \mathscr{I} \cap \mathscr{G}_S^{\delta}$, there

are no edges between nodes in N and nodes in $\tilde{\mathscr{I}}$ in the graph $\mathscr{G}_S^{\delta}$. This implies that

$$\|G^{\tilde{\mathscr{I}},N}\|_{\max} \leq \delta \leq C\delta_p.$$

Introduce a blockwise diagonal matrix $D = \mathrm{diag}\big(G^{\tilde{\mathscr{I}},\tilde{\mathscr{I}}}, G^{N,N}\big)$. It is seen that

$$\begin{aligned} II &= |(G^{\mathscr{I},\mathscr{I}})^{-1}(1,1) - D^{-1}(1,1)| \leq \|(G^{\mathscr{I},\mathscr{I}})^{-1} - D^{-1}\| \\ &\leq \|(G^{\mathscr{I},\mathscr{I}})^{-1}\|\|D^{-1}\|\|G^{\mathscr{I},\mathscr{I}} - D\| \lesssim c_0^{-2}\|G^{\mathscr{I},\mathscr{I}} - D\| \\ &= c_0^{-2}\left\|\begin{bmatrix} 0 & G^{\tilde{\mathscr{I}},N} \\ G^{N,\tilde{\mathscr{I}}} & 0 \end{bmatrix}\right\| \leq c_0^{-2}\|G^{\tilde{\mathscr{I}},N}\| \leq c_0^{-2} g\|G^{\tilde{\mathscr{I}},N}\|_{\max} = O(\delta_p). \end{aligned} \tag{4.58}$$

Combining (4.57)-(4.58), we prove (4.56).

Now suppose $\mathscr{I} \subset \mathscr{G}_S^{\delta}$. We've shown that $|A_{j|\mathscr{I}} - A^0_{j|\mathscr{I}}| = o(\delta_p)$. It suffices to show that the difference between $B^0 = G_0^{j,F} - G_0^{j,N}(G_0^{N,N})^{-1}G_0^{N,F}$ and $B = G^{j,F} - G^{j,N}(G^{N,N})^{-1}G^{N,F}$ is negligible. In fact, by similar argument in (4.57) we have $\|(G^{\mathscr{I}^{(j)},\mathscr{I}^{(j)}})^{-1} - (G_0^{\mathscr{I}^{(j)},\mathscr{I}^{(j)}})^{-1}\| = o(\delta_p)$. Suppose w.l.o.g that $F\cup\{j\}$ are the first several indices of $\mathscr{I}^{(j)}$ where $\mathscr{I}^{(j)} = F\cup\{j\}\cup N$, then we know the inverse of $\tilde{B} = G^{F\cup\{j\},F\cup\{j\}} - G^{F\cup\{j\},N}(G^{N,N})^{-1}G^{N,F\cup\{j\}}$ is the upper left block of $(G^{\mathscr{I}^{(j)},\mathscr{I}^{(j)}})^{-1}$, and B is a submatrix of $\tilde{B}$. We can define $\tilde{B}^0$ similarly where B^0 is a submatrix of $\tilde{B}^0$. By some simple algebra, we get $\|B - B^0\| = o(\delta_p)$

Proof of Lemma 4.5.4 Recall that S is the support set of β and $|S| = s_p$. It is seen that

$$\begin{aligned} \big\|\big(G^{\mathscr{I},\mathscr{I}} - (G^{\delta})^{\mathscr{I},\mathscr{I}}\big)\beta^{\mathscr{I}}\big\|_{\infty} &= \big\|\big(G^{\mathscr{I},\mathscr{I}\cap S} - (G^{\delta})^{\mathscr{I},\mathscr{I}\cap S}\big)\beta^{\mathscr{I}\cap S}\big\|_{\infty} \\ &\leq \|G^{\mathscr{I},S} - (G^{\delta})^{\mathscr{I},S}\|_{\infty}\|\beta^{\mathscr{I}\cap S}\|_{\infty} \\ &\leq a\tau_p \cdot \|G^{\mathscr{I},S} - (G^{\delta})^{\mathscr{I},S}\|_{\infty}. \end{aligned}$$

Therefore, to show the claim, it suffices to show that

$$\|G^{\mathscr{I},S} - (G^{\delta})^{\mathscr{I},S}\|_{\infty} \leq C\left([\log(p)]^{-(1-\gamma)} + s_p\|G - G_0\|_{\max}\right). \tag{4.59}$$

For any $1 \leq i \leq p$, we define $I_i = \{1 \leq j \leq p : |G(i,j)| \leq \delta\}$. Then,

$$\begin{aligned}
\|G^{\mathscr{I},S} - (G^{\delta})^{\mathscr{I},S}\|_{\infty} &\leq \max_{1\leq i\leq p} \sum_{j\in S} |G(i,j) - G^{\delta}(i,j)| = \max_{1\leq i\leq p} \sum_{j\in S\cap I_i} |G(i,j)| \\
&\leq \max_{1\leq i\leq p} \sum_{j\in S\cap I_i} |G_0(i,j)| + \max_{1\leq i\leq p} \sum_{j\in S\cap I_i} |G_0(i,j) - G(i,j)| \\
&\leq \max_{1\leq i\leq p} \sum_{j\in S\cap I_i} |G_0(i,j)| + s_p\|G - G_0\|_{\max}.
\end{aligned}$$

Therefore, to show (4.59), it suffices to show that for any $1 \leq i \leq p$,

$$\sum_{j\in S\cap I_i} |G_0(i,j)| \leq C[\log(p)]^{-(1-\gamma)}. \tag{4.60}$$

We now show (4.60). For any $j \in I_i$, $|G_0(i,j)| \leq |G(i,j)| + \|G - G_0\|_{\max} \leq \delta + \|G - G_0\|_{\max}$, where $\delta \leq C\delta_p = Cb/\log(p)$ by (4.15) and $\|G - G_0\|_{\max} = o(1/\log(p))$ by Lemma 4.2.2. Hence, $|G_0(i,j)| \leq b_1/\log(p)$ whenever $j \in I_i$, where $b_1 > 0$ is a constant. We have

$$\begin{aligned}
\sum_{j\in S\cap I_i} |G_0(i,j)| &\leq \sum_{j\in I_i} |G_0(i,j)|^{\gamma} |G_0(i,j)|^{1-\gamma} \\
&\leq b_1^{1-\gamma}[\log(p)]^{-(1-\gamma)} \sum_{j\in I_i} |G_0(i,j)|^{\gamma} \\
&\leq b_1^{1-\gamma}[\log(p)]^{-(1-\gamma)} \cdot C_0,
\end{aligned}$$

where we have used the assumption $G_0 \in \mathscr{M}_p(g,\gamma,c_0,C_0)$ in the last inequality. This proves (4.60).

CHAPTER 5

DIAGONALLY DOMINANT PRINCIPAL COMPONENT ANALYSIS

The *approximate low-rankness* is a popular structural assumption on covariance matrices. It assumes that a $p \times p$ covariance matrix $\mathbf{\Sigma}$ decomposes into

$$\mathbf{\Sigma} = \boldsymbol{L} + \boldsymbol{A}, \qquad \text{where} \quad \text{rank}(\boldsymbol{L}) = K \ll p, \quad \text{and } \boldsymbol{A} \text{ is a "nice" matrix.} \tag{5.1}$$

Equivalently, it introduces a latent factor model on any random vector X whose covariance matrix is $\mathbf{\Sigma}$, where $\boldsymbol{A}$ is the "residual covariance matrix" after the effects of latent variables are removed. Such a decomposition is not unique and varies with the meaning of a "nice" $\boldsymbol{A}$. One can impose different requirements on $\boldsymbol{A}$ to facilitate different applications. In the classical factor models for econometrics and finance, $\boldsymbol{A}$ is assumed to be a diagonal matrix [80] or a sparse matrix [81], to enforce that the idiosyncratic noise accounts for little cross-sectional risk. In large-scale multiple testing, it is often assumed that the covariance matrix of test statistics has the above decomposition with $\boldsymbol{A}$ being a diagonal matrix [82] or having a small Frobenius norm [83]. The motivation there is development of factor-adjusted multiple testing procedures, to make it legitimate to use conventional multiple testing methods on the post-factor-removal data. In image processing, a similar decomposition on image matrices was proposed [84], where $\boldsymbol{A}$ is assumed to be sparse, for the purpose of capturing details of images.

In this chapter, [1] we explore a new type of *approximate low-rankness* of Σ where

$$\text{Each diagonal entry of } \boldsymbol{A} \text{ is large compared with other entries in the same row.} \tag{5.2}$$

Translated to the latent variable representation, it means, after the effects of latent variables are removed, the *correlation* matrix of "residual" variables have uniformly small off-diagonal entries. One motivation of imposing this condition is to take into account the varying scale of the diagonal

1. The work presented in this chapter is joint with Tracy Ke and Lingzhou Xue, and is under revision of Journal of Computational and Graphic Statistics.

elements of $\boldsymbol{A}$. Most aforementioned approximate low-rank decompositions first perform PCA on $\boldsymbol{\Sigma}$ (or an empirical version of it) to remove the first a few principal components, and then conduct operations on the remaining matrix. It is often observed that the diagonal elements of the remaining matrix has considerable variations in magnitude. To deal with it requires careful adjustment on the post-PCA operations, such as adaptive thresholding [85]. On the contrary, we impose the requirement (5.2) directly in the decomposition (5.1), in hopes of improving the PCA factors and easing the post-PCA operations. Another motivation of adopting the assumption (5.2) is to guarantee that $\boldsymbol{A}^{-1}$ is well-behaved. In many applications such as portfolio management and linear discriminant analysis, $\boldsymbol{A}^{-1}$ plays a key role. In the decomposition (5.1), forcing $\boldsymbol{A}$ to be a strictly diagonal matrix ensures both $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$ are well-behaved, but this requirement is often too restrictive, and (5.2) is a natural relaxation. We note that imposing the common sparsity assumption on $\boldsymbol{A}$ does not even guarantee positive definiteness. Despite remedies such as increasing the threshold or projection to the positive semi-definite cone [86], these approaches still don't guarantee that $\boldsymbol{A}^{-1}$ is a "nice" matrix.

5.1 Problem and methods

To formulate (5.2) mathematically, we define the set of "symmetric c-diagonally-dominant" matrices, for any $c > 0$:

$$\mathscr{SDD}_c^+ = \Big\{ \boldsymbol{A} = (a_{ij})_{p\times p} : \boldsymbol{A}^T = \boldsymbol{A},\, a_{jj} \geq c \sum_{i:i\neq j} |a_{ji}| \ \text{ for all } 1 \leq j \leq p \Big\}. \tag{5.3}$$

For $c = 1$, it reduces to the usual definition of diagonally-dominant matrices, and we omit the subscript and write $\mathscr{SDD}_1^+ = \mathscr{SDD}^+$. Given a $p \times p$ positive semi-definite matrix $\boldsymbol{S}$, we introduce an optimization problem:

$$\min_{(\boldsymbol{L},\boldsymbol{A})} \|\boldsymbol{S} - \boldsymbol{L} - \boldsymbol{A}\|_F, \qquad \text{subject to} \quad \text{rank}(\boldsymbol{L}) \leq K,\ \boldsymbol{L} = \boldsymbol{L}^T,\ \boldsymbol{A} \in \mathscr{SDD}_c^+, \tag{5.4}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm. We call it *Diagonally-Dominant Principal Component Analysis (DD-PCA)*. In this chapter, we are primarily interested in $c = 1$; discussions of $c \leq 1$ are deferred to Section 5.4.

The definition of DD-PCA is a nonconvex optimization with a rank constraint. Similar to solving other rank constrained optimizations in matrix completion, one can either solve a convex relaxation of (5.4) or develop an iterative algorithm that converges to a local minimum of (5.4). These ideas generate several variants of DD-PCA, as detailed in Section 5.4. Among those variants, one is of particular interest, which we call *One-step DD-PCA*:

- PCA: Obtain the K leading eigenvalues and eigenvectors of $\boldsymbol{S}$, denoted as $\lambda_1 \geq \ldots \geq \lambda_K \geq 0$ and $\xi_1, \ldots, \xi_K \in \mathbb{R}^p$. Let $\boldsymbol{L} = \sum_{k=1}^K \lambda_k \xi_k \xi_k^T$.
- Projection to $\mathscr{SDD}^+$: Initialize $\boldsymbol{A}^{(0)} = \boldsymbol{S} - \boldsymbol{L}$ and $\boldsymbol{J}^{(0)} = \boldsymbol{0}$. For $t = 1, 2, \ldots,$
  - Run the MRT algorithm [87][2] to project $[\boldsymbol{A}^{(t-1)} - \boldsymbol{J}^{(t-1)}]$ into the diagonally-dominant cone. Let $\boldsymbol{G}^{(t)}$ be the projected matrix.
  - Update $\boldsymbol{A}^{(t)} = \frac{1}{2}[\boldsymbol{G}^{(t-1)} + (\boldsymbol{G}^{(t-1)})^T]$ and $\boldsymbol{J}^{(t)} = \boldsymbol{J}^{(t-1)} + (\boldsymbol{G}^{(t)} - \boldsymbol{A}^{(t-1)})$.
  - If $\|\boldsymbol{J}^{(t)} - \boldsymbol{J}^{(t-1)}\|_F \leq \varepsilon$, stop and output $\boldsymbol{A} = \boldsymbol{A}^{(t)}$.

This method is obtained by running one outer-loop iteration in the iterative algorithm to be introduced in Section 5.4, explaining the name of *one-step DD-PCA*. It has the same philosophy as the one-step Huber estimator [88] and one-step LLA implementation of non-convex penalized linear regressions [89, 90]. It provides an approximate solution to (5.4), which is much faster to compute than solving (5.4) exactly.

Exploring the approximate low-rank structures is a powerful strategy for big data analysis. The classical PCA has motivated many statistical methods. Similarly, DD-PCA and one-step DD-PCA can also serve as building blocks for statistical methodology development. We exemplify it in two

2. The MRT algorithm computes the unique projection of a $p \times p$ matrix to the convex polyhedral cone consisting of all diagonally dominant matrices. It has a complexity of $O(p^2 \log(p))$. See Section 5.4.

statistical problems: the first is estimating a large covariance matrix, and the second is testing of the global null hypothesis in multiple testing.

Estimation of large covariance matrices is a popular topic in statistical literatures [86]. At the heart of it is two fundamental questions: (a) What structural assumption is appropriate? (b) How to evaluate the methods in real applications?

We adopt the structural assumption that the true covariance matrix $\mathbf{\Sigma}$ has an approximate low-rank decomposition with $\boldsymbol{A} \in \mathscr{SDD}_c^+$. This is a special type of factor covariance structures that are commonly used in econometrics [91], finance [80], genetics [92] and many other fields. Our work is unique in the diagonal dominance assumption on $\boldsymbol{A}$. Intuitively, it is a natural relaxation of assuming $\boldsymbol{A}$ is diagonal, and it implies that, after the effects of latent factors are removed, the "residual variables" are almost *uncorrelated*. Compared with existing covariance matrix estimators which assumes $\boldsymbol{A}$ is sparse (e.g., [93, 6]), this diagonal dominance structure facilitates simultaneous estimation of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$. In factor covariance structures, the singular values of the low rank matrix are much larger than $\|\boldsymbol{A}\|$, so the error of estimating $\mathbf{\Sigma}$ is dominated by the error of recovering the low-rank part. If our goal is merely to estimate $\mathbf{\Sigma}$, we do not gain much from exploring the diagonal dominance structure of $\boldsymbol{A}$. However, if we are also interested in estimating $\mathbf{\Sigma}^{-1}$, the error of estimating $\boldsymbol{A}^{-1}$ will play a key role. Note that there always exists a matrix $\boldsymbol{B} \in \mathbb{R}^{n\times K}$ such that $\boldsymbol{L} = \boldsymbol{B}\boldsymbol{B}^T$. It follows from the matrix inverse formula [94] that

$$\mathbf{\Sigma}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_K + \boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1}.$$

Suppose we have obtained a good estimator $\widehat{\mathbf{\Sigma}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T + \widehat{\boldsymbol{A}}$ by fitting some factor covariance structure on the data. Even though $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$ is small, it is still possible that $\|\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1}\|$ is large so that $\widehat{\mathbf{\Sigma}}^{-1}$ is far from being a good estimator of $\mathbf{\Sigma}^{-1}$. Fortunately, exploring the diagonal dominance structure greatly mitigates this issue, thanks to an appealing feature of the diagonally-

dominant cone $\mathscr{SDD}_c^+$ [94]:

$$\|\boldsymbol{A}^{-1}\| \leq \frac{c}{c-1}\|[\text{diag}(\boldsymbol{A})]^{-1}\|, \qquad \text{for any } \boldsymbol{A} \in \mathscr{SDD}_c^+, \text{ where } c > 1.$$

Therefore, if we enforce $\widehat{\boldsymbol{A}} \in \mathscr{SDD}_c^+$ in fitting the factor covariance structure, for a constant $c > 1$, then $\|\widehat{\boldsymbol{A}}^{-1}\|$ won't explode, preventing ill behaviors of $\widehat{\boldsymbol{\Sigma}}^{-1}$. To this end, we propose a new covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{ddpca}$ using the solution of DD-PCA or one-step DD-PCA. We demonstrate in numerical studies: $\widehat{\boldsymbol{\Sigma}}_{ddpca}$ has comparable performance with other factor-based covariance matrix estimators (e.g., [6]), but the new estimator is tuning free once K is specified, so is more convenient to use. At the same time, it facilitates the use of estimating $\boldsymbol{\Sigma}^{-1}$ by $\widehat{\boldsymbol{\Sigma}}_{ddpca}^{-1}$, and its performance on estimating $\boldsymbol{\Sigma}^{-1}$ is much better than inverting other factor-based covariance matrix estimators.

In real applications, estimating the covariance matrix is rarely the ultimate goal. Often, it serves as an intermediate step for downstream tasks. We demonstrate the usefulness of our covariance estimator by evaluating its performance in two downstream tasks, portfolio management and linear discriminant analysis. In the former, an estimate of the covariance matrix is needed to obtain Markowitz's optimal portfolio weights; in the latter, it is used to compute Fisher's LDA classifier. Note that what is actually plugged into these downstream tasks is the *inverse* of estimated covariance matrix. As we have argued, the main advantage of our method is in estimating $\boldsymbol{\Sigma}^{-1}$ by $\widehat{\boldsymbol{\Sigma}}_{ddpca}^{-1}$, a perfect match to these applications. This is supported by encouraging real data results. It is worthwhile mentioning that our approach is different from the approach of plugging in an existing precision matrix estimator (e.g., the graphical lasso [95]). These methods assume $\boldsymbol{\Sigma}^{-1}$ is sparse, while we assume a factor-type structure on $\boldsymbol{\Sigma}$. For portfolio data, adopting a factor-type covariance structure is the common practice. For classification problems, there also exist a lot of real data sets on which the factor-type structure is appropriate [92].

DD-PCA is also useful to multiple testing. A fundamental challenge of multiple testing is how to deal with complicated correlations. One popular approach to modeling the data correlation is

to assume the covariance matrix of Z-statistics has a latent factor structure as in (5.1). By decomposing the sample covariance matrix $\boldsymbol{S}$ in a similar fashion, one can remove the effects of latent factors and apply standard multiple testing methods to the post-factor-removal data (citations). Different decompositions of $\boldsymbol{S}$ result in different factor-removal procedures, and a good factor removal procedure should guarantee the legitimate use of standard multiple testing methods on post-factor-removal data. We notice that most standard multiple testing procedures (e.g., extreme value test for the global null hypothesis, Benjamini-Hochberg method for false discovery rate control) perform well when the data are weakly correlated. This motivates us to use DD-PCA to decompose the sample covariance matrix to get a new factor-removal procedure. We combine this factor removal approach with the Higher Criticism (HC) test [7] for testing the global null. It gives rise to a new test statistic DD-HC, which improves the original HC and significantly outperforms popular tests for global null when the covariance matrix has a latent factor structure.

5.2 Estimating large covariance matrices by DD-PCA

Let $X \in \mathbb{R}^p$ be a multivariate random vector with a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, where p is presumably much larger than n. We adopt a factor model:

$$X(j) = \sum_{k=1}^{K} b_k(j)W_k + Z(j), \qquad 1 \le j \le p, \tag{5.5}$$

where $W_1, \ldots, W_K$ are unobserved random variables (factors), $b_k \in \mathbb{R}^p$ is a nonrandom vector containing the loadings of the k-th factor, and $Z \in \mathbb{R}^p$ is a random vector independent of the factors such that

$$\boldsymbol{A} \equiv \mathrm{Cov}(Z) \in \mathscr{S}\mathscr{D}\mathscr{D}^+. \tag{5.6}$$

Given *iid* data $X_1, \ldots, X_n \in \mathbb{R}^p$, we are interested in estimating $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

By model (5.5)-(5.6), the covariance matrix of X has a decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{B}\mathrm{Cov}(W)\boldsymbol{B}^T + \boldsymbol{A}, \qquad \text{where} \quad \mathrm{rank}\big(\boldsymbol{B}\mathrm{Cov}(W)\boldsymbol{B}^T\big) = K \quad \text{and} \quad \boldsymbol{A} \in \mathscr{S}\mathscr{D}\mathscr{D}^+.$$

It has the low rank plus diagonal dominance structure. We propose the following estimator: Let $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$ be the sample covariance matrix. Take $\boldsymbol{S}$ as the input to the one-step DD-PCA algorithm in Section 5.1 and let $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ be the output. We estimate $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}}_{ddpca} = \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}}, \qquad \text{where } (\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}}) \text{ is the output of one-step DD-PCA.} \tag{5.7}$$

We then estimate $\boldsymbol{\Sigma}^{-1}$ by the inverse of $\widehat{\boldsymbol{\Sigma}}_{ddpca}$. Here, $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ can be replaced by the output of other variants of DD-PCA (see Section 5.4). They give similar numerical performance, so we stick to one-step DD-PCA for computational convenience.

Different from existing covariance estimation methods under factor structures, our approach imposes the diagonal dominance constraint on $\boldsymbol{A}$. We now compare it with methods that impose the sparsity constraint on $\boldsymbol{A}$. One popular method is POET [6]. Let $\boldsymbol{S} = \sum_{k=1}^{p}\lambda_k\xi_k\xi_k^T$ be the eigen-decomposition of $\boldsymbol{S}$, where λ_k and ξ_k are the k-th eigenvalue and eigenvector, respectively. The POET estimator is

$$\widehat{\boldsymbol{\Sigma}}_{poet} = \widehat{\boldsymbol{L}}_* + \mathscr{T}_a(\widehat{\boldsymbol{A}}_*), \qquad \text{where} \quad \widehat{\boldsymbol{L}}_* = \sum_{k=1}^{K}\lambda_k\xi_k\xi_k^T, \quad \widehat{\boldsymbol{A}}_* = \sum_{k=K+1}^{p}\lambda_k\xi_k\xi_k^T. \tag{5.8}$$

Here, $\mathscr{T}_a(\cdot)$ can be any entry-wise adaptive thresholding operator [96, 85]. [6] suggest using the hard-thresholding operator applied to a "correlation matrix" associated with $\widehat{\boldsymbol{A}}_*$, i.e.,

$$\mathscr{T}_a(\widehat{\boldsymbol{A}}_*) = \widehat{\boldsymbol{D}}^{\frac{1}{2}}H_a\Big(\widehat{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{A}}_*\widehat{\boldsymbol{D}}^{\frac{1}{2}}\Big)\widehat{\boldsymbol{D}}^{\frac{1}{2}} \qquad \text{where} \quad \widehat{\boldsymbol{D}} = \text{diag}(\widehat{\boldsymbol{A}}_*), \tag{5.9}$$

where $H_a(\cdot)$ is the entry-wise hard-thresholding at the threshold $a > 0$. Then, an estimate of $\boldsymbol{\Sigma}^{-1}$ is obtained by $\widehat{\boldsymbol{\Sigma}}_{poet}^{-1}$.

Figure 5.1 gives a simulation example. Fix $(p, n, K) = (2000, 200, 3)$. We generate data from the model (5.5), where the factors $\{W_k(i) : 1 \le k \le K, 1 \le i \le n\}$ are drawn *iid* from $N(0,1)$, the factor loadings $\{b_k(j) : 1 \le k \le K, 1 \le j \le p\}$ are generated *iid* from $N(0,1)$, and the noise vectors $Z_1, \ldots, Z_n$ are drawn *iid* from a multivariate normal $N(\boldsymbol{0}, \boldsymbol{A})$, where $A(i,j) = 0.5^{|i-j|+1}$ for $i \neq j$

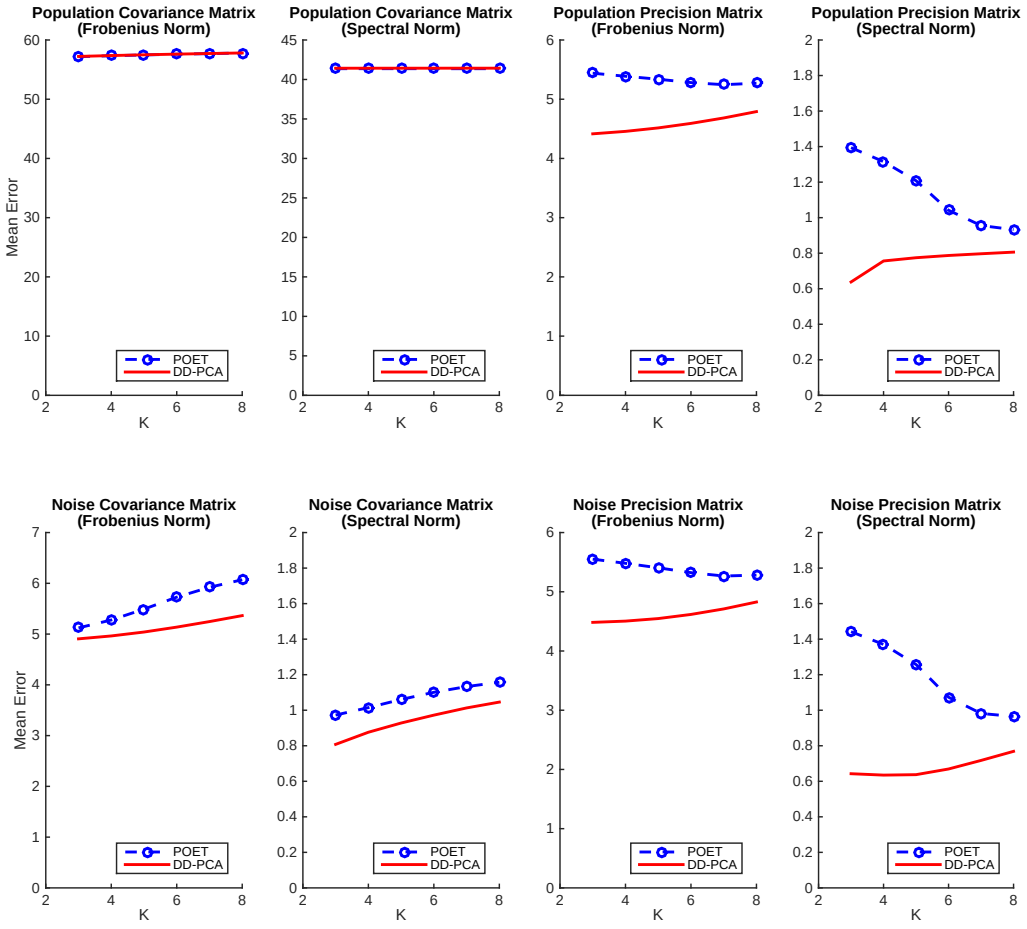


Figure 5.1: Comparison of our method with POET on estimating $\boldsymbol{\Sigma}$ (covariance matrix), $\boldsymbol{\Sigma}^{-1}$ (precision matrix), $\boldsymbol{A}$ (noise covariance matrix) and $\boldsymbol{A}^{-1}$ (noise precision matrix).

and 1 otherwise. For both methods, K is unknown and treated as a tuning integer. POET has an additional tuning threshold a, which is selected by cross-validation (default procedure in the *poet* package).[3]

In the top four panels of Figure 5.1, we show the average estimation errors on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ over 100 repetitions. Since K is unknown, we implement both methods for the true $K = 3$ and misspecified $K \in \{4,5,\ldots,8\}$.[4] For estimating $\boldsymbol{\Sigma}$, the two methods give very similar performance. This is not surprising. Since the eigenvalues of the low-rank part are much larger than $\|\boldsymbol{A}\|$, the

3. This default procedure guarantees that $\widehat{\boldsymbol{\Sigma}}_{poet}$ is invertible.

4. We don't include the results of $K \in \{1,2\}$, as the errors are much larger.

error of estimating $\mathbf{\Sigma}$ is dominated by the error of recovering the low-rank part. Our method and POET has the same low rank part (the $\widehat{\boldsymbol{L}}$ from one-step DD-PCA and the $\widehat{\boldsymbol{L}}_*$ in (5.8) are indeed the same), so they have similar errors on estimating $\mathbf{\Sigma}$. From the bottom left two panels of Figure 5.1, we can see that our method does a better job on estimating $\boldsymbol{A}$, especially, the spectral norm error is 10-20% smaller. However, this improvement is almost negligible compared with the errors on recovering the low-rank part. We conclude that our method and POET have similar performance on estimating $\mathbf{\Sigma}$. Still, our method has an advantage: It has no tuning threshold and is more convenient to use.

How about the performance on estimating $\mathbf{\Sigma}^{-1}$? The top right two panels of Figure 5.1 clearly suggest that our method has a significant advantage. When $K = 3$, the spectral norm error of our method is only one half of the error of POET. Interestingly, the performance of POET improves with an overshooting K; but even for $K = 8$, its spectral norm error is still 10% larger than the error of our method. For the Frobenius norm error, our estimator also outperforms POET for all choices of K. This phenomenon is due to that $\widehat{\boldsymbol{A}}$ plays a dominating role when we compute the inverse of $\widehat{\mathbf{\Sigma}}$, while the low-rank part has a negligible effect, so the advantage of our method on recovering $\boldsymbol{A}$ becomes prominent. This is illustrated in the bottom right two panels of Figure 5.1. We recall that $\widehat{\boldsymbol{A}}$ from one-step DD-PCA and $\widehat{\boldsymbol{A}}_*$ as in (5.8). The Frobenius/spectral norm of $(\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1})$ is significantly smaller than the Frobenius/spectral norm of $(\widehat{\boldsymbol{A}}_*^{-1} - \boldsymbol{A}^{-1})$. Additionally, by comparing the top right two panels with the bottom right two panels, we can see that the error of estimating $\mathbf{\Sigma}^{-1}$ is almost determined by the error of estimating $\boldsymbol{A}^{-1}$.

This numerical example delivers two messages: First, compared with competitive factor-based methods, the major advantage of our method is on estimating $\mathbf{\Sigma}^{-1}$ by $\widehat{\mathbf{\Sigma}}^{-1}$. Second, such an advantage is driven by the better accuracy on recovering $\boldsymbol{A}^{-1}$. Below, we explain them using linear algebra.

Without loss of generality, in model (5.5), we assume the covariance matrix of W equals to the

identify matrix. Then, $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{A}$. By matrix inverse formula,

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_K + \boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1}.$$

Suppose we construct an estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T + \widehat{\boldsymbol{A}}$ from fitting a factor-type covariance structure. Then, $\widehat{\boldsymbol{\Sigma}}^{-1}$ (if it exists) has a similar decomposition:

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\boldsymbol{A}}^{-1} - \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}(\boldsymbol{I}_K + \widehat{\boldsymbol{B}}^T\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}})^{-1}\widehat{\boldsymbol{B}}^T\widehat{\boldsymbol{A}}^{-1}.$$

By some basic linear algebra, we can derive the following proposition:

Proposition 5.2.1. *Let $\widehat{\boldsymbol{A}}^{-\frac{1}{2}}\widehat{\boldsymbol{B}} = \sum_{k=1}^{K}\hat{\sigma}_k\hat{\eta}_k\hat{h}_k'$ be the singular value decomposition of $\widehat{\boldsymbol{A}}^{-\frac{1}{2}}\widehat{\boldsymbol{B}}$, where $\hat{\sigma}_k > 0$ is the k-th singular value and $\hat{\eta}_k \in \mathbb{R}^p$ and $\hat{h}_k \in \mathbb{R}^K$ are the corresponding left and right singular vectors. Then,*

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\boldsymbol{A}}^{-1} - \widehat{\boldsymbol{A}}^{-\frac{1}{2}}\left(\sum_{k=1}^{K}\frac{1}{\hat{\sigma}_k^{-2}+1}\hat{\eta}_k\hat{\eta}_k'\right)\widehat{\boldsymbol{A}}^{-\frac{1}{2}} \tag{5.10}$$

By (5.10), the error of recovering the low-rank part only affects the matrix in the brackets. For $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$ obtained in factor-based methods, nonzero eigenvalues of $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T$ are much larger than $\|\widehat{\boldsymbol{A}}\|$, so $\widehat{\sigma}_k$'s are all very large. Then, the matrix in the brackets can hardly bring in a large error in $\widehat{\boldsymbol{\Sigma}}^{-1}$. The error in $\widehat{\boldsymbol{\Sigma}}^{-1}$ mainly comes from the error in $\widehat{\boldsymbol{A}}^{-1}$.

We further investigate the error in $\widehat{\boldsymbol{A}}^{-1}$. Note that

$$\|\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1}\| \le \|\widehat{\boldsymbol{A}}^{-1}\|\|\boldsymbol{A}^{-1}\|\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|. \tag{5.11}$$

To achieve a small $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$ by imposing structural assumptions on $\boldsymbol{A}$ is not too difficult. However, it typically does not prevent $\|\widehat{\boldsymbol{A}}^{-1}\|$ from exploding. For example, if $\widehat{\boldsymbol{A}}$ is obtained from entry-wise thresholding, we need a comparably large threshold to control $\|\widehat{\boldsymbol{A}}^{-1}\|$, but unfortunately we cannot let the threshold be too large as it significantly increases $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$. It turns out that, if we restrict $\widehat{\boldsymbol{A}} \in \mathscr{SDD}_c^+$ for a constant $c > 1$, then it is automatically guaranteed that $\|\widehat{\boldsymbol{A}}^{-1}\|$ has a

nice bound. As a property of diagonally-dominant matrices [94], for $c > 1$,

$$\lambda_{\min}(\widehat{\boldsymbol{A}}) \geq \min_{1\leq j\leq p}\Big\{a_{jj} - \sum_{i:i\neq j}|a_{ji}|\Big\} \geq \min_{1\leq j\leq p}\Big\{a_{jj} - c^{-1}a_{jj}\Big\} \geq \frac{c-1}{c}\min_{1\leq j\leq p} a_{jj}.$$

It follows that

$$\|\widehat{\boldsymbol{A}}^{-1}\| \leq \frac{c}{c-1}\|[\mathrm{diag}(\boldsymbol{A})]^{-1}\|. \tag{5.12}$$

This explains why the constraint of $\widehat{\boldsymbol{A}} \in \mathscr{SDD}_c^+$ helps significantly reduce the errors in $\widehat{\boldsymbol{A}}^{-1}$ and (ultimately) the errors in $\widehat{\boldsymbol{\Sigma}}^{-1}$.

The above argument applies to $c > 1$. In our method, $c = 1$. Sometimes, we may even have to use $c < 1$, so that the assumption $\boldsymbol{A} \in \mathscr{SDD}_c^+$ is not too restrictive (see Section 5.4). For $c \leq 1$, we do not have a solid argument as (5.12), but a similar phenomenon is observed in numerical studies.

Below, we use two real applications to further demonstrate that exploring the diagonal dominance factor structures is a useful strategy.

5.2.1 Application to portfolio management

Given a collection of p assets, portfolio management aims to determine the weights allocated to each asset. It is often desirable to construct the *minimum risk portfolio*, where the asset weights $\boldsymbol{w}^* = (w_1^*, \ldots, w_p^*)$ are determined by

$$\boldsymbol{w}_* = \mathrm{argmin}_{\boldsymbol{w}^T\mathbf{1}=1}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}, \qquad \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}\text{: positive definite asset covariance matrix.}$$

In practice, $\boldsymbol{\Sigma}$ is unknown. We first obtain an estimate $\widehat{\boldsymbol{\Sigma}}$ using asset returns $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^p$ during a period of n days, then we estimate the weights by

$$\widehat{\boldsymbol{w}}_* = \mathrm{argmin}_{\boldsymbol{w}^T\mathbf{1}=1}\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}.$$

This optimization has an explicit solution:

$$\widehat{\boldsymbol{w}}_* = (\mathbf{1}^T \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{1})^{-1} (\widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{1}). \tag{5.13}$$

Since what we actually need is $\widehat{\boldsymbol{\Sigma}}^{-1}$, exploring the low-rank plus diagonal dominance structure is a potentially useful strategy.

We compare our method with POET on real data. We collected the daily returns of stocks in S&P 100 index from January 1st 2006 to December 31st 2016. After removing companies that were listed after 2006, there are 80 stocks in total. On the first trading day of each month, we created two portfolios from (5.13), where $\widehat{\boldsymbol{\Sigma}}$ is estimated using daily returns for the proceeding 12 months ($n = 252$) by our method and by POET, respectively. We set $K = 3$ for both methods. The threshold in POET is chose by cross-validation (we use the default cross-validation procedure in *poet* package). On the last trading day of the same month, we measure the actual risk of each portfolio by

$$R(\widehat{\boldsymbol{w}}_*) = \frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{y}_t^T \widehat{\boldsymbol{w}}^*)^2,$$

where T is the number of trading days in this month ($T = 21$ for most months) and $\boldsymbol{y}_t \in \mathbb{R}^{80}$ contains the stock returns on day t of the month.

Define $r = (R_{poet} - R_{ddpca})/R_{ddpca}$; note that a positive r indicates that the portfolio created using our method is superior to that of POET. Figure 5.2 displays the histogram of r over 120 months in our data range. It suggests that our method improves POET by 9.5% on average and 14.7% in the median.

5.2.2 Application to linear discriminant analysis

In binary classification, given feature vectors $X_1, \ldots, X_n \in \mathbb{R}^p$ and training labels $\ell_1, \ldots, \ell_n \in \{1, 2\}$, we aim to construct a linear classifier. In the classical regime where p is fixed as the training sample size grows, Fisher's LDA is an effective linear classifier. In the modern high dimensional settings where $p \gg n$, it has been well understood that feature screening is necessary before one applies

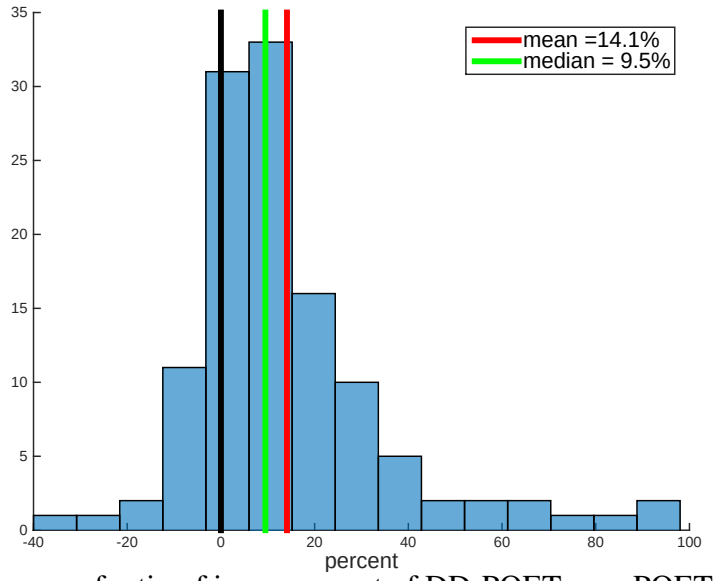


Figure 5.2: Histogram of ratio of improvement of DD-POET over POET over 120 months.

Fisher's LDA [91, 97], and that it is desirable to plug in a good estimate of the inverse covariance matrix that explores structural assumptions [98]. Recently, [99] proposed a linear classifier that uses an estimate of inverse covariance matrix in both the screening step and LDA step, and they showed that this classifier is rate-optimal under a multivariate normal model with even extremely weak signal strength. This classifier was later applied to several large real classification problems with superior results [100]. We shall combine our covariance matrix estimator with this classifier to see whether exploring the low-rank plus diagonal-dominance structure is helpful.

Given an estimate $\widehat{\Omega}$ of the inverse covariance matrix and a threshold $t > 0$, the classifier has four steps [100]:

1. Calculate the feature-wise t-score: For $1 \leq j \leq p$, let $Z(j) = [\bar{X}_1(j) - \bar{X}_2(j)]/(n \cdot s_j)$, where $\bar{X}_1(j)$ and $\bar{X}_2(j)$ are the within-class sample means of feature j and $s_j > 0$ is the pooled standard deviation of feature j. Write $Z = (Z(1), \ldots, Z(p))^T$.

2. Apply the Innovated Transformation [99] to get $\tilde{Z} = \widehat{\Omega} Z$.

3. Feature-wise thresholding: For $1 \leq j \leq p$, let $w(j) = \mathrm{sgn}(\tilde{Z}(j)) \cdot 1\{|\tilde{Z}(j)| \geq t\}$. Write $w = (w(1), w(2), \ldots, w(p))^T$.

4. Classification by LDA. Given a test feature vector $\widetilde{X} \in \mathbb{R}^p$, for $1 \le j \le p$, normalize $\widetilde{X}(j)$ to $\widetilde{X}^*(j) = [\tilde{X}(j) - \frac{1}{2}(\bar{X}_1(j) + \bar{X}_2(j))]/s_j$, where $(\bar{X}_1(j), \bar{X}_2(j), s_j)$ are the same as in Step 1. Write $\widetilde{X}^* = (\widetilde{X}^*(1), \ldots, \widetilde{X}^*(p))^T$. We classify the test sample to class 1 if $w^T \widehat{\Omega} \widetilde{X}^* > 0$ and to class 2 otherwise.

In this classifier, the matrix $\widehat{\Omega}$ plays two roles: First, it is used in the Innovated Transformation, so different $\widehat{\Omega}$ leads to different feature rankings. Second, it is used in the LDA step, so $\widehat{\Omega}$ also affects the classification boundary.

We compare the classification performance of plugging in three versions of $\widehat{\Omega}$: The first is $\widehat{\Sigma}^{-1}_{ddpca}$, the second is $\widehat{\Sigma}^{-1}_{poet}$, and the last is $[\mathrm{diag}(\boldsymbol{S})]^{-1}$, where $\boldsymbol{S}$ is the sample covariance matrix. We note that the last approach is indeed the method FAIR [91]. The above classifier also requires a threshold $t > 0$. To minimize the effects of selecting t, for each $1 \le k \le p$, we set the threshold such that k features are retained and investigate the classifier error. This generates an error curve for each method as k ranges from 1 to p.

We consider two datasets: the lung cancer dataset [101] and the breast cancer dataset [102]. They were downloaded from `http://blog.nus.edu.sg/staww/softwarecode/`. For both datasets, we conducted a processing by ranking all features by the feature-wise t-score and retaining p_0 top-ranked features, where p_0 is a number that is for sure larger than the true number of useful features (but $p_0 \ll p$).

dataset	sample size	dimension	p_0
Lung cancer	181	12,533	100
Breast cancer	276	22,215	1000

The lung cancer dataset was analyzed in various papers [103, 91]. The estimated the number of useful features by these methods is around 30, so we confidently set $p_0 = 100$. The breast cancer dataset is a more difficult one and requires a lot more retained features. [104] analyzed the dataset under a clustering framework and suggested that the number of useful features is around 500, so we set $p_0 = 1000$. We also tried other choices of p_0 (e.g., $p_0 = 200$ for lung cancer data and $p_0 = 2000$ for breast cancer data), and the results are similar.

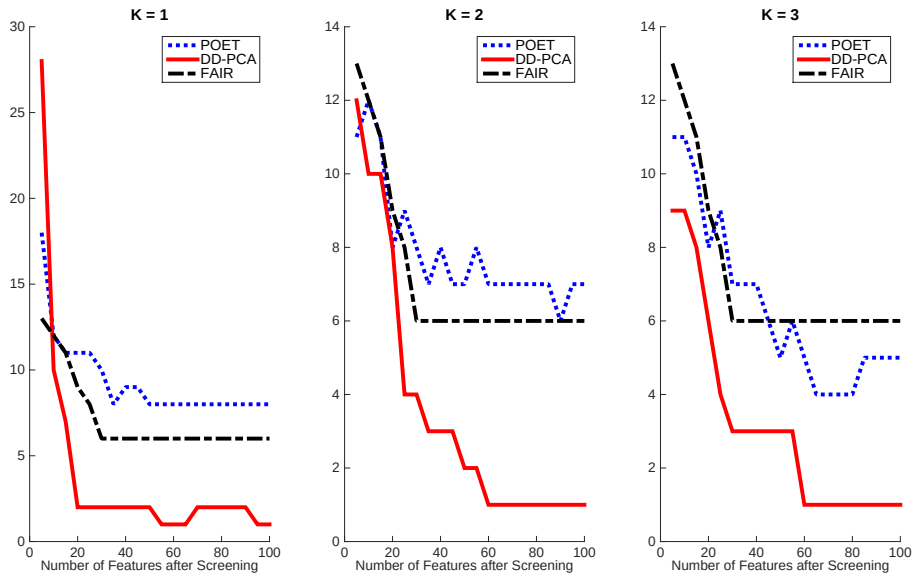


Figure 5.3: Misclassification errors on lung cancer data ($n = 181$).

We evaluate the classification performance by a 5-fold cross-validation procedure with stratified sampling. In detail, we randomly divide samples from class 1 into five folds and do the same to samples from class 2; we then re-combine them to five folds, such that the fraction of class 1 is the same across all folds. Next, we successively leave out each fold, train the classifier on remaining samples, and compute the test error on leave-out samples. The misclassification error reported is the average over 5 folds.

Figure 5.3 displays the results on lung cancer dataset. POET and DD-PCA have a tuning integer K, and we tried $K \in \{1,2,3\}$. The results suggest that, as long as more than 10 features are retained, the classifier powered by DD-PCA uniformly outperforms the other two. Especially, for $K \in \{2,3\}$, the error keeps as low as $1/181$ once the number of retained features exceeds 60. The performance of POET is slightly worse than FAIR for $K \in \{1,2\}$, and slightly better for $K = 3$. We emphasize that the estimated inverse covariance matrix affect both the feature ranking and LDA; therefore, even when the same number of retained features is the same, the actual retained features are different across different methods.

Figure 5.4 displays the results on breast cancer dataset. For both POET and DD-PCA, $K \in \{4,5\}$ is favored to $K = 3$. When $K = 4$, as the number of retained features is in the interval

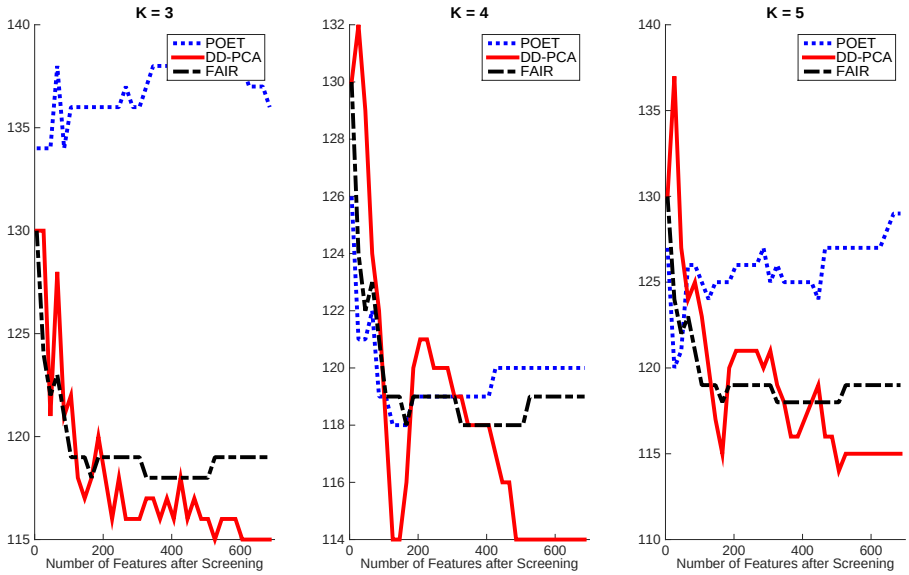


Figure 5.4: Results for breast cancer data ($n = 276$)

of $[500, 700]$, DD-PCA achieves the smallest error of $114/276$. In all three panels, the lowest attainable error of DD-PCA is smaller than those of POET and FAIR.

5.3 Detecting sparse mixtures by DD-PCA

The global detection is a problem of great interest in multiple testing [105, 7, 106]. Let $X_1, \ldots, X_p$ be the z-scores from p tests, where p is presumably large. We assume

$$X \sim \mathcal{N}_p(\mu, \Sigma), \tag{5.14}$$

where μ contains the true effects of these tests and Σ captures the dependence among the z-scores. We are interested in testing

$$H_0 : \mu = 0, \qquad v.s. \qquad H_1 : \mu \neq 0, \text{ and } \mu \text{ is sparse.} \tag{5.15}$$

When Σ is a diagonal matrix, this problem has been studied extensively in the literature. Various tests were proposed, including the χ^2 test, maximum entry test, Higher-Criticism test [7],

Berk-Jones test [107], etc.. When Σ is not a diagonal matrix, this problem becomes much more challenging. One may still use these tests by ignoring the off-diagonals of Σ; unfortunately, this is valid only if the off-diagonal entries of Σ are uniformly small. There are very few literatures about this testing problem under a more general Σ. One pioneer work is [108], where they consider this problem for a class of Σ which have polynomial decays in the off-diagonals. We are primarily interested in a setting where Σ satisfies the assumptions (5.1)-(5.2). In such cases, Σ may not have polynomial decays in the off-diagonals, hence the test in [108] is no longer directly applicable.

For simplicity, we assume Σ is known. In real applications, Σ can be estimated from data if multiple independent copies of X are available. Let

$$(\boldsymbol{L}, \boldsymbol{A}) = \text{Solution of DD-PCA (5.4) by plugging in } \boldsymbol{S} = \Sigma.$$

Note that $\boldsymbol{L}$ has a rank K. Let $\sum_{k=1}^{K} \nu_k \eta_k \eta_k^T$ be the eigen-decomposition of $\boldsymbol{L}$, and write $\boldsymbol{R} = \Sigma - \boldsymbol{L}$. Using elementary probability, we can write the model (5.14) equivalently as

$$X = \mu + \sum_{k=1}^{K} w_k \eta_k + z, \quad w_k \sim \mathcal{N}(0, \nu_k),\ z \sim \mathcal{N}_p(0, \boldsymbol{R}),\ w_1, \ldots, w_K, z \text{ are independent.}$$

These $w_1, \ldots, w_K$ are latent variables that account for most of the heavy dependence. Furthermore, from (5.4), $\boldsymbol{R}$ is approximately diagonally dominant. If we are able to "remove" the latent variables, we are left with a "nice" problem with a covariance matrix $\boldsymbol{R}$. Observing that μ is sparse, we estimate the realized value of $(w_1, \ldots, w_K)$ by regressing X on $\eta_1, \ldots, \eta_K$ using robust regression such as L_1 regression. Let $(\hat{w}_1, \ldots, \hat{w}_K)$ be the estimator. We now have

$$X - \sum_{k=1}^{K} \hat{w}_k \eta_k \approx \mu + \mathcal{N}_p(0, \boldsymbol{R}). \tag{5.16}$$

Comparing it with (5.14), we find that the heavy dependence has been greatly reduced. Since $\boldsymbol{R}$ is approximately diagonally-dominant, we can apply existing tests that are designed for a diagonal Σ.

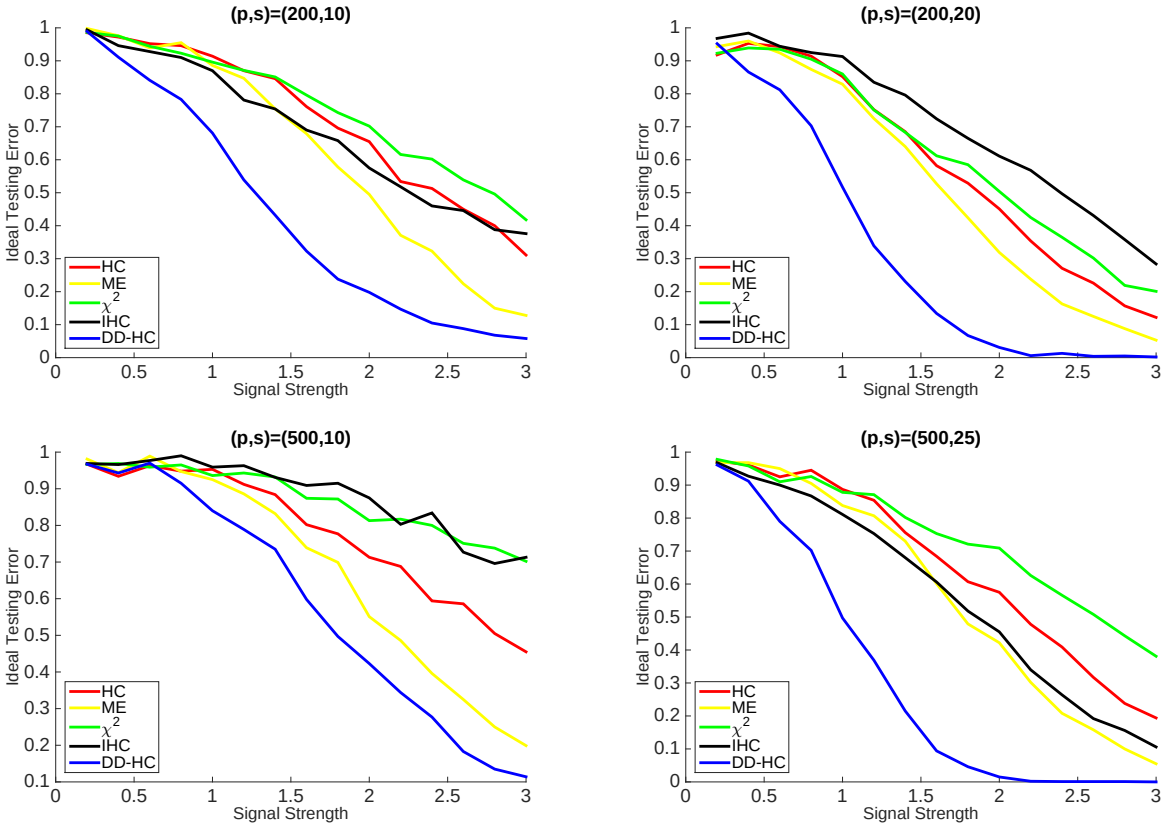


Figure 5.5: Ideal testing error (with the best cut-off value of the test statistics).

We combine the above idea with the Higher Criticism (HC) test [7]. The HC test computes the marginal p-values from a null distribution $X_j \sim \mathcal{N}(0, \Sigma_{jj})$ and combines these p-values into a single test statistic HC_p^*. Motivated by (5.16), we modify the HC test by computing the marginal p-values from $X_j - \sum_{k=1}^{K} \hat{w}_k \eta_{kj} \sim \mathcal{N}(0, R_{jj})$. This gives rise to a new test statistic.

- Obtain $(\boldsymbol{L}, \boldsymbol{A})$ from DD-PCA and obtain $\eta_1, \ldots, \eta_K$, the eigenvectors of $\boldsymbol{L}$.
- Regress X on $(\eta_1, \ldots, \eta_K)$ using a robust regression to obtain coefficients $(\hat{w}_1, \ldots, \hat{w}_K)$.
- Compute the marginal p-values $\pi_j = 2\bar{\Phi}\big(R_{jj}^{-1/2}(X_j - \sum_{k=1}^{K} \hat{w}_k \eta_{kj})\big)$, for $j = 1, 2, \ldots, p$, where $\bar{\Phi}(\cdot)$ is the tail distribution function of a standard normal.
- Sort the p-values: $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(p)}$.

- The test statistic is $HC_p^* = \max_{1\le j\le p/2} HC_{p,j}$, where

$$HC_{p,j} = \frac{\sqrt{p}[(j/p) - \pi_{(j)}]}{\sqrt{\pi_{(j)}(1-\pi_{(j)})}}, \qquad 1 \le j \le p.$$

We call it the DD-HC test. Same as the HC, the null distribution of this test statistic can be obtained from simulations.

Numerical performance of DD-HC: Given (p, s, τ), we let $\mu_j = \tau \cdot 1\{1 \le j \le s\}$ and $\mathbf{\Sigma} = FF^T + \Omega$, where F is a $p \times 2$ matrix whose entries are *i.i.d* drawn from $\mathcal{N}(0, 1/2)$, and $\Omega_{i,j} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. We then generate data using (5.14). The parameters (s, τ) control the the sparsity and signal strength in μ, respectively. We compare our test with the χ^2-test (test statistic: $\|X\|^2$), maximum test (test statistic: $\max_{1\le j\le p} |X_j|$), the HC test, and the Innovated HC test [108]. For all the tests, we evaluate the sum of type I and type II errors associated with the optimal cut-off value using 1000 repetitions. Figure 5.5 displays the results for different values of (p, s, τ). It suggests that our method has a significant advantage over the other tests in the current settings.

5.4 Algorithms for DD-PCA

This section solves both convex and nonconvex optimization for DD-PCA. Section 5.4.1 studies the efficient projection onto $\mathscr{SDD}_c^+$, which is the set of "symmetric c-diagonally-dominant" matrices. Section 5.4.2 proposes a three-block ADMM with provable theoretical guarantees to solve the convex relaxation of DD-PCA, and Section 5.4.3 proposes an iterative projection algorithm to directly solve the nonconvex optimization of DD-PCA. Section 5.4.4 gives a comparison between the convex and nonconvex approaches.

5.4.1 *Efficient projection onto $\mathscr{SDD}_c^+$*

We can write $\mathscr{SDD}_c^+ = \mathscr{S} \cap \mathscr{DD}_c^+$, where $\mathscr{S}$ is the set of symmetric matrices and $\mathscr{DD}_c^+$ is the set of c diagonally dominant matrix with nonnegative diagonal entries, i.e.

$$\mathscr{DD}_c^+ = \{\boldsymbol{A} = (a_{ij})_{p\times p} : a_{jj} \geq c \sum_{i:i\neq j} |a_{ji}| \text{ for all } j\} \tag{5.17}$$

It is not difficult to see that both $\mathscr{DD}_c^+$ and $\mathscr{SDD}_c^+$ are closed and convex polyhedral cones.

To solve DD-PCA, we shall obtain the (Euclidean) projection of a matrix $\boldsymbol{A}$ onto the convex cone $\mathscr{SDD}_c^+$ or $\mathscr{DD}_c^+$, denoted by $\mathscr{P}_{\mathscr{SDD}_c^+}(\boldsymbol{A})$ or $\mathscr{P}_{\mathscr{DD}_c^+}(\boldsymbol{A})$. The following Mendoza-Raydan-Tarazaga algorithm, as summarized in Algorithm 5.4.1, computes the efficient projection $\mathscr{P}_{\mathscr{DD}_c^+}(\boldsymbol{A})$. Following Theorem 2.1 of [87], we have the theoretical guarantee that $\boldsymbol{X}$ obtained by Algorithm 5.4.1 is the unique projection of $\boldsymbol{A}$ onto $\mathscr{DD}_c^+$. Note that the computational complexity of Algorithm 5.4.1 is $O(p^2\log(p))$.

Algorithm 5.4.1. **Mendoza-Raydan-Tarazaga Algorithm**

Given a $p \times p$ matrix $\boldsymbol{A}$, where the jth row of $\boldsymbol{A}$ is denoted by $\mathbf{a}_j$. For $1 \leq j \leq p$, the jth row of the projection $\boldsymbol{X}$, denoted by $\mathbf{x}_j$, is given by

- If $a_{jj} \geq \sum_{l:l\neq j} |a_{jl}|$, then $\mathbf{x}_j = \mathbf{a}_j$.
- If $-\sum_{l:l\neq j} |a_{jl}| \leq a_{jj} < 0$ and $|a_{jj}| > |a_{jl}|$ for all $l \neq j$, or $a_{jj} < -\sum_{l:l\neq j} |a_{jl}|$, then $\mathbf{x}_j = \mathbf{0}$.
- If $-\sum_{l:l\neq j} |a_{jl}| \leq a_{jj} < 0$ and $|a_{jj}| \leq |a_{jl}|$ for some $l \neq j$, or $0 \leq a_{jj} < \sum_{l:l\neq j} |a_{jl}|$, then $\mathbf{x}_j$ is generated as follows:

  1. Sort $|\mathbf{a}_j|$, excluding a_{jj}, in the ascending order, and denote the reordered vector as e. Note that $e_j = a_{jj}$ and $|e_i| \leq |e_l|$ for all $i < l, i \neq j, l \neq j$.
  2. For $m \neq j$, compute $d_m = \sum_{l=m}^{p} |e_l| \cdot I_{\{j\neq l\}} - e_j$ and $\bar{d}_m = d_m/(p-m+1) \cdot I_{\{m<j\}} + d_m/(p-m+2) \cdot I_{\{m>j\}}$

3. Solve $m^\star$ as the smallest integer among $m = 1,\ldots,p$ such that $m \neq j$, $|e_m| > 0$ and $|e_m| \geq \bar{d}_m$

4. Solve $\mathbf{x}_j = (x_{j1},\ldots,x_{jp})$ such that $x_{jj} = a_{jj} + \bar{d}_{m^*}$; $x_{ji} = (a_{ji} - \bar{d}_{m^*})^+$ if $a_{ji} \geq 0$ for $i \neq j$; $x_{ji} = -(a_{ji} + \bar{d}_{m^*})^-$ if $a_{ji} < 0$ for $i \neq j$, where $(z)^+ = \max\{z,0\}$ and $(z)^- = -\min\{z,0\}$.

As for $\mathscr{P}_{\mathscr{S}\mathscr{D}\mathscr{D}^+}(\boldsymbol{A})$, [87] applied Dykstra's alternating projection algorithm between $\mathscr{D}\mathscr{D}^+$ and $\mathscr{S}$ to obtain the projection on $\mathscr{S}\mathscr{D}\mathscr{D}^+$. The algorithm is summarized in Algorithm 5.4.2, and more details can be found in [87].

Algorithm 5.4.2. **Efficient Projection onto** $\mathscr{S}\mathscr{D}\mathscr{D}^+$

Given a $p \times p$ matrix $\boldsymbol{A}$,

- Let $\boldsymbol{G}^{(0)} = \boldsymbol{A}$ and $\boldsymbol{I}^{(0)} = \mathbf{0}$
- For $t = 1,2,\ldots$
  - $\boldsymbol{G}^{(t)} = \mathscr{P}_{\mathscr{D}\mathscr{D}^+}\left(\frac{1}{2}(\boldsymbol{G}^{(t-1)} + (\boldsymbol{G}^{(t-1)})^T) - \boldsymbol{I}^{(t-1)}\right)$
  - $\boldsymbol{I}^{(t)} = \boldsymbol{G}^{(t)} - \left(\frac{1}{2}(\boldsymbol{G}^{(t-1)} + (\boldsymbol{G}^{(t-1)})^T) - \boldsymbol{I}^{(t-1)}\right)$
- Stop if the convergence criterion is met.

When $c \neq 1$, MRT algorithm can't be directly used. In this case, we obtain $\mathscr{P}_{\mathscr{D}\mathscr{D}_c^+}(\boldsymbol{A})$ through Quadratic Programming (QP). The key observation is that the problem can be separated as p independent row-wise projection. For each $1 \leq j \leq p$, the jth row projection can be written as

$$\min_{v_1,\ldots,v_p} \sum_{i=1}^{p} (a_{ji} - v_i)^2 \qquad \text{s.t. } v_j \geq c \sum_{i:i\neq j} |v_i| \tag{5.18}$$

and the solution $(v_1,\ldots,v_p)$ would be the jth row of $\mathscr{P}_{\mathscr{D}\mathscr{D}_c^+}(\boldsymbol{A})$. We can reformulate (5.18) as

$$\min_{\delta_1,\ldots,\delta_p} \sum_{i=1}^{p} \delta_i^2 \qquad \text{s.t. } a_{jj} - \delta_j \geq c \sum_{i:i\neq j} |a_{ji} - \delta_i| \tag{5.19}$$

It's easy to see that for $i \neq j$, we should let $\text{sign}(\delta_i) = \text{sign}(a_{ji})$ and $|\delta_i \leq a_{ji}|$, and hence $|a_{ji} - \delta_i| = |a_{ji}| - |\delta_i|$. Without loss of generality, we assume $a_{ji} \geq 0$ for all $i \neq j$ so we can restrict $\delta_i \geq 0$ for all $i \neq j$. Then (5.19) becomes

$$\min_{\delta_1,\ldots,\delta_p} \sum_{i=1}^{p} \delta_i^2 \qquad \text{s.t. } a_{jj} - \delta_j \geq c \sum_{i:i\neq j} (a_{ji} - \delta_i), \quad a_{ji} \geq \delta_i \geq 0 \text{ for all } i \neq j \tag{5.20}$$

which is a QP problem and can be solved using standard solver.

5.4.2 Convex relaxation and ADMM

This subsection solves the convex relaxation of (5.4) by replacing nonconvex rank constraints with convex nuclear norm constraints. To be specific, we consider the convex optimization:

$$\min_{(\boldsymbol{L},\boldsymbol{A})} \frac{1}{2}\|\boldsymbol{S} - \boldsymbol{L} - \boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{L}\|_* \quad \text{subject to} \quad \boldsymbol{A} \in \mathscr{SDD}_c^+. \tag{5.21}$$

where λ is a tuning parameter to strike a balance between the approximation error and the low rank. A large λ would encourage the solution $\hat{\boldsymbol{L}}$ to be low rank, whereas a smaller λ would lead to relatively smaller approximation error but higher rank in $\hat{\boldsymbol{L}}$.

We introduce a new variable $\boldsymbol{E}$ and rewrite the optimization problem as follows:

$$\min_{(\boldsymbol{L},\boldsymbol{A},\boldsymbol{E})} \frac{1}{2}\|\boldsymbol{E}\|_F^2 + \lambda\|\boldsymbol{L}\|_* + \mathscr{I}_{\boldsymbol{A}\in\mathscr{SDD}_c^+} \quad \text{subject to} \quad \boldsymbol{L} + \boldsymbol{A} + \boldsymbol{E} = \boldsymbol{S}.$$

The objective function would be separable in three blocks, subject to an equality constraint. Now, we define the following augmented Lagrange function:

$$\mathscr{L}_\rho(\boldsymbol{L},\boldsymbol{A},\boldsymbol{E},\boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{E}\|_F^2 + \lambda\|\boldsymbol{L}\|_* + \mathscr{I}_{\boldsymbol{A}\in\mathscr{SDD}_c^+} + \frac{\rho}{2}\|\boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}-\boldsymbol{S}\|_F^2 + \langle\boldsymbol{\Lambda}, \boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}-\boldsymbol{S}\rangle$$

where $\boldsymbol{\Lambda}$ is the Lagrange multiplier associated with the equality constraint, and ρ is a given penalty

parameter. The proposed three-block ADMM proceeds as follows till convergence:

$$\begin{aligned}
\boldsymbol{L} \text{ step}: \quad \boldsymbol{L}^{(t)} &= \arg\min_{\boldsymbol{L}} \mathscr{L}_\rho(\boldsymbol{L}, \boldsymbol{A}^{(t-1)}, \boldsymbol{E}^{(t-1)}, \boldsymbol{\Lambda}^{(t-1)}) \\
\boldsymbol{A} \text{ step}: \quad \boldsymbol{A}^{(t)} &= \arg\min_{\boldsymbol{A}} \mathscr{L}_\rho(\boldsymbol{L}^{(t)}, \boldsymbol{A}, \boldsymbol{E}^{(t-1)}, \boldsymbol{\Lambda}^{(t-1)}) \\
\boldsymbol{E} \text{ step}: \quad \boldsymbol{E}^{(t)} &= \arg\min_{\boldsymbol{E}} \mathscr{L}_\rho(\boldsymbol{L}^{(t)}, \boldsymbol{A}^{(t)}, \boldsymbol{E}, \boldsymbol{\Lambda}^{(t-1)}) \\
\boldsymbol{\Lambda} \text{ step}: \quad \boldsymbol{\Lambda}^{(t)} &= \boldsymbol{\Lambda}^{(t-1)} + \rho(\boldsymbol{A}^{(t)} + \boldsymbol{L}^{(t)} + \boldsymbol{E}^{(t)} - \boldsymbol{S})
\end{aligned}$$

Each subproblem can be efficiently solved. In the $\boldsymbol{L}$ step, we solve $\boldsymbol{L}^{(t)}$ from

$$\begin{aligned}
&\min_{\boldsymbol{L}} \quad \lambda\|\boldsymbol{L}\|_* + \frac{\rho}{2}\|\boldsymbol{L} + \boldsymbol{A}^{(t-1)} + \boldsymbol{E}^{(t-1)} - \boldsymbol{S}\|_F^2 + \langle \boldsymbol{\Lambda}^{(t-1)}, \boldsymbol{L} + \boldsymbol{A}^{(t-1)} + \boldsymbol{E}^{(t-1)} - \boldsymbol{S}\rangle \\
&\min_{\boldsymbol{L}} \quad \frac{1}{2}\|\boldsymbol{L} + \boldsymbol{A}^{(t-1)} + \boldsymbol{E}^{(t-1)} - \boldsymbol{S} + \rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2 + \rho^{-1}\lambda\|\boldsymbol{L}\|_*
\end{aligned}$$

Thus, we have $\boldsymbol{L}^{(t)} = \mathscr{D}_{\rho^{-1}\lambda}\left(\boldsymbol{S} - \boldsymbol{A}^{(t-1)} - \boldsymbol{E}^{(t-1)} - \rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$, where $\mathscr{D}_\tau$ is the singular value thresholding operator given by $\mathscr{D}_\tau(\boldsymbol{\Omega}) = \boldsymbol{U}s_\tau(\boldsymbol{D})\boldsymbol{V}^T$ for any singular value decomposition $\boldsymbol{\Omega} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, and s_τ denotes the soft-thresholding operator given by $s_\tau(x) = \text{sgn}(x)\max(|x| - \tau, 0)$.

In the $\boldsymbol{A}$ step, we may use Algorithm 5.4.2 to obtain the projection on $\mathscr{SDD}^+$ as follows:

$$\begin{aligned}
\boldsymbol{A}^{(t)} &= \arg\min_{\boldsymbol{A}} \mathscr{I}_{\boldsymbol{A}\in\mathscr{SDD}_c^+} + \frac{\rho}{2}\left(\|\boldsymbol{A} + \boldsymbol{L}^{(t)} + \boldsymbol{E}^{(t-1)} - \boldsymbol{S} + \rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2\right) \\
&= \mathscr{P}_{\mathscr{SDD}_c^+}\left(\boldsymbol{S} - \boldsymbol{L}^{(t)} - \boldsymbol{E}^{(t-1)} - \rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right).
\end{aligned}$$

Alternatively, we may follow the proximal-gradient-based ADMM [109] to solve the $\boldsymbol{A}$ step. See Section 4 of [109] for more details about the proximal method.

In the $\boldsymbol{E}$ step, it is straightforward to solve

$$\begin{aligned}\boldsymbol{E}^{(t)} &= \arg\min_{\boldsymbol{E}} \frac{1}{2}\|\boldsymbol{E}\|_F^2 + \frac{\rho}{2}\left(\|\boldsymbol{E}+\boldsymbol{L}^{(t)}+\boldsymbol{A}^{(t)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2\right)\\ &= \arg\min_{\boldsymbol{E}} \left\|\boldsymbol{E}+\frac{\rho}{\rho+1}\left(\boldsymbol{L}^{(t)}+\boldsymbol{A}^{(t)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)\right\|_F^2\\ &= \frac{\rho}{\rho+1}\left(\boldsymbol{S}-\boldsymbol{A}^{(t)}-\boldsymbol{L}^{(t)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)\end{aligned}$$

Hence, the proposed three-block ADMM can be summarized in Algorithm 5.4.3.

Algorithm 5.4.3. **Three-Block ADMM for Solving DD-PCA**

Given a sample covariance matrix $\boldsymbol{S}$, do

- Let $\boldsymbol{A}^{(0)} = \boldsymbol{E}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \boldsymbol{0}$
- For $t = 1, 2, \ldots$
  - $\boldsymbol{L}^{(t)} = \mathscr{D}_{\rho^{-1}\lambda}\left(\boldsymbol{S}-\boldsymbol{A}^{(t-1)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$ where $\mathscr{D}_\tau(\boldsymbol{\Omega})$ is the singular value thresholding operator given by $\mathscr{D}_\tau(\boldsymbol{\Omega}) = \boldsymbol{U}s_\tau(\boldsymbol{D})\boldsymbol{V}^T$ for any singular value decomposition $\boldsymbol{\Omega} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, and s_τ denotes the soft-thresholding operator given by $s_\tau(x) = \operatorname{sgn}(x)\max(|x|-\tau, 0)$.
  - $\boldsymbol{A}^{(t)} = \mathscr{P}_{\mathscr{S}\mathscr{D}\mathscr{D}_c^+}\left(\boldsymbol{S}-\boldsymbol{L}^{(t)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$
  - $\boldsymbol{E}^{(t)} = \frac{\rho}{\rho+1}\left(\boldsymbol{S}-\boldsymbol{A}^{(t)}-\boldsymbol{L}^{(t)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$
  - $\boldsymbol{\Lambda}^{(t)} = \boldsymbol{\Lambda}^{(t-1)} + \rho\left(\boldsymbol{A}^{(t)}+\boldsymbol{L}^{(t)}+\boldsymbol{E}^{(t)}-\boldsymbol{S}\right)$
- Stop if the convergence criterion is met.

Although three-block ADMM does not necessarily converge in general [110], DD-PCA belongs to a class of regularized least squares decomposition problem. For this class of regularized problems, the global convergence of the proposed three-block ADMM is always guaranteed [111].

5.4.3 *Iterative projection algorithm*

In the sequel, we introduce an iterative projection algorithm that directly tackles the nonconvex optimization in DD-PCA. The key observation is that we attempt to find a matrix $\boldsymbol{L}^*$ in the set $\mathscr{L}_K = \{\boldsymbol{L} : \text{rank}(\boldsymbol{L}) = K\}$ that is closest to the set $\mathscr{M}_{\boldsymbol{S}} = \{\boldsymbol{S} - \boldsymbol{A} : \boldsymbol{A} \in \mathscr{S}\mathscr{D}\mathscr{D}_c^+\}$. Inspired by [112], a natural approach would be to iteratively project $(\boldsymbol{S} - \boldsymbol{L})$ onto $\mathscr{S}\mathscr{D}\mathscr{D}_c^+$ to update $\boldsymbol{A}$ and then to project $\boldsymbol{S} - \boldsymbol{A}$ onto $\mathscr{L}_K$ to update $\boldsymbol{L}$. To reduce the computational cost, we replace the projection onto $\mathscr{S}\mathscr{D}\mathscr{D}_c^+$ by the projection onto $\mathscr{D}\mathscr{D}_c^+$, followed by symmetrization. Algorithm 5.4.4 summarizes the details.

Algorithm 5.4.4. **Iterative Projection Algorithm for Solving DD-PCA**

Given a sample covariance matrix $\boldsymbol{S}$ and integer k, do

- Let $\boldsymbol{A}^{(0)} = \boldsymbol{0}$
- For $t = 1, 2, \ldots$
  - $\boldsymbol{L}^{(t)} = \mathscr{P}_{\mathscr{L}_K}(\boldsymbol{S} - \boldsymbol{A}^{(t-1)})$
  - $\tilde{\boldsymbol{A}}^{(t)} = \mathscr{P}_{\mathscr{D}\mathscr{D}_c^+}(\boldsymbol{S} - \boldsymbol{L}^{(t)})$
  - $\boldsymbol{A}^{(t)} = \left(\tilde{\boldsymbol{A}}^{(t)} + (\tilde{\boldsymbol{A}}^{(t)})^T\right)/2$
- Stop if the convergence criterion is met.

In Algorithm 5.4.4, we need to calculate $\mathscr{P}_{\mathscr{L}_K}$ and $\mathscr{P}_{\mathscr{D}\mathscr{D}_c^+}$. The calculation of $\mathscr{P}_{\mathscr{D}\mathscr{D}_c^+}$ is given in Section 5.4.1. The calculation of $\mathscr{P}_{\mathscr{L}_K}$ is given as follows: for any symmetric matrix $\boldsymbol{A}$, we write its eigenvalue decomposition as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_p|$. Hence, the best rank-$K$ approximation is given by $\mathscr{P}_{\mathscr{L}_K}(\boldsymbol{A}) = \boldsymbol{Q}_K\boldsymbol{\Lambda}_K\boldsymbol{Q}_K^T$ where $\boldsymbol{Q}_K$ contains the first K columns of $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}_K = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$.

To use Algorithm 5.4.4, we need to estimate the rank K if it is unknown. A simple estimate of K is to look at the eigenvalues of $\boldsymbol{S}$ to pick the K such that there is a significant gap in magnitude between the first K eigenvalues and the remaining ones. In Section 5.5, we investigate the robustness of the iterative projection algorithm to the estimation of K.

5.4.4 Discussion

This section presents both convex approaches and nonconvex approaches to solve DD-PCA. The convex approaches do not require the knowledge of the rank K of the low rank matrix $\boldsymbol{L}$, and the global convergence of the proposed ADMM (e.g., Algorithm 5.4.3) is guaranteed. However, its convergence rate could be slow in practice. The nonconvex approaches, on the other hand, can be much faster in terms of convergence. The per-iteration cost for Algorithm 5.4.4 is $O(p^2 \max\{\log(p), K\})$, compared to $O(p^3)$ for Algorithm 5.4.3. But the convergence guarantee of Algorithm 5.4.4 is still an open question.

The rigorous convergence analysis of the iterative projection algorithm is difficult due to the non-convexity of the set $\mathscr{L}_K$. The existing result (e.g., [113]) proves the local linear convergence of the alternating projections for two closed sets if the two sets intersect *transversally* at the converging point. We conjecture that such condition would hold for most cases in our setting, therefore the convergence would be guaranteed. In practice, our algorithms are stable and always converge to a valid solution in simulations.

As for stopping criteria, at each step one can check if $\|\boldsymbol{A}^{(t+1)} - \boldsymbol{A}^{(t)}\|$ or $\|\boldsymbol{L}^{(t+1)} - \boldsymbol{L}^{(t)}\|$ is below a given threshold ε. Another possible stopping criteria is iteration time reaching N, where N is a pre-specified integer, which could help reduce computational cost if N is set small when p is relatively large.

5.5 Simulation studies

This section investigates several numerical properties of DD-PCA in simulation studies.

Experiment 1: solving DD-PCA. We investigate the performance of Algorithm 5.4.3 (an ADMM algorithm) and Algorithm 5.4.4 (an iterative projection algorithm) for DD-PCA. Fixing (p, K) and $\sigma > 0$, we first generate a rank-K matrix $\boldsymbol{L} = XX^T$ where X is a $p \times k$ matrix whose entries are i.i.d drawn from $\mathscr{N}(0, 1/p)$. We then generate a matrix $\boldsymbol{A}_0$ with entries sampled i.i.d from $\mathscr{N}(0, 1/p^2)$ and set $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_0^T + \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix whose j-th diagonal is

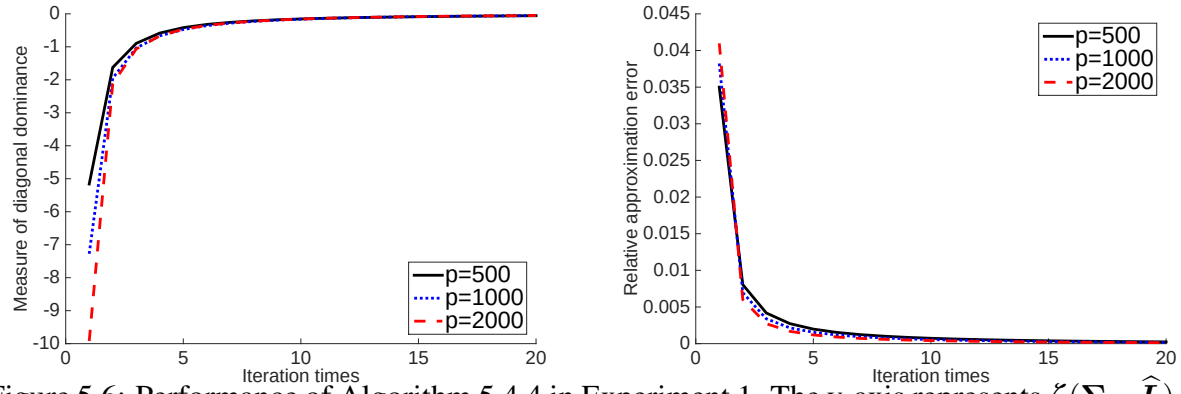


Figure 5.6: Performance of Algorithm 5.4.4 in Experiment 1. The y-axis represents $\zeta(\boldsymbol{\Sigma}-\widehat{\boldsymbol{L}})$ (left panel) and $\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$ (right panel).

equal to $\sum_{i:i\neq j}|A_0(j,i)+A_0(i,j)|-2A_0(j,j)$ for $1\leq j\leq p$; it follows that $\boldsymbol{A}$ is a diagonally dominant matrix. We then generate a $p\times p$ symmetric matrix $\boldsymbol{E}$ whose upper triangular entries are sampled i.i.d from $\mathcal{N}(0,\sigma^2/p)$. Last, let $\boldsymbol{S}=\boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}$.

First, we study Algorithm 5.4.3, which is an ADMM algorithm. Fix $\sigma=1$. We consider $p=500,1000,2000$, and set $K=0.05\cdot p$. The tuning parameter in the algorithm is set as $\lambda=3$, and we look at the solution $(\widehat{\boldsymbol{L}},\widehat{\boldsymbol{A}})$ after 50 iterations. The results are displayed in Table 5.1. For all three settings, the algorithm exactly recovers the true rank of $\boldsymbol{L}$, however, the convergence of $(\widehat{\boldsymbol{L}},\widehat{\boldsymbol{A}})$ is relatively slow. As we shall see below, the performance of Algorithm 5.4.3 is not as good as the iterative projection algorithm—Algorithm 5.4.4, but Algorithm 5.4.3 is theoretically more tractable.

Table 5.1: Performance of Algorithm 5.4.3 in Experiment 1.

Dimension p	rank($\boldsymbol{L}$)	rank($\widehat{\boldsymbol{L}}$)	$\frac{\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F}{\|\boldsymbol{S}\|_F}$	$\frac{\|\widehat{\boldsymbol{L}}-\boldsymbol{L}\|_F}{\|\boldsymbol{L}\|_F}$	$\frac{\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_F}$
500	25	25	0.264	0.166	0.340
1000	50	50	0.269	0.163	0.286
2000	100	100	0.274	0.160	0.243

Next, we study Algorithm 5.4.4, the iterative projection algorithm. We run the algorithm for 20 iterations and measure the relative approximation error $\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$. The results are shown in Figure 5.7. In the left panel, K/p is fixed as 0.05 and the noise level σ varies from

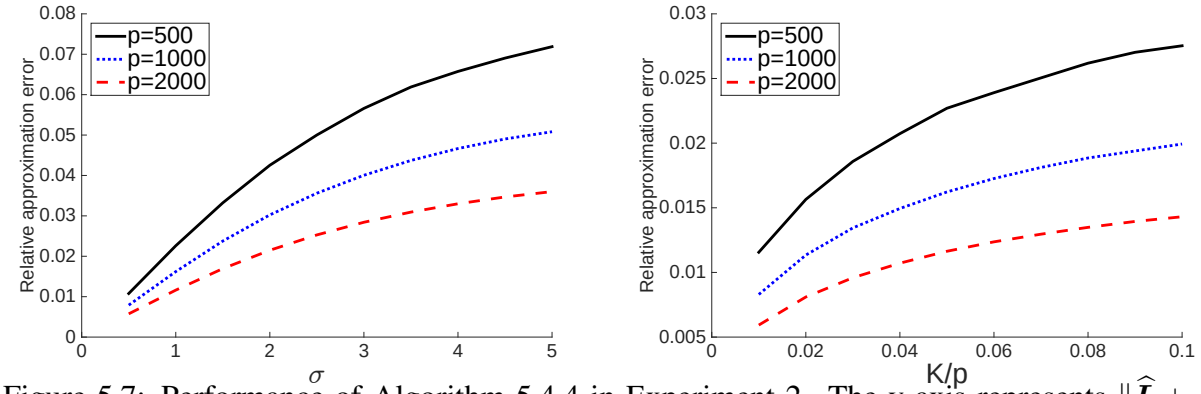


Figure 5.7: Performance of Algorithm 5.4.4 in Experiment 2. The y-axis represents $\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$, and the x-axis represents σ (left panel) and K/p (right panel), respectively.

0.5 to 5. In the right panel, σ is fixed to be 1 and K/p varies from 0.01 to 0.1. For each value of p, the relative approximation error increases, as both σ and K increase. For the same values of σ and K/p, a larger p comes with a smaller relative approximation error. Furthermore, if we compare the results with those in Table 5.1, Algorithm 5.4.4 has a better practical performance than Algorithm 5.4.3.

Experiment 2: Necessity of DD-PCA. If $\boldsymbol{\Sigma}$ truly satisfies the assumption of "low-rank plus diagonal dominance", it is a natural question to know whether one can simply apply PCA and robust PCA [84] to get a diagonally dominant $\boldsymbol{A}$. Unfortunately, this is often not the case. Let us consider applying PCA to a $\boldsymbol{\Sigma}$ which has the decomposition $\boldsymbol{\Sigma}=\boldsymbol{L}_0+\boldsymbol{A}_0$ such that $\operatorname{rank}(\boldsymbol{L}_0)=K$ and $\boldsymbol{A}_0$ is diagonally dominant. Let λ_k and ξ_k be the k-th eigenvalue and eigenvector, respectively, $1\le k\le p$. We construct $\boldsymbol{L}=\sum_{k=1}^{K}\lambda_k\xi_k\xi_k^T$ and $\boldsymbol{A}=\boldsymbol{\Sigma}-\boldsymbol{L}$. We can only hope $\boldsymbol{A}$ is diagonally dominant when $\boldsymbol{A}$ and $\boldsymbol{A}_0$ are entrywise close to each other, or equivalently, when $\|\boldsymbol{L}-\boldsymbol{L}_0\|_{\max}$ is small ($\|\cdot\|_{\max}$ is the entrywise max norm). However, from the literatures on perturbation analysis of PCA, it requires strong conditions to guarantee that $\|\boldsymbol{L}-\boldsymbol{L}_0\|_{\max}$ is small [49]. In particular, when K is moderately large, these conditions may be violated. Similarly, robust PCA cannot produce a diagonally dominant $\boldsymbol{A}$ in general. Therefore, it is necessary to develop new algorithms that are specifically designed for DD-PCA. In Figure 5.8, we present a numerical example, where the output $\boldsymbol{A}$ from our DD-PCA algorithm is much more "diagonally dominant" than the $\boldsymbol{A}$ con-

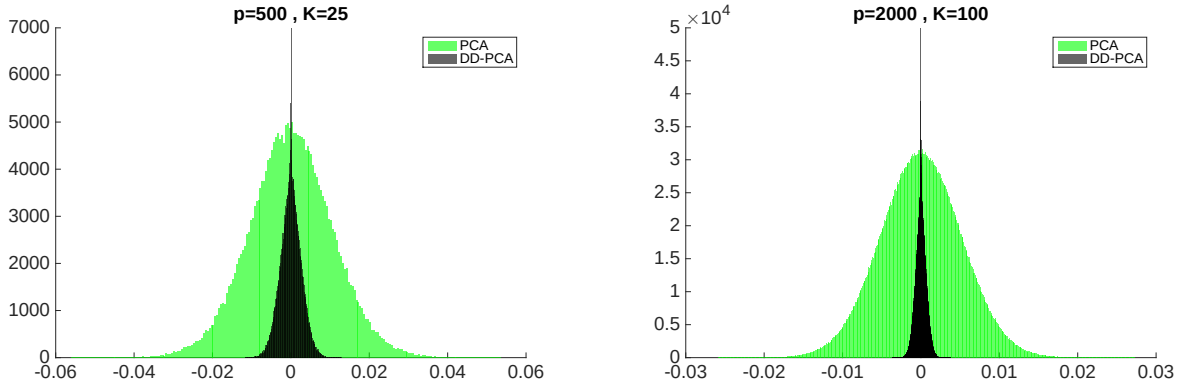


Figure 5.8: Comparison of the output $\boldsymbol{A}$ from DD-PCA and from PCA, where the histogram of $\{a_{ij}/[a_{ii}a_{jj}]^{1/2} : 1 \leq i \neq j \leq p\}$ is displayed. In both panels, the input $\boldsymbol{\Sigma}$ is generated as in Experiment 1 in Section 5.5.

structed from PCA.

Experiment 3: Robustness to the misspecification of K. We use the same setup as in Experiment 1 and investigate the performance of Algorithm 5.4.4 with a misspecified K. Consider two settings where $(p,K) = (500,25)$ and $(p,K) = (2000,100)$, respectively. For each setting, we plug $K = k$ in the algorithm and take the solution after ten iterations. Figure 5.9 shows the relative difference between $\widehat{\boldsymbol{L}}$ and $\boldsymbol{L}$ for various choices of k. It suggests that as long as $k \geq K$, the performance of the algorithm is very stable. Hence, in practice, we recommend that the users pick a relatively large k when the true K is hard to estimate.

Experiment 4: Application to covariance matrix estimation We expand the numerical study in Section 5.2 and investigate the performance of DD-POET on more simulation settings. Given $K = 3$ and $p \in \{100,300,500\}$, we generate data in the same way as in the numerical example of Section 5.2. First, we compare the performance of DD-POET and POET. For both methods, we use the true $K = 3$. POET has an additional threshold, which we set as the ideal one that minimizes the estimation error (the ideal threshold varies as we change the error measure). The results are contained in Column 6 and Column 10 of Table 5.2, where, in all settings, DD-POET has a comparable performance as POET with an ideal threshold, and in some settings, DD-POET

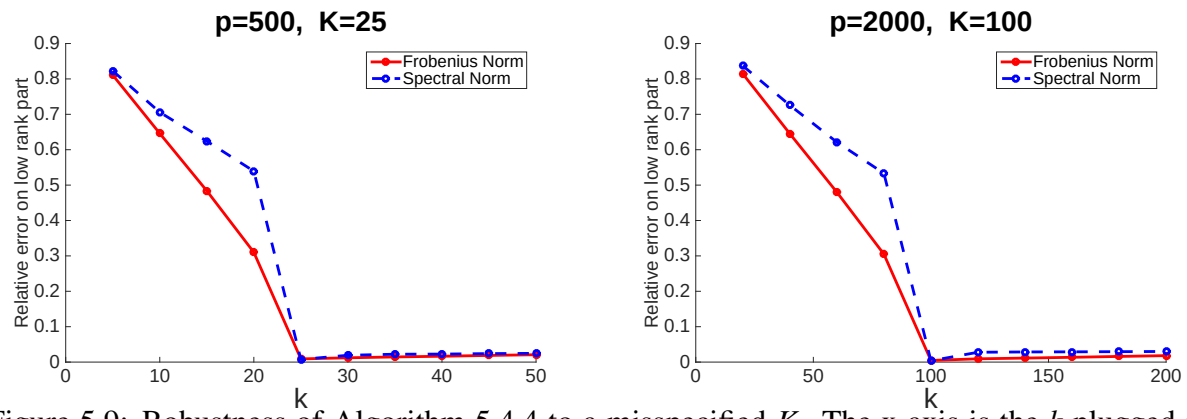


Figure 5.9: Robustness of Algorithm 5.4.4 to a misspecified K. The x-axis is the k plugged into the algorithm, and the y-axis is $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\|/\|\boldsymbol{L}\|$, where $\|\cdot\|$ is either the matrix Frobenius norm or the spectral norm.

is even better. The ideal threshold for POET is not practically feasible, and it is unclear how to set the threshold in a data-driven fashion; however, DD-POET is tuning free once K is given. Second, we investigate the performance of DD-POET when we plug in $K = k$ with $k \in \{1, 2, \ldots, 6\}$; see Table 5.2. If k is misspecified but $k \geq K$, the estimation errors remain relatively stable; if $k < K$, the performance deteriorates. It suggests that an overshooting of K is better than an undershooting. This is consistent with the observations made by [6].

Table 5.2: Estimation errors of DD-POET and its robustness to a misspecified K.

(p, K)	Target	Norm	k						POET*
			1	2	3	4	5	6	$(k=3)$
(100,3)	$\boldsymbol{\Sigma}_u$	Frobenius	48.26	27.52	3.28	3.46	3.63	3.83	**3.24**
		Spectral	22.51	16.80	**0.80**	1.00	1.08	1.14	0.85
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	9.10	7.50	**3.02**	3.17	3.34	3.56	3.65
		Spectral	1.34	1.34	**0.61**	0.72	0.83	0.93	0.64
(300,3)	$\boldsymbol{\Sigma}_u$	Frobenius	95.00	56.96	6.22	6.23	6.26	6.32	**6.04**
		Spectral	32.52	26.04	**0.82**	0.86	0.90	0.93	0.90
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	41.50	17.33	**5.68**	5.66	5.66	5.68	6.37
		Spectral	27.75	7.16	0.66	0.66	0.65	0.64	**0.64**
(500,3)	$\boldsymbol{\Sigma}_u$	Frobenius	126.50	75.74	8.38	8.35	8.35	8.35	**7.99**
		Spectral	38.10	30.86	**0.84**	0.87	0.89	0.91	0.95
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	25.87	18.19	**7.66**	7.61	7.57	7.55	8.26
		Spectral	9.23	1.76	0.69	0.68	0.68	0.67	**0.64**

* POET is implemented with an ideal threshold.

CHAPTER 6

ESTIMATION AND INFERENCE FOR ZERO-INFLATED SEMI-CONTINUOUS DATA

In many real world scenarios, the data we collected often has a significant portion of 0's with the remaining part being positive. Such data, usually called zero-inflated semi continuous data, are pretty common in the time related or cost related applications. Examples include air contaminant study [114], marine science [115], medical cost [116], pharmaceutical research [117] and so on. In the big data era, zero inflated data is prevalent in technology industry. For example, the customer expenditure data collected by e-commerce companies typically exhibits such pattern: a notable proportion of users buy nothing (zero expenditure) and a few "big customers" purchase high price items which drives the positive part to be right skewed. For this example, one may want to estimate either the average customer expenditure or the average expenditure that a customer could have done had he or she decided to purchase anything. These applications demand specific methods that can be tailored to the zero-inflated nature of the data.

In this chapter, [1] we propose several machine learning algorithms to estimate some quantities of interest with covariate information employed. We don't pose any parametric assumptions and our framework can be combined with any base learners. We first introduce our model and method in the one sample case, then we discuss the two sample case where we aim to estimate heterogeneous treatment effect when data in both treatment group and control group are zero-inflated semi-continuous. The algorithms we developed are motivated by state-of-the-art methods in the causal inference literature.

6.1 Related work

Suppose we observe i.i.d data $Y_1, Y_2, \ldots, Y_n$ drawn from a distribution that is a mixture between a point mass at 0 and a distribution supported on positive values. There has been a line of work for

1. The work presented in this chapter is joint with Hao Jiang and has been submitted to KDD 2019.

zero-inflated semi-continuous data that focus on estimating the population mean of Y [118, 119, 115, 120, 121, 114, 116]. These work typically assume a parametric model for the data generation and deals with point estimate as well as related inference problems. For example, a common distribution they assume for the data, so called Delta distribution in the literature, is a mixture between a point mass at zero and a log normal distribution. These assumptions are quite strong: essentially they require the sample points to be independent and identically distributed following some parametric form with unknown parameters. The inference problem then boils down to the inference of several parameters which can be handled using classical statistical tools.

The drawback of aforementioned approaches lies in the fact that they fail to incorporate the information of covariates which are available and often useful in most applications. For example, in customer expenditure data we can often have access to the customer personal information as well as some expenditure related features, so we may seek to estimate mean expenditure or related quantities for each individual. In associated A/B testing experiments, we are often interested in estimating heterogeneous treatment effect to better understand causal mechanisms and to personalize treatment regimes. These observations motivate us to build models for zero-inflated semi-continuous responses with covariates information included.

Another line of work in the literature considers adding covariates into the model. These work typically assume a parametric form for response which depends on covariates through a linear combination of those covariates. Such models include Tobit models [122, 123, 124, 125], two-parts models [126, 127], sample selection models [128, 129, 130], Compound Poisson exponential dispersion models [131, 132] and Ordinal threshold models [133]. A detailed comparison of those models can be found in the survey paper [134]. The success of those statistical models relies on the correct specification of data distribution and most of their methods are not robust to model misspecification. In particular, they compress the covariates information through a linear form which is often inadequate and hence limit the performance in many real world applications.

6.2 Model and methods

Suppose we have i.i.d observations $(X_i, Y_i)_{i=1}^n$, where $X_i \in \mathscr{X}$ encodes covariates information and $Y_i \in [0,\infty)$ is the observed outcome. Let $W_i = \mathbf{1}_{\{Y_i>0\}}$ which is a binary random variable. We posit the existence of potential outcome $Z_i \in (0,\infty)$, such that $Y_i = W_i Z_i$. The framework here is closely related to the potential outcome model (Neyman-Rubin model) for observational data. In addition, we assume the following two conditions

- ZISC-unconfoundedness: $W \perp\!\!\!\perp Z \mid X$
- ZISC-overlap: $0 < e(x) < 1$ for all x in the support of X, where $e(x) = \mathbb{P}\{W = 1|X = x\}$ is the ZISC propensity score function.

Here ZISC stands for Zero-Inflated Semi-Continuous. These assumptions are analogous to the standard assumptions, unconfoundedness and overlap, in potential outcome model. We would like to point out that most commonly used parametric models for zero-inflated semi-continuous data, like zero-inflated log normal or zero-inflated gamma model [135], satisfy these conditions.

In many cases, the target of interest is the conditional *positive* mean $\mu_+(x) = \mathbb{E}[Y \mid Y > 0, X = x]$ instead of the canonical conditional mean $\mu(x) = \mathbb{E}[Y \mid X = x]$. In our framework, we can write it as

$$\mu_+(x) = \mathbb{E}[Z \mid W = 1, X = x] = \mathbb{E}[Z \mid X = x]$$

where we make use of the ZISC-unconfoundedness assumption. In the customer expenditure application, $\mu_+(x)$ measures the average amount of money customer would have spent, had the customer decided to buy anything. Efficient estimation for $\mu_+(x)$ is not straightforward and we will discuss some machine learning algorithm in the following section.

In practice, the positive data may have heavy tails. In this case we could perform a log transformation of the positive response. This is related to estimating $\mathbb{E}[\log(Y) \mid Y > 0, X = x] = \mathbb{E}[\log(Z) \mid X = x]$ and the algorithms we proposed can be readily extended to the scenario here. We can then apply the exponential function on top of it to estimate $\exp(\mathbb{E}[\log(Z) \mid X = x])$ which is

less than or equal to $\mathbb{E}[Z \mid X = x]$ by Jensen's inequality, but in some cases this trick which aims at a surrogate target could yield a more stable estimate for $\mu_+(x)$. We examine this log transformation trick in Section 6.3.

6.2.1 One sample setting: mean estimation

In this section, we introduce several machine learning techniques for estimating $\mu_+(x)$. We then briefly discuss how to estimate $\mu(x)$ but that should not be the focus here.

Conditional mean regression and transformed outcome regression A simple method is to ignore those samples with response 0 and run regression on the rest of data using any supervised learning algorithm. We call this method *Conditional Mean Regression*. In this case, the algorithm tries to estimate $\mathbb{E}[Y \mid X = x, W = 1] = \mu_+(x)$ directly and the variance of response is $\mathrm{Var}(Y \mid X = x, W = 1) = \mathrm{Var}(Z \mid X = x)$.

The drawback of conditional mean regression is that it only uses a fraction of data so the effective sample size is small. This is particularly problematic in the sparse response setting. For example, in the customer expenditure data usually a large proportion (95 to 99 percent) of responses are 0, in which case the conditional mean regression is highly inefficient as it discards most of the data. On population level, the expected effective sample size is $\mathbb{E}[e(X)]$.

An alternative approach relies on estimating the ZISC propensity score function. If we define the transformed outcome as $\tilde{Y} = Y/e(X)$, it's easy to see that $\tilde{Y}$ is an unbiased estimate of $\mu_+(x)$ conditional on $X = x$. In fact, we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{Y} \mid X = x\right] &= \frac{\mathbb{E}[WZ \mid X = x]}{e(x)} \\
&= \frac{\mathbb{E}[W \mid X = x]\,\mathbb{E}[Z \mid X = x]}{e(x)} \\
&= \mathbb{E}[z \mid X = x] = \mu_+(x)
\end{aligned}$$

In practice, $e(X)$ is usually not known and one can use any classification algorithm that produces

probability estimate on $\{(X_i, W_i)\}_{i=1}^n$ to get an estimated ZISC propensity score $\hat{e}(X)$ then plug it in. With these transformed outcomes, any supervised learning algorithm can be employed to fit the regression function. This method is called *Transformed Outcome Regression*. The problem with this method is that the variance of the response is potentially large, especially when ZISC propensity score is near 0. In fact, provided $e(x)$ is known we have

$$\begin{aligned}
& \operatorname{Var}\left(\tilde{Y} | X=x\right)=\frac{\operatorname{Var}(W Z | X=x)}{e^2(x)} \\
= & \frac{\mathbb{E}\left[W^2 Z^2 \mid X=x\right]-\mathbb{E}[W Z \mid X=x]^2}{e^2(x)} \\
= & \frac{e(x) \mathbb{E}\left[Z^2 \mid X=x\right]-e^2(x) \mu_{+}^2(x)}{e^2(x)} \\
= & \frac{\operatorname{Var}(Z \mid X=x)+\mu_{+}^2(x)(1-e(x))}{e(x)}
\end{aligned}$$

which is larger than $\operatorname{Var}(Z \mid X=x)$ unless $e(x)=1$. When $e(x)$ is not known yet estimated by some classification algorithms, the bias is typically introduced and the variance would become even higher.

Comparing these two approaches, we see there is a trade-off between effective sample size and signal to noise ratio of the data. Conditional mean regression has larger signal to noise ratio as the variance of the response is smaller than transformed outcome regression. On the other hand, transformed outcome regression has larger effective sample size as it uses all the samples in the data set. In our simulation, we found that conditional mean regression often outperforms transformed outcome regression in many scenarios.

Zero-inflated causal forest Motivated by method in [136], we propose to adapt causal forest in our setting, and we call this variant *Zero-Inflated Causal Forest* (ZICF). ZICF is an ensemble of zero-inflated causal trees. The construction of each zero- inflated causal tree goes as follows: suppose the tree structure is given, for each terminal node we estimate $\mu_{+}(x)$ using average of *positive* responses fallen in that region. This is the key difference between zero-inflated causal tree

and classical regression tree, as in regression tree the estimate within a node would be the average of *all* responses. As for tree node splitting we use the same splitting rule as in classical causal tree, that is, we maximize the sum of squares of $\hat{\mu}_+(x_i)$ for i in region $\mathscr{J}$ for which we consider splitting. The ensemble of those trees are the same as in [136], where each tree is based on a random sub-sample drawn from the data without replacement and the final estimate is the average of all tree estimate, i.e. $\hat{\mu}_+(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{\mu}_{+,b}(x)$ where $\hat{\mu}_{+,b}(x)$ is the estimate given by the bth tree.

According to [136], we can show that under some mild conditions the ZICF estimate $\hat{\mu}_+(x)$ is asymptotically Gaussian and unbiased. That is,

$$(\hat{\mu}_+(x) - \mu_+(x)) / \sqrt{\mathrm{Var}(\hat{\mu}_+(x))} \to_d \mathscr{N}(0,1)$$

where $\to_d$ denotes converge in distribution. Moreover, we can accurately estimate the asymptotic variance using infinitesimal jackknife for random forests [137]. To be specific, let $N_{ib} \in \{0,1\}$ indicate whether or not the ith training data was used for the bth tree. The variance estimator is

$$\hat{V}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}[\hat{\mu}_{+,b}(x), N_{ib}]^2$$

where s is the sub sample size. This estimator is consistent in the sense that $\hat{V}(x)/\mathrm{Var}(\hat{\mu}_+(x)) \to_p 1$ where $\to_p$ denotes convergence in probability.

Those theoretical results apply for a version of ZICF, which we call honest ZICF, where the tree structure is independent of the leaf prediction given the data. This version is particularly suited for inference, see [136, 138] for more details.

Other machine learning methods for estimating $\mu_+(x)$ Recognizing the resemblance between our model assumptions and those made by potential outcome model in causal inference, we can try to adapt the state-of-the-art machine learning methods designed for heterogeneous treatment effect estimation to the estimation problem we are interested here. The key observation is that if

we input $\{(X_i, Y_i, W_i)\}_{i=1}^n$ to those algorithms, the resulting conditional average treatment effect estimate is exactly what we want as it aims to estimate $\mathbb{E}[Y \mid W = 1, X = x] - \mathbb{E}[Y \mid W = 0, X = x] = \mathbb{E}[Y \mid W = 1, X = x] = \mu_+(x)$. Causal forest discussed in the previous section is one example that can be tailored to our setting. Some other examples can be found in, say, [139].

Connection to missing data and semi-supervised learning As we discussed, the estimation of $\mu_+(x)$ under our model setting is closely related to the problem of heterogeneous treatment effect estimation for observational data under potential outcome model. Actually, our problem can be regarded as a special case under potential outcome model where the outcome for control group is always 0. Causal inference under potential outcome model is essentially a *missing data* problem, and we can take the same view here for our problem. We can think of those observations with zero response as 'missing' data where their potential outcome Z is unobserved. With ZISC-unconfoundedness assumption we know the missing mechanism is *missing at random*, i.e. the missing probability can be modeled as a function of observed data X [2]. Therefore, any method that deals with missing data where data are missing at random could potentially be used here.

Some reader may view the problem as a semi-supervised learning problem, i.e. we have some labeled data (those observations with positive response) and a large amount of unlabeled data (those observations with zero response). In classical semi-supervised learning setting, the choice of labeled sample is independent of X, i.e. there is no selection bias. In contrast, the mechanism of choosing samples to label in our setting does depend on X and such mechanism could be potentially informative. Hence, straightforward application of traditional semi-supervised learning algorithm without care may lead to less efficient and often biased estimate.

Conditional mean $\mu(x)$ estimation In principle, any supervised learning algorithm can be employed to run regression using all data points in order to estimate $\mu(x)$. Due to the sparsity of response, some algorithm may lead to undesirable results. For instance, regression tree for such

2. Note that the missing mechanism here is not *missing completely at random* (MCAR) on which most imputation methods are based.

data may end up with certain regions having estimation 0, yet we know $\mu(x)$ is strictly positive under the ZISC-overlap assumption. An alternative approach is to first estimate $\mu_+(x)$ using any method, such as ZICF, then weight the estimation by estimated ZISC propensity score since $\mu(x) = \mu_+(x)e(x)$. The potential drawback of this approach is that the error in estimating $\mu_+(x)$ and $e(x)$ might propagate through the multiplication and render the variance of final estimator large.

The last thing we would like to mention is that sometimes we are interested in prediction rather than mean estimation. If sparse predictions are preferable, a simple approach is to first run a classification on $\{(X_i, W_i)\}_{i=1}^n$ to decide whether to output 0, and if instead positive result is desired we can output $\hat{\mu}_+(x)$ which can be estimated using any method discussed above. Alternatively, one can output the pair $(\hat{e}(x), \hat{\mu}_+(x))$ which is often more informative than sparse predictions.

6.2.2 Two sample setting: heterogeneous treatment effect estimation

In this section, we consider a two group setting where we have a treatment group and control group, and we are interested in estimating heterogeneous treatment effect and related quantities. Suppose we have i.i.d observations $(X_i, Y_i, T_i)_{i=1}^n$ where $T_i \in \{0,1\}$ is the treatment assignment. We adopt the potential outcome model and posit the existence of $(Y_i(0), Y_i(1))$ corresponding to the outcome if T_i had been equal to 0 or 1, respectively, such that $Y_i = Y_i(T_i)$. Again for the zero-inflated semi-continuous response, we assume $Y_i(0) = W_i(0)Z_i(0)$ and $Y_i(1) = W_i(1)Z_i(1)$. Moreover, we suppose the following two standard assumptions in the causal inference literature

- Unconfoundedness: $T \perp\!\!\!\perp (Y(0), Y(1)) \,\big|\, X$
- Overlap: $0 < \pi(x) < 1$ for all x in the support of X, where $\pi(x) = \mathbb{P}\{T = 1 | X = x\}$ is the propensity score function.

The unconfoundedness assumption is particularly designed for observational data but it also applies to the randomized trial case where treatment assignment T is independent of everything

else. Readers shall not confuse unconfoundedness assumption with ZISC-unconfoundedness assumption, as the latter refers to distributional properties for $Y(0)$ and $Y(1)$ in the context here.

Depending on specific applications, the target of interest could be different. There are two potential interesting quantity here. One is Conditional Average Treatment Effect (CATE) given by

$$\begin{aligned} D(x) &= \mathbb{E}[Y(1) - Y(0) \mid X = x] \\ &= \mathbb{E}[W(1)Z(1) - W(0)Z(0) \mid X = x] \end{aligned}$$

The other one, which we call Conditional Average Implicit Treatment Effect (CAITE), is given by

$$D_I(x) = \mathbb{E}[Z(1) - Z(0) \mid X = x]$$

CAITE measures the average treatment effect for each individual, had the response for that individual turned out to be positive both with treatment and under control. This quantity could serve as a measure metric in A/B testing specifically for zero-inflated semi-continuous data.

Sometimes other metrics are of interest as well. For example, the conditional average propensity effect $\mathbb{E}[W(1) - W(0) \mid X = x]$ measures the change of the probability of making a purchase in customer expenditure application. We do not discuss these metrics in this paper but they may be worth considering in certain scenarios.

CATE estimation Here we review some state-of-the-art methods for estimating CATE. Those methods can be modified to estimate CAITE as well and we will discuss it in next section.

A straightforward way to estimate $D(x)$ is to estimate the conditional mean for treatment group $\mu^1(x)$ and that for control group $\mu^0(x)$ separately, using the data in treatment group and in control group respectively. The final estimator is obtained as $\hat{D}(x) = \hat{\mu}^1(x) - \hat{\mu}^0(x)$. This is called T-learner in [140].

Another method, called S-learner in [140], is to include treatment indicator T as a feature without any special role, and run any supervised machine learning algorithm on the combined

data set to estimate $\mu(x,t) = \mathbb{E}[Y \mid X = x, T = t]$. Then the CATE estimator is given by $\hat{D}(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0)$.

An alternative approach is to use X-learner [140] which often yields better results than the other two especially when two groups are unbalanced. The main steps of X-learner are outlined as follows. First we estimate $\mu^1(x)$ and $\mu^0(x)$ using treatment group data and control group data, respectively. Then we get pseudo-treatment effect using imputed responses, i.e. $\tilde{D}_i = Y_i(1) - \hat{\mu}^0(X_i)$ for i in treatment group and $\tilde{D}_i = \hat{\mu}^1(X_i) - Y_i(0)$ for i in control group. We treat these $\tilde{D}_i$ as responses to get estimated $\hat{D}_1(x)$ with treatment group data, and similarly get $\hat{D}_0(x)$ with control group data. Our final estimator $\hat{D}(x)$ is a weighted combination of $\hat{D}_1(x)$ and $\hat{D}_0(x)$, i.e.

$$\hat{D}(x) = g(x)\hat{D}_0(x) + (1 - g(x))\hat{D}_1(x)$$

where $g \in [0,1]$ is a weight function. Some sensible choices of g are $g = 1$, $g = 0$ and $g(x) = \hat{T}(x)$.

CAITE estimation For estimating $D_I(x)$, we can employ T-learner, S-learner or X-learner with some careful adjustment.

The use of T-learner is straightforward: one can use any method discussed in Section 6.2.1 to estimate $\mu_+^1(x)$ as well as $\mu_+^0(x)$, then obtain $\hat{D}_I(x) = \hat{\mu}_+^1(x) - \hat{\mu}_+^0(x)$.

For S-learner, one can combine all data with treatment indicator as added feature to estimate quantity $\mu_+(x,t) = \mathbb{E}[Z(t) \mid X = x, T = t]$, then the CAITE estimate is $\hat{D}_I(x) = \hat{\mu}_+(x,1) - \hat{\mu}_+(x,0)$.

As for X-learner, we need to do some modifications as $Z_i(0)$ or $Z_i(1)$ are only observable for some cases in both groups. To be specific, we first estimate $\mu_+^1(x)$ and $\mu_+^0(x)$ using treatment group data and control group data respectively using any methods discussed in Section 6.2.1. In the second step, for treatment group we obtain pseudo-implicit treatment effect $\tilde{D}_{Ii} = Z_i(1) - \hat{\mu}_+^0(X_i)$ for $i \in \mathscr{A}_1 = \{i : W_i(1) = 1\}$, then get estimated $\hat{D}_{I1}(x)$ using those values as responses. Similarly, we get pseudo-implicit treatment effect for the control group data points in set $\mathscr{A}_0 = \{i : W_i(0) = 1\}$, and obtain estimated $\hat{D}_{I0}(x)$. The final estimate is then a weighted combination of $\hat{D}_{I1}(x)$ and

$\hat{D}_{I0}(x)$.

Discussion Estimation of heterogeneous treatment effect is a long standing problem in causal inference literature. For zero-inflated responses, the semi-continuous nature of the data render the problem even harder as the effective sample size is often small for both groups. Depending on particular applications, both CATE and CAITE could be the target of interest. The concept of CAITE is specific to the zero-inflated data and as far as we know we are the first to propose flexible machine learning methods designed for such data structures. Even for CATE estimation, the direction application of some classical approaches may not be desirable and is often less efficient as a consequence of ignoring the 'spikiness' structure of the response. Fortunately, we find that meta learners like T, S and X learners are flexible enough to do a decent job in CATE or CAITE estimation with some necessary modifications added.

The choice of T, S or X learner are often application specific and is not easy in general. Here we try to make some brief comparison which hopefully could cast some insights. T learner is conceptually simple and easy to implement so it could be a good starting point or serve as a baseline. The fact that T learner first estimates $\mu(x)$ or $\mu_+(x)$ separately for both groups then takes the difference may lead to inflated variance of the final estimator. S learner is also relatively straightforward and in some cases it gives best results among the three as we shall see in Section **??**. However, S learner on top of certain base approaches like random forest tends to underestimate the target especially when the number of features is large. This can be seen in its estimation step where the estimate is given by the difference of predictions when the treatment assignment value is changed from 1 to 0. If the treatment assignment feature is not considered important by the base learner, the change of values for this feature is likely to result in mild change in the estimate. Hence S learner often yields estimator with small variance but noticeable downward bias. X learner seems to be more complex than the other two, although the idea is quite intuitive: it's trying to impute the unobserved (implicit) treatment effect and then apply machine learning algorithms on top of those imputed values. In our experiments, X learner appears to dominate T learner in most cases and it

seems to be more stable and robust. In addition, it's particularly suited for unbalanced group case as explained in [140]. Therefore, among the three approaches X learner is in a better position to be the off-the-shelve tool for estimating heterogeneous treatment effect.

In the two sample setting, besides estimation one could also be interested in quantifying the uncertainty of the estimate, i.e. constructing confidence intervals for the target of interest. In the associated A/B testing experiment often the question of interest is whether the treatment makes any difference for an experimental unit. This problem is essentially a hypothesis testing of whether CATE or CAITE is equal to 0 for any particular sample point (equivalently one can construct a confidence interval and check if it covers 0 or not). How to tackle those inference problems with finite sample guarantee remains an open question, and we leave it for future work.

6.3 Empirical Study

6.3.1 One sample scenario: $\mu_+(x)$ estimation

In this section, we design simulations for the one sample scenario to evaluate the performance of different methods for estimating conditional positive mean $\mu_+(x)$. Suppose the sample size is n and dimension of covariate is p. Our data generation model is as follows: for $1 \leq i \leq n$

$$X_i \overset{\text{i.i.d}}{\sim} \mathscr{D}_X \qquad W_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(e(X_i))$$
$$\log(Z_i) \overset{\text{ind.}}{\sim} \mathscr{N}(f(X_i), 1) \qquad Y_i = W_i \cdot Z_i$$

Here $\mathscr{D}_X$ is the distribution of the feature vectors X_i. We use the same data generation mechanism in [139]: we draw odd-numbered features independently from standard Gaussian distribution, and draw even-numbered features independently from a Bernoulli distribution with probability 0.5. As for ZISC-propensity score function $e(x)$, we assume it can be represented as a linear function with

logistic-link. To be specific, we set

$$e(x) = \exp(\eta(x))/(1+\exp(\eta(x))) \text{ where } \eta(x) = x^T\beta + \beta_0$$

We set each coordinate of β to be equal to $1/\sqrt{p}$, and set $\beta_0 = -3$ so $\mathbb{E}[e(X_i)]$ is near 0.1 which is close to the practical setting.

The conditional positive mean $\mu_+(x)$ is $\exp(f(x)+1/2)$ in our setting. Here we consider various forms of $f(x)$ as suggested by [139]. They represent cases of both univariate and multivariate; both linear and nonlinear; both additive and interactive. They are

$$\begin{aligned}
f_1(x) &= \mathbf{1}_{x_1>1} \qquad f_2(x) = x_1 \\
f_3(x) &= x_1 + x_3 + x_5 + x_7 + x_8 + x_9 - 2 \\
f_4(x) &= 4\mathbf{1}_{x_1>1}\mathbf{1}_{x_3>0} + 4\mathbf{1}_{x_5>1}\mathbf{1}_{x_7>0} + 2x_8x_9 \\
f_5(x) &= x_2x_4x_6 + 2x_2x_4(1-x_6) + 3x_2(1-x_4)x_6 \\
&\quad +4x_2(1-x_4)(1-x_6) + 5(1-x_2)x_4x_6 \\
&\quad +6(1-x_2)x_4(1-x_6) + 7(1-x_2)(1-x_4)x_6 \\
&\quad +8(1-x_2)(1-x_4)(1-x_6) \\
f_6(x) &= x_1^2 + x_2 + x_3^2 + x_4 + x_5^2 + x_6 + x_7^2 + x_8 + x_9^2 \\
f_7(x) &= \frac{1}{2}(f_3(x) + f_5(x)) \\
f_8(x) &= \sin(\pi x_1 x_2) + 2x_3^2 + x_4 + \frac{1}{2}x_5x_6
\end{aligned}$$

We center and scale each of the eight functions so they all have mean close to 0 and roughly the same variance with respect to $\mathscr{D}_X$. We compare the performance of following five algorithms: (1) Conditional Mean Regression (CMR). We use classical random forest as base learner. (2) CMR with log transformation (Log-CMR). Here we follow the discussion in Section 6.2 to do log transformations on positive responses before applying CMR, then exponential the output to obtain final estimate. (3) Zero-Inflated Causal Forest (ZICF). The implementation of this algorithm relies

on the *grf* R package [141]. (4) ZICF with log transformation (Log-ZICF). The transformation steps are the same as discussed before. (5) Transformed Outcome Regression (TOR). Here we use logistic regression to estimate ZISC propensity score, then use random forest as supervised learning algorithm on transformed responses.

In our simulation, we set $p = 20$ and $n = 500$. For each trial, we generate the data and train aforementioned methods on it, then we evaluate the performance of each algorithm on test data ($n_{\text{test}} = 500$) and record the Mean Squared Error (MSE) for estimating $\mu_+(x)$. We tried 100 trials for each of the eight f functions and plotted boxplot of MSE for each case. The results are shown in Figure 6.1. The MSE for transformed outcome regression turned out to be much larger than the other four for all settings we tried so we didn't show it in the boxplot, but it may behave well in other settings.

According to Figure 6.1, we see that all four algorithms have comparable performance with each other, and ZICF has the best performance in terms of lowest distribution of errors for all scenarios except Scenario 6. The improvement of ZICF compared to the other three is significant especially in Scenario 1 and 2. In general, the log-transformation trick would yield slightly worse result, but exception exists: in Scenario 3 Log-CMR is slightly better than CMR. The benefit of log-transformation trick lies in the fact that it often leads to less variable result, as shown in Scenario 7 and 8 where the variance of MSE is much smaller for log-transformed algorithms than their untransformed counterpart.

6.3.2 Two sample scenario: CAITE estimation

In this section, we generate synthetic data to compare the performance of T-learner, S-learner and X-learner for estimating CAITE in the two group experiment case. We consider randomized trials setting where the treatment assignment is independent of everything else. The generation

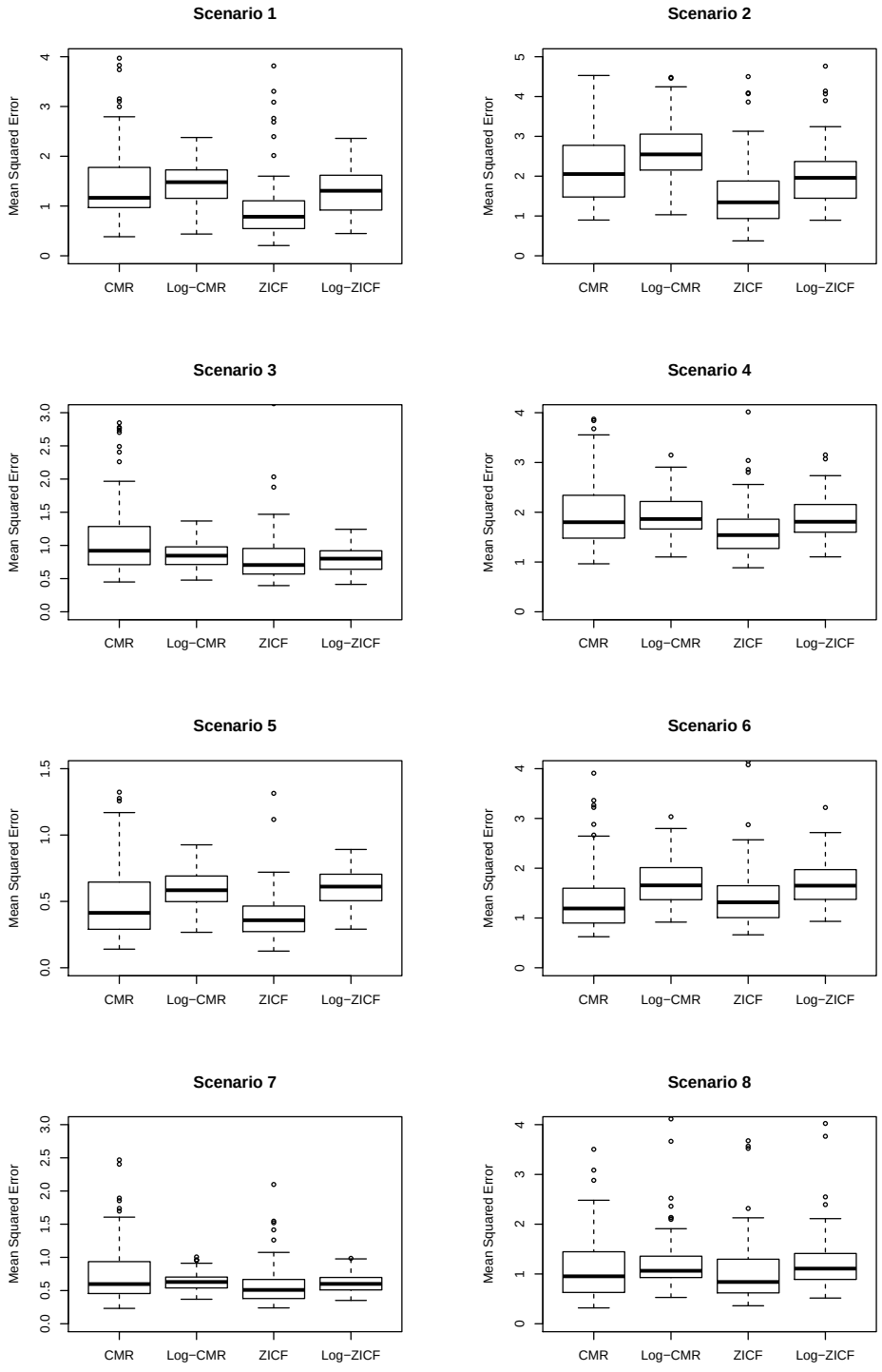


Figure 6.1: Boxplots of mean squared error across eight simulated experiments. The methods are: CMR = conditional mean regression; Log-CMR = conditional mean regression with log-transformation; ZICF = zero-inflated causal forest; Log-ZICF = zero-inflated causal forest with log-transformation. The result of TOR (transformed outcome regression) is much worse than the other four so it's not displayed in the figure. Each boxplot is based on results across 100 trials.

mechanism is as follows: for $1 \leq i \leq n$

$$
\begin{aligned}
&X_i \overset{\text{i.i.d}}{\sim} \mathscr{D}_X \qquad T_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(0.5) \\
&W_i(1), W_i(0) \overset{\text{ind.}}{\sim} \text{Bernoulli}(e(X_i)) \\
&\log(Z_i(1)) \overset{\text{ind.}}{\sim} \mathscr{N}\left(g(X_i) + \tau(X_i), 1\right) \\
&\log(Z_i(0)) \overset{\text{ind.}}{\sim} \mathscr{N}\left(g(X_i) - \tau(X_i), 1\right) \\
&Y_i = W_i(T_i) \cdot Z_i(T_i)
\end{aligned}
$$

Here $\mathscr{D}_X$ and $e(x)$ are the same as in section 6.3.1. As for function g and τ, we tried eight scenarios where in each scenario g and τ are one of the eight functions in section 6.3.1. The allocation of function form is given in Table 6.1. It's easy to see that CAITE at $X = x$ is equal to $\exp(g(x) + \tau(x) + 1/2) - \exp(g(x) - \tau(x) + 1/2)$.

In our simulation, we set $p = 20$ and varied training sample size from 500 to 10,000. For each setting, we generate the data, train different algorithms and evaluate their performance on test data with size 10^5 units. We tried 20 trials and recorded the average MSE for CAITE estimation. We use ZICF as base learners for T-learner, S-learner and first step of X-learner, where we employ classical random forest in the second step of X-learner and use equal weights for final output. Results are shown in Figure 6.2.

From Figure 6.2, we see that X-learner in general performs the best among the three algorithms and it dominates T-learner for most settings we tried. S-learner can be considered "winner" in Scenario 5 and 8 but it performs worst in Scenario 2 and 7. As training size increases, the MSE for all three algorithms decreases which is expected. In general, we would recommend X-learner as the default choice for CAITE estimation.

scenarios	1	2	3	4	5	6	7	8
$g(x)$	$f_6(x)$	$f_3(x)$	$f_4(x)$	$f_1(x)$	$f_8(x)$	$f_4(x)$	$f_5(x)$	$f_7(x)$
$\tau(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$

Table 6.1: Specification for eight simulation scenarios.

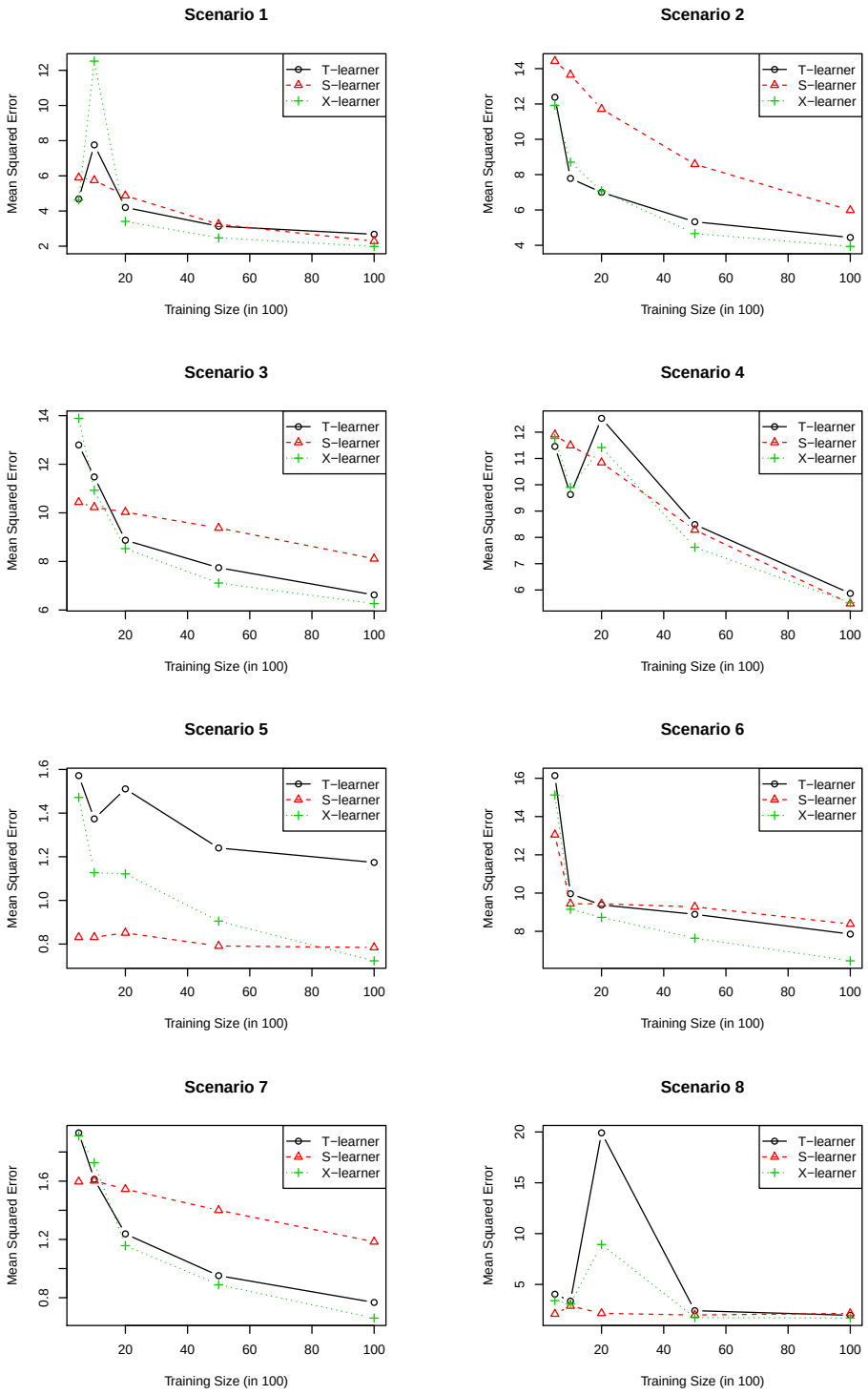


Figure 6.2: Comparison of T-learner, S-learner and X-learner across eight simulation scenarios. For details of generation mechanism, see Table 6.1.

6.3.3 Real data application

We study the donation data set DONOR which is adapted from KDD Cup 1998 data. DONOR can be accessed via *regclass* package in R software, and it contains donation records as well as donor's information from a national veterans organization. After some data cleaning and preprocessing, we are left with $p = 42$ features and $n = 14387$ samples. The response is the amount a past donor donates in response to the 97NK mail solicitation, whereas the donor features include age, gender, total gift amount, most recent donation amount, etc. This is a typical case of zero-inflated semi-continuous response, as most past donors do not donate anything for this event while a small proportion of past donors donate large amount of money. For this data, The proportion of zero donations among past donors is 0.74 and the average amount of donations of those who made positive donations is 15.69.

One sample scenario We consider estimating conditional positive mean $\mu_+(x)$ for the donation data. Since we don't know the true $\mu_+(x)$ for each donor, we instead use prediction error to quantify the accuracy of different methods. For this purpose, we split the data into training set and test set where we train the model on the training data and calculate the prediction error on the test data. In particular, each time we randomly select 80% of sample points as the training set denoted by T_1 and use the remaining samples as the test set denoted by T_2. After training models on T_1, we use the trained model to estimate $\mu_+(x_i)$ for each x_i in the test set T_2 that is paired with positive response y_i. We then calculate the root mean squared errors (RMSE) as $\sqrt{\sum_{i\in T_2, y_i>0}(y_i - \hat{\mu}_+(x_i))^2/n_1}$ where n_1 is the number of positive responses in the test set T_2. By repeating this procedure for 50 times, we obtained the boxplot for RMSE in Figure 6.3. We see that all methods have comparable performance, and ZICF has the lowest average RMSE 8.49 compared to that of CMR which is 8.57. It seems that log transformation on the response does not help in this case. The result suggests ZICF is a promising approach for this data set, and we expect to see better results (lower RMSE) for all methods if more careful feature transformations are performed before the data is fed into those machine learning algorithms.

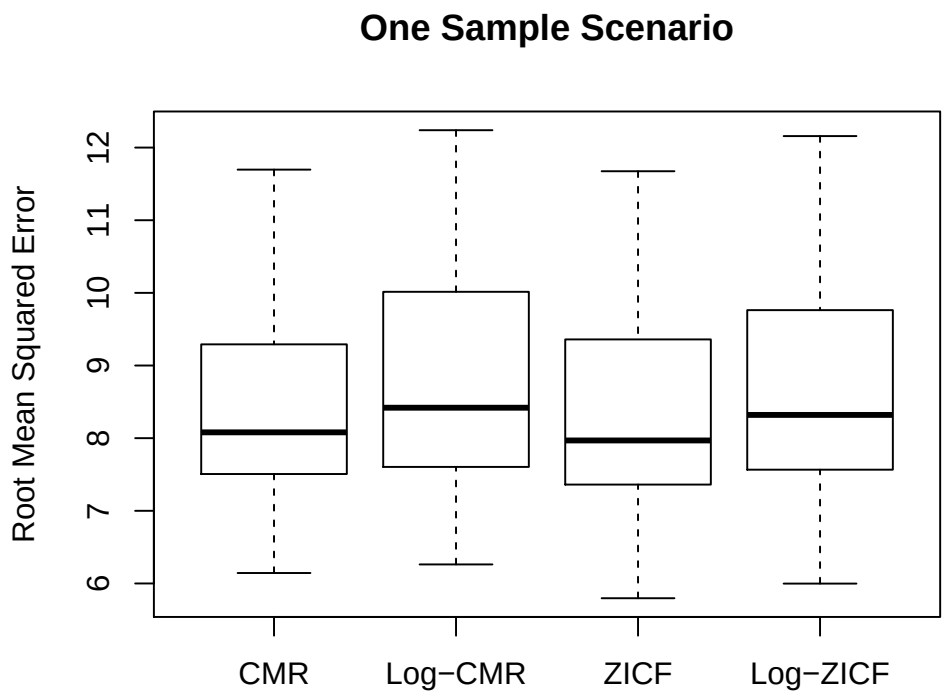


Figure 6.3: Boxplots of RMSE for estimating $\mu_+(x)$ on DONOR data. The results are based on 50 trials.

Two sample scenario We consider estimating CAITE in the two group setting for the donation data. For this purpose, we implement an A/A testing where we randomly choose half of data points in group one and leave the remaining samples in group two with no treatment assigned on either group. Since we do nothing on either group, the (implicit) treatment effect is supposed to be zero. Again we split the data into training set and test set, with group splitting and model training done on training set and estimation performed on the test set. We calculate RMSE for each method on the test set where the estimation is compared against the true CAITE which is 0. We repeat the procedure for 50 times and the results are shown in Figure 6.4. It's clear that S-learner is the winner with average RMSE 0.14 much better than the other two, whereas X-learner performs better than the T- learner. S-learner performs so well in the null case here partly because it tends

to underestimate CAITE. In practice, we suggest users to try both S-learner and X-learner in any A/B testing experiment before reaching to a conclusion.

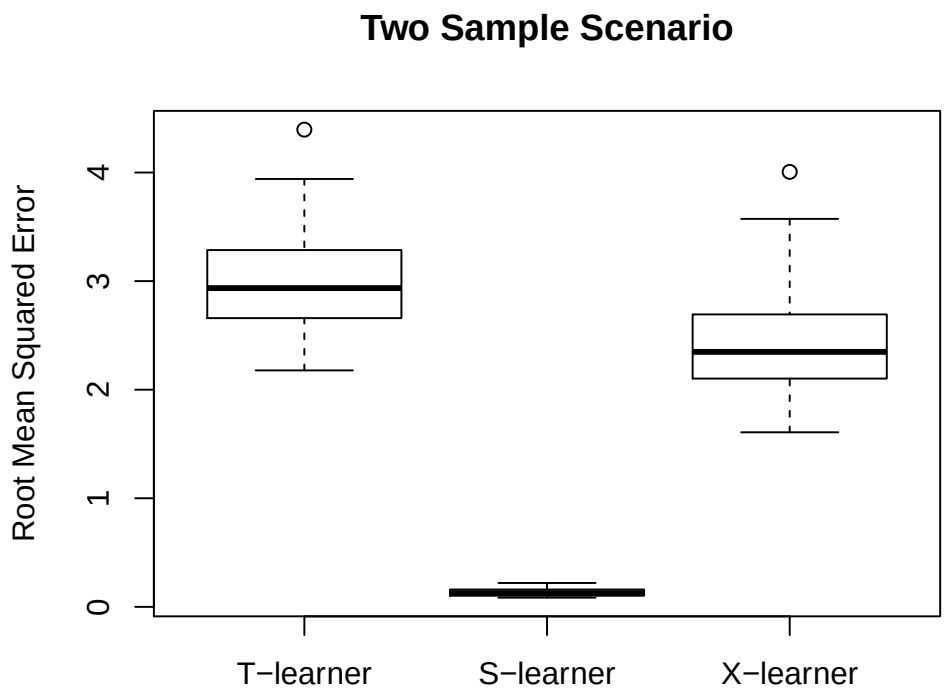


Figure 6.4: Boxplots of RMSE for CAITE estimation on DONOR data. The results are based on 50 trials.

CHAPTER 7

DISCUSSION

In this thesis we studied problems including selective inference, isotonic regression, variable ranking and matrix decomposition. We now summarize the results and discuss potential future research directions for each setting.

In Chapter 2 we discussed how to do selective inference for high dimensional linear model with group sparsity and gave detailed instructions on each step for forward stepwise regression, iterative hard thresholding and group lasso. A potential future direction is regarding how to extend our results to the case when noise level σ is unknown. It's also interesting to study the problem of constructing exact confidence intervals instead of the conservative one presented in this thesis.

Chapter 3 discussed one dimensional isotonic regression and derived finite sample convergence bounds as well as data adaptive confidence bands using new technical tools. An interesting future direction is to extend current results to multi-dimensional setting. Another possible direction is to establish similar results for other shape constrained nonparametric regression like convex regression.

In Chapter 4 we considered variable ranking problem and proposed new method that can take care of the correlation structures among covariates. We discussed its extension to the generalized linear model case and found it is promising via empirical study. It is of interest to study the theoretical properties of GLM version of our method. Another open question is how to control false discovery rate (FDR) based on scores we proposed.

In Chapter 5 we proposed a new matrix decomposition which decomposes a covariance matrix into a low rank part plus a diagonally dominant part. We developed several algorithms for achieving this decomposition and discussed some applications of our method. An interesting problem is to study the convergence properties of our non-convex algorithm and also shed some light on the existence as well as uniqueness of such decomposition. We would also love to see more applications of our method.

In Chapter 6 we cast a new framework to deal with zero-inflated semi-continuous data and

proposed several machine learning algorithms for estimating quantities of interest in both one sample scenario and two sample setting. A potential future direction is to quantify the uncertainly of our estimator and construct valid confidence intervals. We leave this problem to future work.

BIBLIOGRAPHY

[1] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, pp. 4655–4666, 2007. [Online]. Available: http://dx.doi.org/10.1109/TIT.2007.909108

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[3] P. Jain, N. Rao, and I. S. Dhillon, "Structured sparse regression via greedy hard-thresholding," *CoRR*, vol. abs/1602.06042, 2016. [Online]. Available: http://arxiv.org/abs/1602.06042

[4] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 49–67, 2006.

[5] J. R. Loftus and J. E. Taylor, "Selective inference in regression models with groups of variables," 2015, arXiv:1511.01478.

[6] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B*, vol. 75, pp. 603–680, 2013.

[7] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist.*, pp. 962–994, 2004.

[8] J. D. Lee and J. E. Taylor, "Exact post model selection inference for marginal screening," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 136–144.

[9] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, "Exact post-selection inference with the lasso," *arXiv preprint arXiv:1311.6238*, 2013.

[10] M. Yu, M. Kolar, and V. Gupta, "Statistical inference for pairwise graphical models using score matching," in *Advances in Neural Information Processing Systems*, 2016, pp. 2829–2837.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[12] F. Yang, R. F. Barber, P. Jain, and J. Lafferty, "Selective inference for group-sparse linear models," in *Advances in Neural Information Processing Systems*, 2016, pp. 2469–2477.

[13] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 433–440.

[14] S. Mosci, S. Villa, A. Verri, and L. Rosasco, "A primal-dual algorithm for group sparse regularization with overlapping groups," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 2604–2612.

[15] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *Information Theory, IEEE Transactions on*, vol. 55, pp. 1872–1882, 2009.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[17] Y. Yang and H. Zou, *gglasso: Group Lasso Penalized Learning Using A Unified BMD Algorithm*, 2014, r package version 1.3. [Online]. Available: https://CRAN.R-project.org/package=gglasso

[18] R. E. Barlow, D. J. Bartholomew, J. Bremner, and H. D. Brunk, *Statistical inference under*

order restrictions: The theory and application of isotonic regression. Wiley New York, 1972.

[19] T. Robertson, F. T. Wright, and R. L. Dykstra, "Order restricted statistical inference," 1988.

[20] F. Yang, R. F. Barber *et al.*, "Contraction and uniform convergence of isotonic regression," *Electronic Journal of Statistics*, vol. 13, pp. 646–677, 2019.

[21] S. Van de Geer, "Estimating a regression function," *The Annals of Statistics*, pp. 907–924, 1990.

[22] Y. Wang and K. Chen, "The l2risk of an isotonic estimate," *Communications in Statistics-Theory and Methods*, vol. 25, pp. 281–294, 1996.

[23] M. Meyer and M. Woodroofe, "On the degrees of freedom in shape-restricted regression," *Annals of Statistics*, pp. 1083–1104, 2000.

[24] S. Van de Geer, "Hellinger-consistency of certain nonparametric maximum likelihood estimators," *The Annals of Statistics*, pp. 14–44, 1993.

[25] C.-H. Zhang, "Risk bounds in isotonic regression," *The Annals of Statistics*, vol. 30, pp. 528–555, 2002.

[26] S. Chatterjee, A. Guntuboyina, B. Sen *et al.*, "On risk bounds in isotonic and other shape restricted regression problems," *The Annals of Statistics*, vol. 43, pp. 1774–1800, 2015.

[27] C. Gao, F. Han, and C.-H. Zhang, "Minimax risk bounds for piecewise constant models," *arXiv preprint arXiv:1705.06386*, 2017.

[28] H. D. Brunk, *Estimation of isotonic regression*. University of Missouri-Columbia, 1969.

[29] F. T. Wright, "The asymptotic behavior of monotone regression estimates," *The Annals of Statistics*, vol. 9, pp. 443–448, 1981.

[30] E. Cator, "Adaptivity and optimality of the monotone least-squares estimator," *Bernoulli*, vol. 17, pp. 714–735, 2011.

[31] L. Dümbgen, "Optimal confidence bands for shape-restricted curves," *Bernoulli*, vol. 9, pp. 423–449, 2003.

[32] U. Grenander, "On the theory of mortality measurement: part ii," *Scandinavian Actuarial Journal*, vol. 1956, pp. 125–153, 1956.

[33] B. P. Rao, "Estimation of a unimodal density," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 23–36, 1969.

[34] P. Groeneboom, "Estimating a monotone density," *Department of Mathematical Statistics*, pp. 1–14, 1984.

[35] L. Birgé, "Estimating a density under order restrictions: Nonasymptotic minimax risk," *The Annals of Statistics*, pp. 995–1012, 1987.

[36] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Probability Theory and Related Fields*, vol. 97, pp. 113–150, 1993.

[37] C. Carolan and R. Dykstra, "Asymptotic behavior of the grenander estimator at density flat regions," *Canadian Journal of Statistics*, vol. 27, pp. 557–566, 1999.

[38] F. Balabdaoui, H. Jankowski, M. Pavlides, A. Seregin, and J. Wellner, "On the grenander estimator at zero," *Statistica Sinica*, vol. 21, art. no. 873, 2011.

[39] H. Jankowski, "Convergence of linear functionals of the grenander estimator under misspecification," *The Annals of Statistics*, vol. 42, pp. 625–653, 2014.

[40] C. Durot, V. N. Kulikov, and H. P. Lopuhaä, "The limit distribution of the l_∞-error of grenander-type estimators," *The Annals of Statistics*, pp. 1578–1608, 2012.

[41] J. Rice, "Bandwidth choice for nonparametric regression," *The Annals of Statistics*, pp. 1215–1230, 1984.

[42] T. Gasser, L. Sroka, and C. Jennen-Steinmetz, "Residual variance and residual pattern in nonlinear regression," *Biometrika*, pp. 625–633, 1986.

[43] M. Drton and C. Klivans, "A geometric interpretation of the characteristic polynomial of reflection arrangements," *Proceedings of the American Mathematical Society*, vol. 138, pp. 2873–2887, 2010.

[44] S. Chatterjee, A. Guntuboyina, B. Sen *et al.*, "On matrix estimation under monotonicity constraints," *Bernoulli*, vol. 24, pp. 1072–1100, 2018.

[45] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth, "Isotonic regression in general dimensions," *arXiv preprint arXiv:1708.09468*, 2017.

[46] H. Deng and C.-H. Zhang, "Isotonic regression in multi-dimensional spaces and graphs," *arXiv preprint arXiv:1812.08944*, 2018.

[47] L. M. Bregman, "The method of successive projection for finding a common point of convex sets(theorems for determining common point of convex sets by method of successive projection)," *Soviet Mathematics*, vol. 6, pp. 688–692, 1965.

[48] S.-P. Han, "A successive projection method," *Mathematical Programming*, vol. 40, pp. 1–14, 1988.

[49] Z. T. Ke and F. Yang, "Covariate assisted variable ranking," *arXiv preprint arXiv:1705.10370*, 2017.

[50] G. Chamberlain and M. Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, vol. 51, pp. 1281–1304, 1983.

[51] J. Fan, Y. Liao, and M. Mincheva, "High dimensional covariance matrix estimation in approximate factor models," *Ann. Statist.*, vol. 39, art. no. 3320, 2011.

[52] G. Connor and R. A. Korajczyk, "A test for the number of factors in an approximate factor model," *J. Finance*, vol. 48, pp. 1263–1291, 1993.

[53] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genet*, vol. 3, art. no. e161, 2007.

[54] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC bioinformatics*, vol. 7, art. no. 359, 2006.

[55] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Inform.*, vol. 13, pp. 51–60, 2002.

[56] L. Wasserman and K. Roeder, "High dimensional variable selection," *Ann. Statist.*, vol. 37, art. no. 2178, 2009.

[57] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, pp. 217–288, 2011.

[58] A. M. Frieze and M. Molloy, "Splitting an expander graph," *J. Algorithms*, vol. 33, pp. 166–172, 1999.

[59] P. Ji and J. Jin, "Ups delivers optimal phase diagram in high-dimensional variable selection," *Ann. Statist.*, vol. 40, pp. 73–103, 2012.

[60] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 76, pp. 217–242, 2014. [Online]. Available: http://dx.doi.org/10.1111/rssb.12026

[61] S. van de Geer, P. BÃijhlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal

confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, pp. 1166–1202, 06 2014. [Online]. Available: http://dx.doi.org/10.1214/14-AOS1221

[62] J. Jin, C.-H. Zhang, and Q. Zhang, "Optimality of graphlet screening in high dimensional variable selection," *J. Mach. Learn. Res.*, vol. 15, pp. 2723–2772, 2014.

[63] Z. T. Ke, J. Jin, and J. Fan, "Covariate assisted screening and estimation," *Ann. Statist.*, vol. 42, pp. 2202–2242, 2014.

[64] J. Fan, X. Han, and W. Gu, "Estimating false discovery proportion under arbitrary covariance dependence," *J. Amer. Statist. Assoc.*, vol. 107, pp. 1019–1035, 2012.

[65] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 75, pp. 603–680, 2013.

[66] H. Wang, "Factor profiled sure independence screening," *Biometrika*, vol. 99, pp. 15–28, 2012.

[67] J. J. Chen, C.-A. Tsai, S. Tzeng, and C.-H. Chen, "Gene selection with multiple ordering criteria," *BMC Bioinformatics*, vol. 8, art. no. 74, 2007.

[68] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[69] J. Jin and Z. T. Ke, "Rare and weak effects in large-scale inference: methods and phase diagrams," *Statist. Sinica*, vol. 26, pp. 1–34, 2016.

[70] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM J. Numer. Anal.*, vol. 7, pp. 1–46, 1970.

[71] J. Jin, Z. T. Ke, and W. Wang, "Phase transitions for high dimensional clustering and related problems," *Ann. Statist. (to appear)*, 2016.

[72] Z. T. Ke and M. Wang, "A new svd approach to optimal topic estimation," *arXiv:1704.07016*, 2017.

[73] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 70, pp. 849–911, 2008.

[74] X. Wang and C. Leng, "High dimensional ordinary least squares projection for screening variables," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 2015.

[75] G. Li, H. Peng, J. Zhang, and L. Zhu, "Robust rank correlation based screening," *Ann. Statist.*, vol. 40, pp. 1846–1877, 2012.

[76] R. R. Nayak, M. Kearns, R. S. Spielman, and V. G. Cheung, "Coexpression network based on natural variation in human gene expression reveals gene interactions and functions," *Genome Res.*, vol. 19, pp. 1953–1962, 2009.

[77] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: beyond the linear model," *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, 2009.

[78] J. Fan and R. Song, "Sure independence screening in generalized linear models with np-dimensionality," *Ann. Statist.*, vol. 38, pp. 3567–3604, 2010.

[79] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, pp. 1–25, 1982. [Online]. Available: http://www.jstor.org/stable/1912526

[80] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, vol. 33, pp. 3–56, 1993.

[81] G. Chamberlain and M. Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, vol. 51, pp. 1281–1304, 1983.

[82] J. T. Leek and J. D. Storey, "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 18 718–18 723, 2008.

[83] J. Fan, X. Han, and W. Gu, "Estimating false discovery proportion under arbitrary covariance dependence," *Journal of the American Statistical Association*, vol. 107, pp. 1019–1035, 2012.

[84] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the Association for Computing Machinery*, vol. 58, pp. 11:1–37, 2011.

[85] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, vol. 106, pp. 672–684, 2011.

[86] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, pp. C1–C32, 2016.

[87] M. Mendoza, M. Raydan, and P. Tarazaga, "Computing the nearest diagonally dominant matrix," *Numerical linear algebra with applications*, vol. 5, pp. 461–474, 1998.

[88] P. J. Bickel, "One-step huber estimates in the linear model," *Journal of the American Statistical Association*, vol. 70, pp. 428–434, 1975.

[89] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models (with discussion)," *Annals of statistics*, vol. 36, art. no. 1509, 2008.

[90] J. Fan, L. Xue, and H. Zou, "Strong oracle optimality of folded concave penalized estimation," *Annals of statistics*, vol. 42, art. no. 819, 2014.

[91] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, art. no. 2605, 2008.

[92] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, art. no. 904, 2006.

[93] J. Fan, Y. Liao, and M. Mincheva, "High dimensional covariance matrix estimation in approximate factor models," *Annals of statistics*, vol. 39, art. no. 3320, 2011.

[94] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[95] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, pp. 432–441, 2008.

[96] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, pp. 177–186, 2009.

[97] D. Donoho and J. Jin, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 14 790–14 795, 2008.

[98] T. Cai, W. Liu, and X. Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, vol. 106, pp. 594–607, 2011.

[99] Y. Fan, J. Jin, Z. Yao *et al.*, "Optimal classification in sparse gaussian graphic model," *The Annals of Statistics*, vol. 41, pp. 2537–2571, 2013.

[100] S. Huang, J. Jin, Z. Yao *et al.*, "Partial correlation screening for estimating large precision matrices, with applications to classification," *The Annals of Statistics*, vol. 44, pp. 2018–2057, 2016.

[101] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer research*, vol. 62, pp. 4963–4967, 2002.

[102] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, pp. 671–679, 2005.

[103] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 6567–6572, 2002.

[104] J. Jin, W. Wang *et al.*, "Influential features pca for high dimensional clustering," *The Annals of Statistics*, vol. 44, pp. 2323–2359, 2016.

[105] R. J. Simes, "An improved bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, pp. 751–754, 1986.

[106] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, vol. 89, pp. 82–93, 2011.

[107] L. Jager and J. A. Wellner, "Goodness-of-fit tests via phi-divergences," *The Annals of Statistics*, vol. 35, pp. 2018–2053, 2007.

[108] P. Hall and J. Jin, "Innovated higher criticism for detecting sparse signals in correlated noise," *The Annals of Statistics*, vol. 38, pp. 1686–1732, 2010.

[109] S. Ma, L. Xue, and H. Zou, "Alternating direction methods for latent variable gaussian graphical model selection," *Neural computation*, vol. 25, pp. 2172–2198, 2013.

[110] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of admm for multi-block convex minimization problems is not necessarily convergent," *Mathematical Programming*, vol. 155, pp. 57–79, 2016.

[111] T. Lin, S. Ma, and S. Zhang, "Global convergence of unmodified 3-block admm for a class of convex minimization problems," *arXiv preprint arXiv:1505.04252*, 2015.

[112] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *Advances in Neural Information Processing Systems*, 2014, pp. 1107–1115.

[113] D. Drusvyatskiy, A. Ioffe, and A. Lewis, "Transversality and alternating projections for nonconvex sets," *Foundations of Computational Mathematics*, vol. 15, pp. 1637–1651, 2015.

[114] W. Owen and T. DeRouen, "Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants," *Biometrics*, pp. 707–719, 1980.

[115] M. Pennington, "Efficient estimators of abundance, for fish and plankton surveys," *Biometrics*, pp. 281–286, 1983.

[116] X.-H. Zhou and W. Tu, "Confidence intervals for the mean of diagnostic test charge data containing zeros," *Biometrics*, vol. 56, pp. 1118–1125, 2000.

[117] L. H. Moulton and N. A. Halsey, "A mixture model with detection limits for regression analyses of antibody response to vaccine," *Biometrics*, pp. 1570–1578, 1995.

[118] J. Aitchison, "On the distribution of a positive random variable having a discrete probability mass at the origin," *Journal of the american statistical association*, vol. 50, pp. 901–908, 1955.

[119] J. Aitchison and J. A. Brown, "The lognormal distribution with special reference to its uses in economics," 1957.

[120] D. Fletcher, "Confidence intervals for the mean of the delta-lognormal distribution," *Environmental and Ecological Statistics*, vol. 15, pp. 175–189, 2008.

[121] M. A. C. Rosales, *The robustness of confidence intervals for the mean of delta distribution*. Western Michigan University, 2009.

[122] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica: journal of the Econometric Society*, pp. 24–36, 1958.

[123] T. Amemiya, "Tobit models: A survey," *Journal of econometrics*, vol. 24, pp. 3–61, 1984.

[124] P. M. Robinson, "On the asymptotic properties of estimators of models containing limited dependent variables," *Econometrica: Journal of the Econometric Society*, pp. 27–41, 1982.

[125] J. L. Powell, "Symmetrically trimmed least squares estimation for tobit models," *Econometrica: journal of the Econometric Society*, pp. 1435–1460, 1986.

[126] N. Duan, W. G. Manning, C. N. Morris, and J. P. Newhouse, "A comparison of alternative models for the demand for medical care," *Journal of business & economic statistics*, vol. 1, pp. 115–126, 1983.

[127] J. Grytten, D. Holst, and P. Laake, "Accessibility of dental services according to family income in a non-insured population," *Social Science & Medicine*, vol. 37, pp. 1501–1508, 1993.

[128] J. Heckman, "Shadow prices, market wages, and labor supply," *Econometrica: journal of the econometric society*, pp. 679–694, 1974.

[129] J. J. Heckman, "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)," 1977.

[130] W. P. Van de Ven and B. M. Van Praag, "The demand for deductibles in private health insurance: A probit model with sample selection," *Journal of econometrics*, vol. 17, pp. 229–252, 1981.

[131] B. Jorgensen, "Exponential dispersion models," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 127–162, 1987.

[132] ——, *The theory of dispersion models*. CRC Press, 1997.

[133] A. Saei, J. Ward, and C. McGilchrist, "Threshold models in a methadone programme evaluation," *Statistics in Medicine*, vol. 15, pp. 2253–2260, 1996.

[134] Y. Min and A. Agresti, "Modeling nonnegative data with clumping at zero: a survey," *Journal of the Iranian Statistical Society*, vol. 1, pp. 7–33, 2002.

[135] E. D. Mills, "Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data," 2013.

[136] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2017.

[137] S. Wager, T. Hastie, and B. Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *The Journal of Machine Learning Research*, vol. 15, pp. 1625–1651, 2014.

[138] E. Scornet, G. Biau, J.-P. Vert *et al.*, "Consistency of random forests," *The Annals of Statistics*, vol. 43, pp. 1716–1741, 2015.

[139] S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani, "Some methods for heterogeneous treatment effect estimation in high dimensions," *Statistics in medicine*, vol. 37, pp. 1767–1787, 2018.

[140] S. Künzel, J. Sekhon, P. Bickel, and B. Yu, "Meta-learners for estimating heterogeneous treatment effects using machine learning," *arXiv preprint arXiv:1706.03461*, 2017.

[141] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *arXiv preprint arXiv:1610.01271*, 2016.