UNIVERSITY OF CALIFORNIA, SAN DIEGO

Inference for High-dimensional Left-censored Linear Model and High-dimensional Precision Matrix

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

in

Mathematics

by

Jiaqi Guo

Committee in charge:

Professor Jelena Bradic, Chair
Professor Ery Arias-Castro
Professor Ivana Komunjer
Professor Loki Natarajan
Professor Rayan Saab

2018

The dissertation of Jiaqi Guo is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2018

DEDICATION

To my family, who supports and believes in me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor, Professor Jelena Bradic, whose teaching and support guided me throughout the five years of my Ph.D. study at UC San Diego. Professor Bradic introduced me to statistics as a subject and the research in high-dimensional statistics. More importantly, she trained me to become an independent researcher. I cannot owe her more for her patience and encouragement, as well as her care for me at difficult times. Without her, this thesis would not exist.

I would like to thank Professor Arias-Castro and Professor Saab for their teachings. Their courses and seminars in various topics not only provided me with solid foundations in statistical learning, but also expanded my perspective with relevant research topics. I am also grateful to Professor Komunjer and Professor Natarajan for serving as my committee members and offering meaningful discussions.

To the professors I have worked with as a teaching assistant and all my previous students, thank you for perfecting my teaching skills. To the Mathematics Department, I appreciate the resources provided, especially the departmental server `euler.ucsd.edu`, which greatly accelerated my computational experiments. In addition, thanks to Wilson Cheung, Holly Proudfoot and Scott Rollans for the help and assistance in the past. I also want to thank my friends and fellow graduate students in the department. Special thanks to Michelle, Chengjie, Ching Wei, Hanbo, Jingwen, Yuchao, Kuang, Francois, Stephan, Andrew and Selene for all the funs, all the delicious foods, the interesting conversations and the ideas shared.

Outside of the UC San Diego community, I have had the honor to work with Ben Wilson at Viasat, Luyen Le at Target, and Christopher Nam and Kari Torkkola at Amazon. I would like to thank them for hosting me as an intern and helping me gain industry experience. I would also like to note how great it has been living in San Diego, whose always-perfect weather and gorgeous beaches have cheered me up numerous times.

Last but not least, I would like to extend my deepest gratitude and appreciation to my

family. I am truly indebted and thankful for everything my parents have offered me. They taught me to be kind and compassionate, independent and strong, and supported me to pursue my dreams. Special thanks also goes to my grandmother, whose optimism and wisdom of life inspire me every day.

Chapter 1, in full, is a version of the paper “Generalized M-estimators for high-dimensional Tobit I models”. The dissertation author is the principal investigator of this material. The material is under revision for publication.

Chapter 2, in full, is a version of the paper “High-dimensional covariate effects on left-censored quantile event times". The dissertation author is one of the principal investigators of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Chapter 3, in full, is a version of the paper “Testing generalized hypotheses for high-dimensional precision matrix". The dissertation author is the principal investigator of this material. The manuscript is being prepared to be submitted to a major statistics journal.

VITA

2013	B. S. in Applied Mathematics and B. A. in German Studies *summa cum laude*, Emory University
2013-2018	Graduate Teaching Assistant, University of California, San Diego
2015	C. Phil. in Mathematics, University of California, San Diego
2018	Ph. D. in Mathematics, University of California, San Diego

PUBLICATIONS

Bradic, Jelena, and Guo, Jiaqi. "Generalized M-estimators for high-dimensional Tobit I models", *Electronic Journal of Statistics*, Under Revision, 2015.

Bradic, Jelena, Guo, Jiaqi, and Li, Hanbo. "High-dimensional covariate effects on left-censored quantile event times", *Manuscript in Preparation*, 2018.

Liu, Yuchao, and Guo, Jiaqi. "Distribution-free, size adaptive submatrix detection with acceleration", *Statistics and Computing*, Under Review, 2018.

Bradic, Jelena, and Guo, Jiaqi. "Testing generalized hypotheses for high-dimensional precision matrix", *Manuscript in Preparation*, 2018.

ABSTRACT OF THE DISSERTATION

Inference for High-dimensional Left-censored Linear Model and High-dimensional Precision Matrix

by

Jiaqi Guo

Doctor of Philosophy in Mathematics

University of California, San Diego, 2018

Professor Jelena Bradic, Chair

In the first two chapters, we consider inference for high-dimensional left-censored linear models. Left-censored data arises from measurement limits in scientific devices and social science data. We consider the problem of constructing confidence intervals for the parameters in left-censored linear models. In Chapter 1, we present smoothed estimating equations (SEE) and smoothed robust estimating equations(SREE) frameworks that are adaptive to censoring level and are more robust to misspecification of the error distribution. In Chapter 2, we study inference problem for parameters in high-dimensional left-censored quantile regression model. We modify the quantile loss to accommodate the left-censored nature of the problem, by extending the idea

of redistribution of mass. Furthermore, applying the de-biasing technique to the initial estimator leads to an improved estimator suitable for high-dimensional inference under left-censored quantile regression setting. For both problems, asymptotic properties have been investigated.

In Chapter 3, we devise a projection pursuit testing procedure for generalized hypotheses on high-dimensional precision matrix. We illustrate the procedure under specific examples of hypotheses: testing for row sparsity, minimum signal strength, bandedness and generalized bandedness. We demonstrate the performance of the testing procedure through extensive numerical experiments, and present the findings for two real datasets.

Chapter 1

Generalized M-estimation for High-dimensional Left-censored Linear Model

1.1 Introduction

Left-censored data is a characteristic of many datasets. In physical science applications, observations can be censored due to limits in the measurements. For example, if a measurement device has a value limit on the lower end, the observations are recorded with the minimum value, even though the actual result is below the measurement range. In fact, many of the HIV studies have to deal with difficulties due to the lower quantification and detection limits of viral load assays [SCG+14]. In social science studies, censoring may be implied in the nonnegative nature or defined through human actions. Economic policies such as minimum wage and minimum transaction fee result in left-censored data, as quantities below the thresholds will never be observed. At the same time, with advances in modern data collection, high-dimensional data where the number of variables, p, exceeds the number of observations, n, are becoming more

and more commonplace. HIV studies are usually complemented with observations about genetic signature of each patient, making the problem of finding the association between the number of viral loads and the gene expression values extremely high dimensional.

In this chapter, we present a generalized M-estimation scheme for high-dimensional left-censored linear model, also known as Tobit I model, which is first presented in [Tob58]. We begin with an introduction of the model and its areas of applications, along with our major contributions in the novel methodology. Following that we summarize related work in the literature. Finally, we present Smoothed Estimating Equations (SEE) and Smoothed Robust Estimating Equations (SREE) frameworks, together with their theoretical properties.

1.1.1 Contributions

In general, we cannot develop p-values from the high-dimensional observations without further restrictions on the data generating distribution. A standard way to make progress is to assume that the model is selected consistently, for example in [ZY06, FL01], i.e., that the regularized estimator accurately selects the correct set of features. The motivation behind model selection consistency is that, given sparsity of the model at hand, it effectively implies that one can disregard all of the features whose coefficients are equal to zero. An immediate consequence is that p-values are now well defined for the small selected set of variables; see for example [BFW11]. Such results heavily rely on assumptions named "irrepresentative condition" and variants thereof, including but not limited to the minimal signal strength, see [VDGB$^+$09]. Thus, if we were to know that such conditions hold, p-value construction would follow standard literature of what are essentially low-dimensional problems. Many early applications of regularized methods effectively impose conditions similar to the irrepresentable condition, and then rely solely on the results of the regularized estimator. However, such restrictions can make it challenging to discover strong but unexpected significant signals. The SEE and SREE frameworks address these challenges. It is shown that valid p-values can be well defined for all of the features in the model through

development of robust, bias-corrected estimator that yields valid asymptotic inference regardless of whether or not irrepresentable-type conditions are assumed.

Classical approaches to inference in left-censored models, include maximum likelihood approaches as in [Ame73], consistent estimators of the asymptotic covariance matrix as in [Pow84], bayesian methods as in [Chi92], and maximum entropy principles as in [GJP97]. These methods perform well in applications with a small number of covariates (smaller than the sample size), but quickly break down as the number of covariates increases.

The current framework explores the use of ideas from the high-dimensional literature to improve the performance of these classical methods with many covariates. It is based on the family of de-biased estimators introduced by [ZZ14], which allow for optimal inference in high dimensions by building an estimator that corrects for the regularization bias. Bias-corrected estimators are related to one-step M-estimators in that they improve on an initial estimator by following a Newton-Raphson updating rule, see [Bic75]; however, they differ from the classical one-step M-estimators in that their initial step is not consistent and direct estimator of the asymptotic variance does not exist.

1.1.2 Related Work

From a technical point of view, our main contribution is an asymptotic normality theory enabling statistical inference in high-dimensional Tobit I models. Results by [Pow86a], [Pow86b] and [NP90] have established asymptotic properties in low-dimensional setting where the number of features is fixed, while [Son11] and [ZBW$^+$14] developed distribution free and rank-based tests. [MvdG16] offered a penalized version of Powell's estimator (penalized CLAD). Robustness properties of sample-selection models in low-dimensions were studied in [ZGR16].

A growing literature, including [VdGBR$^+$14], [ZZ14], [RSZ$^+$15] and [RWG$^+$16], has considered the use of regularized algorithms for performing inference in high-dimensional regression models. These papers use the bias correction method, and report confidence intervals

and p-values for testing feature significance. Meanwhile, [BCK14, BCK13], [ZKL14] and [JM14b] use robust approaches to estimate the asymptotic variance, and then use related bias correction step to remove the effect of regularization.

Several papers use one-step methods for eliminating the bias of regularized estimates. In removing the bias of the regularized estimates, we follow most closely the approach of [VdGBR$^+$14], which proposes bias correction estimator for least squares losses, and obtain valid confidence intervals. Other related approaches include those of [JM14b] and [NL17], which build different variance estimates to determine a more robust bias correction step; however, these papers only focus on least squares losses (more importantly they do not extend naively to non-smooth or non-differentiable loss functions). [BCK14] and [ZKL14] discuss one-step approaches for quantile inference; however, the tools and techniques heavily depend on the convexity of the quantile loss. It is worth mentioning that the double-robust approach of [BCCW17], which proposes a powerful inference method for quantile regression, is based on leveraging principles of doubly-robust scores and their estimating equations.

1.1.3 Content

In Section 1.2, we introduce the smoothed estimating equations (SEE) for left-censored linear models. In Section 1.3, we present the main result on confidence regions. In Section 1.4, we introduce robust and left-censored Mallow's, Schweppe's and Hill-Ryan's estimators and present their theoretical analysis. Section 1.5 provides numerical results on simulated data sets. In Section 1.6, we include discussions and conclusions for this work. We defer more general results for confidence regions, as well as the Bahadur representation of the SEE estimator, to Section 1.7. In addition, Section 1.8 and 1.9 consist of technical details and proofs.

1.2 Inference in Left-censored Regression

We begin by introducing a general modeling framework followed by highlighting the difficulty for directly applying existing inferential methods (such as de-biasing, score and Wald) to the models with left-censored observations. Finally, we propose a new mechanism, named smoothed estimating equations, to construct semi-parametric confidence regions in high-dimensions.

1.2.1 Left-censored Linear Model

We consider the problem of confidence interval construction where we observe a vector of responses $Y = (y_1, \ldots, y_n)$ and their censoring level $c = (c_1, \ldots, c_n)$ together with covariates $X_1, \ldots X_p$. The type of statistical inference under consideration is regular in the sense that it does not require model selection consistency. A characterization of such inference is that it does not require a uniform signal strength in the model. Since ultra-high dimensional data often display heterogeneity, we advocate a robust confidence interval framework. We begin with the following latent regression model:

$$y_i = \max\left\{c_i, x_i\boldsymbol{\beta}^* + \varepsilon_i\right\},$$

where the response Y and the censoring level c are observed, and the vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is unknown. Observe that the censoring mechanism considered here is fixed and non-random. This model is often called the semi-parametric censored regression model, whenever the distribution of the error ε is not specified. We assume that $\{\varepsilon_i\}_{i=1}^n$ are independent across i, and are independent of x_i. Matrix $X = [X_1, \cdots, X_p]$ is the $n \times p$ design matrix, where x_i's are i.i.d. random variables centered to have variance one element-wise and $\max_{i,j} |X_{ij}| \leq K$. When X_{ij} follows an unbounded continuous distribution, we can easily use truncation arguments to satisfy the bound above; this can be efficiently done for a wide class of sub-gaussian distributions for example. We also denote $S_{\boldsymbol{\beta}} := \{j | \boldsymbol{\beta}_j \neq 0\}$ as the active set of variables in $\boldsymbol{\beta}$ and its cardinality by $s_{\boldsymbol{\beta}} := |S_{\boldsymbol{\beta}}|$. We

restrict the study to constant-censored model, also called Type-I Tobit model, where entries of the censoring vector c are the same. Without loss of generality, we focus on the zero-censored model,

$$y_i = \max\left\{0, x_i\boldsymbol{\beta}^* + \varepsilon_i\right\}. \tag{1.1}$$

1.2.2 Smoothed Estimating Equations (SEE)

Smoothed Estimating Equations framework takes a general approach to the problem of designing robust and semi-parametric inference for left-censored linear models, and is motivated by the principles of estimating equations. Although estimating equations have been studied in many previous works, the smoothed estimating equations (SEE) framework presented in the following tailors to the high-dimensional and censored scenario. In addition, the method is simple enough to apply more generally to non-smooth loss functions. We begin by observing that the true parameter vector $\boldsymbol{\beta}^*$ satisfies the population system of equations

$$\mathbb{E}\Big[\Psi(\boldsymbol{\beta}^*)\Big] = 0. \tag{1.2}$$

for some function $\Psi(\boldsymbol{\beta})$ often taking the form of $\Psi(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta})$ for a class of suitable functions ψ_i. Observe that for left-censored models ε rarely, if ever, follows a specific distribution. A particular example of interest, that allows error misspecifications, is

$$\psi_i(\boldsymbol{\beta}) = \operatorname{sign}\left(y_i - \max\{0, x_i\boldsymbol{\beta}\}\right) w_i^{\top}(\boldsymbol{\beta}) \tag{1.3}$$

where $w_i(\boldsymbol{\beta}) = x_i\,\mathbb{1}\{x_i\boldsymbol{\beta} > 0\}$. The motivation comes from the renowned least absolute deviation l_1 loss. The advantage of the function ψ_i above is that it naturally bounds the effects of outliers; large values of the residuals $y_i - \max\{0, x_i\boldsymbol{\beta}\}$ are down-weighted using l_1 distance. In fact, we work with Ψ resulting from this specific choice of ψ_i function later in the analysis. Nevertheless,

the SEE framework has a much broader spectrum, see Remark 1 below. Other functions Ψ can be applied as well. Another example of a function Ψ that has semi-parametric advantage is a variant of a trimmed least squares loss, where the vanilla quadratic loss is multiplied by an indicator function as follows $\mathbb{I}\{y_i - x_i\boldsymbol{\beta} > 0, x_i\boldsymbol{\beta} > 0\}$.

However, with the appropriate choice of Ψ, solving estimating equations $\Psi(\boldsymbol{\beta}) = 0$, although practically desirable, still has several drawbacks, even in low-dimensional setting. In particular, for semi-parametric estimation and inference in model (1.1), the function Ψ is non-monotone as the loss is non-differentiable and non-convex. Hence, the system above has multiple roots resulting in an estimator that is ill-posed, and additionally presents significant theoretical challenges. Instead of solving the system (1.2) directly, we augment it by observing that, for a suitable choice of the matrix $\boldsymbol{\Upsilon} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\beta}^*$ also satisfies the system of equations

$$\mathbb{E}[\Psi(\boldsymbol{\beta}^*)] + \boldsymbol{\Upsilon}[\boldsymbol{\beta}^* - \boldsymbol{\beta}] = 0. \tag{1.4}$$

For certain choices of the matrix $\boldsymbol{\Upsilon}$, we aim to avoid both non-convexity and huge dimensionality of the system of equations (1.2). To avoid difficulties with non-smooth functions Ψ, we propose to consider a matrix $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)$, where the matrix $\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)$ is defined as

$$\boldsymbol{\Upsilon}(\boldsymbol{\beta}) = \mathbb{E}_X\left[\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta})\right],$$

for a smoothed vector $S(\boldsymbol{\beta})$ defined as

$$S(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} \Phi(\boldsymbol{\beta}, x) f_\varepsilon(x) dx.$$

The unknown error distribution smooths the function Ψ, and acts as a kernel smoother function. In the above display $\Psi(\boldsymbol{\beta}^*) = \Phi(\boldsymbol{\beta}^*, \varepsilon)$, for a suitable function $\Phi = n^{-1}\sum_{i=1}^n \phi_i$ and $\phi_i : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$, whereas f_ε denotes the density of the model error (1.1). Additionally, $\mathbb{E}_X$ denotes expectation

with respect to the random measure generated by the vectors $X_1, \ldots, X_n$.

Following Ψ as in (1.3), the respective smoothed score function that we will be working with is

$$S(\boldsymbol{\beta}^*) = n^{-1} \sum_{i=1}^{n} \left[1 - 2P_{\varepsilon}\left(y_i - x_i\boldsymbol{\beta}^* \leq 0\right)\right] \left(w_i(\boldsymbol{\beta}^*)\right)^{\top}, \tag{1.5}$$

where P_{ε} denotes the probability measure generated by the errors ε in (1.1). Smoothed score typically depends on the unknown density of the error terms and the unknown parameter of interest. For practical purposes, we will propose a suitable estimate of the function (1.5) – for homoscedastic errors ε_i, the unknown cumulative distribution function above can easily be estimated using empirical distribution function. With this choice of the smoothed loss, we obtain an information matrix as follows $\nabla_{\boldsymbol{\beta}^*} S(\boldsymbol{\beta}^*) = 2f_{\varepsilon}(0)n^{-1}\sum_{i=1}^{n} w_i(\boldsymbol{\beta}^*)^{\top} w_i(\boldsymbol{\beta}^*)$. We then proceed to define the matrix $\boldsymbol{\Upsilon}$ as

$$\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*) = 2f_{\varepsilon}(0)\mathbb{E}_X \left[n^{-1} \sum_{i=1}^{n} w_i(\boldsymbol{\beta}^*)^{\top} w_i(\boldsymbol{\beta}^*) \right] := 2f_{\varepsilon}(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*). \tag{1.6}$$

We note that the matrix above is inspired by the linearization of non-differentiable losses, and is in particular very different from the Hessian or the Jacobian matrix typically employed for inference. Throughout the text, we denote the inverse of $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$ as $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$, which is assumed to exist. In addition, we have $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) := n^{-1}\sum_{i=1}^{n} w_i(\boldsymbol{\beta})^{\top} w_i(\boldsymbol{\beta})$. To infer the parameter $\boldsymbol{\beta}^*$, we need to efficiently solve the SEE equation (1.4). We can observe that solving SEE equations (1.4) requires inverting the matrix $\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)$, as we are looking for a solution $\boldsymbol{\beta}$ that satisfies

$$\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)\boldsymbol{\beta} = \boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)\boldsymbol{\beta}^* + \mathbb{E}\Psi(\boldsymbol{\beta}^*).$$

For low-dimensional problems, with $p \ll n$, this can be done efficiently by considering an initial

estimate $\widehat{\boldsymbol{\beta}}$ and a sample plug-in estimate $\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\beta}})$ of $\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)$,

$$\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\beta}}) = 2n^{-1}\widehat{f}_{\varepsilon}(0)\sum_{i=1}^{n} w_i(\widehat{\boldsymbol{\beta}})^{\top} w_i(\widehat{\boldsymbol{\beta}}) \tag{1.7}$$

and a sample estimate of $\mathbb{E}\Psi(\boldsymbol{\beta}^*)$, denoted with $\Psi(\widehat{\boldsymbol{\beta}})$ and a suitable density estimate $\widehat{f}_{\varepsilon}(0)$. However, when $p \gg n$, this is highly inefficient. Instead, it is better to directly estimate $\boldsymbol{\Upsilon}^{-1}(\boldsymbol{\beta}^*) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)/2f_{\varepsilon}(0)$. Let $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ be an estimate of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$ (see Section 1.2.3 for discussion). Then, we proceed to solve SEE equations approximately, by defining the SEE estimator as

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\Psi(\widehat{\boldsymbol{\beta}})/2\widehat{f}_{\varepsilon}(0).$$

Remark 1. The proposed SEE can be viewed as a high-dimensional extension of inference from estimating equations. Although a left-censored linear model is considered, the proposed SEE methodology applies more broadly. For example, this framework includes loss functions based on ranks or non-convex loss functions for the fully observed data. For instance, the method in [VdGBR$^+$14] is based on inverting KKT conditions might not directly apply for the non-convex loss functions (e.g., Cauchy loss) or rank loss functions (e.g., log-rank loss). Recent methods of [NNLL15] do not apply to non-differentiable estimating equations (see Section 2.1 where a twice-differentiable assumption is imposed).

1.2.3 Estimation of the Scale in Left-Censored Models

The methodology for estimating each row of the matrix $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$ is introduced in this section. For further analysis, it is useful to define $W(\boldsymbol{\beta})$ as a matrix composed of row vectors $w_i(\boldsymbol{\beta})$; $W(\boldsymbol{\beta}) = A(\boldsymbol{\beta})X$, where $A(\boldsymbol{\beta}) = \operatorname{diag}(\mathbb{I}(X\boldsymbol{\beta} > 0)) \in \mathbb{R}^n \times \mathbb{R}^n$. The methodology is

motivated by the following observation:

$$\tau_j^{-2}\mathbf{\Gamma}_{(j)}(\boldsymbol{\beta}^*)^\top\mathbf{\Sigma}(\boldsymbol{\beta}^*)=\mathbf{e}_j,$$

where $\mathbf{\Gamma}_{(j)}(\boldsymbol{\beta}^*)=\left[-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_1,\cdots,-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_{j-1},1,-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_{j+1},\cdots,-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_p\right]$ and

$$\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}):=\underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\,\mathbb{E}\left\|W_j(\boldsymbol{\beta})-W_{-j}(\boldsymbol{\beta})\boldsymbol{\gamma}\right\|_2^2/n$$

as well as $\tau_j^2:=n^{-1}\mathbb{E}\left\|W_j(\boldsymbol{\beta}^*)-W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_2^2$. This motivates us to consider the following as an estimator for the inverse $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)$. Let $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$ and $\widehat{\tau}_j^2$ denote the estimators of $\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)$ and τ_j^2 respectively. We will show that a simple plug-in Lasso type estimator is sufficiently good for construction of confidence intervals. We propose to estimate $\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)$, with the following l_1 penalized plug-in least squares regression,

$$\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})=\underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\left\{n^{-1}\left\|W_j(\widehat{\boldsymbol{\beta}})-W_{-j}(\widehat{\boldsymbol{\beta}})\boldsymbol{\gamma}\right\|_2^2+2\lambda_j\|\boldsymbol{\gamma}\|_1\right\}. \tag{1.8}$$

Notice that this regression does not trivially share all the nice properties of the penalized least squares, as in this case the rows of the design matrix are not independent and identically distributed. An estimate of τ_j^2 can then be defined through the estimate of the residuals $\boldsymbol{\zeta}_j^*:=W_j(\boldsymbol{\beta}^*)-W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)$. Throughout this paper we assume that $\boldsymbol{\zeta}_j^*$ has sub-exponential distribution, and we denote $\|\mathbf{\Gamma}_{(j)}(\boldsymbol{\beta}^*)\|_0=s_j$ for $j=1,\cdots,p$, where $\|\cdot\|_0$ denotes the number of nonzero entries in the vector. We propose the plug-in estimate for $\boldsymbol{\zeta}_j^*$ as $\widehat{\boldsymbol{\zeta}}_j=W_j(\widehat{\boldsymbol{\beta}})-W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$, and a bias corrected estimate of τ_j^2 defined as

$$\widehat{\tau}_j^2(\lambda_j)=n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top\widehat{\boldsymbol{\zeta}}_j+\lambda_j\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_1. \tag{1.9}$$

Observe that the naive estimate $n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top \widehat{\boldsymbol{\zeta}}_j$ does not suffice due to the bias carried over by the penalized estimate $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$. Lastly, the matrix estimate of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$, much in the same spirit as [ZZ14] is defined with

$$\boldsymbol{\Omega}_{jj}(\widehat{\boldsymbol{\beta}}) = \widehat{\tau}_j^{-2}, \qquad \boldsymbol{\Omega}_{j,-j}(\widehat{\boldsymbol{\beta}}) = -\widehat{\tau}_j^{-2}\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}), \qquad j = 1,\ldots,p. \tag{1.10}$$

The proposed scale estimate can be considered as the censoring adaptive extension of the graphical lasso estimate of [VdGBR$^+$14].

1.2.4 Density Estimation

Whenever the model considered is homoscedastic, i.e., ε_i are identically distributed with a density function f_ε (denoted whenever possible with f), a novel density estimator designed to be adaptive to the left-censoring in the observations is used. For a positive bandwidth sequence $\widehat{h}_n$, we define the density estimator of $f_\varepsilon(0)$ as

$$\widehat{f}(0) = \widehat{h}_n^{-1} \sum_{i=1}^{n} \frac{\mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)\, \mathbb{I}(0 \le y_i - x_i\widehat{\boldsymbol{\beta}} \le \widehat{h}_n)}{\sum_{i=1}^n \mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)}. \tag{1.11}$$

Of course, more elaborate smoothing schemes for the estimation of $f(0)$ could be devised for this problem, but there seems to be no a priori reason to prefer an alternate estimator.

Remark 2. We will show that a choice of the bandwidth sequence satisfying

$$h_n^{-1} = o(\sqrt{n/(s\log p)})$$

suffices. However, we also propose an adaptive choice of the bandwidth sequence and consider

$\widehat{h}_n = \mathscr{o}(1)$, such that with $u_i := y_i - x_i\widehat{\boldsymbol{\beta}}$,

$$\widehat{h}_n = c\left\{s_{\widehat{\boldsymbol{\beta}}}\log p/n\right\}^{-1/3}\text{median}\left\{u_i : u_i > \sqrt{\log p/n},\ x_i\widehat{\boldsymbol{\beta}} > 0\right\},$$

for a constant $c > 0$. Here, $s_{\widehat{\boldsymbol{\beta}}}$ denotes the size of the estimated set of the non-zero elements of the initial estimator $\widehat{\boldsymbol{\beta}}$, i.e., $s_{\widehat{\boldsymbol{\beta}}} = \|\widehat{\boldsymbol{\beta}}\|_0$.

1.2.5 Confidence Intervals

Following the SEE principles, the solution to the equations is defined as an estimator,

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\Psi(\widehat{\boldsymbol{\beta}})/2\widehat{f}(0). \tag{1.12}$$

For the presentation of our coverage rates of the confidence interval (1.15) and (1.16), we start with the Bahadur representation. Lemmas 1 - 6 in Section 1.9 enable us to establish the following decomposition for the introduced one-step estimator $\widetilde{\boldsymbol{\beta}}$,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) + \Delta, \tag{1.13}$$

where the vector Δ represents the residual component. We show that the residual vector's size is small uniformly and that the leading term is asymptotically normal. The theoretical guarantees required from an initial estimator $\widehat{\boldsymbol{\beta}}$ is presented below.

Condition (I): *An initial estimate $\widehat{\boldsymbol{\beta}}$ is such that the following three properties hold. There exists a sequence of positive numbers r_n and d_n such that $r_n, d_n \to 0$ when $n \to \infty$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathscr{O}_P(r_n)$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathscr{O}_P(d_n)$ and $\|\widehat{\boldsymbol{\beta}}\|_0 = t = \mathscr{O}_P(s_{\boldsymbol{\beta}^*})$.*

One particular choice of such estimator can be l_1 penalized CLAD estimator studied in

[MvdG16]

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathscr{B}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|Y - \max\{0, X\boldsymbol{\beta}\}\|_1 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{1.14}$$

which satisfies the Condition **(I)** with $d_n = s_{\boldsymbol{\beta}^*}\sqrt{\log p/n}$, $r_n^2 = s_{\boldsymbol{\beta}^*} \log p/n$ and $\|\widehat{\boldsymbol{\beta}}\|_0 = \mathscr{O}_P(s_{\boldsymbol{\beta}^*} \times \lambda_{\max}(X^\top X)/n)$, under the suitable conditions. However, other choices are also allowed. It is worth noting that the above condition does not assume model selection consistency of the initial estimator and the methodology does not rely on having a unique solution to the problem (1.14); any local minima suffices as long as the prediction error is bounded accordingly.

With the normality result of the proposed estimator $\widetilde{\boldsymbol{\beta}}$ (as shown in Theorem 10, Section 1.7), we are now ready to present the confidence intervals. Fix α to be in the interval $(0,1)$, and let z_α denote the $(1-\alpha)$th standard normal percentile point. Let $\mathbf{c}$ be a fixed vector in $\mathbb{R}^p$. Based on the results of Section 1.7, the standard studentized approach leads to a $(1-2\alpha)100\%$ confidence interval for $\mathbf{c}^\top\boldsymbol{\beta}^*$ of the form

$$\mathrm{I}_n = \left(\mathbf{c}^\top\widetilde{\boldsymbol{\beta}} - a_n, \mathbf{c}^\top\widetilde{\boldsymbol{\beta}} + a_n \right), \tag{1.15}$$

where $\widetilde{\boldsymbol{\beta}}$ is defined in (1.12) and

$$a_n = z_\alpha \sqrt{\mathbf{c}^\top \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\mathbf{c}} \Big/ 2\sqrt{n}\widehat{f}(0) \tag{1.16}$$

with $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ as defined in (1.10), $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$ as defined in (1.7) and $\widehat{f}(0)$ as defined in (1.11). In the above, for $\mathbf{c} = \mathbf{e}_j$, the above confidence interval provides a coordinate-wise confidence interval for each β_j, $1 \le j \le p$. Notice that the above confidence interval is robust in a sense that it is asymptotically valid irrespective of the distribution of the error term ε.

1.3 High-dimensional Asymptotics

Within this section, we present the theoretical results using a specific initial estimator. However, the methodology has a much broader spectrum of applications. More details on the preliminary theoretical results, as well as more general results than the ones presented below, can be found in Section 1.7 in the later text. We begin with a set of very mild model error assumptions.

1.3.1 Theoretical Background

There has been considerable work in understanding the theoretical properties of high-dimensional one-step bias correction estimators. The convergence and consistency properties of least squares based methods have been studied by, among others, [BRT09], [MY09] and [NYWR09]. Meanwhile, their sampling variability has been analyzed by [VdGBR$^+$14]. However, to the best of our knowledge, Theorem 1 is the first result establishing conditions under which one-step estimators are asymptotically unbiased and normal in high-dimensional Tobit I models.

Probably the closest existing result is that of [BCK14] and [ZKL14], which showed that high-dimensional quantile models can be successfully de-biased for the purpose of confidence intervals construction. However, it is worth noting that their procedures do not adapt to censoring, and their de-biased methods cannot be applied to fixed, left-censored models. Observe that the optimal Hessian matrix we have developed depends on the level of censoring and an initial estimate, whereas procedures in the above mentioned work do not: the post-lasso estimation in [BCK14] relies on the score vector being a convex function of unknown parameters, and the Hessian matrix in [ZKL14] depends merely on features. However, under convexity condition, left-censored models cannot be solved non-parametrically (without knowing the density function of the model error). Of course a surrogate score vector may be developed, but then it remains unclear if efficient attainment of optimal bias-variance decomposition can be achieved. Although the methods of [BCK14] and [ZKL14] may appear qualitatively similar to the current work in the

common choice of LAD loss, they cannot be used for valid inference in left-censored models.

The non-smooth losses have been studied extensively by [BCK13] as well as [BCCW17] who showed that rates slower than that of smooth counterparts should be expected for many inferential problems; in particular rates are slower than those needed for estimation alone. However, it is important to note that in all approaches the de-biasing step consists of a non-smooth score and smooth variance estimate. In the current setting, however, we have non-smooth score as well as non-smooth Hessian matrix (treated as parameters of the unknown). We identify that such departure in structure of the problem requires new concentration of measure as well as contracting principles regarding indicator functions: a step not needed in the mentioned literature. Even in low dimensions, such results are of independent interest, as they provide a unique Bahadur representation for left-censored semi-parametric method. Instead of using projections for Hessian estimation, inference for Tobit models is usually performed in terms of bootstrap sampling. High-dimensional inference with bootstrap, however, have proven to be unreliable and inconsistent (unless done after bias correction step). As observed by [KP16], estimators resulting from direct bootstrap in high dimensions can exhibit surprising properties even in simple situations.

Finally, an interesting question for further theoretical study is to understand the optimal scaling of the sparsity for Tobit models. Size of the model sparsity can be treated as a robustness parameter. It would be of considerable interest to develop methods that adapt to the size of the model sparsity and achieve uniform rates of testing.

1.3.2 Main Results

Condition (E): *The error distribution F has median 0, and is everywhere continuously differentiable, with density f, which is bounded above, $f_{\max} < \infty$, and below, $f_{\min} > 0$. Furthermore, $f(\cdot)$ is also Lipschitz continuous, $|f(t_1) - f(t_2)| \leq L_0 \cdot |t_1 - t_2|$, for some $L_0 > 0$. Define function $G_i(z, \boldsymbol{\beta}, r) = \mathbb{E}[\mathbb{1}(|x_i\boldsymbol{\beta}| \leq \|x_i\| \cdot z)\|x_i\|^r]$. In addition, $G_i(z, \boldsymbol{\beta}, r) \leq K_1 \cdot z$, if $0 \leq z < \xi$, $r = 0, 1, 2,$*

for some positive K_1 and ξ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq \xi$.

We require the error density function to be with bounded first derivative. This excludes densities with unbounded first moment, but includes a class of distributions much larger than the Gaussian. Moreover, this assumption implies that $x_i\boldsymbol{\beta}$ are distributed much like the error ε_i, for $\boldsymbol{\beta}$ close to $\boldsymbol{\beta}^*$ and $x_i\boldsymbol{\beta}$ close to the censoring level 0. Last condition in particular implies that $\mathbb{P}(|x_i\boldsymbol{\beta}| \leq z) = o(z)$ for all $\boldsymbol{\beta}$ close to $\boldsymbol{\beta}^*$. This condition does not exclude deterministic components of the vector x_i, nor components which have discrete distributions; only the linear combination $x_i\boldsymbol{\beta}$ must have a Lipschitz continuous distribution function near zero. Therefore, implying $\mathbb{P}(|x_i\boldsymbol{\beta}^*| = 0) = 0$. For fixed designs, this condition implies $|x_i\boldsymbol{\beta}^*| \geq k_0$, for $k_0 > 0$.

Apart from the condition on the error distribution, we need conditions on the censoring level of the model (1.1) for further analysis.

Condition (C): *There exist constants $C_2 > 0$ and $\phi_0 > 0$, such that for all $\boldsymbol{\beta}$ satisfying $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^C_{\boldsymbol{\beta}^*}}\|_1 \leq 3\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S_{\boldsymbol{\beta}^*}}\|_1$, $\left\|\max\{0, X\boldsymbol{\beta}^*\} - \max\{0, X\boldsymbol{\beta}\}\right\|_2^2 \geq C_2\|X(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2$, and $n\phi_0^2\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S_{\boldsymbol{\beta}^*}}\|_1^2 \leq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \mathbb{E}[X^\top X](\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S_{\boldsymbol{\beta}^*}}$. Additionally, $v_n = \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\beta}^*))$ is also strictly positive, with $1/v_n = \mathscr{O}(1)$ and assume $\max_j \boldsymbol{\Sigma}_{jj}(\boldsymbol{\beta}^*) = \mathscr{O}(1)$.*

The censoring level c_i has a direct influence on the constant C_2. In general, higher values for c_i increase the number of censored data. The bounds for the coverage probability (see Theorem 1 and Theorem 6) do not depend on the censoring level c_i. The fact that the censoring level does not directly appear in the results should be understood in the sense that the percentage of the censored data is important, not the censoring level. Note that the compatibility factor ϕ_0 does not impose any restrictions on the censoring of the model, i.e., it is the same as the one introduced for linear models [BRT09]. Observe that this condition does not impose distribution of W to be Gaussian or continuous. However, it requires that $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$, the population covariance matrix, is at least invertible, a condition unavoidable even in linear models.

In order to establish theoretical results on the improved one-step estimator, we also need to control the scale estimator in the precision matrix estimation, which requires the following

condition. The condition is not uncommon, and can also be found in [VdGBR$^+$14, BCK14].

Condition (Γ): *Parameters* $\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)$ *for all* $j = 1,\ldots,p$ *are bounded and such that* $\left|\{k : \boldsymbol{\gamma}^*_{(j),k}(\boldsymbol{\beta}^*) \neq 0\}\right| \leq s_j$ *for some* $s_j \leq n$. *Function* $\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta})$ *is Lipschitz continuous for all* $\boldsymbol{\beta}$ *satisfying condition* **(C)**.

With the conditions above, we present our main result. More generalized results for initial estimators satisfying Condition **(I)** are presented in Theorem 10 and 11 in Section 1.7.

Theorem 1. *Let* $\widehat{\boldsymbol{\beta}}$ *be defined as in* (1.14) *with a choice of the tuning parameter*

$$\lambda = A_2 K\left(\sqrt{2\log(2p)/n} + \sqrt{\log p/n}\right)$$

for a constant $A_2 > 16$ *and independent of* n *and* p. *Assume that* $\bar{s}(\log p)^{1/2}/n^{1/4} = o(1)$, *for* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee s_\Omega$ *with* $s_\Omega = \max_j s_j$. *Suppose that conditions* **(E)**,**(C)** *and* **(Γ)** *hold. Moreover, let* $\lambda_j = C\sqrt{\log p/n}$ *for a constant* $C > 1$.

(i) Then, for $j = 1,\ldots,p$

$$\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1 = \mathcal{O}_P\left(\frac{1}{\phi_0^2 C_2} s_j\sqrt{\log p/n}\right). \tag{1.17}$$

(ii) For $j = 1,\ldots,p$ *and* $\boldsymbol{\zeta}^*$ *and* $\widehat{\boldsymbol{\zeta}}$

$$\left|\widehat{\boldsymbol{\zeta}}_j^\top \widehat{\boldsymbol{\zeta}}_j/n - \mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*/n\right| = \mathcal{O}_P\left(K^2 s_j\sqrt{\log(p\vee n)/n}\right).$$

(iii) Let $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ *defined in* (1.10). *Then, for* $\widehat{\tau}_j^2$ *as in* (1.9), *we have* $\widehat{\tau}_j^{-2} = \mathcal{O}_P(1)$. *Moreover,*

$$\left\|\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})_j - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_j\right\|_1 = \mathcal{O}_P\left(K^2 s_j^{3/2}\sqrt{\log(p\vee n)/n}\right).$$

(iv) Let $\widetilde{\boldsymbol{\beta}}$ *be defined as in* (1.12) *with* $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ *defined in* (1.10), $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$ *defined in* (1.7) *and* $\widehat{f}(0)$

as defined in (1.11). *Then, for* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee s_\Omega$ *with* $s_\Omega = \max_j s_j$, *the size of the residual term in* (1.13) *is*

$$\|\Delta\|_\infty = \mathscr{O}_P\left(\frac{\bar{s}^2 \log(p \vee n)}{n^{1/2}} \bigvee \frac{s_{\boldsymbol{\beta}^*}(\log(p \vee n))^{3/4}}{n^{1/4}}\right).$$

(v) Assume that $\bar{s}(\log p)^{3/4}/n^{1/4} = o(1)$, *for* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee s_\Omega$ *with* $s_\Omega = \max_j s_j$. *Let* I_n *and* a_n *be defined in* (1.15) *and* (1.16). *Then, for all vectors* $\mathbf{c} = \mathbf{e}_j$ *and any* $j \in \{1, \ldots, p\}$, *when* $\bar{s}, n, p \to \infty$ *we have*

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\mathbf{c}^\top \boldsymbol{\beta}^* \in I_n\right) = 1 - 2\alpha.$$

A few comments are in order. Part (i) of Theorem 1 implies that the proposed estimator and confidence intervals have distinct limiting behaviors with varying magnitude of the censoring level. In particular, (i) implies that $\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1$ inherits the rates available for fully observed linear models whenever C_2 is bounded away from zero. Additionally, if all data is censored, i.e., whenever C_2 converges to zero at a rate faster than λ_j, the estimation error will explode. These results agree with the asymptotic results on consistency in left-censored and low-dimensional models; however, they provide additional details through the exact rates of censoring that is allowed. For example, $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 < n^{-1/4}$ is sufficient for optimal inferential rates, and the asymptotic result above matches those of fully observed linear models. In this sense, our results are also efficient.

Part (ii) provides easy to verify sufficient conditions for the consistency of a class of semi-parametric estimators of the precision matrix for censored regression models. This result highlights specific rate of convergence (see Theorem 1 for more details). Part (iii) establishes properties of the graphical lasso estimate with data matrix that depends on $\widehat{\boldsymbol{\beta}}$. In comparison to linear models, the established rate is slower for a factor of s_j, whereas in comparison to the

results of Section 3 of [VdGBR+14] (see Theorem 3.2 therein), we avoid a strict condition of bounded parameter spaces.

Observe that Part (iv) is a special case of general theory presented in the Supplementary document. There we show that a large class of initial estimates suffices.

For the case of low-dimensional problems with $s = \mathscr{O}(1)$ and $p = \mathscr{O}(1)$, we observe that whenever the initial estimator of rate r_n, is in the order of $n^{-\varepsilon}$, for a small constant $\varepsilon > 0$, then

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = U + \Delta. \tag{1.18}$$

with

$$U = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)$$

and $\|\Delta\|_\infty = \mathscr{O}_P(n^{-2\varepsilon})$. In particular, for a consistent initial estimator, i.e. $r_n = \mathscr{O}(n^{-1/2})$ we obtain that $\|\Delta\|_\infty = \mathscr{O}_P(n^{-1/4})$.

For high-dimensional problems with s and p growing with n, for all initial estimators of the order r_n such that $r_n = \mathscr{O}(s^a_{\boldsymbol{\beta}^*}(\log p)^b/n^c)$ and $t = \mathscr{O}(s_{\boldsymbol{\beta}^*})$ we obtain that

$$\|\Delta\|_\infty = \mathscr{O}_P\left(\bar{s}^{(2a+3)/4}(\log p)^{(1+b)/2}/n^{c/2}\right)$$

whenever $\bar{s}(\log p)^{1/4}/n^{1/4} = \mathscr{O}(1)$, where $\bar{s} = s \vee s_\Omega$. Classical results on inference for left-censored data, with $p \ll n$, only imply that the error rates of the confidence interval is $\mathscr{o}_P(1)$; instead, we obtain a precise characterization of the residual term size.

Remark 3. *In particular, for the special case where the initial estimate is penalized CLAD estimate, we show*

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right]_{jj}^{-\frac{1}{2}} U_j \xrightarrow[n,p,\bar{s}\to\infty]{d} \mathscr{N}\left(0, \frac{1}{4f(0)^2}\right).$$

We obtain that the confidence interval I_n is asymptotically valid and that the coverage errors are of the order $\mathcal{O}\left(\bar{s}(\log p)^{3/4}/n^{1/4}\right)$, whenever $\bar{s}(\log p)^{1/4}/n^{1/4} = \mathcal{O}(1)$.

Moreover, with $p \ll n$ the rates above match the optimal rates of inference for the absolute deviation loss (see e.g. [ZP96]), indicating that our estimator is asymptotically efficient in the sense that the censoring asymptotically disappears even for $p \geq n$.

The condition $\bar{s}^4 \log^3 p \ll n$ is also similar to the results in [BCK13] obtained for $p \gg n$. While it is unclear the orthogonal moments approach therein is applicable for fixed-censored model, the rate condition required for quantile procedure is $s^3 \log^3(p) \ll n$, for known density and $s^4 \log^4(p) \ll n$, for unknown density (see Comment 3.3 and equation (ii) therein).

Lastly, observe that the result above is robust in the sense that it holds regardless of the particular distribution of the model error (1.1), and holds in a uniform sense. Thus, the confidence intervals are honest. In particular, the confidence interval I_n does not suffer from the problems arising from the non–uniqueness of $\boldsymbol{\beta}^*$ (see Theorem 11 in Section 1.7).

1.4 Left-censored Mallow's, Schweppe's and Hill-Ryan's One-step Estimators

Statistical models are seldom believed to be complete descriptions of how real data are generated; rather, the model is an approximation that is useful, if it captures essential features of the data. Good robust methods perform well, even if the data deviates from the theoretical distributional assumptions. The best known example of this behavior is the outlier resistance and transformation invariance of the median. Several authors have proposed one-step and k-step estimators to combine local and global stability, as well as a degree of efficiency under target linear model [Bic75]. There have been considerable challenges in developing good robust methods for more general problems.

We present here a family of robust generalized M-estimators (GM estimators) that stabilize estimation in the presence of "unusual" design or model error distributions. Observe that (1.1) rarely follows distribution with light tail. Namely, model (1.1) can be reparametrized as $y_i = z_i(\boldsymbol{\beta}^*)\boldsymbol{\beta}^* + \xi_i$, where $z_i(\boldsymbol{\beta}^*) = x_i \mathbb{I}\{x_i\boldsymbol{\beta}^* + \varepsilon_i \geq 0\}$ and $\xi_i = \varepsilon_i \mathbb{I}\{x_i\boldsymbol{\beta}^* + \varepsilon_i \geq 0\}$. Hence ξ_i will often have skewed distribution with heavier tails, and it is in this regard important to design estimators that are robust. We introduce Mallow's, Schweppe's and Hill-Ryan's estimators for left-censored models.

1.4.1 Smoothed Robust Estimating Equations (SREE)

In this section, we propose a robust generalized population estimating equations

$$\mathbb{E}[\Psi^r(\boldsymbol{\beta})] = 0 \tag{1.19}$$

with $\Psi^r = n^{-1}\sum_{i=1}^n \psi_i^r(\boldsymbol{\beta})$ and

$$\psi_i^r(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^n q_i w_i^\top(\boldsymbol{\beta})\ \psi\Big(v_i\big(y_i - \max\{0, x_i\boldsymbol{\beta}\}\big)\Big), \tag{1.20}$$

where ψ is an odd, nondecreasing and bounded function. Throughout we assume that the function ψ either has finitely many jumps, or is differentiable with bounded first derivative. Notice that when $q_i = 1$ and $v_i = 1$, with ψ being the sign function, we have $\psi_i^r = \psi_i$ of previous section. Moreover, observe that for the weight functions $q_i = q(x_i)$ and $v_i = v(x_i)$, both functions of $\mathbb{R}^p \to \mathbb{R}^+$, the true parameter vector $\boldsymbol{\beta}^*$ satisfies the robust population system of equations above. Appropriate weight functions q and v are chosen for particular efficiency considerations. Points with high leverage are considered "dangerous", and should be downweighted by the appropriate choice of the weights v_i. Additionally, if the design has "unusual" points, the weights q_i's serve to downweight their effects in the final estimator, hence making generalized M-estimators robust

to the outliers in the model error and the model design.

We augment the system (1.19) similarly as before, and consider the system of equations

$$\mathbb{E}[\Psi^r(\boldsymbol{\beta}^*)] + \boldsymbol{\Upsilon}^r[\boldsymbol{\beta}^* - \boldsymbol{\beta}] = 0, \tag{1.21}$$

for a suitable choice of the robust matrix $\boldsymbol{\Upsilon}^r \in \mathbb{R}^{p\times p}$. Ideally, most efficient estimation can be achieved, when the matrix $\boldsymbol{\Upsilon}^r$ is close to the matrix that linearizes the smoothed score function of the robust equations (1.19).

To avoid difficulties with non-smoothness of ψ, we propose to work with a matrix $\boldsymbol{\Upsilon}^r$ that is smooth enough and robust simultaneously. To that end, observe $\Psi^r(\boldsymbol{\beta}^*) = \Phi^r(\boldsymbol{\beta}^*, \varepsilon)$ for a suitable function $\Phi^r = n^{-1}\sum_{i=1}^n \phi_i^r$ and $\phi_i^r : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$. We consider a smoothed version of the Hessian matrix, and work with $\boldsymbol{\Upsilon}^r = \boldsymbol{\Upsilon}^r(\boldsymbol{\beta}^*)$ for

$$\boldsymbol{\Upsilon}^r(\boldsymbol{\beta}^*) = \mathbb{E}_X\left[\nabla_{\boldsymbol{\beta}^*}\int_{-\infty}^{\infty}\Phi^r(\boldsymbol{\beta}^*, \varepsilon) f_\varepsilon(x)dx\right],$$

where f_ε denotes the density of the model error (1.1). To infer the parameter $\boldsymbol{\beta}^*$, we adapt a one-step approach in solving the empirical counterpart of the population equations above. The empirical equations are named as *Smoothed Robust Estimating Equations* or SREE in short. For a preliminary estimate, we solve an approximation of the robust system of equations above, and search for the $\boldsymbol{\beta}$ that solves

$$\Psi^r(\widehat{\boldsymbol{\beta}}) + \boldsymbol{\Upsilon}^r(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 0.$$

The particular form of the matrix $\boldsymbol{\Upsilon}^r(\boldsymbol{\beta}^*)$ depends on the choice of the weight functions q and v and the function ψ. In particular, for the left-censored model (1.1),

$$\nabla_{\boldsymbol{\beta}^*}\mathbb{E}_\varepsilon[\Psi^r(\boldsymbol{\beta}^*)] = n^{-1}\sum_{i=1}^n q_i \nabla_{\boldsymbol{\beta}^*}\mathbb{E}_\varepsilon\left[\psi\left(v_i(y_i - \max\{0, x_i\boldsymbol{\beta}^*\})\right)\right], \tag{1.22}$$

leads to the following form

$$\boldsymbol{\Upsilon}^r(\boldsymbol{\beta}^*) = \mathbb{E}_X\left[n^{-1}\sum_{i=1}^{n} q_i v_i \psi'(v_i r_i(\boldsymbol{\beta}^*)) x_i^\top w_i(\boldsymbol{\beta}^*)\right],$$

whenever the function ψ is differentiable. We denote $\psi'(v_i r_i(\boldsymbol{\beta})) := \partial\psi(v_i r_i(\boldsymbol{\beta}))/\partial\boldsymbol{\beta}$, where $r_i(\boldsymbol{\beta}) := y_i - \max\{0, x_i\boldsymbol{\beta}\}$). In case of non-smooth ψ, ψ' should be interpreted as $g' = \partial g/\partial\boldsymbol{\beta}$, for $g(\boldsymbol{\beta}) = \mathbb{E}_\varepsilon[\psi(v_i r_i(\boldsymbol{\beta}))]$. For example, if $\psi(\cdot) = \text{sign}(\cdot)$, then $g(\boldsymbol{\beta})$ is equal to $1 - 2P(r_i(\boldsymbol{\beta}) \leq 0)$ and $g'(\boldsymbol{\beta}^*) = 2f_{\varepsilon_i}(0)\,\mathbb{I}(x_i\boldsymbol{\beta}^* > 0)$.

1.4.2 Left-censored Mallow's, Hill-Ryan's and Schweppe's Estimator

Here we provide specific definitions of new robust one-step estimates. We begin by defining a robust estimate of the precision matrix, i.e., $\{\boldsymbol{\Upsilon}^r\}^{-1}(\boldsymbol{\beta}^*)$. We design a robust estimator that preserves the "downweight" functions q and v as to stabilize the estimation in the presence of contaminated observations. For further analysis, it is useful to define the matrix $\widetilde{W}(\boldsymbol{\beta}) = Q^{1/2}W(\boldsymbol{\beta})$ and

$$Q = \text{diag}(\mathbf{q} \circ \mathbf{d}) \in \mathbb{R}^{n\times n},$$

where $\circ$ denotes entry-wise multiplication, also known as the Hadamard product, with $\mathbf{q} = [q(x_1), q(x_2), \cdots, q(x_n)]^\top \in \mathbb{R}^n$ and

$$\mathbf{d} = \left[\psi'(v_1 r_1(\boldsymbol{\beta}^*)), \quad \psi'(v_2 r_2(\boldsymbol{\beta}^*)), \quad \cdots, \quad \psi'(v_n r_n(\boldsymbol{\beta}^*))\right]^\top \in \mathbb{R}^n$$

for $r_i(\boldsymbol{\beta}^*) = y_i - \max\{0, x_i\boldsymbol{\beta}^*\}$. When function ψ does not have first derivative, we replace $\psi'(v_i r_i(\boldsymbol{\beta}^*))$ with $n^{-1}\sum_{i=1}^n [\mathbb{E}\psi(v_i r_i(\boldsymbol{\beta}^*))]'$. With this notation, we have

$$\widetilde{W}_j(\boldsymbol{\beta}^*) = Q^{1/2}A(\boldsymbol{\beta}^*)X_j,$$

and $\boldsymbol{\Upsilon}^r(\boldsymbol{\beta}^*) = n^{-1}\mathbb{E}\left[\widetilde{W}(\boldsymbol{\beta}^*)^\top \widetilde{W}(\boldsymbol{\beta}^*)\right]$ takes the form of a weighted covariance matrix. Hence, to estimate the inverse $\{\boldsymbol{\Upsilon}^r\}^{-1}(\boldsymbol{\beta}^*)$, we project columns onto the space spanned by the remaining columns. For $j = 1, \ldots, p$, we define the vector $\widetilde{\boldsymbol{\theta}}_{(j)}(\boldsymbol{\beta})$ as follows,

$$\widetilde{\boldsymbol{\theta}}_{(j)}(\boldsymbol{\beta}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\, \mathbb{E}\left\|\widetilde{W}_j(\boldsymbol{\beta}) - \widetilde{W}_{-j}(\boldsymbol{\beta})\boldsymbol{\theta}\right\|_2^2 / n. \tag{1.23}$$

Also, we assume the vector $\widetilde{\boldsymbol{\theta}}_{(j)}(\boldsymbol{\beta}^*)$ is sparse with $\widetilde{s}_j := \|\widetilde{\boldsymbol{\theta}}_{(j)}(\boldsymbol{\beta}^*)\|_0 \leq s_\Omega$. Thus, we propose the following as a robust estimate of the scale

$$\widetilde{\boldsymbol{\Omega}}_{jj}(\widehat{\boldsymbol{\beta}}) = \widetilde{\mathscr{I}}_j^{-2}, \qquad \widetilde{\boldsymbol{\Omega}}_{j,-j}(\widehat{\boldsymbol{\beta}}) = -\widetilde{\mathscr{I}}_j^{-2}\widetilde{\boldsymbol{\theta}}_{(j)}(\widehat{\boldsymbol{\beta}}), \tag{1.24}$$

with

$$\widetilde{\boldsymbol{\theta}}_{(j)}(\widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ n^{-1}\left\|\widetilde{W}_j(\widehat{\boldsymbol{\beta}}) - \widetilde{W}_{-j}(\widehat{\boldsymbol{\beta}})\boldsymbol{\theta}\right\|_2^2 + 2\lambda_j\|\boldsymbol{\theta}\|_1 \right\},$$

and the normalizing factor

$$\widetilde{\mathscr{I}}_j^2 = n^{-1}\left\|\widetilde{W}_j(\widehat{\boldsymbol{\beta}}) - \widetilde{W}_{-j}(\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{\theta}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_2^2 + \lambda_j\|\widetilde{\boldsymbol{\theta}}_{(j)}(\widehat{\boldsymbol{\beta}})\|_1.$$

Remark 4. Estimator (1.24) is a high-dimensional extension of Hampel's ideas of approximating the inverse of the Hessian matrix in a robust way, by allowing data specific weights to trim down the effects of the outliers. Such weights can be stabilizing estimation in the presence of high proportion of censoring.[Hil77] compared the efficiency of the Mallow's and Schweppe's estimators to several others and found that they dominate in the case of linear models in low-dimensions.

Lastly, we arrive at a class of robust one-step generalized M-estimators,

$$\check{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) \left(n^{-1} \sum_{i=1}^{n} q_i w_i^{\top}(\widehat{\boldsymbol{\beta}}) \; \psi\Big(v_i \big(y_i - \max\{0, x_i \widehat{\boldsymbol{\beta}}\} \big) \Big) \right). \tag{1.25}$$

We propose a one-step left-censored Mallow's estimator for left-censored high-dimensional regression by setting the weights to be $v_i = 1$, and

$$q_i = \min \left\{ 1, b^{\alpha/2} \left(\Big(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \Big) \boldsymbol{\Omega}_{\widehat{S},\widehat{S}}(\widehat{\boldsymbol{\beta}}) \Big(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \Big)^{\top} \right)^{-\alpha/2} \right\},$$

for constants $b > 0$ and $\alpha \geq 1$, with

$$\bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}})$$

and $\widehat{S} = \{ j : \widehat{\boldsymbol{\beta}}_j \neq 0 \}$. Extending the work of [CH93], it is easy to see that Mallow's one-step estimator with $\alpha = 1$ and $b = \chi^2_{\widehat{s},0.95}$ quantile of chi-squared distribution with $\widehat{s} = |\widehat{S}|$ improves a breakdown point of the initial estimator to nearly 0.5, by providing local stability of the precision matrix estimate.

Similarly, the one-step left-censored Hill-Ryan estimator is defined with

$$v_i = q_i = 1 / \left\| \boldsymbol{\Omega}_{\widehat{S},\widehat{S}}(\widehat{\boldsymbol{\beta}}) (w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}})) \right\|_2, \tag{1.26}$$

and the one-step left-censored Schweppe's estimator with the same q_i as the left hand side of (1.26), but $v_i = 1/q_i$. Note that these are not the only choices of Hill-Ryan and Schweppe's type estimators.

Another family of one-step estimators defined for Tobit-I models, for which we can use the framework above, is the class of adaptive Huber's one-step estimators, where $v_i = 1$ and $q_i = 1$, and the function ψ takes the form of a first order derivative of a Huber loss function.

However, it is unclear what the benefit of such loss would be for left-censored data, as the nice convexity property of traditional least squares is no longer available regardless.

The purpose of this paper is to explore the behavior of the different types of one-step estimators for left-censored regression model through studying their higher order asymptotic properties. This provides a unified synthesis of results as well as new results and insights. We will show that the effect of the initial estimate persists asymptotically, only if it is of least squares type. We also show that the one-step robust estimate has fast convergence rates, and leads to a class of robust confidence intervals and tests.

1.4.3 Theoretical Results

Similar to the concise version of Bahadur representation presented in (1.13) for the standard one-step estimator with $q_i = 1$ and $v_i = 1$, we also have the expression for robust generalized M-estimator,

$$\sqrt{n}\left(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = U^{\mathrm{r}} + \Delta^{\mathrm{r}}, \tag{1.27}$$

but now with the leading term of a different form

$$U^{\mathrm{r}} = \frac{1}{2f(0)}\{\boldsymbol{\Sigma}^{\mathrm{r}}\}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \psi\left(v_i\left(y_i - \max\{0, x_i\boldsymbol{\beta}^*\}\right)\right)(w_i(\boldsymbol{\beta}^*))^{\top}.$$

Next, we show that the leading component has asymptotically normal distribution, and that the residual term is of smaller order. To facilitate presentation, we present results below with an initial estimator being penalized CLAD estimator (1.14) with the choice of tuning parameter as presented in Theorem 1. We introduce the following condition.

Condition (rΓ): *Parameters $\boldsymbol{\theta}^*_{(j)}(\boldsymbol{\beta}^*)$ for all $j = 1, \ldots, p$ are bounded and such that $\left|\{k : \boldsymbol{\theta}^*_{(j),k}(\boldsymbol{\beta}^*) \neq 0\}\right| \leq \widetilde{s}_j$ for some $s_j \leq n$. Function $\boldsymbol{\theta}^*_{(j)}(\boldsymbol{\beta})$ is Lipschitz continuous for all*

$\boldsymbol{\beta}$ *satisfying condition* **(C)**. *In addition, let* q_i *and* v_i *be functions such that* $\max_i |q_i| \leq M_1$ *and* $\max_i |v_i| \leq M_2$ *for positive constants* M_1 *and* M_2 *and* $\mathbb{E}[\psi(\varepsilon_i v_i)] = 0$. *Moreover, let* ψ *be such that* $\psi(z) < \infty$ *and* $0 < \psi'(z) < \infty$.

We will show that for the proposed set of weight functions, the above condition holds. Boundedness of the function ψ' allows for error distributions with unbounded moments, and provides necessary robustness to the possible outliers in the model error. For the leading term of the Bahadur representation (1.27), we obtain the following result.

Theorem 2. *Assume that* $\bar{s}\log^{1/2}(p)/n^{1/4} = o(1)$, *with* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee \widetilde{s}_\Omega$ *and* $\widetilde{s}_\Omega = \max_j \widetilde{s}_j$. *Let Conditions* **(C)**, **(rΓ)** *and* **(E)** *hold and let* $\lambda_j = C\sqrt{\log p/n}$ *for a constant* $C > 1$. *Then,*

$$\left[\widetilde{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Upsilon}^r}(\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\right]_{jj}^{-\frac{1}{2}} U_j^r \xrightarrow[n,p,s_{\boldsymbol{\beta}^*}\to\infty]{d} \mathcal{N}(0,1).$$

For the residual term of the decomposition (1.27) we have the following statement.

Theorem 3. *Let Conditions* **(C)**, **(rΓ)** *and* **(E)** *hold and let* $\lambda_j = C\sqrt{\log p/n}$ *for a constant* $C > 1$. *Assume that* $\bar{s}\log^{1/2}(p)/n^{1/4} = o(1)$, *for* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee \widetilde{s}_\Omega$ *with* $\widetilde{s}_\Omega = \max_j \widetilde{s}_j$. *Then,*

$$\|\Delta^r\|_\infty = \mathcal{O}_P\left(\frac{\bar{s}^2\log(p\vee n)}{n^{1/2}} \bigvee \frac{s_{\boldsymbol{\beta}^*}(\log(p\vee n))^{3/4}}{n^{1/4}}\right).$$

Remark 5. The estimation procedure described above is based on the initial estimator $\widehat{\boldsymbol{\beta}}$ taken to be penalized CLAD. However, it is possible to show that a large family of sparsity encouraging estimator suffices. In particular, suppose that the initial estimator $\bar{\boldsymbol{\beta}}$ is such that $\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \gamma_n$, and let for simplicity $s_{\boldsymbol{\beta}^*} = s$. Then results of Theorem 3 extend to hold for the confidence interval defined as $\bar{I}_n = (\mathbf{c}^\top\widetilde{\boldsymbol{\beta}} - a_n, \mathbf{c}^\top\widetilde{\boldsymbol{\beta}} + a_n)$ with a_n as in (1.29). In particular, the error rates are of the order of

$$(\gamma_n^{1/2}t^{1/4} \vee \gamma_n t^{1/2})t^{1/2}(\log p)^{1/2} + \sqrt{n}s\widetilde{s}_\Omega^{3/2}\lambda_j\gamma_n^2 + \sqrt{n}\widetilde{s}_\Omega^{3/2}\lambda_j\gamma_n.$$

When $s = \mathcal{O}(1)$ and $s_j = \mathcal{O}(1)$, and all $\sqrt{n}\lambda_j = \mathcal{O}(1)$, previous result implies that the initial

estimator needs only to converge at a rate of $\mathscr{o}(n^{-\varepsilon})$ for a small $\varepsilon > 0$.

With the results above, we can now construct a $(1-2\alpha)100\%$ confidence interval for $\mathbf{c}^\top \boldsymbol{\beta}$ of the form

$$\mathrm{I}_n^{\mathrm{r}} = \left(\mathbf{c}^\top \breve{\boldsymbol{\beta}} - \breve{a}_n, \mathbf{c}^\top \breve{\boldsymbol{\beta}} + \breve{a}_n \right), \tag{1.28}$$

where $\breve{\boldsymbol{\beta}}$ is defined in (1.25), $\mathbf{c} = \mathbf{e}_j$ for some $j \in \{1, 2, \ldots, p\}$,

$$\breve{a}_n = z_\alpha \sqrt{\mathbf{c}^\top \widetilde{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Upsilon}^r}(\widehat{\boldsymbol{\beta}}) \widetilde{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) \mathbf{c}} \Big/ \sqrt{n}, \tag{1.29}$$

with the robust covariance estimate that we define as

$$\widehat{\boldsymbol{\Upsilon}^r}(\widehat{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} q_i v_i \psi'(v_i(y_i - x_i^\top \widehat{\boldsymbol{\beta}})) x_i^\top w_i(\widehat{\boldsymbol{\beta}}).$$

Remark 6. Constants M_1 and M_2 change with a choice of the robust estimator. For the Mallow's and Hill-Ryan's, by Lemma 5 in Section 1.7,

$$\left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right)^\top \boldsymbol{\Omega}_{\widehat{S},\widehat{S}}(\widehat{\boldsymbol{\beta}}) \left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right) > C \left\| w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right\|_2^2 \geq 0.$$

Thus, the coverage probability of Mallow's and Hill-Ryan's estimator is the same as that of the M-estimator. However, the coverage of the Schweppe's estimator is slightly slower, as result of Lemma 1 and Lemma 5 in Section 1.7 imply

$$\begin{aligned}
&\left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right)^\top \boldsymbol{\Omega}_{\widehat{S},\widehat{S}}(\widehat{\boldsymbol{\beta}}) \left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right) \\
&\quad \leq \left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \left(w_{i,\widehat{S}}(\widehat{\boldsymbol{\beta}}) - \bar{w}_{\widehat{S}}(\widehat{\boldsymbol{\beta}}) \right) + \mathscr{o}_P(1) \\
&\quad \leq \left\| x_{i,\widehat{S}} \right\|_2^2 / \lambda_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{\beta}^*) \right) = \mathscr{O}_P(s_{\boldsymbol{\beta}^*}).
\end{aligned}$$

Together with Theorem 6 in Section 1.7, we observe now a rate that is slower by a factor of $s_{\boldsymbol{\beta}^*}$,

i.e., the leading term is of the order of $\mathscr{O}\left(s^2_{\boldsymbol{\beta}^*}(\log(p\vee n))^{3/4}n^{-1/4}\right)$.

Theorem 4. *Under Conditions of Theorems 2 and 3, we have for Mallow's and Hill-Ryan's estimator*

$$\|\Delta^r\|_\infty = \mathscr{O}_P\left(\frac{s_{\boldsymbol{\beta}^*}(\log(p\vee n))^{3/4}}{n^{1/4}}\bigvee\frac{\bar{s}^2\log(p\vee n)}{n^{1/2}}\right),$$

whereas for the Schweppe's estimator

$$\|\Delta^r\|_\infty = \mathscr{O}_P\left(\frac{s^2_{\boldsymbol{\beta}^*}(\log(p\vee n))^{3/4}}{n^{1/4}}\bigvee\frac{\bar{s}^3\log(p\vee n)}{n^{1/2}}\right).$$

Remark 7. This result implies that the residual term sizes depend on the type of weight functions chosen. Due to the particular left-censoring, the ideal weights measuring concentration in the error or design depend on the unknown censoring. Hence, we approximate ideal weights with plug-in estimators, and therefore obtain rates of convergence that are slightly slower than those of non-robust estimators. This implies that the robust confidence intervals require larger sample size to achieve the nominal level.

Corollary 5. *Under Conditions of Theorem 2 and 3, for all vectors* $\mathbf{c} = \mathbf{e}_j$ *and any* $j \in \{1,\ldots,p\}$*, when* $\bar{s}, n, p \to \infty$ *and all* $\alpha \in (0,1)$ *we have that (i) whenever the interval is constructed using Mallow's or Hill-Ryan's estimator and* $\bar{s}(\log(p\vee n))^{3/4}/n^{1/4} = o(1)$*, the respective confidence intervals have asymptotic coverage* $1-\alpha$*; (ii) whenever the interval is constructed using Schweppe's estimator and* $\bar{s}^2(\log(p\vee n))^{3/4}/n^{1/4} = o(1)$*, the respective confidence intervals have asymptotic coverage of* $1-\alpha$*.*

1.5 Numerical Results

In this section, we present a number of numerical experiments from both high-dimensional, $p \gg n$, and low-dimensional, $p \ll n$, simulated settings.

We implemented the proposed estimator in a number of different model settings. Specifically, we vary the following parameters of the model. The number of observations, n, is taken to be 300, while p, the number of parameters, is taken to be 40 or 400. The error of the model, ε, is generated from a number of distributions including: standard normal, Student's t with 4 degrees of freedom, Beta distribution with parameters $(2,3)$ and Weibull distribution with parameters $(1/2,1/5)$. In the case of the non-zero mean distributions, we center the observations before generating the model data. The parameter $s_{\boldsymbol{\beta}^*}$, the sparsity of $\boldsymbol{\beta}^*$, $\#\{j : \boldsymbol{\beta}^*_j \neq 0\}$, is taken to be 3, with all signal parameters taken to be 1 and located as the first three coordinates. The $n \times p$ design matrix, X, is generated from a multivariate Normal distribution $\mathcal{N}(\mu, \boldsymbol{\Sigma})$. The mean μ is chosen to be vector of zero, and the censoring level c is chosen to fix censoring proportion at 25%. The covariance matrix, $\boldsymbol{\Sigma}$, of the distribution that X follows, is taken to be the identity matrix or the Toeplitz matrix such that $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$ for $\rho = 0.4$. In each case, we generated 100 samples from one of the settings described above and for each sample we calculated the 95% confidence interval. The complete algorithm is described in Steps 1-4 below. We note that the optimization problem required to obtain the penalized CLAD estimator is not convex. Nevertheless, it is possible to write (1.14) as linear program within the compact set $\mathscr{B}$, and solve accordingly [Pow84],

$$\begin{aligned}
&\underset{\substack{\boldsymbol{\beta}\in\mathscr{B}\\ \mathbf{u}^+,\mathbf{u}^-\geq 0\\ \mathbf{v}^+,\mathbf{v}^-\geq 0\\ \boldsymbol{\beta}^+,\boldsymbol{\beta}^-\geq 0}}{\text{minimize}} && \left\{ n^{-1}\sum_{i=1}^{n}\left(\mathbf{u}_i^+ + \mathbf{u}_i^-\right) + \lambda \sum_{j=1}^{p}\left(\boldsymbol{\beta}_j^+ + \boldsymbol{\beta}_j^-\right)\right\} \\
&\text{subject to} && \mathbf{u}_i^+ - \mathbf{u}_i^- = y_i - \mathbf{v}_i^+, \text{ for } 1 \leq i \leq n \\
& && \mathbf{v}_i^+ - \mathbf{v}_i^- = \sum_{j=1}^{p} X_{ij}\left(\boldsymbol{\beta}_j^+ - \boldsymbol{\beta}_j^-\right), \text{ for } 1 \leq i \leq n.
\end{aligned}$$

In addition, as our theory indicates, we allow for any initial estimator with desired convergence rate. Penalized CLAD is one example thereof.

1. The penalization factor λ is chosen by the one-standard deviation rule of the cross validation, $\widehat{\lambda} = \arg\min_{\lambda \in \{\lambda^1, \ldots, \lambda^m\}} \mathrm{CV}(\lambda)$. We move λ in the direction of decreasing regularization until it ceases to be true that $\mathrm{CV}(\lambda) \leq \mathrm{CV}(\widehat{\lambda}) + \mathrm{SE}(\widehat{\lambda})$. Standard error for the cross-validation curve, $\mathrm{SE}(\widehat{\lambda})$, is defined as a sample standard error of the K fold cross-validation statistics $\mathrm{CV}_1(\lambda), \ldots, \mathrm{CV}_K(\lambda)$. They are calibrated using the censored LAD loss as

$$\mathrm{CV}_k(\lambda) = n_k^{-1} \sum_{i \in F_k} \left| y_i - \max\{0, x_i \widehat{\boldsymbol{\beta}}^{-k}(\lambda)\} \right|,$$

   with $\widehat{\boldsymbol{\beta}}^{-k}(\lambda)$ denoting the CLAD estimator computed on all but the k-th fold of the data.

2. The tuning parameter λ_j in each penalized l_2 regression, is chosen by the one standard deviation rule (as described above). In more details, λ_j is in the direction of decreasing regularization until it ceases to be true that $\mathrm{CV}^j(\lambda_j) \leq \mathrm{CV}^j(\widehat{\lambda}_j) + \mathrm{SE}^j(\widehat{\lambda}_j)$ for $\widehat{\lambda}_j$ as the cross-validation parameter value. The cross-validation statistic is here defined as

$$\mathrm{CV}_k^j(\lambda) = n_k^{-1} \sum_{i \in F_k} \left(W_{ij}(\widehat{\boldsymbol{\beta}}) - W_{ij}(\widehat{\boldsymbol{\beta}}) \widehat{\gamma}_{(j)}^{-k}(\lambda_j) \right)^2,$$

   with $\widehat{\gamma}_j^{-k}(\lambda_j)$ denoting estimators (1.8) computed on all but the k-th fold of the data. This choice leads to the conservative confidence intervals with wider than the optimal length. Theoretically guided optimal choice is highly complicated and depends on both design distribution and censoring level concurrently. Nevertheless, we show that one-standard deviation choice is very reasonable.

3. Whenever the density of the error term is unknown, we estimate $f(0)$, using the proposed estimator (1.11), with a constant $c = 10$. We compute the above estimator by splitting the

sample into two parts: the first sample is used for computing $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ and the other sample is to compute the estimate $\widehat{f}(0)$. Optimal value of h is of special independent interest; however, it is not the main objective of this work.

4. Obtain $\widetilde{\boldsymbol{\beta}}$ by plugging $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ and $\widehat{f}(0)$ into (1.12) with λ and λ_j as specified in the steps above.

The summary of the results is presented across dimensionality of the parameter vector. The *Low-Dimensional Regime with SEE Estimator* are summarized in Table 1.1 and Figures 1.1 and 1.2. The *High-Dimensional Regime* are summarized in Table 1.3 and Figures 1.5 and 1.6. We report average coverage probability across the signal and noise variables independently, as the signal variables are more difficult to cover when compared to the noise variables.

We consider a number of challenging settings. Specifically, the censoring proportion is kept relatively high at 25%, and our parameter space is large with $p = 400$ and $n = 300$. In addition, we consider the case of error distribution being Student with 4 degrees of freedom, which is notoriously difficult to deal with in left-censored problems. For the four error distributions, the observed coverage probabilities are approximately the same.

We also note that symmetric distributions are very difficult to handle in left-censored models. However, when errors were symmetric (Normal), the coverage probabilities were extremely close to the nominal ones. The simulation cases evidently show that our method is robust to asymmetric distributions and does not lose efficiency when the errors are symmetric.

Lastly, to investigate smoothed robust estimating equations (SREE) empirically, we preserve the previous high-dimensional settings with standard normal and Student's t_4 error distributions respectively. However, to illustrate the robustness of the estimator, we artificially create outliers in the design matrix X, and perform Mallow's type SREE estimating procedures with the perturbed $\widetilde{X}$. Within each iteration, after generating X from $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ accordingly, we randomly select 10% of the columns, and then randomly perturb 10% of the entries in X by adding

Table 1.1: Coverage Probability for Low-Dimensional Regime with Smoothed Estimating Equations (SEE) Estimator

Distribution of the error term	Simulation Setting			
	Toeplitz design		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.97	0.98	0.95	0.94
Student	0.97	1	0.97	0.98
Beta	0.94	1	0.98	0.97
Weibull	0.98	0.98	0.94	0.98

Table 1.2: Coverage Probability for Low-Dimensional Regime with Powell Estimator as in [Pow84]

Distribution of the error term	Simulation Setting			
	Toeplitz design		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.96	0.97	0.95	0.98
Student	0.95	0.98	0.96	0.96
Beta	0.94	0.99	0.91	0.99
Weibull	0.99	0.99	0.91	0.98

twice the quantity of the maximum entry in X, i.e. $\widetilde{X}_{ij} = X_{ij} + 2 \times \max_{ij} X_{ij}$. Such perturbations create a considerate proportion of outliers in the design. The results are summarized in Table 1.4 and Figures 1.7 and 1.8. As coverages under various scenarios are close to the nominal level, the results show that the SREE estimator is robust to high leverage points.

1.6 Discussion and Conclusion

SEE and SREE frameworks enrich regular high-dimensional inferential methods with censoring and robust options. While a censoring option adds to the capacity of an existing inferential methods extending them to non-convex problems in general, a robust option has the potential to open a new direction. Usually, inferential methods have been aiming to create efficient methods with asymptotically exact or pivotal properties in a class of specific models. However,

Table 1.3: Coverage Probability for High-Dimensional Regime with Smoothed Estimating Equations (SEE) Estimator

Distribution of the error term	Simulation Setting			
	Toeplitz design		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.92	0.96	0.97	0.95
Student	0.96	0.98	0.96	0.98
Beta	1	1	0.96	0.97
Weibull	0.95	1	0.87	0.97

Table 1.4: Coverage Probability for High-Dimensional Regime with Smoothed Robust Estimating Equations (SREE) estimator

Distribution of the error term	Simulation Setting			
	Toeplitz design		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.89	0.99	0.90	0.97
Student	0.92	0.96	0.90	0.99

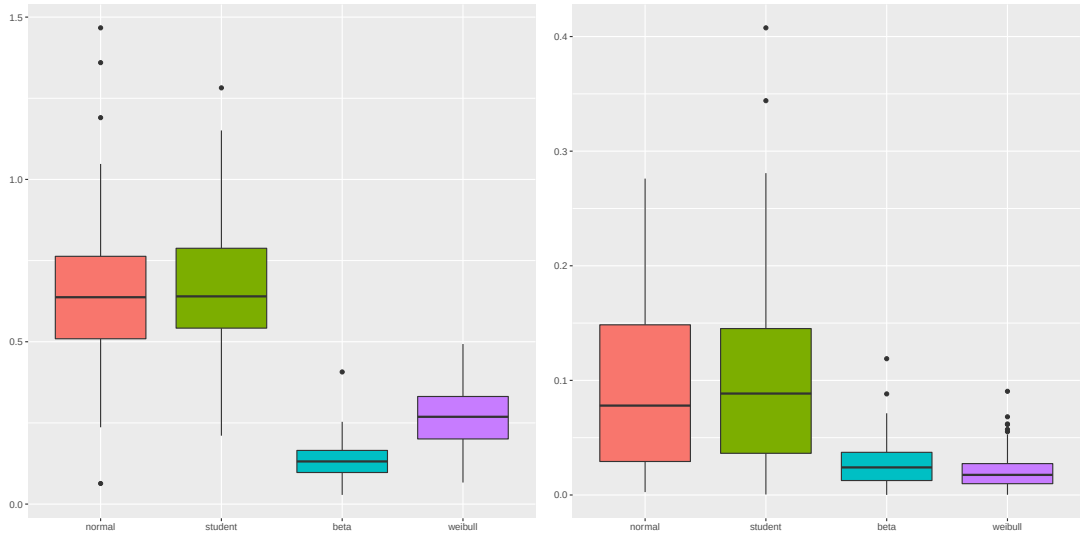

Figure 1.1: SEE estimator $p \ll n$ and Toeplitz Design with $\rho = 0.4$. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

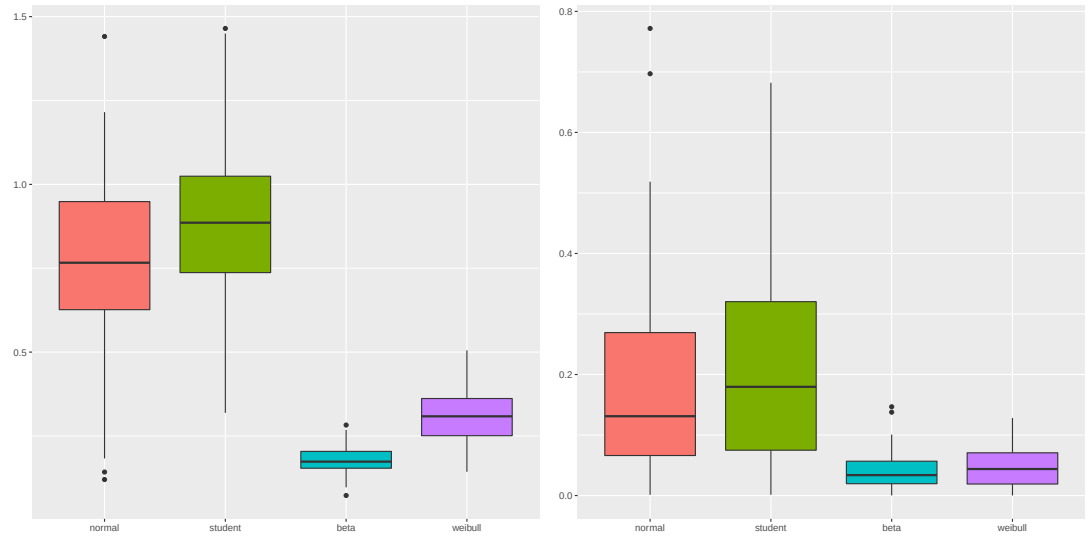

Figure 1.2: SEE estimator $p \ll n$ and Identity Design. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

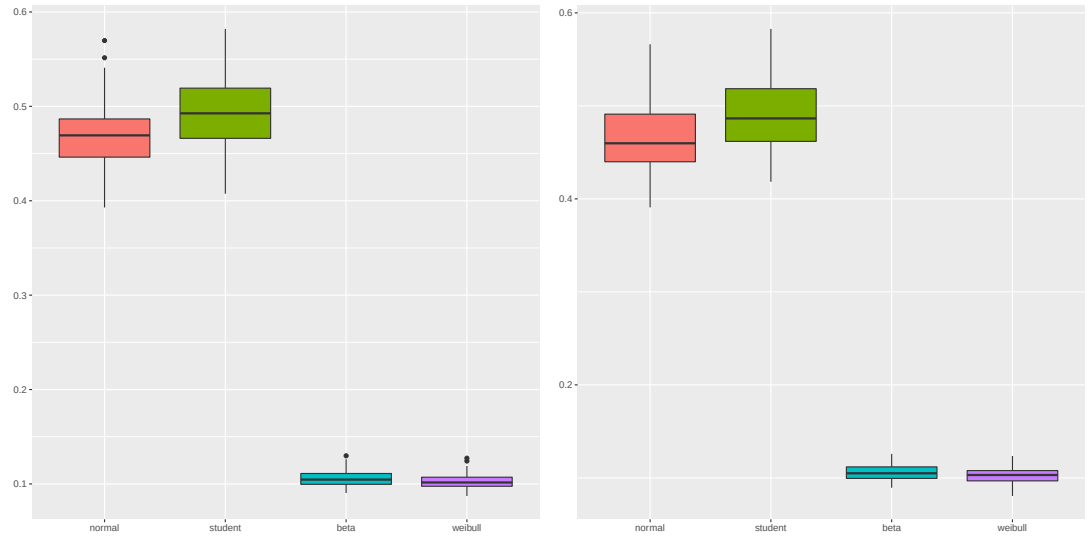

Figure 1.3: Powell estimator under $p \ll n$ and Toeplitz Design with $\rho = 0.4$. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

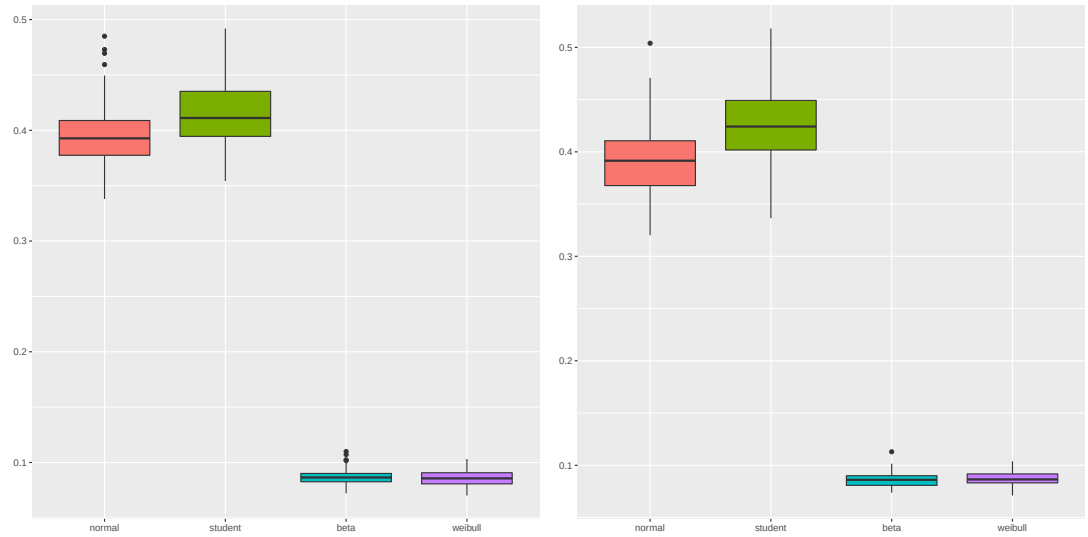

Figure 1.4: Powell estimator under $p \ll n$ and Identity Design. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

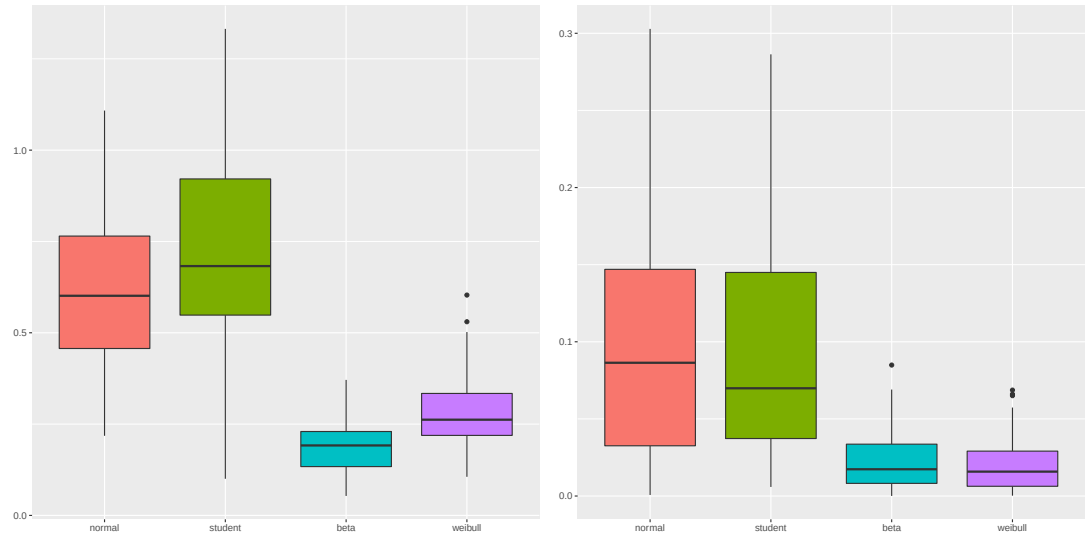

Figure 1.5: SEE estimator $p \gg n$ and Toeplitz Design with $\rho = 0.4$. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

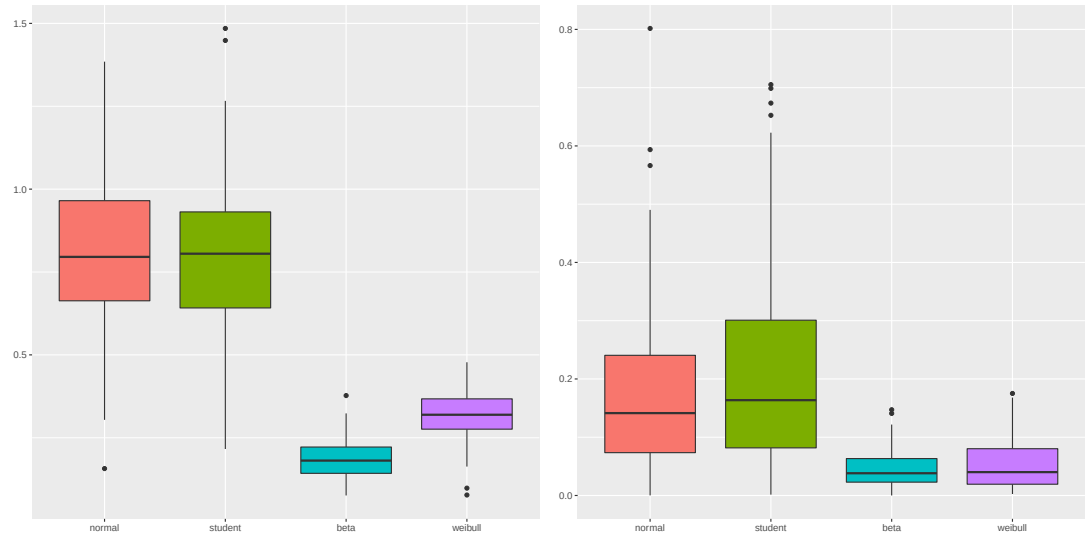

Figure 1.6: SEE estimator $p \gg n$ and Identity Design. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

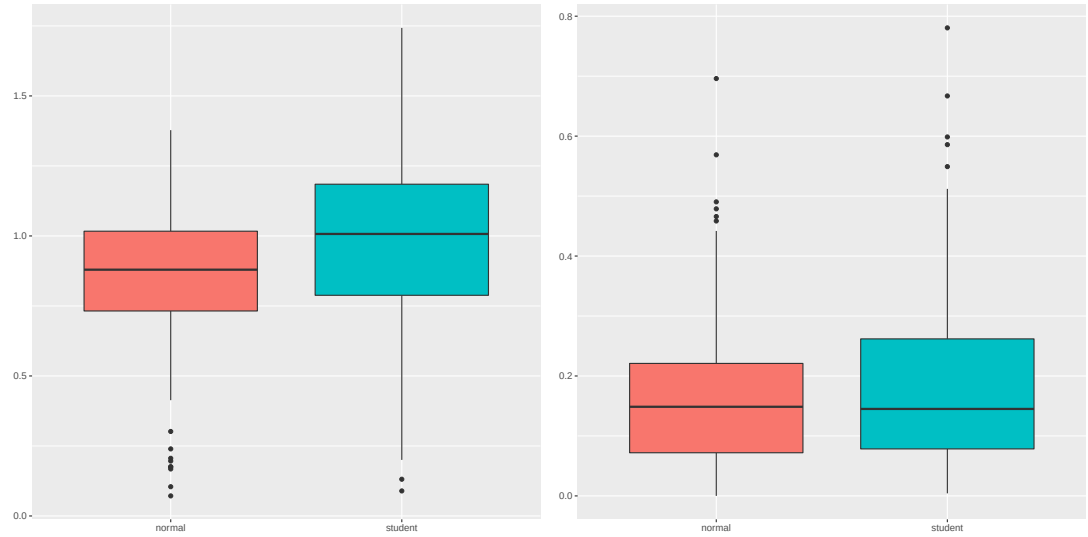

Figure 1.7: SREE estimator $p \gg n$ and Identity Design. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

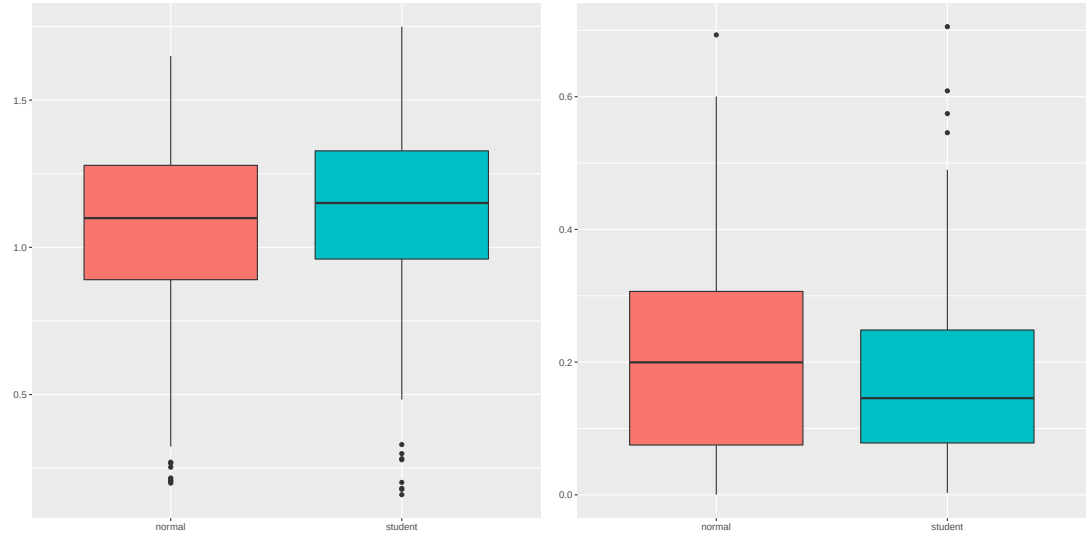

Figure 1.8: SREE estimator $p \gg n$ and Toeplitz Design. Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables.

sometimes the nature of the data collection process has determined that a significant noise is inevitable for some observations, or that portions of the observations have been corrupted by an adversary. In big and high-dimensional data setting, such cases may occur naturally. When the cost of error is too large to bear, it may be wise to consider an alternative that can improve upon the inferential accuracy in a stepwise manner. With one-step robust estimators, one can often successfully iterate the estimate, and identify misleading observations. Therefore, limiting the effect of poor data quality.

Many different loss functions and penalty functions, including non-convex ones, may be incorporated into this framework for the purpose of achieving correct inferential tools. A novel theory is provided, with emphasis on diverging dimensions and left-censoring. Future work will be devoted to how to better utilize longitudinal and heterogeneous observations.

There are many one-step estimators based on a suitable choice of loss function or estimating equations, some of which have proved to work well, especially when the dimension is reasonably high. The proposed method allows for left-censoring, non-smooth, non-convex losses and/or non-monotone equations, and complements the existing methods in these domains. The

method achieves rates comparable the ones of efficient methods (with full observations), and the accompanying analysis provides tight control over both Type I and Type II error rates, which makes it a practically useful and efficient alternative.

1.7 General Results

In this section, we present the general results along with theoretical considerations. Statements and proofs of Lemmas 1 - 6 and Theorems 6 - 11 are included.

We begin theoretical analysis with the following decomposition of (1.12)

$$\begin{aligned}
&\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right) \\
&\quad=\frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)+\frac{1}{2f(0)}\left(\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) \\
&\qquad+\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)+\frac{1}{2f(0)}\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\psi_i(\widehat{\boldsymbol{\beta}})-n^{-1}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)\right). \qquad (1.30)
\end{aligned}$$

We can further decompose the last factor of the last term in (1.30) as

$$n^{-1}\sum_{i=1}^{n}\psi_i(\widehat{\boldsymbol{\beta}})-n^{-1}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)=\mathbb{G}_n(\widehat{\boldsymbol{\beta}})-\mathbb{G}_n(\boldsymbol{\beta}^*)+n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\psi_i(\widehat{\boldsymbol{\beta}})-\psi_i(\boldsymbol{\beta}^*)\right],$$

where

$$\mathbb{G}_n(\boldsymbol{\beta})=n^{-1}\sum_{i=1}^{n}\left[\psi_i(\boldsymbol{\beta})-\mathbb{E}\psi_i(\boldsymbol{\beta})\right]. \qquad (1.31)$$

To characterize the behavior of individual terms in the decomposition above, we develop a sequence of results presented below that rely on the conditions that we listed in Section 1.3.

Lemma 1. *Suppose that the Conditions* **(E)** *hold. Consider the class of parameter spaces modeling sparse vectors with at most t non-zero elements,* $\mathscr{C}(r,t)=\{\mathbf{w}\in\mathbb{R}^p \mid ||\mathbf{w}||_2\leq r_n, \sum_{j=1}^{p}\mathbb{1}\{w_j\neq 0\}\leq t\}$ *where* r_n *is a sequence of positive numbers. Then, there exists a fixed constant C (inde-*

pendent of p and n), such that the process $\mu_i(\boldsymbol{\delta}) = \mathbb{I}\{x_i\boldsymbol{\delta} \geq x_i\boldsymbol{\beta}^*\} - \mathbb{I}\{0 \geq x_i\boldsymbol{\beta}^*\}$ *satisfies with probability* $1-\delta$.

$$\sup_{\boldsymbol{\delta}\in\mathscr{C}(r_n,t)} n^{-1}\left|\sum_{i=1}^{n}\mu_i(\boldsymbol{\delta}) - \mathbb{E}[\mu_i(\boldsymbol{\delta})]\right| \leq C\left(\sqrt{\frac{r_n t\sqrt{t}\log(np/\delta)}{n}} \bigvee \frac{t\log(2np/\delta)}{n}\right).$$

The preceding Lemma immediately implies strong approximation of the empirical process with its expected process, as long as r_n, the estimation error, and t, the size of the estimated set of the initial estimator, are sufficiently small. The power of the Lemma 1 is that it holds uniformly for a class of parameter vectors enabling a wide range of choices for the initial estimator.

Next, we present a linearization result useful for further decomposition of the Bahadur representation (1.30).

Lemma 2. *Suppose that the conditions* **(E)** *hold. For all* $\boldsymbol{\beta}$*, such that* $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1 < \xi$*, the following representation holds*

$$n^{-1}\sum_{i=1}^{n}\mathbb{E}\psi_i(\boldsymbol{\beta}) = 2f(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)(\boldsymbol{\beta}^*-\boldsymbol{\beta}) + \mathscr{O}(\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1)(\boldsymbol{\beta}^*-\boldsymbol{\beta}).$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$ *is defined in* (1.6).

Once the properties of the initial estimator are provided, such as Condition **(I)**, Lemma 2 can be used to linearize the population level difference of the functions $\psi_i(\widehat{\boldsymbol{\beta}})$ and $\psi_i(\boldsymbol{\beta}^*)$. Together with Lemma 1, Lemma 2 allows us to overpass the original highly discontinuous and non-convex loss function. Utilizing Lemma 2, Conditions **(I)**-**(C)** and representation (1.30), the Bahadur representation of $\widetilde{\boldsymbol{\beta}}$ becomes

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right) = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) + I_1 + I_2 + I_3 + I_4 \tag{1.32}$$

where

$$I_1 = \sqrt{n}\left(I - \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right),\ I_2 = -\frac{1}{2f(0)}\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\sqrt{n}\cdot\mathscr{O}_P(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$$
$$I_3 = \frac{1}{2f(0)}\left(\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*),\ I_4 = \frac{1}{2f(0)}\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\sqrt{n}\left[\mathbb{G}_n(\widehat{\boldsymbol{\beta}}) - \mathbb{G}_n(\boldsymbol{\beta}^*)\right].$$

We show that the last four terms of the right hand side above, each converges to 0 asymptotically at a faster rate than the first term on the right hand side of (1.32).

The following two lemmas help to establish l_1 column bound of the corresponding precision matrix estimator. The first one provides properties of the estimator $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$ as defined in (1.8). Although this estimator is obtained via Lasso-type procedure, significant challenges arise in its analysis due to dependencies in the plug-in loss function. The design matrix of this problem does not have independent and identically distributed rows. We overcome these challenges by approximating the solution to the oracle one and without imposing any new conditioning of the design matrix.

Lemma 3. *Let* $\lambda_j = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4}(\log p/n)^{1/2}\right) t^{3/4}(\log p/n)^{1/2}\right)$ *for a constant* $C > 1$ *and let Conditions* **(I)**, **(E)**, **(C)** *and* **(Γ)** *hold. Then,*

$$\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1 = \mathscr{O}_P\left(\frac{1}{\phi_0^2 C_2} s_j \lambda_j\right).$$

Remark 8. *The choice of the tuning parameter* λ_j *depends on the* l_2 *convergence rate of the initial estimator* r_n*, and the size of its estimated non-zero set. However, we observe that whenever* r_n *is such that* $r_n \leq t^{-3/4}$ *and the sparsity of the initial estimator is such that* $ts_j\sqrt{\log p/n} < 1$*, then the optimal choice of the tuning parameter is of the order of* $\sqrt{\log p/n}$*. In particular, any initial estimator that satisfies* $r_n < n^{-1/4}$ *is sufficient for optimal rates of inference in a model where* $t \leq n^{1/4}$ *and* $s_j \leq n^{1/4}$*.*

The next result gives a bound on the variance of our $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$ estimator.

Lemma 4. *Let* $\lambda_j = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4}(\log p/n)^{1/2}\right) t^{3/4}(\log p/n)^{1/2}\right)$ *for a constant* $C > 1$ *and let Conditions* **(I)**, **(E)**, **(C)** *and* **(Γ)** *hold. Then, for* $j = 1, \ldots, p$ *and* $\boldsymbol{\zeta}_j^*$ *and* $\widehat{\boldsymbol{\zeta}}_j$

$$\left|\widehat{\boldsymbol{\zeta}}_j^\top \widehat{\boldsymbol{\zeta}}_j/n - \mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*/n\right| = \mathscr{O}_P\left(K^2 s_j \lambda_j\right).$$

Next is the main result on the properties of the proposed matrix estimator $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$.

Lemma 5. *Let the setup of Lemma 4 hold. Let* $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ *be the estimator as in* (1.10)*. Then, for* $\widehat{\tau}_j^2$ *as in* (1.9)*, we have* $\widehat{\tau}_j^{-2} = \mathscr{O}_P(1)$*. Moreover,*

$$\left\|\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})_j - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_j\right\|_1 = \mathscr{O}_P\left(K^2 s_j^{3/2}\lambda_j\right).$$

The one-step estimator $\widetilde{\boldsymbol{\beta}}$ relies crucially on the bias correction step that carefully projects the residual vector in the direction close to the most efficient score. The next result measures the uniform distance of such projection.

Lemma 6. *Let the setup of Lemma 4 hold. There exists a fixed constant C (independent of p and n), such that the process* $\mathbb{V}_n(\boldsymbol{\delta}) = \boldsymbol{\Omega}(\boldsymbol{\delta} + \boldsymbol{\beta}^*)\left[\mathbb{G}_n(\boldsymbol{\delta} + \boldsymbol{\beta}^*) - \mathbb{G}_n(\boldsymbol{\beta}^*)\right]$ *satisfies*

$$\sup_{\boldsymbol{\delta} \in \mathscr{C}(r_n, t)} \|\mathbb{V}_n(\boldsymbol{\delta})\|_\infty \leq C\left(\sqrt{\frac{(r_n t^{1/2} \vee r_n^2 t)t\log(np/\delta)}{n}} \bigvee \frac{t\log(2np/\delta)}{n}\right),$$

with probability $1 - \delta$ *and a constant* K_1 *defined in Condition* (**E**)*.*

Lemma 6 establishes a uniform tail probability bound for a growing supremum of an empirical process $\mathbb{V}_n(\boldsymbol{\delta})$. It is uniform in $\boldsymbol{\delta}$ and it is growing as supremum is taken over p, possibly growing ($p = p(n)$) coordinates of the process. The proof of Lemma 6 is further challenged by the non-smooth components of the process $\mathbb{V}_n(\boldsymbol{\delta})$ itself and the multiplicative nature of the factors within it. It proceeds in two steps. First, we show that for a fixed $\boldsymbol{\delta}$ the term $||\mathbb{V}_n(\boldsymbol{\delta})||_\infty$ is small. In the second step, we devise a new epsilon net argument to control the non-smooth and multiplicative

terms uniformly for all $\boldsymbol{\delta}$ simultaneously. This is established by devising new representations of the process that allow for small size of the covering numbers. In conclusion, Lemma 6 establishes a uniform bound $\|I_4\|_\infty = \mathscr{O}_P\left(r_n^{1/2} t^{3/4} (\log p)^{1/2} \bigvee r_n t (\log p)^{1/2} \bigvee t \log p / n^{1/2}\right)$ in (1.32).

Size of the remainder term in (1.13) is controlled by the results of Lemmas 1-6 and we provide details below.

Theorem 6. *Let* $\lambda_j = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4} (\log p/n)^{1/2}\right) t^{3/4} (\log p/n)^{1/2}\right)$ *for a constant* $C > 1$ *and let Conditions* **(I)**, **(E)**, **(C)** *and* **(Γ)** *hold. With* $s_\Omega = \max_j s_j$,

$$\|\Delta\|_\infty = \mathscr{O}_P\left((r_n^{1/2} t^{1/4} \vee r_n t^{1/2}) t^{1/2} (\log p)^{1/2} \bigvee \sqrt{nt} s_\Omega^{3/2} \lambda_j r_n^2 \bigvee \sqrt{n} s_\Omega^{3/2} \lambda_j r_n\right).$$

We first notice that the expression above requires $t = o(n^{1/2}/\log(p \vee n))$, a condition frequently imposed in high-dimensional inference (see [ZZ14] for example). Then, in the case of low-dimensional problems with $s = \mathscr{O}(1)$ and $p = \mathscr{O}(1)$, we observe that whenever the initial estimator of rate r_n, is in the order of $n^{-\varepsilon}$, for a small constant $\varepsilon > 0$, then $\|\Delta\|_\infty = \mathscr{O}_P(n^{-\varepsilon/2})$. In particular, for a consistent initial estimator, i.e. $r_n = \mathscr{O}(n^{-1/2})$ we obtain that $\|\Delta\|_\infty = \mathscr{O}_P(n^{-1/4})$. For high-dimensional problems with s and p growing with n, for all initial estimators of the order r_n such that $r_n = \mathscr{O}(s_{\boldsymbol{\beta}^*}^a (\log p)^b / n^c)$ and $t = \mathscr{O}(s_{\boldsymbol{\beta}^*})$ we obtain that

$$\|\Delta\|_\infty = \mathscr{O}_P\left(\bar{s}^{(2a+3)/4} (\log p)^{(1+b)/2} / n^{c/2}\right)$$

whenever $\bar{s}(\log p)^{1/4}/n^{1/4} = \mathscr{O}(1)$, where $\bar{s} = t \vee s_\Omega$.

Next, we present the result on the asymptotic normality of the leading term of the Bahadur representation (1.13).

Theorem 7. *Let* $\lambda_j = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4} (\log p/n)^{1/2}\right) t^{3/4} (\log p/n)^{1/2}\right)$ *for a constant* $C > 1$ *and let Conditions* **(I)**, **(E)**, **(C)** *and* **(Γ)** *hold.*

Define $U := \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}^*) = \mathcal{O}_P(\sqrt{n})$. *Furthermore, assume*

$$(r_n^{1/2}t^{1/4} \vee r_n t^{1/2})t^{1/2}(\log p)^{1/2}\bigvee \sqrt{nt}s_\Omega^{3/2}\lambda_j r_n^2 \bigvee \sqrt{n}s_\Omega^{3/2}\lambda_j r_n = o(1).$$

Denote $\bar{s} = t \vee s_\Omega$. *If* $f(0)$, *the density of* ε *at 0 is known,*

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right]_{jj}^{-\frac{1}{2}} U_j \xrightarrow[n,p,\bar{s}\to\infty]{d} \mathcal{N}\left(0, \frac{1}{4f(0)^2}\right).$$

Remark 9. A few remarks are in order. Theorem 7 implies that the effects of censoring asymptotically disappear. Namely, the limiting distribution only becomes degenerate when the censoring rate asymptotically explodes, implying that no data is fully observed. However, in all other cases the limiting distribution is fixed and does not depend on the censoring level.

Density estimation is a necessary step in the semiparametric inference for left-censored models. Below we present the result guaranteeing good qualities of density estimator proposed in (1.11).

Theorem 8. *There exists a sequence* h_n *such that* $h_n = \mathcal{O}(1)$ *and* $\lim_{n\to\infty}\widehat{h}_n/h_n = 1$ *and* $h_n^{-1}(r_n \vee r_n^{1/2}t^{3/4}(\log p/n)^{1/2} \vee t\log p/n) = o(1)$. *Assume Conditions* **(I)** *and* **(E)** *hold, then*

$$\left|\widehat{f}(0) - f(0)\right| = \mathcal{O}_P(1).$$

Together with Theorem 7 we can provide the next result.

Corollary 9. *With the choice of density estimator as in* (1.11)*, under conditions of Theorem 7 and 8, the results of Theorem 7 continue to hold unchanged, i.e.,*

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right]_{jj}^{-\frac{1}{2}} U_j \cdot 2\widehat{f}(0) \xrightarrow[n,p,\bar{s}\to\infty]{d} \mathcal{N}(0,1).$$

Remark 10. *Observe that the result above is robust in the sense that the result holds regardless*

of the particular distribution of the model error (1.1). *Condition* **(E)** *only assumes minimal regularity conditions on the existence and smoothness of the density of the model errors. In the presence of censoring, our result is unique as it allows $p \gg n$, and yet it successfully estimates the variance of the estimation error.*

Combining all the results obtained in previous sections we arrive at the main conclusions.

Theorem 10. *Let* $\lambda_j = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4}(\log p/n)^{1/2}\right) t^{3/4}(\log p/n)^{1/2}\right)$ *for a constant* $C > 1$ *and let Conditions* **(I)**, **(E)**, **(C)** *and* **(Γ)** *hold. Furthermore, assume*

$$(r_n^{1/2}t^{1/4} \vee r_n t^{1/2})t^{1/2}(\log p)^{1/2} \bigvee \sqrt{nt}s_\Omega^{3/2}\lambda_j r_n^2 \bigvee \sqrt{n}s_\Omega^{3/2}\lambda_j r_n = o(1),$$

for $s_\Omega = \max_j s_j$. *Denote* $\bar{s} = t \vee s_\Omega$. *Let* I_n *and* a_n *be defined in* (1.15) *and* (1.16). *Then, for all vectors* $\mathbf{c} = \mathbf{e}_j$ *and any* $j \in \{1, \ldots, p\}$, *when* $n, p, \bar{s} \to \infty$ *we have*

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\mathbf{c}^\top \boldsymbol{\beta}^* \in I_n\right) = 1 - 2\alpha$$

Let $\mathbb{P}_{\boldsymbol{\beta}^*}$ be the distribution of the data under the model (1.1). Then the following holds.

Theorem 11. *Under the setup and assumptions of Theorem 10 when* $n, p, \bar{s} \to \infty$

$$\sup_{\boldsymbol{\beta} \in \mathscr{B}} \mathbb{P}_{\boldsymbol{\beta}}\left(\mathbf{c}^\top \boldsymbol{\beta}^* \in I_n\right) = 1 - 2\alpha.$$

1.8 Proofs of Main Theorems

Proof of Theorem 1. The proof for the result with initial estimator chosen as the penalized CLAD estimator of [MvdG16] follows directly from Lemma 1-6 and Theorem 6-10 with $r_n = s_{\boldsymbol{\beta}^*}^{1/2}(\log p/n)^{1/2}$ and $t = s_{\boldsymbol{\beta}^*}$. □

Proof of Theorems 2, 3 and 4. Due to the limit of space, we follow the line of the proof of

Theorem 7 but only give necessary details when the proof is different. First, we observe that with a little abuse in notation

$$\psi_i(\boldsymbol{\beta}) = w_i^{\top}(\boldsymbol{\beta}) R_i^r, \qquad R_i^r = q_i \psi(-v_i \varepsilon_i)$$

thus it suffices to provide the asymptotic of

$$T_n^r := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i^r = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_1 \, \mathbb{1}\{x_i \boldsymbol{\beta} > 0\} R_i^r.$$

Moreover, observe that R_i^r are necessarily bounded random variables (see Condition (r$\boldsymbol{\Gamma}$). Following similar steps as in Theorem 7 we obtain

$$\mathrm{Var}(T_n^r) \geq n - 2\exp\{-n^2/2\}$$

where in the last step we utilized Hoeffding's inequality for bounded random variables.

Next, we focus on establishing an equivalent of Lemma 2 but now for the robust generalized M-estimator. Observe that

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}_{\varepsilon}[\psi_i^r(\boldsymbol{\beta})] = n^{-1} \sum_{i=1}^{n} x_i^{\top} \, \mathbb{1}\{x_i \boldsymbol{\beta} > 0\} q_i \mathbb{E}_{\varepsilon}\left[\psi\Big(-v_i x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - v_i \varepsilon_i\Big)\right]. \tag{1.33}$$

Moreover, whenever ψ' exists we have

$$\mathbb{E}_{\varepsilon}\left[\psi\Big(-v_i x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - v_i \varepsilon_i\Big)\right] = -v_i x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \int_{-\infty}^{\infty} \psi'(\xi(u)) f(u) du.$$

for $\xi(u) = \alpha(-v_i x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta})) + (1-\alpha)(-v_i u)$ for some $\alpha \in (0,1)$. When ψ' doesn't exist we can decompose ψ into a finite sum of step functions and then apply exactly the same technique on each of the step functions as in Lemma 2. Hence, it suffices to discuss the differentiable case

only. Let us denote the RHS of (1.33) with $\Lambda_n^r(\boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})$, i.e.

$$\Lambda_n^r(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} - \mathbb{I}\{x_i\boldsymbol{\beta} > 0\} q_i v_i x_i^\top x_i \int_{-\infty}^{\infty} \psi'(\xi(u)) f(u) du.$$

Next, we observe that by Condition (r$\boldsymbol{\Gamma}$),

$$\left| \int_{-\infty}^{\infty} \psi'(\xi(u)) f(u) du - \psi'(v_i \varepsilon_i) \right| \leq \sup_x |\psi'(x)| := C_1$$

for a constant $C_1 < \infty$. With that the remaining steps of Lemma 2 can be completed with $\boldsymbol{\Sigma}$ replaced with $\boldsymbol{\Sigma}^r$.

Next, by observing the proofs of Lemmas 3, 4 and 5 we see that the proofs remain to hold under Condition (r$\boldsymbol{\Gamma}$), and with W replaced with $\widetilde{W}$. The constants K appearing in the simpler case will now be KM_1M_2. However, the rates remain the same up to these constant changes.

Next, we discuss Lemma 6. For the case of robust generalized M-estimator $\nu_n(\boldsymbol{\delta})$ of Lemma 6 takes the following form

$$\widetilde{\nu}_n(\boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{\Omega}}(\boldsymbol{\delta} + \boldsymbol{\beta}^*) x_i^\top [f_i(\boldsymbol{\delta}) \widetilde{g}_i(\boldsymbol{\delta}) - f_i(\mathbf{0}) \widetilde{g}_i(\mathbf{0})]$$

with $\widetilde{g}_i(\boldsymbol{\delta}) = q_i \psi(v_i(x_i\boldsymbol{\delta} + \varepsilon_i))$. Moreover, $\mathbb{E}_\varepsilon[f_i(\boldsymbol{\delta})\widetilde{g}_i(\boldsymbol{\delta})] = f_i(\boldsymbol{\delta})\mathbb{E}_\varepsilon[q_i\psi(v_i(x_i\boldsymbol{\delta} + \varepsilon_i))] := \widetilde{w}_i(\boldsymbol{\delta})$. We consider the same covering sequence as in Lemma 6. Then, we observe that a bound equivalent to T_1 of Lemma 6 is also achievable here.

Term T_2 can be handled similarly as in Lemma 6. We illustrate the particular differences only in T_{21} as others follows similarly. Observe that

$$f_i(\boldsymbol{\delta})\widetilde{g}_i(\boldsymbol{\delta}) = \mathbb{I}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\} q_i \psi(v(\varepsilon_i)) + \mathbb{I}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\} q_i v_i x_i \boldsymbol{\delta} \psi'(\xi_{\boldsymbol{\delta}})$$

for $\xi_{\boldsymbol{\delta}} = v_i\varepsilon_i + (1-\alpha)v_i x_i\boldsymbol{\delta}$ for some $\alpha \in (0,1)$. Next, we consider the decomposition

$$f_i(\boldsymbol{\delta})\widetilde{g}_i(\boldsymbol{\delta}) - \mathbb{E}\left[f_i(\boldsymbol{\delta})\widetilde{g}_i(\boldsymbol{\delta})\right] = T^r_{211}(\boldsymbol{\delta}) + T^r_{212}(\boldsymbol{\delta})$$

where

$$T^r_{211}(\boldsymbol{\delta}) = \left(\mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\} - \mathbb{P}(x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*)\right) q_i\psi(v_i\varepsilon_i)$$

and

$$T^r_{212}(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\,\psi'(\xi_{\boldsymbol{\delta}}) - \mathbb{E}\left[\mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\,\psi'(\xi_{\boldsymbol{\delta}})\right]$$

Furthermore, we observe that the same techniques developed in Lemma 6 apply to $T^r_{211}(\boldsymbol{\delta})$ hence we only discuss the case of $T^r_{212}(\boldsymbol{\delta})$. We begin by considering the decomposition $T^r_{212}(\boldsymbol{\delta}) = T^r_{2121}(\boldsymbol{\delta}) + T^r_{2122}(\boldsymbol{\delta})$ with

$$T^r_{2121}(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\left(\psi'(\xi_{\boldsymbol{\delta}}) - \mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}}))\right)$$

and

$$T^r_{2122}(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) - \mathbb{E}\left[\mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\mathbb{E}_\varepsilon\psi'(\xi_{\boldsymbol{\delta}})\right]$$

Let us focus on the last expression as it is the most difficult one to analyze. Observe that we are interested in the difference $T^r_{2122}(\boldsymbol{\delta}) - T^r_{2122}(\widetilde{\boldsymbol{\delta}}_k)$. We decompose this difference into four terms, two related to random variables and two related to the expectations. We handle them separately and observe that because of symmetry and monotonicity of the indicator functions once we can bound the difference of random variables we can repeat the arguments for the expectations.

Hence, we focus on

$$I_1 = \mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\boldsymbol{\delta}\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) - \mathbb{1}\{x_i\widetilde{\boldsymbol{\delta}}_k \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\widetilde{\boldsymbol{\delta}}_k\mathbb{E}_\varepsilon(\psi'(\xi_{\widetilde{\boldsymbol{\delta}}_k})).$$

First due to monotonicity of indicators and (1.57) we have

$$|I_1| \leq I_{11} + I_{12} + I_{13}$$

with

$$\begin{aligned}
I_{11} &= \left(\mathbb{1}\{x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n \geq -x_i\boldsymbol{\beta}^*\} - \mathbb{1}\{x_i\widetilde{\boldsymbol{\delta}}_k \geq -x_i\boldsymbol{\beta}^*\}\right) q_i v_i x_i\widetilde{\boldsymbol{\delta}}_k\mathbb{E}_\varepsilon(\psi'(\xi_{\widetilde{\boldsymbol{\delta}}_k})) \\
I_{12} &= \mathbb{1}\{x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n \geq -x_i\boldsymbol{\beta}^*\}q_i v_i\widetilde{L}_n\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) \\
I_{13} &= \mathbb{1}\{x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n \geq -x_i\boldsymbol{\beta}^*\}q_i v_i x_i\widetilde{\boldsymbol{\delta}}_k\left(\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) - \mathbb{E}_\varepsilon(\psi'(\xi_{\widetilde{\boldsymbol{\delta}}_k}))\right)
\end{aligned}$$

As $\sup\psi' < \infty$, I_{11} can be handled in the same manner as T_{21} of the proof of Lemma 6, whereas $I_{12} = \mathscr{O}_P(\widetilde{L}_n)$. For I_{13} it suffices to discuss the difference at the end of the right hand side of its expression. It is not difficult to see that

$$\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) - \mathbb{E}_\varepsilon(\psi'(\xi_{\widetilde{\boldsymbol{\delta}}_k})) \leq 4Cv_i\widetilde{L}_n \leq 4CM_1\widetilde{L}_n$$

with $C = \sup_x|\psi''(x)|$ for the case of twice differentiable ψ, $C = \sup_y \partial/\partial y|\int_{-\infty}^{y}\psi'(x)dx|$ for the case of once differentiable ψ and $C = f_{\max}$ for the case of non-differentiable functions ψ. Combining all the things together we observe that the rate of Lemma 6 for the case of robust generalized M-estimators is of the order of

$$C\left(\sqrt{\frac{M_3(r_n t^{1/2} \vee K^2M_1^2M_2^2r_n^2t)t\log(2np/\delta)}{n}} \bigvee \frac{t\log(2np/\delta)}{n}\right).$$

with $M_3 = \sup_x |\psi'(x)|$ for once differentiable ψ and $M_3 = f_{\max}$ for non-differentiable ψ.

Now, with equivalents of Lemmas 1-6 are established, we can use them to bound successive terms in the Bahadur representation much like those of Theorem 1. Details are ommitted due to space considerations.

For Theorem 4 in the Main Material, the same line of the proof of Theorem 11 applies, but only replace the matrix $\mathbf{\Sigma}$ with the matrix $\mathbf{\Sigma}^{\mathrm{r}}$. The result of the Theorem then follows from the arguments in Remark 2 in the Main Material. Uniformity of the obtained results is not compromised as the weight functions q_i and v_i only depend on the design matrix. □

Proof of Theorem 6. The proof of the theorem follows from the bounding residual terms in the Bahadur representation (1.32) with the help of Lemma 3 - 6.

Recall in Lemma 6, we showed that

$$\|I_4\|_\infty = \mathscr{O}_P\left((r_n^{1/2}t^{1/4} \vee r_n t^{1/2})t^{1/2}(\log p)^{1/2}\bigvee t\log p/n^{1/2}\right).$$

For the term I_3, we have that

$$\begin{aligned}&\left\|\frac{1}{2f(0)}\left(\mathbf{\Omega}(\widehat{\boldsymbol{\beta}}) - \mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}^*)\right\|_\infty \\ &\qquad \leq \mathscr{O}_P\left(s_\Omega^{3/2}\lambda_j\right),\end{aligned}$$

by applying Hölder's inequality and Hoeffding's inequality along with Lemma 5.

For the term I_2, we have

$$\begin{aligned}&\left\|\frac{1}{2f(0)}\mathbf{\Omega}(\widehat{\boldsymbol{\beta}})\sqrt{n}\cdot\mathscr{O}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_\infty \\ &\quad \leq \frac{\sqrt{nt}}{2f(0)}\left(\left\|\mathbf{\Omega}(\widehat{\boldsymbol{\beta}}) - \mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_1 + \left\|\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_2\right)\mathscr{O}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2) \\ &\quad \leq \mathscr{O}_P\left(\sqrt{nt}s_\Omega^{3/2}\lambda_j r_n^2 \bigvee \sqrt{nt}r_n^2\right),\end{aligned}$$

by Hölder's inequality and Lemma 5, where $\|A\|_\infty$ denotes the max row sum of matrix A, and $\|A\|_1$ denotes the max column sum of matrix A.

Lastly, for the only remainder term in (1.32), I_1, we apply Hölder's inequality and Lemma 5,

$$\begin{aligned}
&\sqrt{n}\left(I-\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\\
&\quad=\sqrt{n}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)-\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\\
&\quad\leq C\sqrt{n}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)-\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right\|_1\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\\
&\quad\leq \mathscr{O}_P\left(\sqrt{n}s_\Omega^{3/2}\lambda_j r_n\right).
\end{aligned}$$

□

Proof of Theorem 7. We begin the proof by noticing that

$$\begin{aligned}
\psi_i(\boldsymbol{\beta}^*)&=\operatorname{sign}(y_i-\max\{0,x_i\boldsymbol{\beta}^*\})(w_i(\boldsymbol{\beta}^*))^\top\\
&=\operatorname{sign}(\max\{0,x_i\boldsymbol{\beta}^*+\varepsilon_i\}-\max\{0,x_i\boldsymbol{\beta}^*\})(w_i(\boldsymbol{\beta}^*))^\top.
\end{aligned}$$

Recollect that by Condition **(E)**, $\mathbb{P}(\varepsilon_i\geq 0)=1/2$. Additionally, we observe that in distribution, the term on the right hand side is equal to $w_i^\top(\boldsymbol{\beta}^*)R_i$, with $\{R_i\}_{i=1}^n$ denoting an i.i.d. Rademarcher sequence defined as $R_i=\operatorname{sign}(-\varepsilon_i)$. Hence, it suffices to analyze the distributional properties of $w_i^\top(\boldsymbol{\beta}^*)R_i$. Moreover, Rademacher random variables are independent in distribution from $w_i(\boldsymbol{\beta}^*)$. Thus, we provide asymptotics of

$$\frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i^\top(\boldsymbol{\beta}^*)R_i.$$

We begin by defining

$$V_i := \frac{1}{\sqrt{n}} W_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i = \frac{1}{\sqrt{n}} X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i$$

and we also define $T_n := \sum_{i=1}^n V_i$. Notice that V_i's are independent from each other, since we assumed that each observation is independent in our design. We have

$$\sum_{i=1}^n \mathbb{E}|V_i|^{2+\delta} = \left(\frac{1}{\sqrt{n}}\right)^{2+\delta} \mathbb{E} \sum_{i=1}^n |X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0)|^{2+\delta} \leq n^{-1-\delta/2} \mathbb{E} \sum_{i=1}^n |X_{ij}|^{2+\delta} \leq n^{-\delta/2} K. \tag{1.34}$$

Moreover, $\text{Var} T_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i \right)^2 - \left(\mathbb{E} X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i \right)^2$. Since R_i are independent from X,

$$\mathbb{E} X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i = \mathbb{E} X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) \cdot \mathbb{E} R_i = 0.$$

In addition, also due to this fact, V_i follows a symmetric distribution about 0. Thus,

$$\text{Var} T_n = \frac{1}{n} \mathbb{E} \sum_{i=1}^n \left(X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i \right)^2 = \frac{1}{n} \mathbb{E} \left(\sum_{i=1}^n X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i \right)^2 \geq \frac{1}{n} \int_{-n}^n t_n^2 f(t_n) dt_n,$$

where with a little abuse in notation we denote the density and distribution of T_n to be $f(t_n)$ and $F(t_n)$. Observe that

$$\frac{1}{n} \mathbb{E} \left(\sum_{i=1}^n X_{ij} \, \mathbb{1}(x_i \boldsymbol{\beta}^* > 0) R_i \right)^2 = \frac{1}{n} \int_{-\infty}^{\infty} t_n^2 f(t_n) dt_n \geq \frac{1}{n} \int_{-n}^n t_n^2 f(t_n) dt_n.$$

Thus,

$$\begin{aligned}\mathrm{Var}T_n &\geq \frac{1}{n}\left(t_n^2 F(t_n)\,\big|_{-n}^{n} - 2\int_{-n}^{n} t_n F(t_n)dt_n\right) \\ &\geq \frac{1}{n}\left(n^2F(n) - n^2F(-n) - 2\int_{-n}^{n} t_n dt_n\right) \\ &= \frac{1}{n}\left(2n^2F(n) - n^2\right) = n\left(2F(n)-1\right)\end{aligned} \tag{1.35}$$

Now combining (1.34) and (1.35), we have $\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\mathbb{E}|V_i|^{2+\delta}}{(\mathrm{Var}T_n)^{1+\frac{\delta}{2}}} = 0$. Thereby, we arrive at the result

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} w_i^{\top}(\boldsymbol{\beta}^*)R_i\right)_j \xrightarrow{d} \mathscr{N}\left(0, \mathrm{Var}T_n\right),$$

with the fact that $\mathrm{Var}T_n = \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} W_{ij}(\boldsymbol{\beta}^*)^2 = \frac{1}{n}\mathbb{E}W_j^{\top}(\boldsymbol{\beta}^*)W_j(\boldsymbol{\beta}^*) = \boldsymbol{\Sigma}(\boldsymbol{\beta}^*)_{jj}$. Also, the covariance

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} w_i^{\top}(\boldsymbol{\beta}^*)R_i\right)_{j_1}\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} w_i^{\top}(\boldsymbol{\beta}^*)R_i\right)_{j_2}\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} W_{ij_1}(\boldsymbol{\beta}^*)W_{ij_2}(\boldsymbol{\beta}^*)\right] = \boldsymbol{\Sigma}(\boldsymbol{\beta}^*)_{j_1j_2}.\end{aligned}$$

Therefore, we have the following conclusion,

$$\left[\frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)\right]_j \xrightarrow{d} \mathscr{N}\left(0, \frac{1}{4f(0)^2}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right)^{\top}\right]_{jj}\right),$$

where $j = 1, \cdots, p$. This gives

$$\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_{jj}\right]^{-\frac{1}{2}}\left[\frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)\right]_j \xrightarrow{d} \mathscr{N}\left(0, \frac{1}{4f(0)^2}\right) \tag{1.36}$$

Notice that for two nonnegative real numbers a and b, it holds that

$$\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{ab}} = \frac{b-a}{\sqrt{ab}(\sqrt{b}+\sqrt{a})}.$$

We first make note of a result in the proof of Theorem 10, that

$$\left\| \widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_{\max} = \mathscr{O}_P(1) \tag{1.37}$$

Let $a = \left[\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\right]_{jj}$ and $b = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_{jj}$. By Condition **(C)**, we have $\sqrt{b}$ is bounded away from zero. Then, $\sqrt{a}$ is also bounded away from zero by (1.37), and so is $\sqrt{ab}(\sqrt{b}+\sqrt{a})$, since we have

$$\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right]_{jj} - \left[\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\right]_{jj} \leq \left\| \widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_{\max} = \mathscr{O}_P(1).$$

The rate above follows from (1.41) in the proof of Theorem 10. Notice the rate is of order smaller than the rate assumption in Theorem 6.

Thus, we can deduce that

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\right]_{jj}^{-\frac{1}{2}} - \left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_{jj}\right]^{-\frac{1}{2}} \leq C \left\| \widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_{\max}.$$

for some finite constant C. Applying Slutsky theorem on (1.36) with the inequality above, the desired result is obtained. □

Proof of Theorem 8. We can rewrite the expression $\widehat{f}(0)$ in (1.11) as

$$\begin{aligned}
\widehat{f}(0) &= \widehat{h}_n^{-1} \frac{\sum_{i=1}^n \mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)\, \mathbb{I}(0 \leq y_i - x_i\widehat{\boldsymbol{\beta}} \leq \widehat{h}_n)}{\sum_{i=1}^n \mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)} \\
&= \widehat{h}_n^{-1} \frac{n^{-1}\sum_{i=1}^n \mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)\, \mathbb{I}(0 \leq y_i - x_i\widehat{\boldsymbol{\beta}} \leq \widehat{h}_n)}{n^{-1}\sum_{i=1}^n \mathbb{P}\{x_i\boldsymbol{\beta}^* > 0\}} \cdot \frac{n^{-1}\sum_{i=1}^n \mathbb{P}\{x_i\boldsymbol{\beta}^* > 0\}}{n^{-1}\sum_{i=1}^n \mathbb{I}(x_i\widehat{\boldsymbol{\beta}} > 0)}.
\end{aligned}$$

Since $\left| n^{-1}\sum_{i=1}^{n}\left[\mathbb{I}\{x_i\widehat{\boldsymbol{\beta}}>0\}-\mathbb{P}\{x_i\boldsymbol{\beta}^*>0\}\right]\right| = \mathscr{O}_P(1)$, we have

$$\widehat{f}(0) \xrightarrow{d} \frac{(\widehat{h}_n n)^{-1}\sum_{i=1}^{n}\mathbb{I}(x_i\widehat{\boldsymbol{\beta}}>0)\,\mathbb{I}(0\le y_i - x_i\widehat{\boldsymbol{\beta}}\le \widehat{h}_n)}{n^{-1}\sum_{i=1}^{n}\mathbb{P}\{x_i\boldsymbol{\beta}^*>0\}}.$$

Using a similar argument and the fact that $\lim_{n\to\infty}\widehat{h}_n/h_n = 1$, we have

$$\widehat{f}(0) \xrightarrow{d} \frac{(h_n n)^{-1}\sum_{i=1}^{n}\mathbb{I}(x_i\widehat{\boldsymbol{\beta}}>0)\,\mathbb{I}(0\le y_i - x_i\widehat{\boldsymbol{\beta}}\le \widehat{h}_n)}{n^{-1}\sum_{i=1}^{n}\mathbb{P}\{x_i\boldsymbol{\beta}^*>0\}}.$$

Now we work on the numerator of right hand side. Specifically, let $\eta_i = y_i - x_i\boldsymbol{\beta}^*$ and $\widehat{\eta}_i = y_i - x_i\widehat{\boldsymbol{\beta}}$, we look at the difference of the quantities below,

$$\begin{aligned}
&(h_n n)^{-1}\left|\sum_{i=1}^{n}\mathbb{I}\{x_i\widehat{\boldsymbol{\beta}}>0\}\,\mathbb{I}\{0\le\widehat{\eta}_i\le\widehat{h}_n\}-\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*>0\}\,\mathbb{I}\{0\le\eta_i\le h_n\}\right| \\
&\quad\le (h_n n)^{-1}\left|\sum_{i=1}^{n}\mathbb{I}\{x_i\widehat{\boldsymbol{\beta}}>0\}\,\mathbb{I}\{0\le\widehat{\eta}_i\le\widehat{h}_n\}-\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*>0\}\,\mathbb{I}\{0\le\widehat{\eta}_i\le\widehat{h}_n\}\right| \\
&\qquad+2(h_n n)^{-1}\left|\sum_{i=1}^{n}\mathbb{I}\{x_i\widehat{\boldsymbol{\beta}}>0\}\,\mathbb{I}\{0\le\eta_i\le h_n\}-\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*>0\}\,\mathbb{I}\{0\le\eta_i\le h_n\}\right| \\
&\qquad+(h_n n)^{-1}\left|\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*>0\}\,\mathbb{I}\{0\le\widehat{\eta}_i\le\widehat{h}_n\}-\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*>0\}\,\mathbb{I}\{0\le\eta_i\le h_n\}\right| \\
&\quad\le \underbrace{3(h_n n)^{-1}\sum_{i=1}^{n}\mathbb{I}\{x_i\boldsymbol{\beta}^*\le x_i(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\}}_{T_1} \\
&\qquad+\underbrace{(h_n n)^{-1}\left|\sum_{i=1}^{n}\Big(\mathbb{I}\{0\le\widehat{\eta}_i\le\widehat{h}_n\}-\mathbb{I}\{0\le\eta_i\le h_n\}\Big)\right|}_{T_2}.
\end{aligned}$$

We begin with term T_1. By Condition **(E)**, we have $\mathbb{E}T_1 = \mathscr{O}(h_n^{-1}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1)$. By Corollary 1, we have

$$T_1 - \mathbb{E}T_1 \le |T_1 - \mathbb{E}T_1| = \mathscr{O}_P\left(h_n^{-1}(r_n^{1/2}t^{3/4}(\log p/n)^{1/2}\vee t\log p/n)\right),$$

which then brings us that T_1 is of order $\mathscr{O}_P(1)$. For term T_2, we work out the expression

$$\begin{aligned}
&\mathbb{I}\{0 \le \widehat{\eta}_i \le \widehat{h}_n\} - \mathbb{I}\{0 \le \eta_i \le h_n\} = \mathbb{I}\{0 \le \widehat{\eta}_i\}\,\mathbb{I}(\widehat{\eta}_i \le \widehat{h}_n\} - \mathbb{I}\{0 \le \eta_i\}\,\mathbb{I}\{\eta_i \le h_n\} \\
&\quad = \mathbb{I}\{0 \le \widehat{\eta}_i\}\left(\mathbb{I}(\widehat{\eta}_i \le \widehat{h}_n\} - \mathbb{I}(\eta_i \le h_n\}\right) + \left(\mathbb{I}\{0 \le \widehat{\eta}_i\} - \mathbb{I}\{0 \le \eta_i\}\right)\mathbb{I}\{\eta_i \le h_n\} \\
&\quad \le \mathbb{I}\{\widehat{\eta}_i \le \widehat{h}_n\} - \mathbb{I}\{\eta_i \le h_n\} + \mathbb{I}\{0 \le \widehat{\eta}_i\} - \mathbb{I}\{0 \le \eta_i\}.
\end{aligned}$$

Next, we notice that for real numbers a and b, we have $\mathbb{I}(a > 0) - \mathbb{I}(b > 0) \le \mathbb{I}(|b| \le |a-b|)$. Thus, we have

$$\begin{aligned}
T_2 &\le (h_n n)^{-1}\left|\sum_{i=1}^{n}\left\{\mathbb{I}(\widehat{\eta}_i \le \widehat{h}_n\} - \mathbb{I}\{\eta_i \le h_n\} + \mathbb{I}\{0 \le \widehat{\eta}_i\} - \mathbb{I}\{0 \le \eta_i\}\right)\right| \\
&\le h_n^{-1}n^{-1}\sum_{i=1}^{n}\mathbb{I}\{|h_n - \eta_i| \le |\widehat{h}_n - h_n| + |\eta_i - \widehat{\eta}_i|\} + h_n^{-1}n^{-1}\sum_{i=1}^{n}\mathbb{I}\{|\eta_i| \le |\widehat{\eta}_i - \eta_i|\} \\
&\le \underbrace{h_n^{-1}n^{-1}\sum_{i=1}^{n}\mathbb{I}\{|h_n - \eta_i| \le |\widehat{h}_n - h_n| + \|x_i\|_\infty\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\}}_{T_{21}} \\
&\quad \underbrace{+ h_n^{-1}n^{-1}\sum_{i=1}^{n}\mathbb{I}\{|\eta_i| \le \|x_i\|_\infty\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1\}}_{T_{22}}
\end{aligned}$$

To bound T_{21}, we use similar techniques as with T_1. Notice that

$$\mathbb{E}T_{21} = h_n^{-1}\mathbb{P}\left(|h_n - \eta_i| \le |\widehat{h}_n - h_n| + \|x_i\|_\infty\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\right)$$

It is easy to see that $|h_n - \eta_i|$ shares the nice property of the density of ε_i. Thus, $\mathbb{E}T_{21}$ is bounded by $\mathscr{O}_P(1)$. Then by Hoeffding's inequality, we have that with probability approaching 1 that T_{21} is of $\mathscr{O}_P(1)$. T_{22} can be bounded in exactly the same steps.

Finally, we are ready to put everything together that

$$(h_n n)^{-1} \left| \sum_{i=1}^{n} \mathbb{I}\{x_i \widehat{\boldsymbol{\beta}} > 0\} \, \mathbb{I}\{0 \le \widehat{\eta}_i \le \widehat{h}_n\} - \sum_{i=1}^{n} \mathbb{I}\{x_i \boldsymbol{\beta}^* > 0\} \, \mathbb{I}\{0 \le \eta_i \le h_n\} \right| = \mathscr{O}_P(1).$$

By applying Slutsky theorem, the result follows directly,

$$\widehat{f}(0) \xrightarrow{d} \frac{\sum_{i=1}^{n} \mathbb{I}\{x_i \boldsymbol{\beta}^* > 0\} \, \mathbb{I}\{0 \le \eta_i \le h_n\}}{n^{-1} \sum_{i=1}^{n} \mathbb{P}\{x_i \boldsymbol{\beta}^* > 0\}}.$$

□

Proof of Corollary 9. By multiplying and dividing the term $f(0)$, we can rewrite the term on the left hand side as

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \right]_{jj}^{\frac{1}{2}} U_j \cdot 2\widehat{f}(0) = \left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \right]_{jj}^{\frac{1}{2}} U_j \cdot 2f(0) \frac{\widehat{f}(0)}{f(0)}.$$

Also, as a result of theorem 8, we have

$$\frac{|\widehat{f}(0) - f(0)|}{f(0)} = |\widehat{f}(0)/f(0) - 1| = \mathscr{O}_P(1),$$

with Condition **(E)** guarantees that $f(0)$ is bounded away from 0. It also indicates that

$$\widehat{f}(0)/f(0) \xrightarrow{d} 1.$$

Finally, we apply Slutsky's Theorem and Theorem 7, we have

$$\left[\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \right]_{jj}^{\frac{1}{2}} U_j \cdot 2\widehat{f}(0) \xrightarrow[n,p,s_{\boldsymbol{\beta}^*} \to \infty]{d} \mathscr{N}(0,1).$$

□

Proof of Theorem 10. The result of Theorem 10 is a simple consequence of Wald's device and results of Corollary 9. The only missing link is an upper bound on

$$\left\| \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_{\max}. \tag{1.38}$$

First, observe that

$$\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) = \underbrace{\left(\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right) \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})}_{T_1} + \underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \left(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \mathbb{I} \right)}_{T_2}.$$

Regarding term T_1, observe that by Lemma 5 it is equal to $\mathscr{o}_P(1)$ whenever $\|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\|_{\max}$ is $\mathscr{O}_P(1)$. This can be seen from the decomposition of $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \mathbb{I}$, which reads,

$$\begin{aligned} \left\| \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \mathbb{I} \right\|_{\max} = & \underbrace{\left\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \left(\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}^*) \right) \right\|_{\max}}_{T_{21}} \\ & + \underbrace{\left\| \left(\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right) \left(\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}^*) \right) \right\|_{\max}}_{T_{22}} \\ & + \underbrace{\left\| \boldsymbol{\Sigma}(\boldsymbol{\beta}^*) \left(\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right) \right\|_{\max}}_{T_{23}} \end{aligned}$$

We notice that

$$
\begin{aligned}
T_{21} = \Bigg\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \Bigg(& n^{-1} \sum_{i=1}^{n} w_i^\top(\widehat{\boldsymbol{\beta}}) w_i(\widehat{\boldsymbol{\beta}}) - n^{-1} \sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) \\
& + n^{-1} \sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) - n^{-1} \mathbb{E} \sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) \Bigg) \Bigg\|_{\max} \\
\leq \Bigg\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) & \left(n^{-1} \sum_{i=1}^{n} \left(w_i(\widehat{\boldsymbol{\beta}}) + w_i(\boldsymbol{\beta}^*) \right)^\top \left(w_i(\widehat{\boldsymbol{\beta}}) - w_i(\boldsymbol{\beta}^*) \right) \right) \Bigg\|_{\max} \qquad (1.39) \\
+ \Bigg\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) & \left(n^{-1} \sum_{i=1}^{n} \left(w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) - \mathbb{E} w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) \right) \right) \Bigg\|_{\max} . \qquad (1.40)
\end{aligned}
$$

For (1.39), we have the following bound

$$
\begin{aligned}
(1.39) &\leq \left\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_\infty \left\| n^{-1} \sum_{i=1}^{n} \left(w_i(\widehat{\boldsymbol{\beta}}) + w_i(\boldsymbol{\beta}^*) \right)^\top \left(w_i(\widehat{\boldsymbol{\beta}}) - w_i(\boldsymbol{\beta}^*) \right) \right\|_{\max} \\
&\leq C s_\Omega^{1/2} n^{-1} \sum_{i=1}^{n} 2K^2 \left(\mathbb{I}(x_i \widehat{\boldsymbol{\beta}} > 0) - \mathbb{I}(x_i \boldsymbol{\beta}^*) \right),
\end{aligned}
$$

for some positive constant C, where $\|A\|_\infty$ denotes the max row sum of matrix A and $\|A\|_{\max}$ denotes the maximum element in the matrix A. By Lemma 1, we can easily bound the term above with $\mathscr{O}_P\left(K^2 s_\Omega^{1/2} (r_n^{1/2} t^{3/4} (\log p/n)^{1/2} \vee t \log p/n) \right)$. For (1.40), we start with the following term,

$$
n^{-1} \sum_{i=1}^{n} \left(W_{ij}(\boldsymbol{\beta}^*) W_{ik}(\boldsymbol{\beta}^*) - \mathbb{E} W_{ij}(\boldsymbol{\beta}^*) W_{ik}(\boldsymbol{\beta}^*) \right).
$$

Applying Hoeffding's inequality on this term, we have that with probability approaches 1, the term is bounded by $\mathscr{O}_P(n^{-1/2})$. Then we bound term (1.40) as following, for some constant C,

$$
\begin{aligned}
(1.40) &\leq \left\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right\|_\infty \left\| n^{-1} \sum_{i=1}^{n} \left(w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) - \mathbb{E} w_i^\top(\boldsymbol{\beta}^*) w_i(\boldsymbol{\beta}^*) \right) \right\|_{\max} \\
&\leq C s_\Omega^{1/2} \max_{j,k} \left\{ n^{-1} \sum_{i=1}^{n} \left(W_{ij}(\boldsymbol{\beta}^*) W_{ik}(\boldsymbol{\beta}^*) - \mathbb{E} W_{ij}(\boldsymbol{\beta}^*) W_{ik}(\boldsymbol{\beta}^*) \right) \right\} = \mathscr{O}_P(1)
\end{aligned}
$$

Term T_{22} can be bounded using Lemma 5 and the results from term T_{21}, and turns out to be of order $\mathscr{O}_P\left(K^4 s_\Omega^{3/2}\lambda_j(r_n^{1/2}t^{3/4}(\log p/n)^{1/2}\vee t\log p/n)\right)$.

Lastly, by Lemma 5, term T_{23} is of order $\mathscr{O}_P\left(K^2 s_\Omega^{3/2}\lambda_j\right)$.

Putting the terms together, we have $\left\|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})-\mathbb{I}\right\|_{\max}$ bounded by

$$\mathscr{O}_P\left((s_\Omega^{1/2}\vee s_\Omega^{3/2}\lambda_j)(r_n^{1/2}t^{3/4}(\log p/n)^{1/2}\vee t\log p/n)\bigvee s_\Omega^{3/2}\lambda_j\right)$$

Thus, $\|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\|_{\max}$ is $\mathscr{O}_P(1)$, and so can T_2 be shown similarly. The expression (1.38) is then bounded as,

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_{\max} \\ &\quad=\mathscr{O}_P\left((s_\Omega^{1/2}\vee s_\Omega^{3/2}\lambda_j)(r_n^{1/2}t^{3/4}(\log p/n)^{1/2}\vee t\log p/n)\bigvee s_\Omega^{3/2}\lambda_j\right)\end{aligned} \tag{1.41}$$

which then completes the proof.

□

Proof of Theorem 11. The result of Theorem 11 holds by observing that Bahadur representations (1.32) remain accurate uniformly in the sparse vectors $\boldsymbol{\beta}\in\mathscr{B}$; hence, all the steps of Theorem 6 apply in this case as well. □

1.9 Proofs of Lemmas

Proof of Lemma 1. Let $\{\widetilde{\boldsymbol{\delta}}_k\}_{k\in[N_\delta]}$ be the centers of the balls of radius $r_n\xi_n$ that cover the set $\mathscr{C}(r_n,t)$. Such a cover can be constructed with $N_\delta\leq\binom{p}{t}(3/\xi_n)^t$, see [VdV00] for example. Furthermore, let $\mathbb{D}_n(\boldsymbol{\delta})=n^{-1}\sum_{i=1}^n[\mu_i(\boldsymbol{\delta})-\mathbb{E}[\mu_i(\boldsymbol{\delta})]]$ and let

$$\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k,r)=\left\{\boldsymbol{\delta}\in\mathbb{R}^p:||\widetilde{\boldsymbol{\delta}}_k-\boldsymbol{\delta}||_2\leq r\ ,\ \mathrm{supp}(\boldsymbol{\delta})\subseteq\mathrm{supp}(\widetilde{\boldsymbol{\delta}}_k)\right\}$$

be a ball of radius r centered at $\widetilde{\boldsymbol{\delta}}_k$ with elements that have the same support as $\widetilde{\boldsymbol{\delta}}_k$. In what follows, we will bound $\sup_{\boldsymbol{\delta}\in\mathscr{C}(r_n,t)}|\mathbb{D}_n(\boldsymbol{\delta})|$ using an ε-net argument. In particular, using the above introduced notation, we have the following decomposition

$$\begin{aligned}\sup_{\boldsymbol{\delta}\in\mathscr{C}(r_n,t)}|\mathbb{D}_n(\boldsymbol{\delta})| &= \max_{k\in[N_\delta]}\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k,r_n\xi_n)}|\mathbb{D}_n(\boldsymbol{\delta})| \\ &\leq \underbrace{\max_{k\in[N_\delta]}|\mathbb{D}_n(\widetilde{\boldsymbol{\delta}}_k)|}_{T_1} + \underbrace{\max_{k\in[N_\delta]}\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k,r_n\xi_n)}|\mathbb{D}_n(\boldsymbol{\delta})-\mathbb{D}_n(\widetilde{\boldsymbol{\delta}}_k)|}_{T_2}.\end{aligned} \tag{1.42}$$

We first bound the term T_1 in (1.42). To that end, let $Z_{ik} = \left(\mu_i(\widetilde{\boldsymbol{\delta}}_k) - \mathbb{E}\left[\mu_i(\widetilde{\boldsymbol{\delta}}_k)\right]\right)$. With a little abuse of notation we use l to denote the density of $x_i\boldsymbol{\beta}^*$ for all i. Observe,

$$\mathbb{E}\left[\mu_i(\boldsymbol{\delta})\right] = \mathbb{P}\left(x_i\boldsymbol{\beta}^* \leq x_i\boldsymbol{\delta}\right) - \mathbb{P}\left(x_i\boldsymbol{\beta}^* \leq 0\right) = w_i(\boldsymbol{\delta}) - w_i(\mathbf{0}),$$

where $w_i(\boldsymbol{\delta}) := \mathbb{P}(x_i\boldsymbol{\beta}^* \leq x_i\boldsymbol{\delta})$, as a function of $\boldsymbol{\delta}$. Then $T_1 = \max_{k\in[N_\delta]}\left|n^{-1}\sum_{i\in[n]}Z_{ik}\right|$. Note that $\mathbb{E}[Z_{ik}] = 0$ and

$$\begin{aligned}\mathrm{Var}[Z_{ik}] &= \mathbb{E}\left[\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right) + \mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right) - 2\,\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right)\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right)\right] \\ &\quad - \left[\mathbb{E}\,\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right) - \mathbb{E}\,\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right)\right]^2 \\ &\overset{(i)}{\leq} \mathbb{E}\left[\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right) + \mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right) - 2\,\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right)\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right)\right] \\ &\quad + 2\mathbb{E}\left[\left(\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right) - \mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right)\right)\mathbb{I}\left(x_i\boldsymbol{\beta}^* \leq 0\right)\right] \\ &\overset{(ii)}{\leq} w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0}) + 2\left|w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})\right| \leq 3\left|w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})\right|,\end{aligned} \tag{1.43}$$

where (i) follows from dropping a negative term, and (ii) follows from taking absolute value within the second expectation. We can apply linearization techniques on the difference of

$w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})$.

$$\left| w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0}) \right| \overset{(iii)}{\leq} \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| l\left(c_i x_i \widetilde{\boldsymbol{\delta}}_k \right) \overset{(iv)}{\leq} \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| K_1 \quad (c_i \in [0,1]),$$

where (iii) follows by the mean value theorem and (iv) from the Condition **(E)**. Hence, we have that almost surely, $|Z_{ik}| \leq C \max_i \left| x_i \widetilde{\boldsymbol{\delta}}_k \right|$ for a constant $C < \infty$. For a fixed k, Bernstein's inequality, see Section 2.2.2 of [VDVW96] for example, gives us

$$\left| n^{-1} \sum_{i \in [n]} Z_{ik} \right| \leq C \left(\sqrt{\frac{K_1 \log(2/\delta)}{n^2} \sum_{i \in [n]} \left| x_i \widetilde{\boldsymbol{\delta}}_k \right|} \bigvee \frac{\log(2/\delta)}{n} \right)$$

with probability $1 - \delta$. Observe that for $\sum_{i \in [n]} \left| x_i \widetilde{\boldsymbol{\delta}}_k \right|$, we have

$$\sum_{i \in [n]} \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| \leq C^2 n \sqrt{\widetilde{\boldsymbol{\delta}}_k^\top X^\top X \widetilde{\boldsymbol{\delta}}_k} \leq C^2 n r_n t^{1/2} \tag{1.44}$$

where the line follows using the Cauchy-Schwartz inequality.

Hence, with probability $1 - 2\delta$ we have for all $\lambda_j \geq A\sqrt{\log p / n}$ that

$$\left| n^{-1} \sum_{i \in [n]} Z_{ik} \right| \leq C \left(\sqrt{\frac{r_n \sqrt{t} \log(2/\delta)}{n}} \bigvee \frac{\log(2/\delta)}{n} \right).$$

Using the union bound over $k \in [N_\delta]$, with probability $1 - 2\delta$, we have

$$T_1 \leq C \left(\sqrt{\frac{r_n \sqrt{t} \log(2N_\delta/\delta)}{n}} \bigvee \frac{\log(2N_\delta/\delta)}{n} \right).$$

Let us now focus on bounding T_2 term. Let $Q_i(\boldsymbol{\delta}) = \mu_i(\boldsymbol{\delta}) - \mathbb{E}\mu_i(\boldsymbol{\delta})$. For a fixed k we have

$$\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)} \left|\mathbb{D}_n(\boldsymbol{\delta}) - \mathbb{D}_n(\widetilde{\boldsymbol{\delta}}_k)\right| \leq \sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)} \left| n^{-1} \sum_{i\in[n]} Q_i(\boldsymbol{\delta}) - Q_i(\widetilde{\boldsymbol{\delta}}_k) \right| := T_{21}.$$

We further simply the expression, with a little abuse of notation,

$$\begin{aligned} Z'_{ik} := Q_i(\boldsymbol{\delta}) - Q_i(\widetilde{\boldsymbol{\delta}}_k) = &\left[\mathbb{I}(x_i\boldsymbol{\delta} \geq x_i\boldsymbol{\beta}^*) - \mathbb{I}(x_i\widetilde{\boldsymbol{\delta}}_k \geq x_i\boldsymbol{\beta}^*) \right] \\ &- \left[\mathbb{E}\,\mathbb{I}(x_i\boldsymbol{\delta} \geq x_i\boldsymbol{\beta}^*) + \mathbb{E}\,\mathbb{I}(x_i\widetilde{\boldsymbol{\delta}}_k \geq x_i\boldsymbol{\beta}^*) \right]. \end{aligned}$$

Then it is clear that $\mathbb{E}Z'_{ik} = 0$ and as shown earlier in $\mathrm{Var}(Z_{ik})$,

$$\mathrm{Var}(Z'_{ik}) \leq 3\left| w_i(\boldsymbol{\delta}) - w_i(\widetilde{\boldsymbol{\delta}}_k) \right| \leq 3K_1 \left| x_i\left(\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k\right) \right|$$

Moreover,

$$\left| x_i(\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k) \right| \leq K||\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k||_2 \sqrt{\left|\mathrm{supp}(\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k)\right|}$$

where K is a constant such that $\max_{i,j}|x_{ij}| \leq K$. Hence,

$$\max_{k\in[N_\delta]} \max_{i\in[n]} \sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)} \left| x_i\boldsymbol{\delta} - x_i\widetilde{\boldsymbol{\delta}}_k \right| \leq r_n\xi_n\sqrt{t}\max_{i,j}|x_{ij}| \leq Cr_n\xi_n\sqrt{t} =: \widetilde{L}_n,$$

The term T_{21} can be bounded in a similar way to T_1 by applying Bernstein's inequality and hence the details are omitted. With probability $1-2\delta$,

$$T_{21} \leq C\left(\sqrt{\frac{\widetilde{L}_n \log(2/\delta)}{n}} \bigvee \frac{\log(2/\delta)}{n} \right)$$

A bound on T_2 now follows using a union bound over $k \in [N_\delta]$. We can choose $\xi_n = n^{-1}$,

which gives us $N_\delta \lesssim \left(pn^2\right)^t$. With these choices, we obtain

$$T \leq C\left(\sqrt{\frac{r_n t\sqrt{t}\log(np/\delta)}{n}} \bigvee \frac{t\log(2np/\delta)}{n}\right),$$

which completes the proof.

□

Proof of Lemma 2. We begin by rewriting the term $n^{-1}\sum_{i=1}^n \psi_i(\boldsymbol{\beta})$, and aim to represent it through indicator functions. Observe that

$$n^{-1}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n x_i^\top \mathbb{1}(x_i\boldsymbol{\beta} > 0)[1 - 2\cdot \mathbb{1}(y_i - x_i\boldsymbol{\beta} < 0)]. \tag{1.45}$$

Using the fundamental theorem of calculus, we notice that if $x_i\boldsymbol{\beta}^* > 0$, $\int_{x_i(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^0 f(\varepsilon_i)d\varepsilon_i = F(0) - F(x_i(\boldsymbol{\beta}-\boldsymbol{\beta}^*)) = \frac{1}{2} - P(y_i < x_i\boldsymbol{\beta})$, where F is the univariate distribution of ε_i. Therefore, with expectation on ε, we can obtain an expression without the y_i.

$$\begin{aligned} n^{-1}\sum_{i=1}^n \mathbb{E}_\varepsilon \psi_i(\boldsymbol{\beta}) &= \left[n^{-1}\sum_{i=1}^n x_i^\top \mathbb{1}(x_i\boldsymbol{\beta} > 0)\cdot 2\int_{x_i(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^0 f(u)du\right] \\ &= \left[n^{-1}\sum_{i=1}^n x_i^\top \mathbb{1}(x_i\boldsymbol{\beta} > 0)\cdot 2f(u^*)x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right] := \Lambda_n(\boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta}), \end{aligned}$$

for some u^* between 0 and $x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta})$, and where we have defined

$$\Lambda_n(\boldsymbol{\beta}) = \left[n^{-1}\sum_{i=1}^n \mathbb{1}(x_i\boldsymbol{\beta} > 0)x_i^\top x_i \cdot 2f(u^*)\right].$$

We then show a bound for $\Delta := \left|\left[\mathbb{E}_X\Lambda_n(\boldsymbol{\beta}) - 2f(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\right]_{jk}\right|$, where we recall $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$ is

defined as earlier, $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*) = n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta}^* > 0)x_i^\top x_i$. By triangular inequality,

$$\Delta \leq \left| n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)x_{ij}x_{ik}\cdot 2(f(u^*) - f(0)) \right| \tag{1.46}$$

$$+ \left| n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)x_{ij}x_{ik}\cdot 2f(0) - n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta}^* > 0)x_{ij}x_{ik}\cdot 2f(0) \right|. \tag{1.47}$$

Notice that $\mathbb{1}(x_i\boldsymbol{\beta} > 0) - \mathbb{1}(x_i\boldsymbol{\beta}^* > 0) \leq \mathbb{1}(x_i\boldsymbol{\beta} \geq 2x_i\boldsymbol{\beta}^*) = \mathbb{1}[x_i\boldsymbol{\beta}^* \leq x_i(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]$. Moreover, the original expresion is also smaller than or equal to $\mathbb{1}\left(|x_i\boldsymbol{\beta}^*| \leq |x_i(\boldsymbol{\beta} - \boldsymbol{\beta}^*)|\right)$. The term (1.47) can be bounded by the design matrix setup and Condition **(E)**,

$$\left| n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)x_{ij}x_{ik}\cdot 2f(0) - n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta}^* > 0)x_{ij}x_{ik}\cdot 2f(0) \right|$$

$$\leq 2f(0)K^2 n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}\left(|x_i\boldsymbol{\beta}^*| \leq \|x_i\|_\infty \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_1\right) \leq 2f(0)K^2\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_1.$$

With the help of Hölder's inequality, $|(1.46)| \leq n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)\|x_i\|_\infty^2 \cdot 2\,|f(u^*) - f(0)|$. By triangular inequality and Condition **(E)** we can further upper bound the right hand side with

$$2\cdot n^{-1}\sum_{i=1}^n \mathbb{E}_X \|x_i\|_\infty^2 \cdot L_0\|x_i\|_\infty \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1.$$

Then we are ready to put terms together and obtain a bound for Δ. Additionally, by the design matrix setup we have

$$\Delta \leq (C + 2f(0))K^3\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1,$$

for $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 < \xi$ and a constant C. Essentially, this proves that Δ is not greater than a constant multiple of the difference between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Thus, we have as $n \to \infty$

$$n^{-1}\sum_{i=1}^n \mathbb{E}\psi_i(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbb{E}_X\mathbb{E}_\varepsilon \psi_i(\boldsymbol{\beta}) = 2f(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \mathscr{O}(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1)(\boldsymbol{\beta}^* - \boldsymbol{\beta}). \tag{1.48}$$

□

Proof of Lemma 3. For the simplicity in notation we fix $j = 1$ and denote $\widehat{\boldsymbol{\gamma}}_{(1)}(\widehat{\boldsymbol{\beta}})$ with $\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}})$. The proof is composed of two steps: the first establishes a cone set and an event set of interest whereas the second proves the rate of the estimation error by certain approximation results.

Step 1. Here we show that the estimation error $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ belongs to the appropriate cone set with high probability. We introduce the loss function $l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} \left(W_{i,1}(\boldsymbol{\beta}) - W_{i,-1}(\boldsymbol{\beta})\boldsymbol{\gamma}\right)^2$. The loss function above is convex in $\boldsymbol{\gamma}$ hence

$$(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \left[\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}} - \nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\right] \geq 0.$$

Let $h^* = \left\| \nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \right\|_\infty$. Let $\boldsymbol{\delta} = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$. KKT conditions provide $\left(\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*+\boldsymbol{\delta}}\right)_j = -\lambda_1 \operatorname{sgn}(\boldsymbol{\gamma}_j^* + \boldsymbol{\delta}_j)$ for all $j \in S_1^c \cap \{\widehat{\boldsymbol{\gamma}}_j \neq 0\}$ with $S_1 = \{j : \boldsymbol{\gamma}^* \neq 0\}$. Moreover, observe that $\boldsymbol{\delta}_j = 0$ for all $j \in S_1^c \cap \{\widehat{\boldsymbol{\gamma}}_j = 0\}$. Then,

$$\begin{aligned}
&(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \left[\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}} - \nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\right] \\
&= \sum_{j \in S_1^c} \boldsymbol{\delta}_j (\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*+\boldsymbol{\delta}})_j + \sum_{j \in S_1} \boldsymbol{\delta}_j (\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*+\boldsymbol{\delta}})_j + \boldsymbol{\delta}^\top (-\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}) \\
&\leq \sum_{j \in S_1^c} \boldsymbol{\delta}_j (-\lambda_1 \operatorname{sgn}(\boldsymbol{\gamma}_j^* + \boldsymbol{\delta}_j)) + \lambda_1 \sum_{j \in S_1} |\boldsymbol{\delta}_j| + h^* \|\boldsymbol{\delta}\|_1 \\
&= \sum_{j \in S_1^c} -\lambda_1 |\boldsymbol{\delta}_j| + \sum_{j \in S_1} \lambda_1 |\boldsymbol{\delta}_j| + h^* \|\boldsymbol{\delta}_{S_1}\|_1 + h^* \|\boldsymbol{\delta}_{S_1^c}\|_1 \\
&= (h^* - \lambda_1) \|\boldsymbol{\delta}_{S_1^c}\|_1 + (\lambda_1 + h^*) \|\boldsymbol{\delta}_{S_1}\|_1.
\end{aligned}$$

Hence on the event $h^* \leq (a-1)/(a+1)\lambda_1$ for a constant $a > 1$, the estimation error $\boldsymbol{\delta}$ belongs to the cone set

$$\mathscr{C}(a, S_1) = \{\mathbf{x} \in \mathbb{R}^{p-1} : \|\mathbf{x}_{S_1^c}\|_1 \leq a \|\mathbf{x}_{S_1}\|_1\} \tag{1.49}$$

Next, we proceed to show that the event above holds with high probability for certain choice of the tuning parameter λ_1. We begin by decomposing

$$h^* \leq \left\|\nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\beta}^*, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\right\|_\infty + \left\|\nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\beta}^*, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} - \nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\right\|_\infty$$

Let $H_1 = \nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\beta}^*, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}$ and let $H_2 = \nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\beta}^*, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} - \nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}$ We begin by observing that $\nabla_{\boldsymbol{\gamma}} l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} = \nabla_{\boldsymbol{\gamma}} l(\boldsymbol{\beta}^*, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$, for

$$\begin{aligned}
\Delta_1 &= -2n^{-1}\left(W_{-1}(\widehat{\boldsymbol{\beta}}) - W_{-1}(\boldsymbol{\beta}^*)\right)^\top W_1(\widehat{\boldsymbol{\beta}}) \\
\Delta_2 &= -2n^{-1}\left(W_{-1}(\boldsymbol{\beta}^*)\right)^\top \left(W_1(\widehat{\boldsymbol{\beta}}) - W_1(\boldsymbol{\beta}^*)\right) \\
\Delta_3 &= -2n^{-1}\left(W_{-1}(\widehat{\boldsymbol{\beta}})\right)^\top \left(W_{-1}(\widehat{\boldsymbol{\beta}}) - W_{-1}(\boldsymbol{\beta}^*)\right)\boldsymbol{\gamma}^* \\
\Delta_4 &= 2n^{-1}\left(W_{-1}(\widehat{\boldsymbol{\beta}}) - W_{-1}(\boldsymbol{\beta}^*)\right)^\top W_{-1}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*
\end{aligned}$$

Next, by Lemma 1 we observe

$$|\Delta_{1,j}| \leq 2K^2 n^{-1}\left|\sum_{i=1}^n \mu_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) - \mu_i(0)\right| = \mathscr{O}_P\left(K^2 r_n^{1/2} t^{3/4}(\log p/n)^{1/2}\bigvee K^2 t\log p/n\right),$$

and similarly $|\Delta_{2,j}| = \mathscr{O}_P\left(K^2 r_n^{1/2} t^{3/4}(\log p/n)^{1/2}\bigvee K^2 t\log p/n\right)$. Then, it is not difficult to see that such assumption provides $\|W_{-1}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*\|_\infty = \mathscr{O}_P(K)$. By Hölder's inequality followed by Lemma 1

$$\begin{aligned}
|\Delta_{3,j}| &\leq 2K^2 n^{-1}\left|\sum_{i=1}^n \left[\mu_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) - \mu_i(0)\right]\right| \\
&= \mathscr{O}_P\left(K^2 r_n^{1/2} t^{3/4}(\log p/n)^{1/2}\bigvee K^3 t\log p/n\right),
\end{aligned}$$

and similarly $|\Delta_{4,j}| = \mathscr{O}_P\left(K^2 r_n^{1/2} t^{3/4}(\log p/n)^{1/2}\bigvee K^2 t\log p/n\right)$. Putting all the terms together

we obtain

$$H_2 = \mathscr{O}_P\left(K^2 r_n^{1/2} t^{3/4} (\log p/n)^{1/2} \bigvee K^2 t \log p/n\right).$$

Next, we focus on the term H_1. Simple computation shows that for all $k = 2, \cdots p$, we have

$$H_{1,k} = -2n^{-1} \sum_{i=1}^{n} u_i$$

for $u_i = X_{ik} \boldsymbol{\zeta}_{1,i}^* \mathbb{I}\{x_i \boldsymbol{\beta}^* > 0\}$. Observe that the sequence $\{u_i\}$ across $i = 1, \cdots, n$, is a sequence of independent random variables. As ε_i and x_i are independent we have by the tower property $\mathbb{E}[r_i] = \mathbb{E}_X\left[X_{ik}\, \mathbb{I}\{x_i \boldsymbol{\beta}^* > 0\} \mathbb{E}_\varepsilon[\boldsymbol{\zeta}_{1,i}^*]\right] = 0$. Moreover, as $\boldsymbol{\zeta}_1^*$ is sub-exponential random vector, by Bernstein's inequality and union bound we have

$$P(\|H_1\|_\infty \geq c) \leq p \exp\left\{-\frac{n}{2}\left(\frac{c^2}{\widetilde{K}^2} \vee \frac{c}{\widetilde{K}}\right)\right\}$$

where $\|u_i\|_{\psi_1} \leq K \|\boldsymbol{\zeta}_{1,i}^*\|_{\psi_1} := \widetilde{K} < \infty$. We pick c to be $(\log p/n)^{1/2}$, then we have with probability converging to 1 that

$$\begin{aligned} h^* &\leq \|H_1\|_\infty + \|H_2\|_\infty \leq (\log p/n)^{1/2} + C_1 r_n^{1/2} t^{3/4} (\log p/n)^{1/2} + C_2 t \log p/n \\ &\leq (a-1)/(a+1)\lambda_1, \end{aligned}$$

for some constant C_1 and C_2. Thus, with λ_1 chosen as

$$\lambda_1 = C\left((\log p/n)^{1/2} \bigvee \left(r_n^{1/2} \bigvee t^{1/4} (\log p/n)^{1/2}\right) t^{3/4} (\log p/n)^{1/2}\right),$$

for some constant $C > 1$, we have that $h^* \leq (a-1)/(a+1)\lambda_1$ with probability converging to 1. More directly, with the condition on the penalty parameter λ_1, this implies that the event for the cone set (1.49) to be true holds with high probability.

Step 2. We begin by a basic inequality

$$l(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})+\lambda_1\|\widehat{\boldsymbol{\gamma}}\|_1 \leq l(\widehat{\boldsymbol{\beta}},\boldsymbol{\gamma}^*)+\lambda_1\|\boldsymbol{\gamma}^*\|_1$$

guaranteed as $\widehat{\boldsymbol{\gamma}}$ minimizes the penalized loss (1.8). Here and below in the rest of the proof we suppress the subscript 1 and $\boldsymbol{\beta}$ in the notation of $W_1(\widehat{\boldsymbol{\beta}})$ and $W_{-1}(\widehat{\boldsymbol{\beta}})$ and use $\widehat{W}$ and $\widehat{W}^-$ instead and similarly $W^* := W_1(\boldsymbol{\beta}^*)$ and $W^{-*} = W_{-1}(\boldsymbol{\beta}^*)$. Rewriting the inequality above we obtain

$$\begin{aligned}
&-2n^{-1}\widehat{W}^\top\widehat{W^-}\widehat{\boldsymbol{\gamma}}+n^{-1}\widehat{\boldsymbol{\gamma}}^\top\widehat{W^-}^\top\widehat{W^-}\widehat{\boldsymbol{\gamma}}\\
&\leq -2n^{-1}\widehat{W}^\top\widehat{W^-}\boldsymbol{\gamma}^*+n^{-1}\boldsymbol{\gamma}^{*\top}\widehat{W^-}^\top\widehat{W^-}\boldsymbol{\gamma}^*-\lambda_1\|\widehat{\boldsymbol{\gamma}}\|_1+\lambda_1\|\boldsymbol{\gamma}^*\|_1
\end{aligned}$$

Observe that $W_{ij}(\widehat{\boldsymbol{\beta}}) = W_{ij}(\boldsymbol{\beta}^*)+X_{ij}[\mu_i(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})-\mu_i(0)]$. Let $\alpha_{ij} = X_{ij}[\mu_i(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})-\mu_i(0)]$. Let $\mathbf{A}$ be a matrix such that $\mathbf{A} = \{\alpha_{ij}\}_{1\leq i\leq n, 1\leq j\leq p}$. From now on we only consider $\mathbf{A}$ to mean $\mathbf{A}_1$ and $\mathbf{A}^-$ to mean $\mathbf{A}_{-1}$. Next, note that $W_i^* = W_i^{-*}\boldsymbol{\gamma}^*+\zeta_i^*$ by the definition of $\boldsymbol{\gamma}^*$ in the node-wise plug-in lasso problem. Together with the above, we observe that $\widehat{W}_i = W_i^{-*}\boldsymbol{\gamma}^*+\zeta_i^*+\mathbf{A}_i := W_i^{-*}\boldsymbol{\gamma}^*+\varepsilon_i^*$. Hence, the basic inequality above becomes,

$$\begin{aligned}
&-2n^{-1}\left(W^{-*}\boldsymbol{\gamma}^*+\boldsymbol{\varepsilon}^*\right)^\top(W^{-*}+\mathbf{A}^-)\widehat{\boldsymbol{\gamma}}+n^{-1}\widehat{\boldsymbol{\gamma}}^\top(W^{-*}+\mathbf{A}^-)^\top(W^{-*}+\mathbf{A}^-)\widehat{\boldsymbol{\gamma}}\\
&\leq -2n^{-1}\left(W^{-*}\boldsymbol{\gamma}^*+\boldsymbol{\varepsilon}^*\right)^\top(W^{-*}+\mathbf{A}^-)\boldsymbol{\gamma}^*+n^{-1}\boldsymbol{\gamma}^{*\top}(W^{-*}+\mathbf{A}^-)^\top(W^{-*}+\mathbf{A}^-)\boldsymbol{\gamma}^*\\
&-\lambda_1\|\widehat{\boldsymbol{\gamma}}\|_1+\lambda_1\|\boldsymbol{\gamma}^*\|_1.
\end{aligned}$$

With reordering the terms in the inequality above, we obtain

$$n^{-1}\left\|W^{-*}\widehat{\boldsymbol{\gamma}}-W^{-*}\boldsymbol{\gamma}^*\right\|_2^2 \leq \delta_1+\delta_2+\delta_3-\lambda_1\|\widehat{\boldsymbol{\gamma}}\|_1+\lambda_1\|\boldsymbol{\gamma}^*\|_1,$$

$$\begin{aligned}
\text{for} \qquad \delta_1 &= 2n^{-1}\varepsilon_1^{*\top}\left(W^{-*}+\mathbf{A}^-\right)\left(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\right),\\
\delta_2 &= 2n^{-1}\boldsymbol{\gamma}^{*\top}W^{-*\top}\mathbf{A}^-\left(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\right),\\
\delta_3 &= n^{-1}\left(\boldsymbol{\gamma}^*+\widehat{\boldsymbol{\gamma}}\right)^\top\left(\mathbf{A}^{-\top}\mathbf{A}^-+2W^{-*\top}A^-\right)\left(\boldsymbol{\gamma}^*-\widehat{\boldsymbol{\gamma}}\right).
\end{aligned}$$

Next, we observe that A_i are bounded, mean zero random variables and hence

$$n^{-1}|\sum_{i=1}^n A_i| = \mathscr{O}_P(n^{-1/2}).$$

Moreover ε_i^* is a sum of sub-exponential and bounded random variables, hence is sub-exponential. Thus, utilizing the above and results of Step 1 we obtain

$$\delta_1 \leq K^2(a+1)\|\widehat{\boldsymbol{\gamma}}_{S_1}-\boldsymbol{\gamma}^*_{S_1}\|_1\mathscr{O}_P(n^{-1/2}),$$

$$\delta_2 \leq K^2(a+1)\|\widehat{\boldsymbol{\gamma}}_{S_1}-\boldsymbol{\gamma}^*_{S_1}\|_1\|\boldsymbol{\gamma}^*_{S_1}\|_1\mathscr{O}_P(n^{-1/2}),$$

Lastly, observe that

$$\delta_3 \leq n^{-1}\boldsymbol{\gamma}^{*\top}\left(\mathbf{A}^{-\top}\mathbf{A}^-+2W^{-*\top}A^-\right)\boldsymbol{\gamma}^*+n^{-1}\widehat{\boldsymbol{\gamma}}^\top\left(\mathbf{A}^{-\top}\mathbf{A}^-+2W^{-*\top}A^-\right)\widehat{\boldsymbol{\gamma}} \tag{1.50}$$

Moreover, as $\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*$ belongs to the cone $C(a,S_1)$ (1.49) by Step 1, by convexity arguments it is easy to see that $\widehat{\boldsymbol{\gamma}}$ belongs to the same cone. Together with Hölder's inequality we obtain

$$\delta_3 \leq 3Kn^{-1}\sum_{i=1}^n W^{-*\top}_{i,S_1}\mathbf{A}^-_{i,S_1}\left[\|\boldsymbol{\gamma}^*_{S_1}\|_2^2+\|\widehat{\boldsymbol{\gamma}}_{S_1}\|_2^2\right]$$

Utilizing Lemma 1 now provides

$$\delta_3 \leq \kappa \left[\|\boldsymbol{\gamma}^*_{S_1}\|_2^2 + \|\widehat{\boldsymbol{\gamma}}_{S_1}\|_2^2 \right]$$

where κ is such that $\kappa = \mathcal{O}_P(K^2 r_n^{1/2} t^{3/4} (\log p/n)^{1/2})$. Moreover, observe that if λ_1 is chosen to be larger than the upper bound of κ. Putting all the terms together we obtain

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \left(W_i^{-*} \widehat{\boldsymbol{\gamma}} - W_i^{-*} \boldsymbol{\gamma}^* \right)^2 &\leq 2\lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_1 + \lambda_1 \|\boldsymbol{\gamma}^*_{S_1}\|_2^2 + \lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1}\|_2^2 - \lambda_1 \|\widehat{\boldsymbol{\gamma}}\|_1 + \lambda_1 \|\boldsymbol{\gamma}^*\|_1 \\ &\leq 3\lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_1 + \lambda_1 \|\boldsymbol{\gamma}^*_{S_1}\|_2^2 + \lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1}\|_2^2 \end{aligned}$$

where the last inequality holds as $|\widehat{\gamma}_j - \gamma^*_j| \geq |\gamma^*_j| - |\widehat{\gamma}_j|$ for $j \in S_1$, and disregarding the negative terms $-\lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1^c}\|_1$.

Moreover, by Condition **(C)** and Step 1 we have that the left hand side is bigger than or equal to $C_2 n^{-1} \sum_{i=1}^{n} \left(X_i^{-} \widehat{\boldsymbol{\gamma}} - X_i^{-} \boldsymbol{\gamma}^* \right)^2$, allowing us to conclude

$$n^{-1} C_2 \left\| X(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right\|_2^2 \leq 3\lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_1 + 2\lambda_1 \|\boldsymbol{\gamma}^*_{S_1}\|_2^2 + \lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_2^2 \tag{1.51}$$

holds with probability approaching one. Let $S = S_{\boldsymbol{\beta}^*}$ for short. Condition **(Γ)** and **(CC)** together imply that now we have

$$(\phi_0^2 C_2 - \lambda_1) \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_2^2 \leq 3\sqrt{s_1}\lambda_1 \|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_2 + 2\lambda_1 \|\boldsymbol{\gamma}^*_{S_1}\|_2^2.$$

Solving for $\|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_2$ in the above inequality we obtain

$$\|\widehat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}^*_{S_1}\|_2 \leq 3\sqrt{s_1}\lambda_1 / (\phi_0^2 C_2 - \lambda_1)$$

The result then follows from a simple norm inequality

$$\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 \leq (a+1)\|\widehat{\boldsymbol{\gamma}}_{S_1}-\boldsymbol{\gamma}^*_{S_1}\|_1 \leq (a+1)\sqrt{s_1}\|\widehat{\boldsymbol{\gamma}}_{S_1}-\boldsymbol{\gamma}^*_{S_1}\|_2$$

and considering an asymptotic regime with $n, p, s_{\boldsymbol{\beta}^*}, s_1 \to \infty$.

□

Proof of Lemma 4 . Recall the definitions of $\widehat{\boldsymbol{\zeta}}_j$ and $\boldsymbol{\zeta}^*_j$. Observe that we have the following inequality,

$$\begin{aligned}\left|\widehat{\boldsymbol{\zeta}}_j^\top\widehat{\boldsymbol{\zeta}}_j/n-\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j/n\right| &\leq \left|n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top\widehat{\boldsymbol{\zeta}}_j-n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j\right|+\left|n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j-n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j\right| \\ &\leq n^{-1}\left\|\widehat{\boldsymbol{\zeta}}_j+\boldsymbol{\zeta}^*_j\right\|_\infty\left\|\widehat{\boldsymbol{\zeta}}_j-\boldsymbol{\zeta}^*_j\right\|_1+\left|n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j-n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}^*_j\right|,\end{aligned}$$

using triangular inequality and Hölder's inequality.

We proceed to upper bound all of the three terms on the right hand side of the previous inequality. First, we observe

$$\left\|\widehat{\boldsymbol{\zeta}}_j+\boldsymbol{\zeta}^*_j\right\|_\infty \leq \left\|W_j(\boldsymbol{\beta}^*)-W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_\infty+\left\|W_j(\widehat{\boldsymbol{\beta}})-W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_\infty. \tag{1.52}$$

Moreover, the conditions imply that $\|W_j(\widehat{\boldsymbol{\beta}})\|_\infty \leq K$ (by the design matrix condition),

$$\|W_{-j}\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\|_\infty \leq K\left(\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\|_1+\mathscr{O}_P(K)\right)$$

and by Lemma 3, for λ_j as defined, the right hand size is $\mathscr{O}_P\left(Ks_j\lambda_j \vee K\right)$. Thus, we conclude $\left\|\widehat{\boldsymbol{\zeta}}_j+\boldsymbol{\zeta}^*_j\right\|_\infty = \mathscr{O}_P\left(K \bigvee Ks_j\lambda_j \bigvee K\right) = \mathscr{O}_P\left(K \vee K \vee Ks_j\lambda_j\right)$.

Its multiplying term can be decomposed as following

$$n^{-1}\left\|\widehat{\boldsymbol{\zeta}}_j-\boldsymbol{\zeta}_j^*\right\|_1 \leq \underbrace{n^{-1}\left\|X_j\circ\left(\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)-\mathbb{I}(X\boldsymbol{\beta}^*>0)\right)\right\|_1}_{i} + \underbrace{n^{-1}\left\|W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})-W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1}_{ii}, \tag{1.53}$$

where $\circ$ denotes entry wise multiplication between two vectors. The reason we have to spend such a great effort in separating the terms to bound this quantity is that we are dealing with a 1-norm here, rather than an infinity-norm, which is bounded easily.

We start with term i. Notice that

$$n^{-1}\left\|X_j\circ\left(\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)-\mathbb{I}(X\boldsymbol{\beta}^*>0)\right)\right\|_1 \leq Kn^{-1}\sum_{i=1}^{n}\left|\mathbb{I}(x_i\widehat{\boldsymbol{\beta}}>0)-\mathbb{I}(x_i\boldsymbol{\beta}^*>0)\right|,$$

by Hölder's inequality and the design matrix condition. Moreover, by Lemma 1 we can easily bound the term above with $\mathscr{O}_P\left(Kr_n^{1/2}t^{3/4}(\log p/n)^{1/2}\bigvee Kt\log p/n\right)$, with r_n and t as defined in Condition **(I)**.

For the term ii, we have

$$\begin{aligned} ii \leq & n^{-1}\left\|X_{-j}\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\circ\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)-X_{-j}\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\circ\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)\right\|_1 \\ & + n^{-1}\left\|X_{-j}\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\circ\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)-X_{-j}\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\circ\mathbb{I}(X\boldsymbol{\beta}^*>0)\right\|_1. \end{aligned}$$

Observe, that the right hand side is upper bounded with

$$\begin{aligned} & K\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1\left\|\mathbb{I}(X\widehat{\boldsymbol{\beta}}>0)\right\|_\infty \\ & +\left\|X_{-j}\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_\infty\left|n^{-1}\sum_{i=1}^{n}\left[\mathbb{I}(x_i\widehat{\boldsymbol{\beta}}>0)-\mathbb{I}(x_i\boldsymbol{\beta}^*>0)\right]\right| \end{aligned}$$

by the design matrix condition. Utilizing Lemma 1, Lemma 3 and Condition ($\mathbf{\Gamma}$) together we obtain

$$ii = \mathscr{O}_P\left(Ks_j\lambda_j\right) + \mathscr{O}_P\left(Kr_n^{1/2}t^{3/4}(\log p/n)^{1/2}\bigvee Kt\log p/n\right),$$

for the chosen λ_j. Combining bounds for the terms i and ii, we obtain

$$n^{-1}\left\|\widehat{\boldsymbol{\zeta}}_j - \boldsymbol{\zeta}_j^*\right\|_1 = \mathscr{O}_P\left(Ks_j\lambda_j\bigvee Kr_n^{1/2}t^{3/4}(\log p/n)^{1/2}\bigvee Kt\log p/n\right)$$

Next, we bound $\left|n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^* - n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*\right|$. If we rewrite the inner product in summation form, we have $\left|n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^* - n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*\right| = n^{-1}\sum_{i=1}^n\left(\zeta_{ij}^{*2} - \mathbb{E}\zeta_{ij}^{*2}\right)$. Notice that $\zeta_{ij}^* = W_{ij}(\boldsymbol{\beta}^*) - W_{i,-j}\boldsymbol{\gamma}_{(j)}^*(\boldsymbol{\beta}^*)$ is a bounded random variable and such that $|\zeta_{ij}^*| = \mathscr{O}_P(K + Ks_j^{1/2})$. We then apply Hoeffding's inequality for bounded random variables, to obtain

$$\left|n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^* - n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*\right| = O_P(K^2s_jn^{-1/2}).$$

□

Proof of Lemma 5 . We begin by first establishing that $\widehat{\tau}_j^{-2} = \mathscr{O}_P(1)$. In the case when the penalty part $\lambda_j\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_1$ happens to be 0, which means $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) = 0$, the worst case scenario is that the regression part, $n^{-1}\left\|W_j(\widehat{\boldsymbol{\beta}}) - W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_2^2$, also results in 0, i.e.

$$0 = W_j(\widehat{\boldsymbol{\beta}}) - W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) \tag{1.54}$$

We show that these terms cannot be equal to zero simultaneously, since this forces $W_j(\widehat{\boldsymbol{\beta}}) = 0$, which is not true. Thus, $\widehat{\tau}_j^{-2}$ is bounded away from 0.

In order to show results about the matrices $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{\Omega}(\boldsymbol{\beta}^*)$, we first provide a bound on the $\widehat{\tau}$ and τ. This is critical, since the magnitude of $\boldsymbol{\Omega}(\cdot)$ is determined by τ. To derive the bound on the τ's, we have to decompose the terms very carefully and put a bound on each one of them.

Recall definitions of $\widehat{\boldsymbol{\zeta}}_j$ and $\boldsymbol{\zeta}_j^*$

$$\widehat{\boldsymbol{\zeta}}_j = W_j(\widehat{\boldsymbol{\beta}}) - W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}), \qquad \boldsymbol{\zeta}_j^* = W_j(\boldsymbol{\beta}^*) - W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}_{(j)}^*(\boldsymbol{\beta}^*).$$

Moreover, by the Karush-Kuhn-Tucker conditions of problem (1.8) we have $\lambda_j \|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\|_1 = n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}})$, which in turn enables a representation

$$\widehat{\tau}_j^2 = n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top \widehat{\boldsymbol{\zeta}}_j + n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}).$$

By definition we have that $\tau_j^2 = n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^*$, for which we have $\widehat{\tau}_j^2$ as an estimate. The τ_j^2 and $\widehat{\tau}_j^2$ carry information about the magnitude of the values in $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$ and $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ respectively. We next break down τ_j^2 and $\widehat{\tau}_j^2$ into parts related to difference between $\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{\gamma}_{(j)}^*(\boldsymbol{\beta}^*)$, which we know how to control. Thus, we have the following decomposition,

$$|\widehat{\tau}_j^2 - \tau_j^2| \leq \underbrace{\left|n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top \widehat{\boldsymbol{\zeta}}_j - \tau_j^2\right|}_{I} + \underbrace{\left|n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right|}_{II}.$$

The task now boils down to bounding each one of the terms I and II , independently. Term I is now bounded by Lemma 4 and is in order of $\mathscr{O}_P\left(K^2 s_j \lambda_j\right)$. Regarding term II, we first point out one result due to the Karush-Kuhn-Tucker conditions of (6),

$$\lambda_j \cdot 1^\top \geq \lambda_j \text{sign}\left(\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right)^\top = n^{-1}\left(W_j(\widehat{\boldsymbol{\beta}}) - W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right)^\top W_{-j}(\widehat{\boldsymbol{\beta}}) = n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}}).$$

For the term II, we then have

$$\left|n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right| \leq \left\|n^{-1}\widehat{\boldsymbol{\zeta}}_j^\top W_{-j}(\widehat{\boldsymbol{\beta}})\right\|_\infty \left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_1 = \mathscr{O}_P\left(s_j^{1/2}\lambda_j \vee s_j\lambda_j^2\right),$$

since by Lemma 3 we have

$$\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}})\right\|_1 \leq \left\|\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1 + \left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1 = \mathscr{O}_P(s_j^{1/2}) + \mathscr{O}_P(s_j\lambda_j).$$

Putting all the pieces together, we have shown that rate

$$\left|\widehat{\tau}_j^2 - \tau_j^2\right| = \mathscr{O}_P\left(K^2 s_j\lambda_j \vee s_j^{1/2}\lambda_j \vee s_j\lambda_j^2\right)$$

As $\widehat{\tau}_j^{-2} = \mathscr{O}_P(1)$ we have $\left|\frac{1}{\widehat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| = \mathscr{O}_P\left(\left|\tau_j^2 - \widehat{\tau}_j^2\right|\right)$. We then conclude

$$\begin{aligned}\left\|\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})_j - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_j\right\|_1 &\leq \widehat{\tau}_j^{-2}\left\|\widehat{\boldsymbol{\gamma}}_{(j)}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1 + \left\|\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\right\|_1\left|\frac{1}{\widehat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| \\ &= \mathscr{O}_P\left(K^2 s_j^{3/2}\lambda_j \vee s_j\lambda_j \vee s_j^{3/2}\lambda_j^2\right)\end{aligned}$$

□

Proof of Lemma 6. For the simplicity of the proof we introduce some additional notation. Let $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, and

$$\nu_n(\boldsymbol{\delta}) = n^{-1}\sum_{i=1}^n \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})\left[\psi_i(\widehat{\boldsymbol{\beta}}) - \psi_i(\boldsymbol{\beta}^*)\right].$$

Observe that $\mathbb{I}\left\{y_i - x_i\widehat{\boldsymbol{\beta}} \leq 0\right\} = \mathbb{I}\left\{x_i\boldsymbol{\delta} \geq \varepsilon_i\right\}$ and hence

$$1 - 2\,\mathbb{I}\{y_i - x_i\widehat{\boldsymbol{\beta}} > 0\} = 2\,\mathbb{I}\left\{y_i - x_i\widehat{\boldsymbol{\beta}} \leq 0\right\} - 1.$$

The term we wish to bound then can be expressed as

$$\mathbb{V}_n(\boldsymbol{\delta}) = \nu_n(\boldsymbol{\delta}) - \mathbb{E}\nu_n(\boldsymbol{\delta})$$

for $\nu_n(\boldsymbol{\delta})$ denoting the following quantity

$$\nu_n(\boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{\Omega}(\boldsymbol{\delta} + \boldsymbol{\beta}^*) x_i^\top \left[f_i(\boldsymbol{\delta}) g_i(\boldsymbol{\delta}) - f_i(\mathbf{0}) g_i(\mathbf{0}) \right]$$

and

$$f_i(\boldsymbol{\delta}) = \mathbb{I}\left\{ x_i \boldsymbol{\delta} \geq -x_i \boldsymbol{\beta}^* \right\}, \qquad g_i(\boldsymbol{\delta}) = 2\,\mathbb{I}\left\{ x_i \boldsymbol{\delta} \geq \varepsilon_i \right\} - 1.$$

Let $\{\widetilde{\boldsymbol{\delta}}_k\}_{k \in [N_\delta]}$ be centers of the balls of radius $r_n \xi_n$ that cover the set $\mathscr{C}(r_n, t)$. Such a cover can be constructed with $N_\delta \leq \binom{p}{t} (3/\xi_n)^t$, see [VdV00] for example. Furthermore, let

$$\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r) = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : ||\widetilde{\boldsymbol{\delta}}_k - \boldsymbol{\delta}||_2 \leq r \,,\ \mathrm{supp}(\boldsymbol{\delta}) \subseteq \mathrm{supp}(\widetilde{\boldsymbol{\delta}}_k) \right\}$$

be a ball of radius r centered at $\widetilde{\boldsymbol{\delta}}_k$ with elements that have the same support as $\widetilde{\boldsymbol{\delta}}_k$. In what follows, we will bound $\sup_{\boldsymbol{\delta} \in \mathscr{C}(r_n, t)} ||\mathbb{V}_n(\boldsymbol{\delta})||_\infty$ using an ε-net argument. In particular, using the above introduced notation, we have the following decomposition

$$\begin{aligned} \sup_{\boldsymbol{\delta} \in \mathscr{C}(r_n, t)} ||\mathbb{V}_n(\boldsymbol{\delta})||_\infty &= \max_{k \in [N_\delta]} \sup_{\boldsymbol{\delta} \in \mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n \xi_n)} ||\mathbb{V}_n(\boldsymbol{\delta})||_\infty \\ &\leq \underbrace{\max_{k \in [N_\delta]} ||\mathbb{V}_n(\widetilde{\boldsymbol{\delta}}_k)||_\infty}_{T_1} + \underbrace{\max_{k \in [N_\delta]} \sup_{\boldsymbol{\delta} \in \mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n \xi_n)} ||\mathbb{V}_n(\boldsymbol{\delta}) - \mathbb{V}_n(\widetilde{\boldsymbol{\delta}}_k)||_\infty}_{T_2}. \end{aligned} \tag{1.55}$$

Observe that the term T_1 arises from discretization of the sets $\mathscr{C}(r_n, t)$. To control it, we will apply the tail bounds for each fixed l and k. The term T_2 captures the deviation of the process in a small neighborhood around the fixed center $\widetilde{\boldsymbol{\delta}}_k$. For those deviations we will provide covering number arguments. In the remainder of the proof, we provide details for bounding T_1 and T_2.

We first bound the term T_1 in (1.55). Let $a_{ij}(\boldsymbol{\beta}) = \mathbf{e}_j^\top \boldsymbol{\Omega}(\boldsymbol{\beta}) x_i^\top$ We are going to decouple

dependence on x_i and ε_i. To that end, let

$$Z_{ijk} = a_{ij}(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)\left(\left(f_i(\widetilde{\boldsymbol{\delta}}_k)g_i(\widetilde{\boldsymbol{\delta}}_k) - \mathbb{E}\left[f_i(\widetilde{\boldsymbol{\delta}}_k)g_i(\widetilde{\boldsymbol{\delta}}_k)|x_i\right]\right) - \left(f_i(\mathbf{0})g_i(\mathbf{0}) - \mathbb{E}\left[f_i(\mathbf{0})g_i(\mathbf{0})|x_i\right]\right)\right)$$

and

$$\begin{aligned}\widetilde{Z}_{ijk} = a_{ij}(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)\left(\mathbb{E}\left[f_i(\widetilde{\boldsymbol{\delta}}_k)g_i(\widetilde{\boldsymbol{\delta}}_k)|x_i\right] - \mathbb{E}\left[f_i(\mathbf{0})g_i(\mathbf{0})|x_i\right]\right)\\ - \mathbb{E}\left[a_{ij}(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)\left(f_i(\widetilde{\boldsymbol{\delta}}_k)g_i(\widetilde{\boldsymbol{\delta}}_k) - f_i(\mathbf{0})g_i(\mathbf{0})\right)\right].\end{aligned}$$

With a little abuse of notation we use f to denote the density of ε_i for all i. Observe that $\mathbb{E}\left[f_i(\boldsymbol{\delta})g_i(\boldsymbol{\delta})|x_i\right] = f_i(\boldsymbol{\delta})\mathbb{P}(\varepsilon_i \leq x_i\boldsymbol{\delta})$. We use $w_i(\boldsymbol{\delta})$ to denote the right hand side of the previous equation. Then

$$T_1 = \max_{k\in[N_\delta]}\max_{j\in[p]}\left|n^{-1}\sum_{i\in[n]}\left(Z_{ijk} + \widetilde{Z}_{ijk}\right)\right| \leq \underbrace{\max_{k\in[N_\delta]}\max_{j\in[p]}\left|n^{-1}\sum_{i\in[n]}Z_{ijk}\right|}_{T_{11}} + \underbrace{\max_{k\in[N_\delta]}\max_{j\in[p]}\left|n^{-1}\sum_{i\in[n]}\widetilde{Z}_{ijk}\right|}_{T_{12}}.$$

Note that $\mathbb{E}[Z_{ijk} \mid \{x_i\}_{i\in[n]}] = 0$. With a little abuse of notation we use l to denote the density of $x_i\boldsymbol{\beta}^*$ for all i.

$$\begin{aligned}\mathrm{Var}[Z_{ijk} \mid \{x_i\}_{i\in[n]}] &\overset{(i)}{\leq} 3a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)\left|w_i(\widetilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})\right|\\ &\overset{(ii)}{\leq} 3a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)f_i(\widetilde{\boldsymbol{\delta}}_k)\left|x_i\widetilde{\boldsymbol{\delta}}_k\right| l\left(\eta_i x_i\widetilde{\boldsymbol{\delta}}_k\right) \quad (\eta_i \in [0,1])\\ &\overset{(iii)}{\leq} 3a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)\left|x_i\widetilde{\boldsymbol{\delta}}_k\right| K_1\end{aligned}$$

where (i) follows similarly as in equation (1.43) in proof of Lemma 1, (ii) follows by the mean value theorem, and (iii) from the assumption that the conditional density is bounded stated in Condition **(E)** and taking the indicator to be 1.

Furthermore, conditional on $\{x_i\}_{i\in[n]}$ we have that almost surely,

$$|Z_{ijk}| \leq 2 \max_{ij} |a_{ij}(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)|.$$

We will work on the event

$$\mathscr{A} = \left\{ \max_{j\in[p]} \left\| \boldsymbol{\Omega}_j(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k) - \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\beta}^*) \right\|_1 \leq C_n \right\} \tag{1.56}$$

which holds with probability at $1-\delta$ using Lemma 5. For a fixed j and k Bernstein's inequality, see Section 2.2.2 of [VDVW96] for example, gives us

$$\left| n^{-1} \sum_{i\in[n]} Z_{ijk} \right| \leq C \left(\sqrt{\frac{K_1 \log(2/\delta)}{n^2} \sum_{i\in[n]} a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k) \left| x_i \widetilde{\boldsymbol{\delta}}_k \right|} \bigvee \frac{\max_{i\in[n], j\in[p]} |a_{ij}(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k)|}{n} \log(2/\delta) \right)$$

with probability $1-\delta$. On the event $\mathscr{A}$

$$\begin{aligned}
\sum_{i\in[n]} a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k) \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| &= \sum_{i\in[n]} \left(\left(\boldsymbol{\Omega}_j(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) \right) x_i^\top + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) x_i^\top \right)^2 \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| \\
&\leq \sum_{i\in[n]} \left(\left\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) x_i^\top \right\|_2^2 + K^2 C_n^2 \right) \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| \\
&\leq \sum_{i\in[n]} K^2 \left(\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)) + C_n^2 \right) \left| x_i \widetilde{\boldsymbol{\delta}}_k \right| \\
&\leq K^2 \left(\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*) + C_n^2 \right) n r_n t^{1/2}
\end{aligned}$$

where the line follows using the Cauchy-Schwartz inequality, equation (1.44) in proof of Lemma 1, and results of Lemma 5. Combining all of the results above, with probability $1-2\delta$ we have

that

$$\left| n^{-1} \sum_{i \in [n]} Z_{ijk} \right| \leq C \left(\sqrt{\frac{C_n^2 r_n \sqrt{t} \log(2/\delta)}{n}} \bigvee \frac{C_n \log(2/\delta)}{n} \right).$$

Using the union bound over $j \in [p]$ and $k \in [N_\delta]$, with probability $1 - 2\delta$, we have

$$T_{11} \leq C \left(\sqrt{\frac{C_n r_n \sqrt{t} \log(2N_\delta p/\delta)}{n}} \bigvee \frac{C_n \log(2N_\delta p/\delta)}{n} \right).$$

We deal with the term T_{12} in a similar way. For a fixed k and j, conditional on the event $\mathscr{A}$ we apply Bernstein's inequality to obtain

$$\left| n^{-1} \sum_{i \in [n]} \widetilde{Z}_{ijk} \right| \leq C \left(\sqrt{\frac{C_n^2 r_n^2 t \log(2/\delta)}{n}} \bigvee \frac{C_n \log(2/\delta)}{n} \right)$$

with probability $1 - \delta$, since on the event $\mathscr{A}$ in (1.56) we have that $\left| \widetilde{Z}_{ijk} \right| \leq C_n \Lambda_{\max}(\mathbf{\Sigma}(\boldsymbol{\beta}^*))$ and

$$\begin{aligned} \mathrm{Var}\left[\widetilde{Z}_{ijk} \right] &\leq \mathbb{E}\left[a_{ij}^2(\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}_k) \left(f_i(\widetilde{\boldsymbol{\delta}}_k) \mathbb{P}(\varepsilon_i \leq x_i \widetilde{\boldsymbol{\delta}}_k) - f_i(0) \mathbb{P}(\varepsilon_i \leq 0) \right)^2 \right] \\ &\leq K^2 \left(\Lambda_{\min}^{-1}(\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*) + C_n^2 \right) \left(3 \left| G_i(r_n, \boldsymbol{\beta}^*, 0) - G_i(0, \boldsymbol{\beta}^*, 0) \right| + f_{\max}^2 r_n t^{1/2} \right)^2 \leq C C_n^2 r_n^2 t \end{aligned}$$

where in the last step we utilized Condition (**E**) with $z = r_n$. The union bound over $k \in [N_\delta]$, and $j \in [p]$, gives us

$$T_{12} \leq C \left(\sqrt{\frac{C_n^2 r_n^2 t \log(2N_\delta p/\delta)}{n}} \bigvee \frac{C_n \log(2N_\delta p/\delta)}{n} \right)$$

with probability at least $1 - 2\delta$. Combining the bounds on T_{11} and T_{12}, with probability $1 - 4\delta$,

we have

$$T_1 \leq C\left(\sqrt{\frac{C_n^2(r_n t^{1/2} \vee r_n^2 t)\log(2N_\delta p/\delta)}{n}} \bigvee \frac{C_n \log(2N_\delta p/\delta)}{n}\right),$$

since $r_n = \mathscr{O}_P(1)$. Let us now focus on bounding T_2 term. Note that $a_{ij}(\boldsymbol{\beta}^* + \boldsymbol{\delta}_k) = a_{ij}(\boldsymbol{\beta}^*) + a'_{ij}(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\delta}_k$ for some $\bar{\boldsymbol{\beta}}_k$ between $\boldsymbol{\beta}^* + \boldsymbol{\delta}_k$ and $\boldsymbol{\beta}^*$. Let

$$W_{ij}(\boldsymbol{\delta}) = a'_{ij}(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\delta}_k\left(f_i(\boldsymbol{\delta})g_i(\boldsymbol{\delta}) - f_i(\mathbf{0})g_i(\mathbf{0})\right),$$

and

$$Q_{ij}(\boldsymbol{\delta}) = a_{ij}(\boldsymbol{\beta}^*)\left(f_i(\boldsymbol{\delta})g_i(\boldsymbol{\delta}) - f_i(\mathbf{0})g_i(\mathbf{0})\right).$$

Let $\mathbb{Q}(\boldsymbol{\delta}) = Q(\boldsymbol{\delta}) - \mathbb{E}[Q(\boldsymbol{\delta})]$. For a fixed j, and k we have

$$\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)} \left|\mathbf{e}_j^\top\left(\mathbb{V}_n(\boldsymbol{\delta}) - \mathbb{V}_n(\widetilde{\boldsymbol{\delta}}_k)\right)\right|$$

is upper bounded with

$$\underbrace{\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)}\left|n^{-1}\sum_{i\in[n]}\mathbb{Q}_{ij}(\boldsymbol{\delta}) - \mathbb{Q}_{ij}(\widetilde{\boldsymbol{\delta}}_k)\right|}_{T_{21}} + \underbrace{\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)}\left|n^{-1}\sum_{i\in[n]}W_{ij}(\boldsymbol{\delta}) - \mathbb{E}\left[W_{ij}(\boldsymbol{\delta})\right]\right|}_{T_{22}}.$$

We will deal with the two terms separately. Let $Z_i = \max\{\varepsilon_i, -x_i\boldsymbol{\beta}^*\}$

$$f_i(\boldsymbol{\delta})g_i(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \geq Z_i\} - \mathbb{1}\left\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\right\}.$$

Observe that the distribution of Z_i is similar to the distribution of $|\varepsilon_i|$ due to the Condition (**E**). Moreover,

$$\left|x_i(\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k)\right| \leq K||\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k||_2\sqrt{\left|\text{supp}(\boldsymbol{\delta} - \widetilde{\boldsymbol{\delta}}_k)\right|}$$

where K is a constant such that $\max_{i,j}|x_{ij}| \leq K$. Hence,

$$\max_{k\in[N_\delta]}\max_{i\in[n]}\sup_{\boldsymbol{\delta}\in\mathscr{B}(\widetilde{\boldsymbol{\delta}}_k, r_n\xi_n)}\left|x_i\boldsymbol{\delta} - x_i\widetilde{\boldsymbol{\delta}}_k\right| \leq r_n\xi_n\sqrt{t}\max_{i,j}|x_{ij}| \leq Cr_n\xi_n\sqrt{t} =: \widetilde{L}_n. \tag{1.57}$$

For T_{21}, we will use the fact that $\mathbb{1}\{a < x\}$ and $\mathbb{P}\{Z < x\}$ are monotone function in x. Therefore,

$$\begin{aligned}
T_{21} \leq n^{-1}\sum_{i\in[n]}\Bigg[\left|a_{ij}(\boldsymbol{\beta}^*)\right|\Big(&\mathbb{1}\left\{Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right\} - \mathbb{1}\left\{-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right\} - \mathbb{1}\left\{Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k\right\}\\
&+ \mathbb{1}\left\{-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right\} - \mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right] + \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right]\\
&+ \mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k\right] - \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right]\Big)\Bigg]
\end{aligned}$$

Furthermore, by adding and substracting appropriate terms we can decompose the right hand side above into two terms. The first,

$$\begin{aligned}
n^{-1}\sum_{i\in[n]}\Bigg[\left|a_{ij}(\boldsymbol{\beta}^*)\right|\Big(&\mathbb{1}\left\{Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right\} - \mathbb{1}\left\{-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right\} - \mathbb{1}\left\{Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k\right\}\\
&+ \mathbb{1}\left\{-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right\} - \mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right] + \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right]\\
&+ \mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k\right] - \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k\right]\Big)\Bigg]
\end{aligned}$$

and the second

$$\begin{aligned}
n^{-1}\sum_{i\in[n]}\Bigg[\left|a_{ij}(\boldsymbol{\beta}^*)\right|\Big(&\mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right] - \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right]\\
&- \mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k - \widetilde{L}_n\right] + \mathbb{P}\left[-x_i\boldsymbol{\beta}^* \leq x_i\widetilde{\boldsymbol{\delta}}_k + \widetilde{L}_n\right]\Big)\Bigg].
\end{aligned}$$

The first term in the display above can be bounded in a similar way to T_1 by applying Bernstein's inequality and hence the details are omitted. For the second term we have a bound

$CC_n\widetilde{L}_n$, since $\left|a_{ij}(\boldsymbol{\beta}^*)\right| \leq K\left(\Lambda_{\min}^{-1/2}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)+C_n\right)$ by the definition of a_{ij} and Lemma 5, and $\mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k+\widetilde{L}_n\right]-\mathbb{P}\left[Z_i \leq x_i\widetilde{\boldsymbol{\delta}}_k-\widetilde{L}_n\right] \leq C\|f_{|\varepsilon_i|}\|_\infty\widetilde{L}_n \leq 2Cf_{\max}\widetilde{L}_n$. In the last inequality we used the fact that $\|f_{|\varepsilon_i|}\|_\infty \leq 2\|f_{\varepsilon_i}\|_\infty$. Therefore, with probability $1-2\delta$,

$$T_{21} \leq C\left(\sqrt{\frac{f_{\max}C_n^2\widetilde{L}_n\log(2/\delta)}{n}}\bigvee\frac{C_n\log(2/\delta)}{n}\bigvee f_{\max}\widetilde{L}_n\right).$$

A bound on T_{22} is obtain similarly to that on T_{21}. The only difference is that we need to bound $a'_{ij}(\bar{\boldsymbol{\beta}}_k)\boldsymbol{\delta}_k$, for $\bar{\boldsymbol{\beta}}_k=\alpha\boldsymbol{\beta}^*+(1-\alpha)(\boldsymbol{\beta}^*+\widetilde{\boldsymbol{\delta}}_k)$ and $\alpha\in(0,1)$, instead of $|a_{ij}(\boldsymbol{\beta}^*)|$. Observe that $a_{ij}(\boldsymbol{\beta})\widehat{\tau}_j^2=-\widehat{\gamma}_{(j),i}$. Moreover, by construction $\widehat{\tau}_j$ is a continuous, differentiable and convex function of $\boldsymbol{\beta}$ and is bounded away from zero by Lemma 5. Additionally, $\widehat{\boldsymbol{\gamma}}_{(j)}$ is a convex function of $\boldsymbol{\beta}$ as a set of solutions of a minimization of a convex function over a convex constraint is a convex set. Moreover, $\widehat{\gamma}_j$ is a bounded random variable according to Lemma 5. Hence, $|a'_{ij}(\boldsymbol{\beta}^*)| \leq K'$, for a large enough constant K'. Therefore, for a large enough constant C we have

$$T_{22} \leq C\left(\sqrt{\frac{f_{\max}r_n^2\xi_n^2\widetilde{L}_n\log(2/\delta)}{n}}\bigvee\frac{\widetilde{L}_n\log(2/\delta)}{n}\bigvee f_{\max}C_n\widetilde{L}_n\right).$$

A bound on T_2 now follows using a union bound over $j\in[p]$ and $k\in[N_\delta]$.

We can choose $\xi_n=n^{-1}$, which gives us $N_\delta \lesssim \left(pn^2\right)^t$. With these choices, the term T_2 is negligible compared to T_1 and we obtain

$$T \leq C\left(\sqrt{\frac{C_n^2(r_nt^{1/2}\vee r_n^2t)t\log(np/\delta)}{n}}\bigvee\frac{C_nt\log(2np/\delta)}{n}\right),$$

which completes the proof.

□

1.10 Acknowledgement

Chapter 1, in full, is a version of the paper “Generalized M-estimators for high-dimensional Tobit I models”. The dissertation author is the principal investigator of this material. The material is under revision for publication.

Chapter 2

Estimation and Inference for High-dimensional Left-censored Quantiles

2.1 Introduction

In this chapter, we present a quantile regression framework for high-dimensional left-censored linear models. Comparing to the generalized M-estimation framework in Chapter 1, the method introduced below is tailored towards quantile regression. Specifically, even though least absolute deviation (LAD) estimator for the left-censored linear model is used as a primary example in the Chapter 1, and LAD estimator is a special case of quantile estimators, a different approach, namely redistribution of mass, was adopted in the initial estimation here. This creates new challenges in estimation and inference of the problem. In return, the optimization problem can be transformed from a nonconvex optimization involving left-censored data into a modified quantile regression, which then greatly relieves computational burden. Since the problem considered in this chapter relates much to Chapter 1, we only include here related work in addition to the literature review in Chapter 1.

2.1.1 Contributions

We develop methodology for the quantile estimation and inference under high-dimensional and left-censoring settings. In details, the work provides a τ-quantile estimator and confidence intervals for high-dimensional left-censored regression, for any $\tau \in (0, 1)$, along with the theoretical guarantees. We modify a quantile regression estimation approach for right-censored data to accommodate the left-censored nature of our problem, and further extend the recently developed de-biasing techniques to derive an improved estimator suitable for high-dimensional inference.

2.1.2 Related Work

Quantile regression, as an robust alternative to ordinary linear regression, has received great attention since its introduction in [KBJ78]. The concept has then been taken to settings with heteroskedastic errors [KBJ82] and non-linear regression model [Obe82]. [Pow86a] first studied censored quantile regression, where the method was first applied under fixed left-censored data setting, with known censoring levels. Despite of the difficulties present in the censored nature of the data, Powell showed that the proposed natural estimator is consistent and asymptotically normal. However, many works, including [KP96], [Fit97], [BH98] and [FW07], have discussed computational burden due to the nonconvexity nature of the minimization objective function involved in Powell's estimator.

Meanwhile, progress has been made in application of survival analysis. Under right-censored data settings, both [KG01] and [Por03] have studied quantile regression with random right-censored data in details. Moreover, [Por03] proposed a recursively reweighted estimator of the regression quantile process, which generalized the Kaplan-Meier estimating scheme. Based on the redistribution of mass idea of [Efr67], the method in [Por03] recursively updates the weight of censored cases. Similarly, motivated by the same idea, [WW12] proposed a method, such that the weights of the censored observations are estimated in a single step. We extend the idea to

high-dimensional left-censored models.

2.1.3 Content

In Section 2.2, the methodology is presented with both procedures for deriving the initial estimator and the de-biased estimator. In Section 2.3, we study the conditions and asymptotic theory of the proposed method. Numerical simulations and a real data application are presented in Section 2.4. Finally, lemmas and their proofs are provided in Section 2.5 and 2.6, and the proofs of theorems are provided in Section 2.7.

2.2 Methodology

We start with the problem setup with model description. Then we lay out the methodology in two parts. In the first subsection, we describe our proposal for initial estimator, and in the second subsection we present the details of bias correction for the initial estimator.

2.2.1 Model Description

We consider the problem in the context of left-censored linear models. Let T_i be an underlying response variable, which is uncensored. We also denote $\mathbf{x}_i$ as our covariates vector of length p. The underlying latent quantile regression model for some quantile $\tau \in (0,1)$ comes in the form of

$$T_i = \mathbf{x}_i \boldsymbol{\beta}^o(\tau) + \varepsilon_i(\tau),\ i = 1, \ldots, n, \tag{2.1}$$

where $\varepsilon_i(\tau)$ is a random error, whose τ-th quantile conditional on $\mathbf{x}_i$ we assume is at 0. Due to left-censoring, however, we only observe the triplet $(\mathbf{x}_i, Y_i, C_i)$, where

$$Y_i = \max(T_i, C_i), \text{ and let } \delta_i = \mathbb{I}(T_i > C_i), \tag{2.2}$$

and $i = 1, \ldots, n$. Together (2.1) and (2.2) specify a left-censored quantile regression model. As C_i is observed, one can always reduce (2.2) to a constant-censored model, also known as Type-I Tobit model, in which the censoring vector is a constant c across i. Our interest lies in obtaining confidence intervals for the quantile coefficient $\boldsymbol{\beta}^o(\tau)$ for various τ, under high-dimensional settings with $p \gg n$. Bearing the high-dimensionality in mind, we denote $S_{\boldsymbol{\beta}^o} = \{j | \boldsymbol{\beta}^o_j \neq 0\}$ as the active set of variables of the coefficients and denote its cardinality by $s_{\boldsymbol{\beta}^o} = |S_{\boldsymbol{\beta}^o}|$.

2.2.2 Initial Estimator

In the case without censoring, quantile regression is carried out with the specific loss function $\rho_\tau(z) = z(\tau - \mathbb{I}\{z < 0\})$, also known as the check function. In the censoring case, however, directly fitting using the quantile loss results in a nonconvex optimization problem. In addition, simply removing the censored observations results in loss of information and bias. With such consideration, we borrow an algorithm from [WW12]. Specifically, we mimicked the "locally weighted censored quantile regression" method, which is based on Efron's redistribution of mass idea. The method assigns different weights on censored data and non-censored data, and avoids discarding all censored data, while maintaining partial information provided by the non-censored ones.

The method redistributes the mass of each censored observation to some point far on left, which is $-\infty$ in the case of left censoring. Note that if $\mathbf{x}_i\boldsymbol{\beta}(\tau) > C_i$ for all $\mathbf{x}_i$, then the left censoring at C_i has no impact on our estimate of τ-quantile. This observation comes from the fact that the quantile regression estimator is only determined by the signs of residuals, in another

word, we only care about the order of the responses.

We now present the initial estimator $\widehat{\boldsymbol{\beta}}$, with the justification of the weights following.

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[w_i(F_0) \rho_\tau(Y_i - \mathbf{x}_i \boldsymbol{\beta}) + (1 - w_i(F_0)) \rho_\tau(Y_i^{-\infty} - \mathbf{x}_i \boldsymbol{\beta}) \right] + \lambda_n \sum_{j=1}^{p} |\boldsymbol{\beta}_j|,$$

where $w_i(F_0)$ is defined as following, F_0 being the distribution of T_i,

$$w_i(F_0) = \begin{cases} 1 & \text{if } \delta_i = 1 \text{ or } F_0(C_i|\mathbf{x}_i) < \tau \\ 1 - \frac{\tau}{F_0(C_i|\mathbf{x}_i)} & \text{if } \delta_i = 0 \text{ and } F_0(C_i|\mathbf{x}_i) > \tau \end{cases}.$$

Notice that the additional penalty term is added, in order to accommodate the high-dimensional setting. To make sense out of the weights, we begin from the objective function of the underlying model under quantile loss,

$$U_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(T_i - \mathbf{x}_i \boldsymbol{\beta}).$$

Taking the derivative, we have the first order estimating equation

$$D_n(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \left(\tau - \mathbb{1}\{T_i - \mathbf{x}_i \boldsymbol{\beta} < 0\} \right).$$

The subgradient condition $D_n(\boldsymbol{\beta}) = \mathscr{O}_p(1)$ depends only on $\mathbb{1}\{T_i - \mathbf{x}_i \boldsymbol{\beta} < 0\}$ for each $\mathbf{x}_i$. Now fix any $\boldsymbol{\beta}$, if an observation is uncensored, then $Y_i = T_i$ is observed, and so is $\mathbb{1}\{T_i - \mathbf{x}_i \boldsymbol{\beta} < 0\}$. For censored observations ($Y_i = C_i > T_i$), if $\mathbf{x}_i \boldsymbol{\beta} > C_i$, we immediately know $T_i < \mathbf{x}_i \boldsymbol{\beta}$. The tricky case is when $\mathbf{x}_i \boldsymbol{\beta} < C_i$, we cannot determine the sign of $T_i - \mathbf{x}_i \boldsymbol{\beta}$. Hence, we look at the expectation

$$\mathbb{E}[\mathbb{1}\{T_i - \mathbf{x}_i \boldsymbol{\beta} > 0\} | T_i < C_i] = \frac{\mathbb{P}(\mathbf{x}_i \boldsymbol{\beta} < T_i < C_i)}{\mathbb{P}(T_i < C_i)},$$

where F_0 is the distribution of T_i. When $\boldsymbol{\beta} = \boldsymbol{\beta}^o(\tau)$,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\{T_i - \mathbf{x}_i\boldsymbol{\beta}^o(\tau) > 0\} | T_i < C_i\right] &= \frac{\mathbb{P}(\mathbf{x}_i\boldsymbol{\beta}^o(\tau) < T_i < C_i)}{\mathbb{P}(T_i < C_i)} \\ &= \frac{F_0(C_i|\mathbf{x}_i) - \tau}{F_0(C_i|\mathbf{x}_i)}.\end{aligned}$$

The observations above motivated us to assign weight $w_i(F_0) = 1$ to the first two scenarios, when we have uncensored or $F_0(C_i|\mathbf{x}_i) < \tau$ observations. Note that at the location $\mathbf{x}_i$, even when a data point is censored, if we believe the quantile of interest is above the censoring level, we still assign full weight to that data. Intuitively, we are only interested in estimating in quantile τ. In terms of a specific data point, our only concern is whether it is above or below the quantile line $\mathbf{x}_i\boldsymbol{\beta}^o$. For censored and ambiguous scenarios which we cannot determine the sign of $T_i - \mathbf{x}_i\boldsymbol{\beta}^o(\tau)$, we assign weight $w_i(F_0) = 1 - \frac{\tau}{F_0(C_i|\mathbf{x}_i)}$. By assigning the complimentary weight to any point below, such as $(\mathbf{x}_i, -\infty)$ or $(\mathbf{x}_i, Y_i^{-\infty})$, the quantile fit remains unaffected. Without loss of generality, we assume fixed censoring level $C_i = 0$ for all i.

Finally, using a consistent plug in estimator $\widehat{F}_n$ for F_0, we have the initial estimator as,

Step 0: Initial estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \Big[& w_i(\widehat{F}_n)\rho_\tau(Y_i - \mathbf{x}_i\boldsymbol{\beta}) \\ & + (1 - w_i(\widehat{F}_n))\rho_\tau(Y_i^{-\infty} - \mathbf{x}_i\boldsymbol{\beta})\Big] + \lambda_n \sum_{j=1}^{p} |\boldsymbol{\beta}_j|.\end{aligned} \tag{2.3}$$

We delay the discussion of the estimator $\widehat{F}_n$ to Condition 1, where we will lay out the requirement on such estimator.

2.2.3 Bias Correction

With our inference objective, the estimator given in (2.3) needs improvement. As we show later, the initial estimator is consistent. However, as other penalized estimators, our initial estimator is also a biased one. Following classical one-step estimation framework, typically an one-step improvement of the following form is considered. With appropriate estimators plugged in as proxies, we have

Step 1: Bias correction

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbf{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{F}_n), \tag{2.4}$$

where the vector $\mathbf{S}_n$ is the score and the matrix $\widehat{\boldsymbol{\Theta}}$ is a proxy to the inverse Hessian matrix $\mathbf{H}^{-1}$. $\mathbf{H}$ is defined as the subgradient of $\mathbf{S}_n$.

We first define $\mathbf{S}_n$, and then provide an explanation for the transition between $\widehat{\boldsymbol{\Theta}}$ and $\mathbf{H}^{-1}$.

$$\mathbf{S}_n(\boldsymbol{\beta}, F) := -\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \left[w_i(F)\psi_\tau(Y_i - \mathbf{x}_i\boldsymbol{\beta}) + (1 - w_i(F))\psi_\tau(Y_i^{-\infty} - \mathbf{x}_i\boldsymbol{\beta}) \right]$$

with $\psi_\tau(z) = \tau - \mathbb{I}\{z < 0\}$ being the differential of $\rho_\tau(z)$. Note that $Y_i^{-\infty} - \mathbf{x}_i\boldsymbol{\beta} < 0$ due to our choice of $Y_i^{-\infty} = -\infty$. Therefore, we have $\psi_\tau(Y_i^{-\infty} - \mathbf{x}_i\boldsymbol{\beta}) = \tau - 1$ for all i, and hence

$$\begin{aligned} \mathbf{S}_n(\boldsymbol{\beta}, F) &= -\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \left[w_i(F)\psi_\tau(Y_i - \mathbf{x}_i\boldsymbol{\beta}) + (1 - w_i(F))(\tau - 1) \right] \\ &= -\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \left[w_i(F)\, \mathbb{I}\{Y_i - \mathbf{x}_i\boldsymbol{\beta} \geq 0\} - (1 - \tau) \right]. \end{aligned} \tag{2.5}$$

Notice that $\mathbf{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{F}_n)$ depends on both the initial estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{F}_n$. This imposes an additional challenge on the theory, which we address later in Lemma 8. As for (2.3) being a consistent estimator, only consistency of the estimator $\widehat{F}_n$ is required. However, for inference a slightly stronger convergence rate requirement on the error of the estimator $\widehat{F}_n$ needs to be

imposed, which is summarized in Condition 1.

As for the Hessian matrix $\mathbf{H}$, we observe that the function ψ_τ is not everywhere differentiable. Hence, we propose to consider another candidate for the subgradient of $\mathbf{S}_n$. We first compute the expectation of the score $\mathbf{S}_n(\boldsymbol{\beta}, F)$, and then compute its gradient. Thus, for the simplicity of notation, the following expectations are taken with respect to T_i given $\mathbf{x}$.

Proposition 12. *Assuming the true distribution F_0, we have*

$$\begin{aligned}\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}, F_0)] &= -\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\Big(\tau - \mathbb{P}(Y_i < \mathbf{x}_i\boldsymbol{\beta}) - \tau\,\mathbb{1}\{F_0(0|\mathbf{x}_i) > \tau\}\,\mathbb{1}\{\mathbf{x}_i\boldsymbol{\beta} \le 0\}\Big) \\ &= \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,(\tau - \mathbb{P}(T_i < \mathbf{x}_i\boldsymbol{\beta})) & \text{if } \mathbf{x}_i\boldsymbol{\beta} > 0 \\ -\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,(\tau - \tau\,\mathbb{1}\{F_0(0|\mathbf{x}_i) > \tau\}) & \text{if } \mathbf{x}_i\boldsymbol{\beta} \le 0 \end{cases} \end{aligned} \tag{2.6}$$

and hence the Hessian

$$\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}, F_0)] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i f_0(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i)\,\mathbb{1}\{\mathbf{x}_i\boldsymbol{\beta} > 0\} \tag{2.7}$$

where f_0 is the density function of T_i.

Remark 11. *Note that $\mathbb{E}[D_n(\boldsymbol{\beta})] = -n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\Big(\tau - \mathbb{P}(T_i < \mathbf{x}_i\boldsymbol{\beta})\Big)$. Comparing to* (2.6)*, we know when $\mathbf{x}_i\boldsymbol{\beta} > 0$, $\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}, F_0)] = \mathbb{E}[D_n(\boldsymbol{\beta})]$, and hence when $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, $\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}_o, F_0)] = 0$ since $\mathbb{E}[D_n(\boldsymbol{\beta}_o)] = 0$. Furthermore, $\mathbb{1}\{F_0(0|\mathbf{x}_i) > \tau\} = \mathbb{1}\{\mathbf{x}_i\boldsymbol{\beta}_o \le 0\}$, if F_0 is strictly increasing, and hence $\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}_o, F_0)] = 0$ when $\mathbf{x}_i\boldsymbol{\beta}_o \le 0$ as well. In summary, at the truth $\boldsymbol{\beta}_o$, the expectation of our score estimator $\mathbb{E}[\mathbf{S}_n(\boldsymbol{\beta}_o, F_0)]$ is indeed zero.*

Note that the matrix $\mathbf{H}(\boldsymbol{\beta})$ is not invertible for general $\boldsymbol{\beta}$ when the number of parameters p exceeds the number of observations n. In fact, with a little abuse of notation, we only assume the existence of $\mathbf{H}^{-1}$, which is layed out as Condition 7 later in the text (here, the expectation is with respect to $\mathbf{x}_i$). In the following section, we describe the details in obtaining the proxy $\widehat{\boldsymbol{\Theta}}$ for $\mathbf{H}^{-1}$.

2.2.4 Inverse Hessian Estimator: Nodewise Lasso

Our Inverse Hessian estimator is inspired by the nodewise lasso method proposed in [VdGBR+14]. For notation simplicity, we first rewrite (2.7),

$$\mathbf{H}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{u}_i^\top\mathbf{u}_i = n^{-1}\mathbf{X}_{\boldsymbol{\beta}}^\top\mathbf{X}_{\boldsymbol{\beta}},$$

where $\mathbf{u}_i := \mathbf{x}_i\,\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0)\sqrt{f_0(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i)}$, $\mathbf{X}_{\boldsymbol{\beta}} = \mathbf{W}_{\boldsymbol{\beta}}\mathbf{X}$, and $\mathbf{W}_{\boldsymbol{\beta}}$ is defined as

$$\mathbf{W}_{\boldsymbol{\beta}} = \operatorname{diag}\left(\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0)\sqrt{f_0(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i)}\right)_{i=1}^{n}. \tag{2.8}$$

That is, $\mathbf{X}_{\boldsymbol{\beta}}$ is a new design matrix with i-th row to be $\mathbf{u}_i$, which can also be treated as the product of weighted matrix $\mathbf{W}_{\boldsymbol{\beta}}$ and $\mathbf{X}$. Note that for fixed data, $(\mathbf{W}_{\boldsymbol{\beta}})_{jj}$ only depends on $\boldsymbol{\beta}$.

Then we carry out nodewise lasso using $\mathbf{X}_{\boldsymbol{\beta}}$. Note that as we use the initial estimator $\widehat{\boldsymbol{\beta}}$ as the plug in for $\mathbf{X}_{\widehat{\boldsymbol{\beta}}}$, we also use a consistent estimator $\widehat{f_n}$ in place for f_0 in (2.8). The discussion of the estimator $\widehat{f_n}$ is delayed later to Condition 2. We have the nodewise lasso scheme as following. For each $j = 1,\cdots,p$, define

$$\widehat{\boldsymbol{\gamma}}_j := \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\left(n^{-1}||(\mathbf{X}_{\widehat{\boldsymbol{\beta}}})_j - (\mathbf{X}_{\widehat{\boldsymbol{\beta}}})_{-j}\boldsymbol{\gamma}||_2^2 + 2\lambda_j||\gamma||_1\right), \tag{2.9}$$

where $(\mathbf{X}_{\widehat{\boldsymbol{\beta}}})_{-j}$ is the design submatrix without the j-th column. Note that (2.9) can be solved using standard lasso regression. We further denote the components of $\widehat{\boldsymbol{\gamma}}_j \in \mathbb{R}^{p-1}$ as $\{\widehat{\boldsymbol{\gamma}}_{j,k} : k = 1,\cdots,p, k\neq j\}$. Then define

$$\widehat{C} := \begin{pmatrix} 1 & -\widehat{\boldsymbol{\gamma}}_{1,2} & \cdots & -\widehat{\boldsymbol{\gamma}}_{1,p} \\ -\widehat{\boldsymbol{\gamma}}_{2,1} & 1 & \cdots & -\widehat{\boldsymbol{\gamma}}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\boldsymbol{\gamma}}_{p,1} & -\widehat{\boldsymbol{\gamma}}_{p,2} & \cdots & 1 \end{pmatrix}$$

and

$$\widehat{D}^2 := \text{diag}\left(\widehat{d}_1^2, \cdots, \widehat{d}_p^2\right),$$

where for $j = 1, \cdots, p$,

$$\widehat{d}_j^2 := n^{-1}||(X_{\widehat{\boldsymbol{\beta}}})_j - (\mathbf{X}_{\widehat{\boldsymbol{\beta}}})_{-j}\widehat{\boldsymbol{\gamma}}_j||_2^2 + \lambda_j||\widehat{\boldsymbol{\gamma}}_j||_1. \tag{2.10}$$

$\widehat{d}_j^2$ serves as an estimate to the noise level of the regression in (2.9). Finally, our proxy $\widehat{\boldsymbol{\Theta}}$ is defined as,

$$\widehat{\boldsymbol{\Theta}} := \widehat{D}^{-2}\widehat{C}. \tag{2.11}$$

In addition, we note that using the KKT conditions, we can show

$$||\mathbf{H}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Theta}}_j^\top - e_j||_\infty \leq \lambda_j/\widehat{d}_j^2, \tag{2.12}$$

and

$$\left(\mathbf{H}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Theta}}^\top\right)_{jj} = 1. \tag{2.13}$$

Finally, we propose the novel High-dimensional Left-censored Quantile Regression in Algorithm 1 and 2.

2.3 Theoretical Considerations

In what follows, we briefly discuss the preliminary theoretical results, along with the conditions required. In the first subsection, we address the requirements for the distribution and density estimators. Then we move on to conditions for acquiring consistency using the initial

Algorithm 1 High-dimensional Left-censored Quantile Regression

procedure INITIAL ESTIMATION
    Obtain an estimator $\widehat{F}_n$
    Plug in $\widehat{F}_n$ into (2.3) and obtain $\widehat{\boldsymbol{\beta}}$
end procedure
procedure ONE-STEP CORRECTION
    Obtain estimator $\widehat{\boldsymbol{\Theta}}$, more details in Algorithm 2
    Plug in initial estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{F}_n$ for $\mathbf{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{F}_n)$ as in (2.5)
    Obtain the one-step improved estimator $\widetilde{\boldsymbol{\beta}}$ as in (2.4)
end procedure

Algorithm 2 Inverse Hessian estimation $\widehat{\boldsymbol{\Theta}}$

Obtain an estimator $\widehat{f}_n$
Plug in initial estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{f}_n$ into (2.8)
for $j = 1, \ldots, p$ **do**
    Obtain $\widehat{\boldsymbol{\gamma}}_j$ and $\widehat{d}_j^2$ as in (2.9) and (2.10) respectively
end for
Obtain $\widehat{\boldsymbol{\Theta}}$ as described in (2.11)

estimator. We are inspired by the consistency result of the penalized censored least absolute deviation estimator in [MvdG16]. Finally, we present the derivation of the normality result for the improved one-step estimator, which follows from the sketch of [BG16]. Under the current context, however, extra challenges surface as both score and inverse Hessian depends on distribution and density estimator in addition to the parameter estimator $\widehat{\boldsymbol{\beta}}$.

2.3.1 Distribution and Density Estimators

We impose the following condition on the choice of distribution estimator.

Condition 1 (Distribution estimator condition)**.** *The estimator $\widehat{F}_n(t|\mathbf{x})$ is a consistent estimator of the conditional distribution of T, $F_0(t|\mathbf{x})$, for all $\mathbf{x}$. More precisely, for any $t \in \mathbb{R}$,*

$$\sup_{\mathbf{x}\in\mathbb{R}^p} \left| \widehat{F}_n(t|\mathbf{x}) - F_0(t|\mathbf{x}) \right| = O_p\left(\delta_{\widehat{F}}\right),$$

where $\delta_{\widehat{F}} \xrightarrow{p} 0$ *as* $n \to \infty$.

Note that the condition essentially only requires $\widehat{F}_n$ to be a consistent estimator. We have selected the classical Kaplan-Meier estimator for analysis later in the paper. Likewise, we also impose a consistency condition on the density estimator $\widehat{f}_n$ as following.

Condition 2 (Density condition)**.** *1. The conditional density function* $f_0(y|\mathbf{x})$ *is a Lipschitz function in* y *with a uniform Lipschitz constant* L *for all* $\mathbf{x}$.

2. *There exists* $M > m > 0$ *such that* $m \leq f_0(y|\mathbf{x}) \leq M$ *for all* y *and* $\mathbf{x}$.

3. *The conditional density estimator* $\widehat{f}_n(y|\mathbf{x})$ *is a consistent estimator of* $f_0(y|\mathbf{x})$. *To be precise,*

$$\int\int \left(\widehat{f}_n(y|\mathbf{x}) - f_0(y|\mathbf{x})\right)^2 d\mu(\mathbf{x})dy = o_p(1),$$

where μ *is a measure on the support of* $\mathbf{x}$.

4. $\lim_{\varepsilon \to 0^+} \mathbb{P}(|\mathbf{x}\boldsymbol{\beta}^o| > \varepsilon) = 1$.

The two conditions above are not restrictive in their nature, though distribution and density estimation in high-dimensional settings remains an active research topic. Nevertheless, we refer one to [HY05], [Efr07] and [IL15] for more discussions on the topic.

2.3.2 Consistency of Initial Estimator

In the section, we present the consistency analysis for the initial estimator. For notational simplicity, throughout this section, $\mathbf{x}$ and $\mathbf{x}_i$ are row vectors. Also, we denote $\widehat{w} = w(\widehat{F})$ and $w^0 = w(F_0)$. We also define the linear function $f_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, the reweighted loss function $\rho_f(\mathbf{x}, y, w) = w\rho_\tau(y - f_{\boldsymbol{\beta}}(\mathbf{x})) + (1-w)\rho_\tau(y^{-\infty} - f_{\boldsymbol{\beta}}(\mathbf{x}))$, the risk $\mathscr{P}\rho_f = \mathbb{E}\rho_f(\mathbf{x}, y, w^0)$, the empirical risk $\mathscr{P}_n\rho_f = \frac{1}{n}\sum_{i=1}^n \rho_f(\mathbf{x}_i, y_i, w_i^0)$ at F_0, and the empirical risk $\widehat{\mathscr{P}}_n\rho_f = \frac{1}{n}\sum_{i=1}^n \rho_f(\mathbf{x}_i, y_i, \widehat{w}_i)$ at $\widehat{F}$.

Then we define f^* as a linear functional such that for all $\mathbf{x}$,

$$f^*(\mathbf{x}) = \underset{a}{\operatorname{argmin}} \mathbb{E}\left[w^0 \rho_\tau(y-a) + (1-w^0)\rho_\tau(y^{-\infty}-a)\middle|\mathbf{x}\right].$$

In order for f^* to be uniquely defined, we need the following censoring condition Condition 3. To see the necessity of this condition, let $\boldsymbol{\beta}^o$ be the true parameter. By the first order property, $\mathbb{E}[w^0 \psi_\tau(y-a) + (1-w^0)(\tau-1)|\mathbf{x}] = 0$. Hence, for all $\mathbf{x}$,

$$\mathbb{E}[w^0 \, \mathbb{1}(y > a)|\mathbf{x}] = 1-\tau. \tag{2.14}$$

By the definition of weight w^0, if $F_0(0|\mathbf{x}) < \tau$, (2.14) means $F_0(a|\mathbf{x}) = \tau$, and hence, $f_0(\mathbf{x}) = f_{\boldsymbol{\beta}^o}(\mathbf{x})$. But if $F_0(0|\mathbf{x}) > \tau$, then any $a < 0$ is a solution to (2.14). However, we require (2.14) to hold for every $\mathbf{x}$. So as long as not for all $\mathbf{x}$, $F_0(0|\mathbf{x}) > \tau$, then because of the linearity of f^*, there exists a unique solution.

Condition 3 (Censoring condition)**.** *Let μ be measure on $\mathscr{X}$. There exists a set $E \subset \mathscr{X}$ such that $\mu(E) > 0$ and $F_0(0|\mathbf{x}) < \tau$ for all $\mathbf{x} \in E$. Furthermore, at the censoring level 0, there exists a constant $0 < M_0 < \tau$ such that $F_0(0|\mathbf{x}) \geq M_0$ for all $\mathbf{x}$.*

Some additional conditions also need to be imposed.

Condition 4 (Error condition)**.** *The conditional error distribution function $\nu_0(t|\mathbf{x})$ is continuously differentiable for all $\mathbf{x}$, and the first derivative $\dot{\nu}_0(t|\mathbf{x})$ satisfies Lipschitz condition with constant L uniformly for all $\mathbf{x}$, and is bounded from above and below. Furthermore, $\dot{\nu}_0(0|\mathbf{x}) > 0$ and $\int_0^\varepsilon (\varepsilon - t) d\nu_0(t|\mathbf{x}) > 0$ for all $\varepsilon > 0$ and $\mathbf{x}$.*

The above condition is our only limitation on the error distribution. Even though we require bounded first derivative for the error density, which excludes densities with unbounded first moment, the condition still allows for a class of distributions much larger than the Gaussian.

Next, we have a condition on the design. First, we denote $\gamma_j := \text{argmin}_\gamma \mathbb{E}\|\mathbf{X}_j - \mathbf{X}_{-j}\gamma\|_n^2$, and then $\mathbf{X}_{-j}\gamma_j$ is the projection of $\mathbf{X}_j$ into $\mathbf{X}_{-j}$ under the inner product $\langle \mathbf{X}_i, \mathbf{X}_j \rangle = \mathbb{E}\mathbf{X}_i^\top \mathbf{X}_j/n$.

Condition 5 (Design matrix condition)**.** *The design matrix X satisfies* $\|\mathbf{X}\|_\infty = \max_{i,j}|X_{i,j}| =$ *O(1), that is, every column* $\|\mathbf{X}_j\|_\infty = O(1)$*. If furthermore, the projection* $\mathbf{X}_{-j}\gamma_j$ *is also bounded for all j, i.e.* $\|\mathbf{X}_{-j}\gamma_j\|_\infty = O(1)$*, we say* $\mathbf{X}$ *is strongly bounded.*

A bounded condition on design matrix entries $\mathbf{X}_{ij}$ is not uncommon in high-dimensional settings [VdGBR$^+$14]. In fact, in many cases, if $\mathbf{X}$ follows an unbounded distribution, one can always approximate its distribution with a truncated one. The following is the same compatibility condition introduced for linear models [BRT09], which is standard condition when applying lasso estimators.

Condition 6 (Compatibility condition)**.** *There exists some* $\phi_0 > 0$ *and all* $\boldsymbol{\beta}$ *satisfying* $||(\boldsymbol{\beta} - \boldsymbol{\beta}^o)_{S^c_{\boldsymbol{\beta}^o}}||_1 \leq 3||(\boldsymbol{\beta} - \boldsymbol{\beta}^o)_{S_{\boldsymbol{\beta}^o}}||_1$ *it holds that*

$$||(\boldsymbol{\beta} - \boldsymbol{\beta}^o)_{S_{\boldsymbol{\beta}^o}}||_1^2 \leq \frac{s_{\boldsymbol{\beta}^o}}{\phi_0^2}(\boldsymbol{\beta} - \boldsymbol{\beta}^o)\mathbb{E}[\mathbf{x}^T\mathbf{x}](\boldsymbol{\beta} - \boldsymbol{\beta}^o).$$

Denoting the excess risk as $\mathscr{E}(f) = \mathscr{P}\rho_f - \mathscr{P}\rho_{f_0}$, and the sum of squares norm as $||f||^2 = \mathbb{E}f^2(\mathbf{x})$, in the linear case, $||f_{\boldsymbol{\beta}}||^2 = \mathbb{E}f^2_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^T\mathbb{E}[\mathbf{x}^T\mathbf{x}]\boldsymbol{\beta}$, we are now ready to present the consistency result.

Theorem 13. *Under Conditions 3 - 6 and define*

$$\lambda(t) = 4K_X\sqrt{\frac{2\log(2p)}{n}} + K_X\sqrt{\frac{32t}{n}}.$$

Then for $\lambda \geq 4\lambda(t)$ *with* $t = 2\log(p)$ *and some constant C, with probability at least* $1 - \log_2(8np^2)/p^2$*,*

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o||_1 \leq \frac{6C\lambda s_{\boldsymbol{\beta}^o}}{\phi_0^2}, \tag{2.15}$$

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)^T\mathbb{E}[\mathbf{x}^T\mathbf{x}](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)\leq\frac{9C^2\lambda^2 s_{\boldsymbol{\beta}^o}}{\phi_0^2}. \tag{2.16}$$

In other words, with $\lambda\asymp\sqrt{\log p/n}$, *we have* $||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1=O_p\left(s_{\boldsymbol{\beta}^o}\sqrt{\frac{\log(p)}{n}}\right)$ *and*

$$n^{-1}\left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)\right\|_2^2=O_p\left(s_{\boldsymbol{\beta}^o}\frac{\log(p)}{n}\right).$$

Corollary 14. *Under the assumption* $s_{\boldsymbol{\beta}^o}=o\left(\sqrt{n/\log(p)}\right)$, *we have consistency for the initial estimator* $\widehat{\boldsymbol{\beta}}$.

2.3.3 Asymptotic Normality of One-step Penalized Estimator

This section entails the delicate details of obtaining the asymptotic normality of the improved one-step estimator, with imposed conditions as well as the preliminary lemmas. We start the analysis with the following decomposition of (2.4),

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^o\right)=&\sqrt{n}\underbrace{\left(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^o\right)}_{\mathrm{I}}-\sqrt{n}\underbrace{\left(\widehat{\boldsymbol{\Theta}}_j\mathbf{S}_n(\boldsymbol{\beta}^o,F_0)\right)}_{\mathrm{N}}\\&-\sqrt{n}\underbrace{\left[\widehat{\boldsymbol{\Theta}}_j\left(\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)-\mathbb{E}\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)\right)-\widehat{\boldsymbol{\Theta}}_j\left(\mathbf{S}_n(\boldsymbol{\beta}^o,F_0)-\mathbb{E}\mathbf{S}_n(\boldsymbol{\beta}^o,F_0)\right)\right]}_{\mathrm{II}},\\&-\sqrt{n}\underbrace{\widehat{\boldsymbol{\Theta}}_j\left(\mathbf{S}_n(\widehat{\boldsymbol{\beta}},\widehat{F}_n)-\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)\right)}_{\mathrm{III}}-\sqrt{n}\underbrace{\left(\widehat{\boldsymbol{\Theta}}_j(\mathbb{E}\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)-\mathbb{E}\mathbf{S}_n(\boldsymbol{\beta}^o,F_0)\right)}_{\Delta}\end{aligned} \tag{2.17}$$

where $\widehat{\boldsymbol{\Theta}}_j$ denotes the j-th row of $\widehat{\boldsymbol{\Theta}}$. With the help of this decomposition, our aim is to show that part (N) converges to a Normal distribution, while the other terms converge to zero at a faster rate. In order to characterize and bound each individual term, we have lemmas for results leading up to Theorem 15 below. However, for the purpose of presentation, we defer the lemmas to Section 2.5.

Finally, we introduce the last condition we impose. One may also refer to this condition

as the restrictive eigenvalue assumption, which requires the population Hessian to be at least invertible. We note that even in linear models without censoring, this is an indispensable condition.

Condition 7. *The smallest eigenvalue* $\Lambda_{\min}$ *of* $\mathbb{E}\left[X_{\boldsymbol{\beta}^o}^T X_{\boldsymbol{\beta}^o}/n\right]$ *is strictly positive and* $1/\Lambda_{\min} = O(1)$.

We are now ready to present the main result.

Theorem 15. *Under Conditions 1 - 7, with* $\lambda \asymp \sqrt{\log p/n}$ *and* $\lambda_j \asymp \sqrt{\log p/n}$, *and define* $s_j := \left\|\boldsymbol{\Theta}_j^0\right\|_0 = \left|\{k \neq j : \boldsymbol{\Theta}_{j,k}^0 \neq 0\}\right|$, *assuming* $Ks_{\boldsymbol{\beta}^o}^2 \log p/n \vee s_{\boldsymbol{\beta}^o}^{1/2} s_j^{1/2} (\log p/n)^{1/4} \vee K\|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^0\|_1 = o(1)$, *where* $K = \sqrt{s_j}$ *and in the strongly bounded case,* $K = 1$. *Let* $I_n = \left(\widetilde{\boldsymbol{\beta}}_j - a_n, \widetilde{\boldsymbol{\beta}}_j + a_n\right)$ $a_n = z_\alpha \sqrt{\widehat{\boldsymbol{\Theta}}_j \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{\Theta}}_j^\top / n}$, *where*

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \left(\widehat{\phi}_i + \widehat{\psi}_i\right)^2,$$

$\widehat{\psi}_i := -\left[w_i(\widehat{F}_n)\, \mathbb{1}\{Y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}} \geq 0\} - (1-\tau)\right]$ *and*

$$\widehat{\phi}_i := \tau\, \mathbb{1}\left(\mathbf{x}_i \widehat{\boldsymbol{\beta}} \leq 0\right) \mathbb{1}\left(Y_i = 0\right) \frac{\mathbb{1}(\widehat{F}_n > \tau)}{\widehat{F}_n^2} \sum_{\substack{l=1 \\ l \neq i}}^n B_{nl}(\mathbf{x}_i) \left(1 - \frac{\mathbb{1}(Y_l = 0)}{\widehat{F}_n}\right).$$

The distribution estimator $\widehat{F}_n$ *is chosen to be the classical Kaplan-Meier estimator,*

$$\widehat{F}_n(t|\mathbf{x}) = \prod_{j=1}^n \left(1 - \frac{1}{\sum_{k=1}^n \mathbb{1}(Y_k \leq Y_j)}\right)^{\eta_j(t)}, \tag{2.18}$$

where $\eta_j(t) = \mathbb{1}(Y_j > t, \delta_j = 1)$. *For* $j \in \{1, \ldots, p\}$, *when* $n, p \to \infty$, *we have*

$$\mathbb{P}\left(\boldsymbol{\beta}_j^o \in I_n\right) = 1 - 2\alpha.$$

Remark 12. *The quantity* s_j *quantifies the sparsity nature of the underlying precision matrix* $\boldsymbol{\Theta}^0$, *which we aim to estimate with* $\widehat{\boldsymbol{\Theta}}$. *This is a standard assumption in high dimensional inference.*

Essentially, it restricts the column $(\mathbf{X}_{\boldsymbol{\beta}^o})_j$ *to be dependent with only* s_j *number of columns in* $(\mathbf{X}_{\boldsymbol{\beta}^o})_{-j}$.

2.4 Numerical Experiments and Application

In this section, we present the application our proposed method in details, along with simulation results under various settings and an application in real data study.

2.4.1 Further Details of Algorithm 1 and 2

We start with the definition of $Y^{-\infty}$. In practice, we have taken

$$Y^{-\infty} := -1000 \times \|Y\|_\infty = -1000 \times \max_i |Y_i|.$$

For the estimator of conditional distribution of T_i, as mentioned earlier, there are options specifically tailored for distribution estimation in high-dimensions, we provide here a possible estimator $\widehat{F}_n$ for line 2 in Algorithm 1 based on the ideas of Kaplan-Meier estimator, which is defined as the following.

$$\widehat{F}_n(t|\mathbf{x}) = \prod_{j=1}^{n} \left(1 - \frac{B_{nj}(\mathbf{x})}{\sum_{k=1}^{n} \mathbb{1}(Y_k \leq Y_j) B_{nk}(\mathbf{x})}\right)^{\eta_j(t)}, \tag{2.19}$$

where $\eta_j(t) = \mathbb{1}(Y_j > t, \delta_j = 1)$. Choosing $B_{nk}(\mathbf{x}) = 1/n$ results in the classical Kaplan-Meier estimator. We also note that the Nadaraya-Watson's type weights for $B_{nk}(\mathbf{x})$ is also a common choice, which is

$$B_{nk}(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}-\mathbf{x}_k}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)}, \tag{2.20}$$

where K is a density kernel function, and $h_n \in \mathbb{R}^+$ is the bandwidth converging to zero as $n \to \infty$. In the simulations, we have opted for the classical Kaplan-Meier estimator for simplicity. In addition, we have the following density estimator for $\widehat{f_n}$ in line 1 in Algorithm 2. For a positive bandwidth sequence $\widehat{h}_n$,

$$\widehat{f_n} = \widehat{h}_n^{-1} \sum_{i=1}^{n} \frac{\mathbb{1}(x_i\widehat{\boldsymbol{\beta}} > 0)\, \mathbb{1}(0 \le Y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}} \le \widehat{h}_n)}{\sum_{i=1}^{n} \mathbb{1}(x_i\widehat{\boldsymbol{\beta}} > 0)}. \tag{2.21}$$

This estimator is inspired by the estimator for error density at 0 presented in [BG16], which translates to an estimation for density of T_i at $\mathbf{x}_i\boldsymbol{\beta}^o(\tau)$. For the choice of $\widehat{h}_n$, we also follow the adaptive choice of the bandwidth sequence thereof. Let $u_i := y_i - x_i\widehat{\boldsymbol{\beta}}$,

$$\widehat{h}_n = c\left\{s_{\widehat{\boldsymbol{\beta}}} \log p/n\right\}^{-1/3} \text{median}\left\{u_i : u_i > \sqrt{\log p/n},\ x_i\widehat{\boldsymbol{\beta}} > 0\right\},$$

for a constant $c > 0$. Here, $s_{\widehat{\boldsymbol{\beta}}}$ denotes the size of the estimated set of the non-zero elements of the initial estimator $\widehat{\boldsymbol{\beta}}$, i.e., $s_{\widehat{\boldsymbol{\beta}}} = \|\widehat{\boldsymbol{\beta}}\|_0$.

An additional note is also in place for line 3 of Algorithm 1. Regarding the computation procedure to obtain the initial estimator, we note that this boils down to a weighted quantile regression problem and is readily solvable using linear programming techniques. The penalty parameter λ in (2.3) is chosen by the minimum of K-fold cross validation statistic, $\text{argmin}_\lambda \sum_{k=1}^{K} \text{CV}_k(\lambda)$, and

$$\text{CV}_k(\lambda) := n_k^{-1} \sum_{i \in F_k} \left[w_i(\widehat{F_n})\rho_\tau(Y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^k) + (1 - w_i(\widehat{F_n}))\rho_\tau(Y_i^{-\infty} - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^k) \right], \tag{2.22}$$

where F_k denotes the k-th fold of the n observations, n_k is the number of observations in F_k, and $\widehat{\boldsymbol{\beta}}^k$ is the parameter coefficients fitted on F_k^c observations. Likewise, the choice of λ_j in line 4 of Algorithm 2 is chosen in the same way, except in the cross validation statistic, the squared error loss is used instead of the weighted quantile loss in (2.22).

Table 2.1: $\tau = 0.4$ Coverage Probability for High-dimensional Left-censored Quantile Regression (HLQR) with True F_0 and True f_0

Distribution of the error term	Simulation Setting for $n = 200$, $p = 300$			
	Toeplitz design $\rho = 0.3$		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.95	0.97	0.95	0.93
Student's	0.95	0.94	0.95	0.92
Beta	0.90	0.93	0.91	0.93
Weibull	0.94	0.97	0.98	0.94

2.4.2 Simulation Data

We are now ready to present the simulation results. The size of the model settings are chosen to be of $n = 200$ for the number of observations, and $p = 300$ for the number of parameters. In addition, the sparsity of the underlying true parameter $\boldsymbol{\beta}^o$, denoted as $s_{\boldsymbol{\beta}^o}$ earlier in the text, is set to be 5. We have also selected four different distributions for the error of the model: standard normal, Student's t with 4 degrees of freedom, Beta distribution with parameters $(2,3)$ and Weibull distribution with parameters $(1,1)$. The design matrix $\mathbf{X}$ is generated from a multivariate Normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$, where μ is chosen to be the zero vector, and the covariance matrix $\mathbf{\Sigma}$ is taken to be the identity matrix or the Toeplitz matrix such that $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$ for $\rho = 0.3$. The two quantiles of interest are chosen to be $\tau = 0.4$ and $\tau = 0.7$. In the case when τ-th quantile of the error is not zero, we subtract off the τ-th quantile of the error distribution from the model. The censoring level c is chosen such that the proportion of the censoring data is set at 10%. We present simulation results for when the true F_0 and f_0 plugged in, and also when we use our proposed rudimentary estimators $\widehat{F}_n$ and $\widehat{f}_n$ as described earlier in the section.

Table 2.1 and 2.2 summarize the average coverage probabilities of the constructed 95% level confidence intervals for obtaining $\tau = 0.4$ and 0.7 quantile regression estimators under various settings. We report the signal and noise parameters separately, as the coverage of the signal ones are known to be more difficult. In conjunction, we have also included box plots of interval

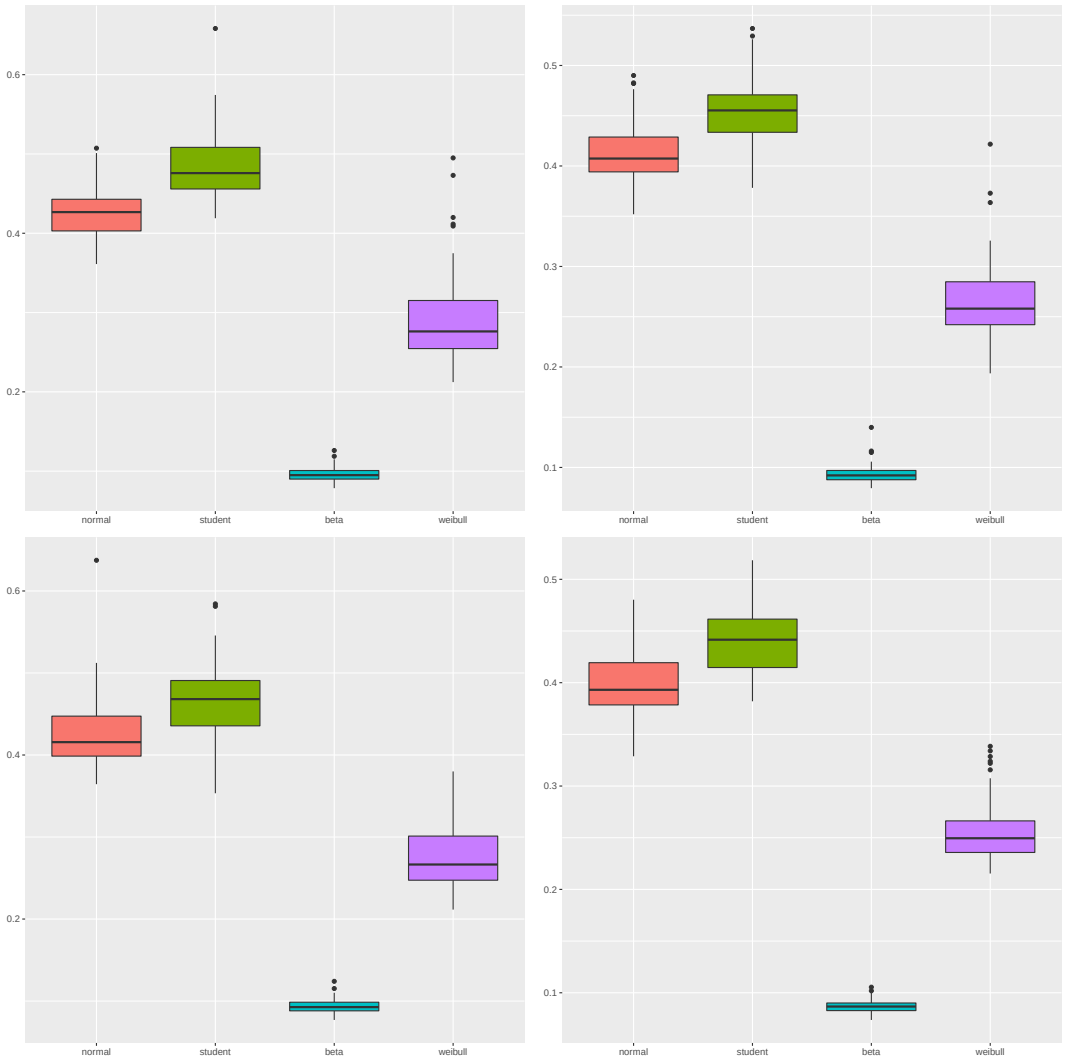

Figure 2.1: $\tau = 0.4$ comparative boxplots of the average interval length (with true F_0 and true f_0). Signal (left) and noise (right) variables, and Toeplitz design with $\rho = 0.3$ (top) and identity design (bottom).

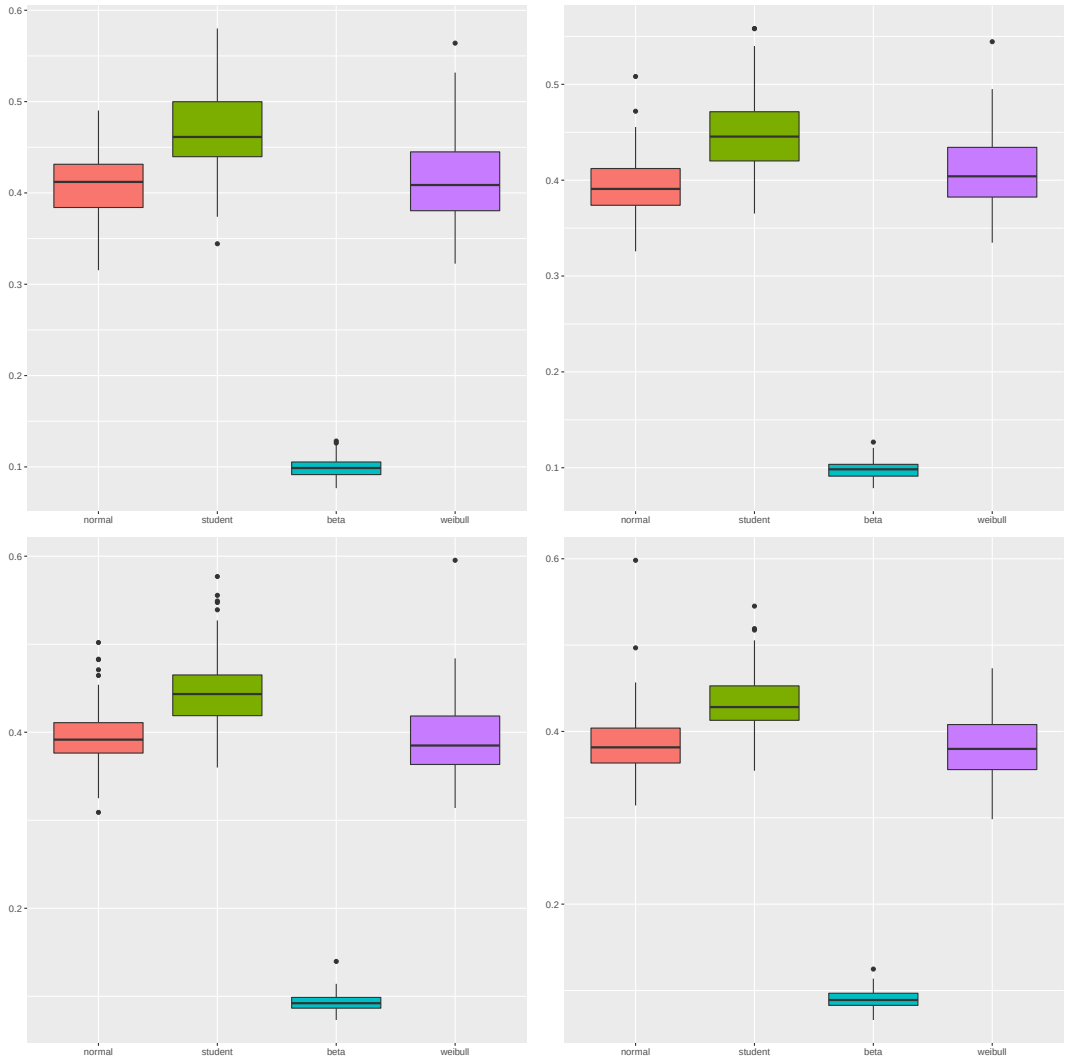

Figure 2.2: $\tau = 0.7$ comparative boxplots of the average interval length (with true F_0 and true f_0). Signal (left) and noise (right) variables, and Toeplitz design with $\rho = 0.3$ (top) and identity design (bottom).

Table 2.2: $\tau = 0.7$ Coverage Probability for High-dimensional Left-censored Quantile Regression (HLQR) with True F_0 and True f_0

Distribution of the error term	Simulation Setting for $n = 200$, $p = 300$, $\tau = 0.7$			
	Toeplitz design $\rho = 0.3$		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.94	0.97	0.92	0.97
Student's	0.91	0.94	0.91	0.95
Beta	0.96	0.99	0.89	0.95
Weibull	0.92	0.94	0.87	0.91

widths under these settings (Figure 2.1 and 2.2). From the results of applying our methodology with true F_0 and true f_0, it is observed that the coverage probabilities are approximately the same and are close to the nominal values. In addition, we noticed that among the four chosen error distributions, our method turns out to be most efficient, in terms of the confidence interval width, when the error distribution is bounded. However, it is observed that our method is sensitive to heavy-tailed distributions, such as the Student's t distribution with degrees of freedom being 4.

The results of plugging in estimators $\widehat{F}_n$ and $\widehat{f}_n$ are summarized in Table 2.3 and 2.4 for the two quantile settings $\tau = 0.4$ and 0.7. In terms of coverage probability, we observe similar results as the ones with true F_0 and f_0, as the probabilities are approximately the same and are close to the nominal values. We notice that the interval widths almost tripled for the cases of error being standard normal and Student's t distribution as seen in Figure 2.3 and 2.4. However, this is not unexpected as we using estimators instead of the true underlying values. With better tailored estimators to the scenario, we believe that the width of the intervals in the two cases can be reduced.

In addition, we have also examined the power of our estimator. Maintaining similar settings as in previous simulations, that is $n = 200$ and $p = 300$, whereas $s_{\boldsymbol{\beta}^o}$ is also set to be 5. We have our null hypothesis for the coefficients being 1 for the signals and 0 for the noises. We test $H_0 : \widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^o$ versus $H_1 : \widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^o + h$. While keeping the significance level at 0.05, we

Table 2.3: $\tau = 0.4$ Coverage Probability for High-dimensional Left-censored Quantile Regression (HLQR) with Estimated $\widehat{F}_n$ and Estimated $\widehat{f_n}$

Distribution of the error term	Simulation Setting for $n = 200$, $p = 300$			
	Toeplitz design $\rho = 0.3$		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.95	0.97	0.97	0.94
Student's	0.98	0.94	0.98	1.00
Beta	0.99	0.95	0.97	0.97
Weibull	0.99	0.92	0.96	0.95

Table 2.4: $\tau = 0.7$ Coverage Probability for High-dimensional Left-censored Quantile Regression (HLQR) with Estimated $\widehat{F}_n$ and Estimated $\widehat{f_n}$

Distribution of the error term	Simulation Setting for $n = 200$, $p = 300$, $\tau = 0.7$			
	Toeplitz design $\rho = 0.3$		Identity design	
	Signal Variable	Noise Variable	Signal Variable	Noise Variable
Normal	0.89	0.99	0.96	0.97
Student's	0.93	0.93	1.00	0.96
Beta	0.96	0.97	0.91	0.96
Weibull	0.95	0.95	0.99	0.96

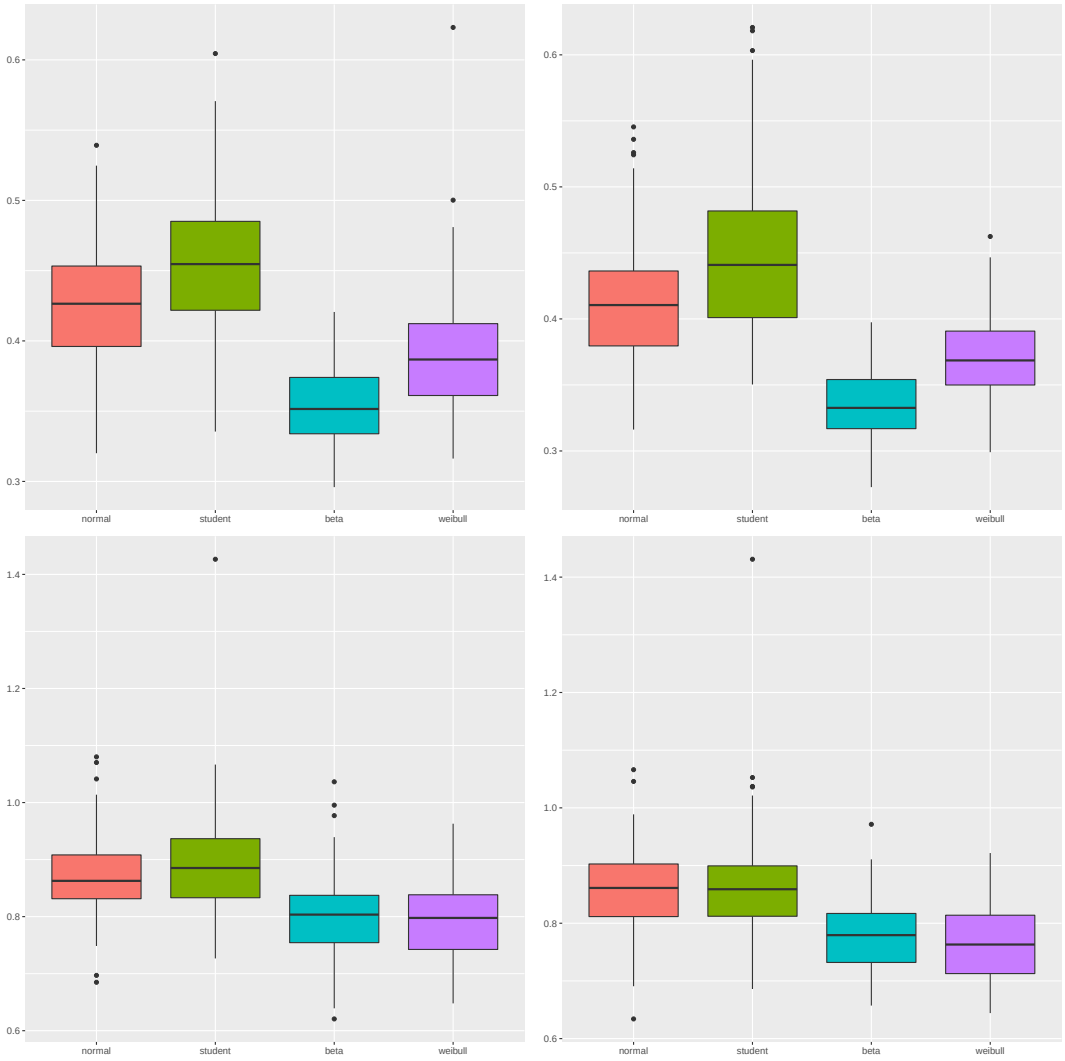

Figure 2.3: $\tau = 0.4$ comparative boxplots of the average interval length (with estimated $\widehat{F}_n$ and true $\widehat{f}_n$). Signal (left) and noise (right) variables, and Toeplitz design with $\rho = 0.3$ (top) and identity design (bottom).

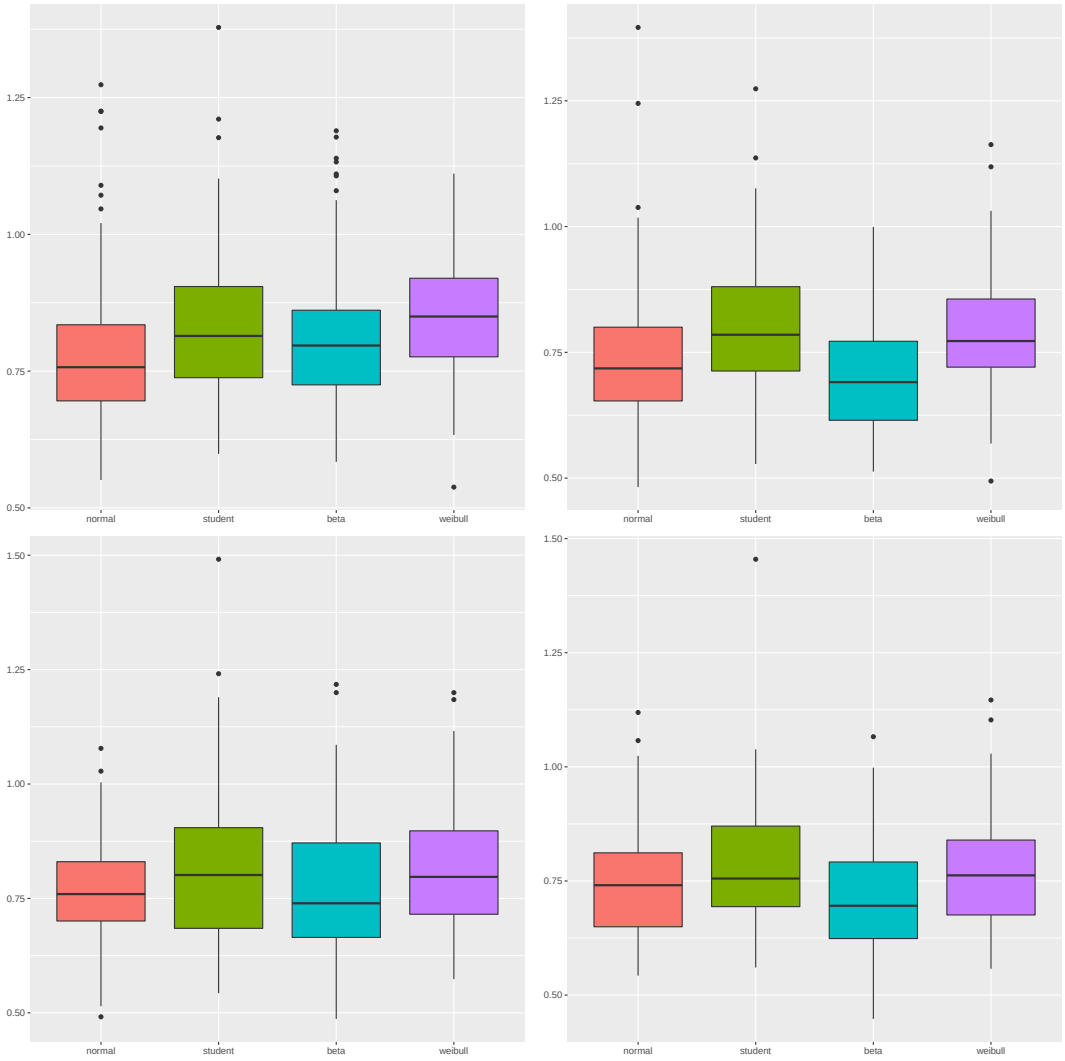

Figure 2.4: $\tau = 0.7$ comparative boxplots of the average interval length (with estimated $\widehat{F}_n$ and true $\widehat{f}_n$). Signal (left) and noise (right) variables, and Toeplitz design with $\rho = 0.3$ (top) and identity design (bottom).

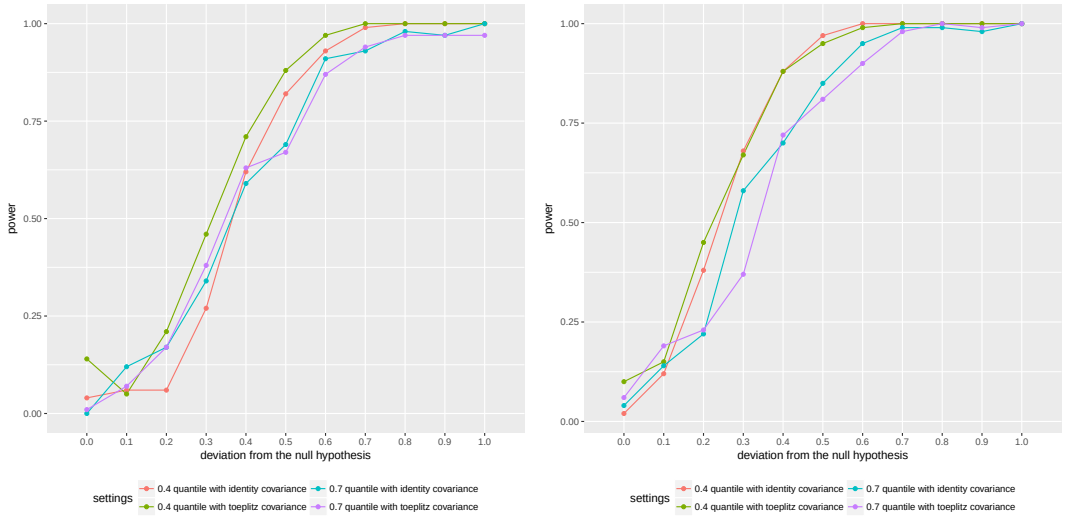

Figure 2.5: Power curve of signal (left) and noise (right) variables under normal errors, H_0 : $\boldsymbol{\beta}_j^o = c$ versus $H_1 : \boldsymbol{\beta}_j^o \neq c$, where the true parameter $\boldsymbol{\beta}_j^o = c + h$. The deviation from the null hypothesis h ranges from 0 to 1.

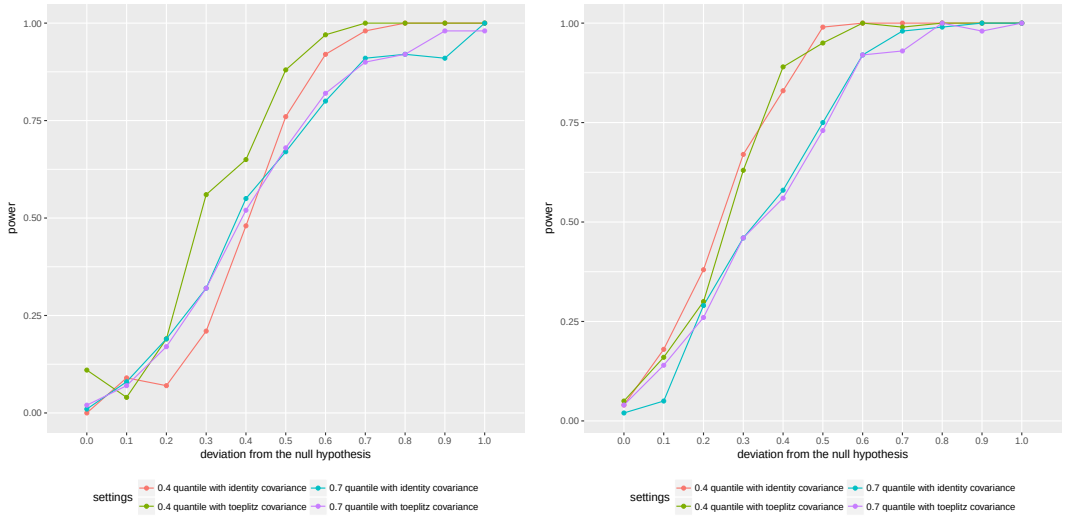

Figure 2.6: Power curve of signal (left) and noise (right) variables under normal errors, H_0 : $\boldsymbol{\beta}_j^o = c$ versus $H_1 : \boldsymbol{\beta}_j^o \neq c$, where the true parameter $\boldsymbol{\beta}_j^o = c + h$. The deviation from the null hypothesis h ranges from 0 to 1.

increase the deviation from the null hypothesis h gradually from 0.1 to 1. We observe that both the signal and noise variables converges to power of 1 quickly for various settings, which testifies the effectiveness of our estimator. The results are summarized in Figure 2.5 and 2.6 below.

2.4.3 Real Data

In this section, we apply our High-dimensional Left-censored Quantile Regression (HLQR) to a microarray dataset of cardiomyopathy in transgenic mice, kindly provided by Professor Mark Segal, who also studied the dataset in [SDC03]. To study human diseases such as chamber dilation and left ventricular conduction delay, a transgenic mouse model of dilated cardiomyopathy was used.

Specifically, [RDK$^+$00] proposed to control a G protein-coupled receptor, designated as Ro1, through an inducible expression system. Thirty mice are used for the study, and are divided into four experimental groups. Six transgenic mice expressed Ro1 for two weeks, which did not show symptoms of disease. Nine other transgenic mice expressed Ro1 for eight weeks, and exhibited cardiomyopathy symptoms. The recovery group consists of seven transgenic mice, whose expression of Ro1 was on for eight weeks and off for four weeks. Finally, the control group is made up of non-transgenic mice expressed Ro1 for eight weeks.

The goal is to identify genes involved in the Ro1 expression changes, which may provide new diagnostic markers for cardiomyopathy. To this end, Affymetrix Mu6500 arrays were used for the study, and the response of interest is Ro1, whereas the predictors are 6,319 microarray gene expressions. The dimensionality of the model is then 30 observations ($n = 30$) and 6,319 features ($p = 6319$). In order to verify the effectiveness of our High-dimensional Left-censored Quantile Regression framework, we artificially created a 10% censoring on the response Ro1 value, and fitted the dataset for five quantiles, $\tau = 0.5, 0.75,$ and 0.9. The regularization parameter in the initial estimator is chosen using a five-fold cross validation procedure as described in (2.22). The gene expressions deemed to be significant by the confidence intervals are summarized in Table

Table 2.5: Gene expressions selected by High-dimensional Left-censored Quantile Regression (HLQR) with 10% censoring in comparison with the ones selected by L_1 norm QR model in [LZ08] (L_1QR) with no censoring

GeneBank	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.9$	
	HLQR	L_1QR	HLQR	L_1QR	HLQR	L_1QR
D31717	$(97.68, 97.92)$	✓	$(97.65, 97.96)$	✓	$(97.61, 97.91)$	✓
U73744	$(20.09, 20.32)$	✓	$(20.08, 20.29)$	✓	$(20.06, 20.31)$	✓
U25708	$(46.61, 46.82)$	✓	$(46.60, 46.83)$	✓	$(46.60, 46.90)$	
AA061310	$(9.07, 9.26)$	✓	$(9.07, 9.22)$		$(9.05, 9.29)$	
M30127	$(-0.04, 0.06)$	✓	$(-0.03, 0.05)$	✓	$(-0.04, 0.06)$	✓
L38971	$(20.36, 20.54)$	✓	$(20.35, 20.54)$		$(20.34, 20.58)$	
Z32675	$(25.07, 25.28)$	✓	$(25.03, 25.15)$		$(25.02, 25.36)$	
W75373	$(41.96, 42.17)$	✓	$(41.94, 42.20)$		$(41.94, 42.16)$	
AA044561	$(0.02, 0.18)$		$(-0.01, 0.28)$	✓	$(-0.05, 0.33)$	
AA111168	$(-0.12, 0.22)$		$(-0.10, 0.17)$		$(-0.13, 0.21)$	✓
M18194	$(-0.04, 0.10)$		$(-0.12, 0.15)$		$(-0.04, 0.09)$	

2.5. We also noticed that the same dataset has also been studied in both [LZ08]. Thereby, we included real data results therein for comparison.

As one can see from Table 2.5, there are quite a few overlaps between the gene expressions selected in [LZ08] and the ones selected by our High-dimensional Left-censored Quantile Regression method, even with 10% of censoring introduced. In addition to merely identifying the significant genes, our methodology is capable of providing a precise confidence interval for the significant gene expressions. Moreover, we notice that the sets of selected genes by models across various quantiles, i.e. $\tau = 0.5, 0.75,$ and 0.9, using our HLQR are more consistent than the sets reported for models with different quantiles from L_1QR. In other words, our methodology tends to agree on a common set of significant gene expressions across models with different quantile levels.

The starkest contrast between the gene expressions reported can be seen in M30127 (Mouse MHC class I tum-transplantation antigen P35B gene), whose importance has been noted consistently across quantiles in L_1QR, whereas our HLQR procedure does not find the expression significant. Instead, we do notice that our resulting confidence interval does suggest the significance of another gene expression M20985 (Mouse MHC class I H2-Qa-Mb1 gene).

The confidence intervals for M20985 is as following $(91.14, 91.32)$ in $\tau = 0.5$, $(91.14, 91.30)$ in $\tau = 0.75$, and $(91.11, 91.35)$ in $\tau = 0.9$. Whereas as of date the M30127 expression's role in the cardiomyopathy development is yet to be determined, [PSA$^+$10] has confirmed that M20985 is part of a locus that confers susceptibility of viral-induced chronic myocarditis. In such case, our methodology has correctly identified a substantial gene candidate for further study of the disease.

Last but not the least, we would like to emphasize on the necessity of considering censoring data cases. In fact, it is difficult to accurately measure absolute expression levels and reliably detect low abundance genes [DKES06]. Thus, we believe our method would be a great asset for researchers analyzing datasets, which have observations with lower detection limit.

2.5 Lemmas

The following result gives a bound on the estimation error of our inverse Hessian estimator $\widehat{\mathbf{\Theta}}_j$ to the underlying population quantity $\mathbf{\Theta}^0_j$.

Lemma 7. *Under Conditions 1 - 7,*

$$\|\widehat{\mathbf{\Theta}}_{\widehat{\boldsymbol{\beta}},j} - \mathbf{\Theta}_{\boldsymbol{\beta}^o,j}\|_1 = O_p(\lambda_j s_j) + O_p(K\sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j}) + O_p(K(\lambda s_{\boldsymbol{\beta}^o} s_j^2/n)^{1/4}) + O_p(\sqrt{s_j \delta_{f,n}} K),$$

where $\delta_{f,n} := n^{-1}\sum_{i=1}^n \left(\widehat{f}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i) - f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)\right)^2$. *For bounded case,* $K = \sqrt{s_j}$, *and* $K = 1$ *in the strongly bounded case.*

Remark 13. *In particular, in the bounded case, if we choose* $\lambda \asymp \sqrt{\log(p)/n}$, $\lambda_j \asymp \sqrt{\log(p)/n}$, $s_j^2 s_{\boldsymbol{\beta}^o}\sqrt{\log(p)/n} = o_p(1)$, $s_{\boldsymbol{\beta}^o} s_j^4 \sqrt{\log(p)/n^3} = o_p(1)$ *and* $s_j\sqrt{\delta_{f,n}} = o_p(1)$, *then*

$$\|\widehat{\mathbf{\Theta}}_{\widehat{\boldsymbol{\beta}},j} - \mathbf{\Theta}_{\boldsymbol{\beta}^o,j}\|_1 = o_p(1).$$

In the strongly bounded case, we only require $\lambda \asymp \sqrt{\log(p)/n}$, $\lambda_j \asymp \sqrt{\log(p)/n}$,

$$s_j s_{\boldsymbol{\beta}^o} \sqrt{\log(p)/n} = o_p(1)$$

and $s_j \delta_{f,n} = o_p(1)$.

Finally, we begin presenting preliminary results for each term in the decomposition (2.17). We start with term (III), which measures the error of the one-step improvement quantity using the estimator $\widehat{F}_n$.

Lemma 8. *Under Condition 1 - 7, for* $\widehat{F}_n$ *chosen to be as in* (2.18)

$$III = -\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \phi_i + O_p\left(\frac{K}{n} + K\left(\frac{\log n}{n}\right)^{3/4}\right),$$

where $K = \sqrt{s_j}$, *and in the strongly bounded case,* $K = 1$, *and*

$$\phi_i := \tau\, \mathbb{I}\left(\mathbf{x}_i \boldsymbol{\beta}^o \le 0\right) \mathbb{I}\left(T_i \le 0\right) \frac{\mathbb{I}(F_0 > \tau)}{F_0^2} \sum_{\substack{l=1 \\ l \neq i}}^{n} B_{nl}(\mathbf{x}_i) \left(\frac{\mathbb{I}(Y_l > 0, \delta_l = 1)}{F_0(Y_l|\mathbf{x})} - \int_{\max\{0, Y_l\}}^{\infty} \frac{\mathrm{d}F_0(s|\mathbf{x})}{F_0^2(s|\mathbf{x})} \right).$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \phi_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\phi^2}{n}\right),$$

where $\sigma_\phi^2 = \mathbb{E}\widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega}_\phi \widehat{\boldsymbol{\Theta}}_j^\top$ *and* $\boldsymbol{\Omega}_\phi := \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \phi_i^2 / n$.

Remark 14. *Lemma 8 implies that an additional normality term results from using the classical Kaplan-Meier estimator as a proxy for the true distribution* F. *Such a term can be understood as the extra variability due to the missing information regarding underlying distribution.*

In the following, we apply linearization on the term (Δ) and then combine the term together with (I), which then gives us the following Lemma. The rationale behind such arrangement is that

the term (Δ) describes the difference in the one-step correction with expectation of score using initial estimator $\widehat{\boldsymbol{\beta}}$, whereas the term (I) is exactly the difference of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^o$.

Lemma 9. *Under Conditions 1 - 7, when* $\left\|\widehat{\boldsymbol{\Theta}}_{\widehat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^o,j}\right\|_1 = o_p(1)$,

$$|I - \Delta| = O_p\left(K\lambda_j \lambda s_{\boldsymbol{\beta}^o}\right) + O_p(K\lambda^2 s^2_{\boldsymbol{\beta}^o}),$$

where $K = \sqrt{s_j}$, *and in the strongly bounded case,* $K = 1$.

For part (II), we have the following lemma, which aims to bound the difference of a empirical process.

Lemma 10. *Under Conditions 1 - 7,*

$$|II| = O_p\left(\sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j / n}\right).$$

Last but not the least, we show the normality of the term $\sqrt{n}\widehat{\boldsymbol{\Theta}}\mathbf{S}_n(\boldsymbol{\beta}^o)$ for part (N). The lemma shows that the leading term of the Bahadur decomposition (2.17) follows a normal distribution.

Lemma 11. *Assuming Conditions 1 - 7,*

$$N = -\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \psi_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\psi^2}{n}\right),$$

where $\psi_i = -\left[w_i(F_0)\, \mathbb{I}\{Y_i - \mathbf{x}_i\boldsymbol{\beta}^o \geq 0\} - (1-\tau)\right]$, *and* $\sigma_\psi^2 = \mathbb{E}\widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega}_\psi \widehat{\boldsymbol{\Theta}}_j^\top$ *and*

$$\boldsymbol{\Omega}_\psi := \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i \psi_i^2 / n.$$

2.6 Proofs of Lemmas

Proof of Lemma 7. Let $\mathbf{w}_{\boldsymbol{\beta}}$ be the diagonal of the weighted matrix $\mathbf{W}_{\boldsymbol{\beta}}$. Denote

$$\mathbf{X}_{\boldsymbol{\beta}^o,j} = \mathbf{X}_{\boldsymbol{\beta}^o,-j}\gamma_{\boldsymbol{\beta}^o,j} + \eta_{\boldsymbol{\beta}^o,j}, \tag{2.23}$$

and

$$\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j} = \underset{\gamma}{\operatorname{argmin}} \|\mathbf{X}_{\widehat{\boldsymbol{\beta}},j} - \mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}\gamma\|_n^2 + 2\lambda_j\|\gamma\|_1,$$

where $\gamma_{\boldsymbol{\beta}^o,j} = \operatorname{argmin}_\gamma \mathbb{E}\|\mathbf{X}_{\boldsymbol{\beta}^o,j} - \mathbf{X}_{\boldsymbol{\beta}^o,-j}\gamma\|_n^2$. Define

$$\eta_j := \mathbf{X}_j - \mathbf{X}_{-j}\gamma_{\boldsymbol{\beta}^o,j}, \tag{2.24}$$

we can rewrite equation (2.23) as

$$\mathbf{W}_{\boldsymbol{\beta}^o}\mathbf{X}_j = \mathbf{W}_{\boldsymbol{\beta}^o}\mathbf{X}_{-j}\gamma_{\boldsymbol{\beta}^o,j} + \mathbf{W}_{\boldsymbol{\beta}^o}\eta_j,$$

and similarly by (2.24), we also have

$$\mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}_j = \mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}_{-j}\gamma_{\boldsymbol{\beta}^o,j} + \mathbf{W}_{\widehat{\boldsymbol{\beta}}}\eta_j. \tag{2.25}$$

By the definition of $\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}$,

$$\begin{aligned} & \|\mathbf{X}_{\widehat{\boldsymbol{\beta}},j} - \mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}\|_n^2 + 2\lambda_j\|\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}\|_1 \\ \leq \; & \|\mathbf{X}_{\widehat{\boldsymbol{\beta}},j} - \mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}\gamma_{\boldsymbol{\beta}^o,j}\|_n^2 + 2\lambda_j\|\gamma_{\boldsymbol{\beta}^o,j}\|_1. \end{aligned}$$

Replacing $\widehat{\mathbf{X}}_{\widehat{\boldsymbol{\beta}},j}$ by (2.25) and rearranging terms, we get

$$\begin{aligned} & \|\mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}-\gamma_{\boldsymbol{\beta}^o,j})\|_n^2+2\lambda_j\|\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}\|_1 \\ \leq & \frac{2}{n}\left(\mathbf{W}^2_{\widehat{\boldsymbol{\beta}}}\eta_j\right)^\top \mathbf{X}_{-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}-\gamma_{\boldsymbol{\beta}^o,j})+2\lambda_j\|\gamma_{\boldsymbol{\beta}^o,j}\|_1 \\ = & \frac{2}{n}\eta_{\boldsymbol{\beta}^0,j}^\top \mathbf{X}_{\boldsymbol{\beta}^o,-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}-\gamma_{\boldsymbol{\beta}^o,j})+2\lambda_j\|\gamma_{\boldsymbol{\beta}^o,j}\|_1+\text{Rem}, \end{aligned}$$

where the remainder $\text{Rem}=(2/n)\left((\mathbf{W}^2_{\widehat{\boldsymbol{\beta}}}-\mathbf{W}^2_{\boldsymbol{\beta}^o})\eta_j\right)^\top \mathbf{X}_{-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}-\gamma_{\boldsymbol{\beta}^o,j})$. Note that by Condition 5, $\|\eta_j\|_\infty \leq \|\mathbf{X}_j\|_\infty+\|\mathbf{X}_{-j}\gamma_{\boldsymbol{\beta}^o,j}\|_\infty = O_p(\sqrt{s_j})$. In the strongly bounded case, we have the projection $\|\mathbf{X}_{\boldsymbol{\beta}^o,-j}\gamma_{\boldsymbol{\beta}^o,j}\|_\infty = O_p(1)$, hence $\|\eta_j\|_\infty = O_p(1)$. In the following, we write $\|\eta_j\|_\infty = O_p(K)$ where $K=\sqrt{s_j}$ in general case, and $K=1$ when data is strongly bounded.

We can bound the remainder term

$$|\text{Rem}| \leq \frac{2}{n}\|(\mathbf{W}^2_{\widehat{\boldsymbol{\beta}}}-\mathbf{W}^2_{\boldsymbol{\beta}^o})\eta_j\|_2\|\mathbf{X}_{-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j}-\gamma_{\boldsymbol{\beta}^o,j})\|_2.$$

Therefore,

$$
\begin{aligned}
& n^{-1}\|(\mathbf{W}^2_{\widehat{\boldsymbol{\beta}}} - \mathbf{W}^2_{\boldsymbol{\beta}^o})\eta_{\boldsymbol{\beta}^0,j}\|_2^2 \\
\leq\ & \frac{1}{n}\|\eta_{\boldsymbol{\beta}^0,j}\|_\infty^2 \sum_{i=1}^n \left(\mathbf{w}^2_{\widehat{\boldsymbol{\beta}},i} - \mathbf{w}^2_{\boldsymbol{\beta}^o,i}\right)^2 \\
=\ & \frac{1}{n}\|\eta_{\boldsymbol{\beta}^0,j}\|_\infty^2 \sum_{i=1}^n \left(\widehat{f}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i)\ \mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}} > 0) - f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)\ \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta}^o > 0)\right)^2 \\
\leq\ & \frac{1}{n}\|\eta_{\boldsymbol{\beta}^0,j}\|_\infty^2 \left\{\sum_{i=1}^n \left(\widehat{f}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i) - f_0(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i)\right)^2 + \sum_{i=1}^n \left(f_0(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i) - f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)\right)^2\right\} \\
& + \frac{1}{n}\|\eta_{\boldsymbol{\beta}^o,j}\|_\infty \sum_{i=1}^n f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)^2 \left(\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}} > 0) - \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta}^o > 0)\right)^2 \\
=\ & \delta_{f,n}O_p(K^2) + \frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|_2^2 O_p(K^2) + \frac{1}{n}\sum_{i=1}^n \left(\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}} > 0) - \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta}^o > 0)\right)^2 O_p(K^2) \\
=\ & O(\delta_{f,n}K^2) + O_p(\lambda^2 s_{\boldsymbol{\beta}^o}K^2) + O_p(K^2)\frac{1}{n}\sum_{i=1}^n \mathscr{B}_i(\widehat{\boldsymbol{\beta}})
\end{aligned}
$$

where $\delta_{f,n} = n^{-1}\sum_{i=1}^n \left(\widehat{f}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i) - f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)\right)^2$ and $\mathscr{B}_i(\boldsymbol{\beta}) = \left(\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0) - \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta}^o > 0)\right)^2$. Observe that for any fixed $\boldsymbol{\beta}$, $\mathscr{B}_i(\boldsymbol{\beta})$ is Bernoulli random variable. Let $\mathscr{P} = \mathbb{P}(\mathscr{B}_i = 1)$. Note that

$$
\max_i |\mathbf{x}_i\boldsymbol{\beta} - \mathbf{x}_i\boldsymbol{\beta}^o| = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}^o\|_\infty \leq \|\mathbf{X}\|_\infty\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1 \leq K_X\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1,
$$

and

$$
\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i\boldsymbol{\beta} - \mathbf{x}_i\boldsymbol{\beta}^o)^2 = n^{-1}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^o)\|_2^2.
$$

Therefore, $\mathscr{P} \leq \mathbb{P}(|\mathbf{x}_i\boldsymbol{\beta}^o| \leq K_X\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1) = O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1)$ by the boundedness of density f_0. By Chernoff inequality,

$$
\left|\frac{1}{n}\sum_{i=1}^n \mathscr{B}_i(\boldsymbol{\beta})\right| = O_p(\mathscr{P}) + O_p\left(\frac{\sqrt{\mathscr{P}(1-\mathscr{P})}}{\sqrt{n}}\right).
$$

Hence, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{B}_i(\widehat{\boldsymbol{\beta}})\right| = O_p(\lambda s_{\boldsymbol{\beta}^o}) + O_p\left(\frac{\sqrt{\lambda s_{\boldsymbol{\beta}^o}}}{\sqrt{n}}\right).$$

Therefore, for any $\delta > 0$,

$$\begin{aligned}|\text{Rem}| = \delta\|\mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j} - \gamma_{\boldsymbol{\beta}^o,j})\|_n^2 + O(\delta_{f,n}K^2) \\ + O_p(\lambda^2 s_{\boldsymbol{\beta}^o}K^2) + O_p(K^2\lambda s_{\boldsymbol{\beta}^o}) + O_p(K^2\sqrt{\lambda s_{\boldsymbol{\beta}^o}/n}).\end{aligned}$$

By the standard arguments, choosing $\lambda_j \asymp \sqrt{\log(p)/n}$, we get

$$\|\mathbf{X}_{\widehat{\boldsymbol{\beta}},-j}(\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j} - \gamma_{\boldsymbol{\beta}^o,j})\|_n^2 = O_p(\lambda_j^2 s_j) + O(\delta_{f,n}K^2) + O_p(\lambda s_{\boldsymbol{\beta}^o}K^2) + O_p(K^2\sqrt{\lambda s_{\boldsymbol{\beta}^o}/n})$$

and

$$\|\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j} - \gamma_{\boldsymbol{\beta}^o,j}\|_1 = O_p(\lambda_j s_j) + O(\sqrt{\delta_{f,n}}K\sqrt{s_j}) + O_p(K\sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j}) + O_p(K(\lambda s_{\boldsymbol{\beta}^o} s_j^2/n)^{1/4}).$$

Using (2.24) again, we get

$$\widehat{d}^2_{\widehat{\boldsymbol{\beta}},j} - d^2_{\boldsymbol{\beta}^o,j} = \underbrace{\mathbf{X}^\top_{\boldsymbol{\beta}^o,j}(\mathbf{X}_{\boldsymbol{\beta}^o,j} - \mathbf{X}_{\boldsymbol{\beta}^o,-j}\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j})/n - d^2_{\boldsymbol{\beta}^o,j}}_{(i)} + \underbrace{\mathbf{X}_j^\top(\mathbf{W}^2_{\widehat{\boldsymbol{\beta}}} - \mathbf{W}^2_{\boldsymbol{\beta}^o})(\mathbf{X}_j - \mathbf{X}_{-j}\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j})/n}_{(ii)}.$$

By Theorem 2.4 in [VdGBR$^+$14], we have $(i) = O_p(\lambda_j\sqrt{s_j})$. For the second part (ii), by Condition 5,

$$\begin{aligned}(ii) &= O_p(K)\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{f}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i)\,\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}} > 0) - f_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i)\,\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta}^o > 0)\right| \\ &= O_p(\sqrt{\delta_{f,n}}K) + O_p(\lambda\sqrt{s_{\boldsymbol{\beta}^o}}K) + O_p(K(\lambda s_{\boldsymbol{\beta}^o}/n)^{1/4}).\end{aligned}$$

Therefore,

$$\left|\widehat{d}^2_{\widehat{\boldsymbol{\beta}},j} - d^2_{\boldsymbol{\beta}^o,j}\right| = O_p(\lambda_j\sqrt{s_j}) + O_p(\sqrt{\delta_{f,n}}K) + O_p(\lambda\sqrt{s_{\boldsymbol{\beta}^o}}K) + O_p(K(\lambda s_{\boldsymbol{\beta}^o}/n)^{1/4}).$$

Combining all previous results,

$$\begin{aligned}
& \|\widehat{\Theta}_{\widehat{\boldsymbol{\beta}},j} - \Theta_{\boldsymbol{\beta}^o,j}\|_1 \\
\leq\ & \|\widehat{\gamma}_{\widehat{\boldsymbol{\beta}},j} - \gamma_{\boldsymbol{\beta}^o,j}\|_1/\widehat{d}^2_{\widehat{\boldsymbol{\beta}},j} + \|\gamma_{\boldsymbol{\beta}^o,j}\|_1\left(1/\widehat{d}^2_{\widehat{\boldsymbol{\beta}},j} - 1/d^2_{\boldsymbol{\beta}^o,j}\right) \\
=\ & O_p(\lambda_j s_j) + O_p(K\sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j}) + O_p(K(\lambda s_{\boldsymbol{\beta}^o} s_j^2/n)^{1/4}) + O_p(\sqrt{s_j\delta_{f,n}}K).
\end{aligned}$$

□

Proof of Lemma 8. We begin with expanding on the following difference,

$$\widehat{\boldsymbol{\Theta}}_j\left(\mathbf{S}_n(\widehat{\boldsymbol{\beta}},\widehat{F}_n) - \mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)\right) = \widehat{\boldsymbol{\Theta}}_j\left.\frac{\partial\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F)}{\partial F}\right|_{F=F_0}\left(\widehat{F}_n - F_0\right) + \frac{1}{2}\left.\frac{\partial^2\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F)}{\partial F^2}\right|_{F=\widetilde{F}}\left(\widehat{F}_n - F_0\right)^2, \tag{2.26}$$

for some $\widetilde{F}$ between $\widehat{F}_n$ and F_0. We then work on rewriting the terms in the summation of

$\mathbf{S}_n(\boldsymbol{\beta}, F)$. Let $\mathbf{S}_n(\boldsymbol{\beta}, F) := n^{-1}\sum_{i=1}^n \mathbf{s}_i(\boldsymbol{\beta}, F)$,

$$
\begin{aligned}
\mathbf{s}_i(\boldsymbol{\beta}, F) &= -\mathbf{x}_i^\top \left[w_i(F)\, \mathbb{I}(Y_i - \mathbf{x}_i\boldsymbol{\beta} \geq 0) + \tau - 1 \right] \\
&= -\mathbf{x}_i^\top \Big[\mathbb{I}(T_i \leq 0) \left(\tau - 1 + \frac{\tau}{F}\, \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} \leq 0)\, \mathbb{I}(F > \tau) \right) \\
&\qquad + \mathbb{I}(T_i > 0)\left(\tau - 1 + \mathbb{I}(T_i \geq \mathbf{x}_i\boldsymbol{\beta})\right) \Big] \\
&= -\mathbf{x}_i^\top \Big[\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} \leq 0) \left(\mathbb{I}(T_i \leq 0)(\tau - 1) + \mathbb{I}(T_i \leq 0)\frac{\tau}{F}\, \mathbb{I}(F > \tau) + \tau\, \mathbb{I}(T_i > 0) \right) \\
&\qquad + \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0)\left(\mathbb{I}(T_i \leq 0)(\tau - 1) + \mathbb{I}(T_i > 0)(\tau - 1) + \mathbb{I}(T_i \geq \mathbf{x}_i\boldsymbol{\beta}) \right) \Big] \\
&= -\mathbf{x}_i^\top \Big[\mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} \leq 0) \left(\tau - \mathbb{I}(T_i \leq 0) + \mathbb{I}(T_i \leq 0)\frac{\tau}{F}\, \mathbb{I}(F > \tau) \right) \\
&\qquad + \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0)\left(\tau - 1 + \mathbb{I}(T_i \geq \mathbf{x}_i\boldsymbol{\beta}) \right) \Big] \\
&= -\mathbf{x}_i^\top \Big[\tau - \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} \leq 0, T_i \leq 0) + \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} \leq 0, T_i \leq 0)\, \frac{\tau}{F}\, \mathbb{I}(F > \tau) \\
&\qquad - \mathbb{I}(\mathbf{x}_i\boldsymbol{\beta} > 0) + \mathbb{I}(T_i \geq \mathbf{x}_i\boldsymbol{\beta}, \mathbf{x}_i\boldsymbol{\beta} > 0) \Big].
\end{aligned}
$$

We derive the first derivative of $\mathbf{S}_n$ with respect to F at F_0,

$$
\begin{aligned}
\left.\frac{\partial \mathbf{S}_n(\widehat{\boldsymbol{\beta}}, F)}{\partial F}\right|_{F=F_0} &= \lim_{\varepsilon \to 0} -\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}} \leq 0, T_i \leq 0\right) \\
&\qquad \times \frac{1}{\varepsilon(F - F_0)} \left(\frac{\mathbb{I}(F_0 + \varepsilon(F - F_0) > \tau)}{F_0 + \varepsilon(F - F_0)} - \frac{\mathbb{I}(F_0 > \tau)}{F_0} \right) \\
&= -\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}} \leq 0, T_i \leq 0\right) \\
&\qquad \times \lim_{\varepsilon \to 0} \frac{1}{\varepsilon(F - F_0)} \left(\frac{\mathbb{I}(F_0 + \varepsilon(F - F_0) > \tau)}{F_0 + \varepsilon(F - F_0)} - \frac{\mathbb{I}(F_0 > \tau)}{F_0 + \varepsilon(F - F_0)} \right. \\
&\qquad\qquad \left. + \frac{\mathbb{I}(F_0 > \tau)}{F_0 + \varepsilon(F - F_0)} - \frac{\mathbb{I}(F_0 > \tau)}{F_0} \right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}} \leq 0\right) \mathbb{I}(T_i \leq 0)\, \frac{\mathbb{I}(F_0 > \tau)}{F_0^2},
\end{aligned}
$$

where the details of taking the limit is as the following.

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{1}{\varepsilon(F-F_0)}\left(\frac{\mathbb{I}(F_0+\varepsilon(F-F_0)>\tau)}{F_0+\varepsilon(F-F_0)}-\frac{\mathbb{I}(F_0>\tau)}{F_0+\varepsilon(F-F_0)}+\frac{\mathbb{I}(F_0>\tau)}{F_0+\varepsilon(F-F_0)}-\frac{\mathbb{I}(F_0>\tau)}{F_0}\right)\\
&=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon(F-F_0)}\left(\frac{\mathbb{I}(F_0+\varepsilon(F-F_0)>\tau)-\mathbb{I}(F_0>\tau)}{F_0+\varepsilon(F-F_0)}-\frac{\varepsilon(F-F_0)}{F_0(F_0+\varepsilon(F-F_0))}\mathbb{I}(F_0>\tau)\right)\\
&=-\frac{\mathbb{I}(F_0>\tau)}{F_0^2},
\end{aligned}$$

since F_0 is bounded away from τ. Likewise, we have the second derivative of $\mathbf{S}_n$ with respect to F at $\widetilde{F}$ as

$$\left.\frac{\partial^2\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F)}{\partial F^2}\right|_{F=\widetilde{F}}=-\frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\tau\,\mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\le 0\right)\mathbb{I}\left(T_i\le 0\right)\frac{\mathbb{I}(\widetilde{F}>\tau)}{\widetilde{F}^3},$$

as for F close to F_0, $\widetilde{F}$ is also bounded away from τ.

Plugging the derivatives into (2.26), we have

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}}_j\left(\mathbf{S}_n(\widehat{\boldsymbol{\beta}},\widehat{F}_n)-\mathbf{S}_n(\widehat{\boldsymbol{\beta}},F_0)\right)=&\underbrace{\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^\top\tau\,\mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\le 0\right)\mathbb{I}\left(T_i\le 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2}\left(\widehat{F}_n-F_0\right)}_{\text{(i)}}\\
&\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^\top\tau\,\mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\le 0\right)\mathbb{I}\left(T_i\le 0\right)\frac{\mathbb{I}(\widetilde{F}>\tau)}{\widetilde{F}^3}\left(\widehat{F}_n-F_0\right)^2}_{\text{(ii)}}.
\end{aligned}$$

Following the framework of Theorem 1 of [LS86] and Theorem 2.3 of [GMCS94] that for the classical Kaplan-Meier estimator $\widehat{F}_n$ as defined in (2.18), we have the following linearization.

$$\widehat{F}_n(0|\mathbf{x})-F_0(0|\mathbf{x})=\frac{1}{n}\sum_{l=1}^{n}\zeta(Y_l,\delta_l,\mathbf{x})+O_p\left(\left(\frac{\log n}{n}\right)^{3/4}\right)=O_p\left(\frac{1}{\sqrt{n}}+\left(\frac{\log n}{n}\right)^{3/4}\right)$$

for some $\boldsymbol{\theta}_i$ between $(\mathbf{x}-\mathbf{x}_i)/h_n$ and $(\mathbf{x}-\mathbf{x}_l)/h_n$, where

$$\zeta(Y_l,\delta_l,\mathbf{x}) = \frac{\mathbb{I}(Y_l>0,\delta_l=1|\mathbf{x})}{F_0(Y_l|\mathbf{x})} - \int_{\max\{0,Y_l\}}^{\infty} \frac{\mathrm{d}F_0(s|\mathbf{x})}{F_0^2(s|\mathbf{x})}.$$

In fact, for $i\neq l$, $\mathbb{I}(T_i\leq 0)\,\zeta(Y_l,\delta_l,\mathbf{x})$ are independent random variables with mean zero and finite variances for any given $\mathbf{x}$.

Replacing the term $\left(\widehat{F}_n - F_0\right)$ with its linearization, and separating the terms of $i=l$ from $i\neq l$, for term (i), we have

$$\begin{aligned}
\text{(i)} = {} & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\boldsymbol{\beta}^o\leq 0\right)\mathbb{I}\left(T_i\leq 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2} B_{nl}(\mathbf{x}_i)\zeta(Y_l,\delta_l,\mathbf{x}_i) \\
& + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \tau\left(\mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\leq 0\right) - \mathbb{I}\left(\mathbf{x}_i\boldsymbol{\beta}^o\leq 0\right)\right) \qquad (2.27)\\
& \qquad \times \mathbb{I}\left(T_i\leq 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2} B_{nl}(\mathbf{x}_i)\zeta(Y_l,\delta_l,\mathbf{x}_i) \\
& + \frac{1}{n^2}\sum_{i=1}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\leq 0\right)\mathbb{I}\left(T_i\leq 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2} B_{ni}(\mathbf{x}_i)\zeta(Y_i,\delta_i,\mathbf{x}_i) \\
& + \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\leq 0\right)\mathbb{I}\left(T_i\leq 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2}\right)\cdot O_p\left(\left(\frac{\log n}{n}\right)^{3/4}\right) \\
= {} & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n} \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \tau\, \mathbb{I}\left(\mathbf{x}_i\boldsymbol{\beta}^o\leq 0\right)\mathbb{I}\left(T_i\leq 0\right)\frac{\mathbb{I}(F_0>\tau)}{F_0^2} B_{nl}(\mathbf{x}_i)\zeta(Y_l,\delta_l,\mathbf{x}_i) \qquad (2.28)\\
& + O_p\left(\frac{K\lambda s_{\boldsymbol{\beta}^o}}{n}\right) + O_p\left(\frac{K}{n^{3/2}}\right) + O_p\left(K\left(\frac{\log n}{n}\right)^{3/4}\right), \qquad (2.29)
\end{aligned}$$

where $K=\sqrt{s_j}$, and in the strongly bounded case, $K=1$. The order in (2.29) results from the condition that $\|\widehat{\boldsymbol{\Theta}}_{\widehat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^o,j}\|_1 = o_p(1)$, and similar arguments as in Lemma 7. For the other

term (ii), we can bound it as following,

$$\begin{aligned}(\text{ii}) &= \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^\top\tau\,\mathbb{1}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\le 0\right)\mathbb{1}\left(T_i\le 0\right)\frac{\mathbb{1}(\widetilde{F}>\tau)}{\widetilde{F}^3}\right)\cdot O_p\left(\frac{1}{n}+\left(\frac{\log n}{n}\right)^{3/2}+\frac{\log^{3/4}n}{n^{4/5}}\right)\\ &= O_p\left(\frac{K}{n}+K\left(\frac{\log n}{n}\right)^{3/2}\right).\end{aligned}$$

For convenience in notations, define random variables ϕ_i as following,

$$\phi_i := \tau\,\mathbb{1}\left(\mathbf{x}_i\boldsymbol{\beta}^o\le 0\right)\mathbb{1}\left(T_i\le 0\right)\frac{\mathbb{1}(F_0>\tau)}{F_0^2}\sum_{\substack{l=1\\ l\neq i}}^{n}B_{nl}(\mathbf{x}_i)\zeta(Y_l,\delta_l,\mathbf{x}_i).$$

Then $\{\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^\top\phi_i\}_{i=1}^n$ are i.i.d. mean zero random variables with finite variance. Thus, by the central limit theorem, (2.28) $\xrightarrow{d}\mathcal{N}\left(0,\sigma_1^2/n\right)$, where $\sigma_1^2=\mathbb{E}\widehat{\boldsymbol{\Theta}}_j\boldsymbol{\Omega}_1\widehat{\boldsymbol{\Theta}}_j^\top$, and $\boldsymbol{\Omega}_1 := \sum_{i=1}^n\mathbf{x}_i^\top\mathbf{x}_i\phi_i^2/n$.

□

Lemma 12 (Preliminary Result for Lemma 9)**.** *By the construction of inverse matrix* $\boldsymbol{\Theta}^0$ *and* $\widehat{\boldsymbol{\Theta}}$*, we have* $1/\widehat{d}_j^2 = O(1)$.

Proof of Lemma 12. First, we note that $\boldsymbol{\Theta}^0_{j,j}=1/d_j^2$, which is a result of the KKT condition following similar arguments as in 2.3.1 of [BG16]. Second, following the proof of lemma 5.3 in [VdGBR$^+$14], we can show $\widehat{d}_j^2 = d_j^2+o_p(1)$. Then the results follows from Condition 7. □

Proof of Lemma 9. We will suppress F_0 in the argument of $\mathbf{S}_n$ for the proof, and start by first examining part of Δ. Denote $\mathbf{H}(\mathbf{b})=[\partial\mathbb{E}\mathbf{S}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}]_{\boldsymbol{\beta}=\mathbf{b}}$,

$$\begin{aligned}\mathbb{E}\mathbf{S}_n(\widehat{\boldsymbol{\beta}})-\mathbb{E}\mathbf{S}_n(\boldsymbol{\beta}^o) &= \mathbf{H}(\mathbf{b})\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\right)\\ &= \mathbf{H}(\widehat{\boldsymbol{\beta}})\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\right)+\left(\mathbf{H}(\mathbf{b})-\mathbf{H}(\widehat{\boldsymbol{\beta}})\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\right).\end{aligned}$$

Thus, we can rewrite Δ as

$$\Delta \;=\; \widehat{\boldsymbol{\Theta}}_j \mathbf{H}(\widehat{\boldsymbol{\beta}})\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)+\widehat{\boldsymbol{\Theta}}_j\big(\mathbf{H}(\mathbf{b})-\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big).$$

Subtracting (Δ) from (I), we have

$$\begin{aligned} I-\Delta &= \widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}^o_j-\widehat{\boldsymbol{\Theta}}_j\mathbf{H}(\widehat{\boldsymbol{\beta}})\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)-\widehat{\boldsymbol{\Theta}}_j\big(\mathbf{H}(\mathbf{b})-\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big) \\ &= \underbrace{\big(e_j^T-\widehat{\boldsymbol{\Theta}}_j\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)}_{\text{(i)}}+\underbrace{\widehat{\boldsymbol{\Theta}}_j\big(\mathbf{H}(\widehat{\boldsymbol{\beta}})-\mathbf{H}(\mathbf{b})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)}_{\text{(ii)}} \end{aligned}$$

Using the KKT condition described in (2.12), we could work out a bound for (i). In more detail,

$$\begin{aligned} \Big|\big(e_j^T-\widehat{\boldsymbol{\Theta}}_j\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)\Big| &\le ||\big(e_j^T-\widehat{\boldsymbol{\Theta}}_j\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)||_\infty||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \\ &\le \frac{\lambda_j}{\widehat{d}_j^2}||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \\ &= O_p(\lambda_j\lambda s_{\boldsymbol{\beta}^o}) \end{aligned}$$

where the last inequality is due to the consistency result of Theorem 13 and the fact that $1/\widehat{d}_j^2$ is bounded, which is shown in Lemma 12. Now for part (ii),

$$\begin{aligned} &\Big|\widehat{\boldsymbol{\Theta}}_j\big(\mathbf{H}(\mathbf{b})-\mathbf{H}(\widehat{\boldsymbol{\beta}})\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)\Big| \\ \le\;& \left|\frac{1}{n}\sum_{i=1}^n\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^T\cdot\mathbf{x}_i\big(\mathbb{I}(\mathbf{x}_i\mathbf{b}>0)f_0(\mathbf{x}_i\mathbf{b}|\mathbf{x}_i)-\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}>0)f_0(\mathbf{x}_i\widehat{\boldsymbol{\beta}}|\mathbf{x}_i)\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\big)\right| \\ \le\;& \left|\frac{L}{n}\sum_{i=1}^n\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^T\Big(\mathbf{x}_i(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)\Big)^2\right|+M\left|\frac{1}{n}\sum_{i=1}^n\widehat{\boldsymbol{\Theta}}_j\mathbf{x}_i^\top\mathbf{x}_i(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)\Big(\mathbb{I}(\mathbf{x}_i\mathbf{b}>0)-\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}>0)\Big)\right| \\ \le\;& L\|\mathbf{X}\widehat{\boldsymbol{\Theta}}_j^\top\|_\infty\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)\|_2/n+MK_X\|\mathbf{X}\widehat{\boldsymbol{\Theta}}_j^\top\|_\infty\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1\frac{1}{n}\sum_{i=1}^n\Big|\mathbb{I}(\mathbf{x}_i\mathbf{b}>0)-\mathbb{I}(\mathbf{x}_i\widehat{\boldsymbol{\beta}}>0)\Big| \\ =\;& O_p(K\lambda^2 s_{\boldsymbol{\beta}^o})+O_p(K\lambda s_{\boldsymbol{\beta}^o})\frac{1}{n}\sum_{i=1}^n\mathscr{B}_i. \end{aligned}$$

When $\|\widehat{\mathbf{\Theta}}_j - \mathbf{\Theta}_j^0\|_1 = o_p(1)$, the term $\|\mathbf{X}\widehat{\Theta}_j^\top\|_\infty$ is $O_p(K)$, where $K = \sqrt{s_j}$ in the bounded case, and $K = 1$ in the strongly bounded case. By similar argument in Lemma 7, $n^{-1}\sum_{i=1}^n \mathscr{B}_i = O_p(\lambda s_{\boldsymbol{\beta}^o})$.

Putting parts of (i) and (ii) together, we have

$$|I - \Delta| = O_p\left(\lambda_j \lambda s_{\boldsymbol{\beta}^o}\right) + O_p(K\lambda^2 s^2_{\boldsymbol{\beta}^o}).$$

□

Proof of Lemma 10. Suppressing the argument F_0 for simplicity of notation, define

$$\begin{aligned}\Xi(\boldsymbol{\beta}) &= \mathbf{\Theta}_j\left[\mathbf{S}_n(\boldsymbol{\beta}) - \mathbf{S}_n(\boldsymbol{\beta}^o)\right] - \mathbf{\Theta}_j\left[\mathbb{E}\mathbf{S}_n(\boldsymbol{\beta}) - \mathbb{E}\mathbf{S}_n(\boldsymbol{\beta}^o)\right] \\ &= \underbrace{\mathbf{\Theta}_j\left[\mathbf{S}_n(\boldsymbol{\beta}) - \mathbf{S}_n(\boldsymbol{\beta}^o)\right]}_{\bar{\xi}_n} - \mathbb{E}\mathbf{\Theta}_j\left[\mathbf{S}_n(\boldsymbol{\beta}) - \mathbf{S}_n(\boldsymbol{\beta}^o)\right],\end{aligned}$$

where the expectation is with respect to response variables T_i and $\mathbf{\Theta}$ is any p by p matrix with $||\mathbf{\Theta}_j|| = O(\sqrt{s_j})$ (s_j is still the j-th row cardinality of $\mathbf{\Theta}^o$. So in another word, $\mathbf{\Theta}$ is any matrix with the same row cardinality as $\mathbf{\Theta}^o$). Then the term (II) is just $\Xi(\widehat{\boldsymbol{\beta}})$ with $\mathbf{\Theta} = \widehat{\mathbf{\Theta}}$. Note that

$$\bar{\xi}_n = \frac{\sqrt{s_j}}{n}\sum_{i=1}^n \underbrace{s_j^{-1/2}\mathbf{\Theta}_j\mathbf{x}_i^T w_i\left[\mathbb{I}(Y_i \geq \mathbf{x}_i\boldsymbol{\beta}^o) - \mathbb{I}(Y_i \geq \mathbf{x}_i\boldsymbol{\beta})\right]}_{\widetilde{\xi}_i}.$$

Now for any i, without loss of generality, assume $\mathbf{x}_i\boldsymbol{\beta} > \mathbf{x}_i\boldsymbol{\beta}^o \geq 0$. Then $\xi_i = \widetilde{\xi}_i/\mathbf{\Theta}_j\mathbf{x}_i^T w_i$ is a Bernoulli random variable

$$\xi_i = \begin{cases} 1, & \text{if } \mathbf{x}_i\boldsymbol{\beta}^o \leq Y_i < \mathbf{x}_i\boldsymbol{\beta} \\ 0, & \text{elsewhere} \end{cases}$$

and $\mathbb{P}(\xi_i = 1) = F_0(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) - F_0(\mathbf{x}_i\boldsymbol{\beta}^o|\mathbf{x}_i) = f_0(\mathbf{x}_i\mathbf{b}|\mathbf{x}_i)\mathbf{x}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^o)$ for some $\mathbf{x}_i\boldsymbol{\beta}^o < \mathbf{x}_i\mathbf{b} < \mathbf{x}_i\boldsymbol{\beta}$. Therefore, $\mathrm{Var}(\xi_i) \leq \mathbb{P}(\xi_i = 1) = O_p(\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1)$ by Condition 2 and 5, and so is the variance of $\widetilde{\xi}_i$

because $||s_j^{-1/2}\mathbf{\Theta}_j(\boldsymbol{\beta})\mathbf{x}_i^T w_i||_\infty$ is bounded. Furthermore, it is easy to see that $\widetilde{\xi}_i$ is a stochastically bounded random variable, say $|\widetilde{\xi}_i| \leq a$ almost surely. Then $\text{Var}(\widetilde{\xi}_i)/a = O_p(\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_1)$ and this holds true for all $\boldsymbol{\beta}$. Invoking Bennett's inequality and the fact $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p(s_{\boldsymbol{\beta}^o}\lambda)$, we have $\Xi(\widehat{\boldsymbol{\beta}}) = O_p(\sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j/n})$, and hence Lemma 10. □

Proof of Lemma 11. We start by rewriting part of term (N), we note that

$$\begin{aligned}\mathbf{S}_n(\boldsymbol{\beta}^o, F_0) &= -\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top \left[w_i(F_0)\, \mathbb{1}\{Y_i - \mathbf{x}_i\boldsymbol{\beta}^o \geq 0\} - (1-\tau)\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top \psi_i\end{aligned}$$

where $\psi_i = -\left[w_i(F_0)\, \mathbb{1}\{Y_i - \mathbf{x}_i\boldsymbol{\beta}^o \geq 0\} - (1-\tau)\right]$. It is easy to show that, for each i,

$$\mathbb{E}[\psi_i|\mathbf{x}_i] = -\left(\tau - \mathbb{P}(Y_i < \mathbf{x}_i\boldsymbol{\beta}^o) - \tau(\mathbb{1}(\mathbf{x}_i\boldsymbol{\beta}^o \leq 0))^2\right) = 0.$$

Furthermore, $|\psi_i| \leq 1$. Then we can apply Lindeberg central limit theorem to random variable $\{\widehat{\mathbf{\Theta}}_j\mathbf{x}_i^\top \psi_i\}_{i=1}^n$. We have

$$\widehat{\mathbf{\Theta}}_j\mathbf{S}_n(\boldsymbol{\beta}^o, F_0) = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_j\mathbf{x}_i^\top \psi_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\psi^2}{n}\right),$$

where $\sigma_2^2 = \mathbb{E}\widehat{\mathbf{\Theta}}_j\left[n^{-1}\sum_{i=1}^n \mathbf{x}_i^\top\mathbf{x}_i\psi_i^2\right]\widehat{\mathbf{\Theta}}_j^\top = \mathbb{E}\widehat{\mathbf{\Theta}}_j\mathbf{\Omega}_\psi\widehat{\mathbf{\Theta}}_j^\top$ and $\mathbf{\Omega}_\psi := \sum_{i=1}^n \mathbf{x}_i^\top\mathbf{x}_i\psi_i^2/n$.

□

Proof of Lemma 13. Assume $f_{\boldsymbol{\beta}}(\mathbf{x}) = a > 0$. Let the distribution function of error at $\mathbf{x}$ be

$\nu_0(t|\mathbf{x}) = \mathbb{P}(\varepsilon \leq t|\mathbf{x})$.

$$
\begin{aligned}
\mathscr{P}\rho_f|\mathbf{x} &= \mathbb{E}[w\rho_\tau(y-a)+(1-w)\rho_\tau(y^{-\infty}-a)|\mathbf{x}] \\
&= \int_a^\infty \left[w(t)\rho_\tau(t-a)+(1-w(t))\rho_\tau(y^{-\infty}-a)\right] dF_0(t|\mathbf{x}) \\
&\quad + \int_0^a \left[w(t)\rho_\tau(t-a)+(1-w(t))\rho_\tau(y^{-\infty}-a)\right] dF_0(t|\mathbf{x}) \\
&\quad + \int_{-\infty}^0 \left[w(t)\rho_\tau(t-a)+(1-w(t))\rho_\tau(y^{-\infty}-a)\right] dF_0(t|\mathbf{x}) \\
&= \int_a^\infty \tau(t-a)dF_0(t|\mathbf{x}) + \int_0^a (\tau-1)(t-a)dF_0(t|\mathbf{x}) \\
&\quad + \int_{-\infty}^0 \left[(1-\frac{\tau}{F_0(0|\mathbf{x})})(\tau-1)(t-a)+\frac{\tau}{F_0(0|\mathbf{x})}(\tau-1)(y^{-\infty}-a)\right] dF_0(t|\mathbf{x}) \\
&= \tau\int_0^\infty t dF_0 - \int_0^a t dF_0 + (\tau F_0(0|\mathbf{x})-\tau-F_0(0|\mathbf{x}))\,a + aF_0(a|\mathbf{x}) \\
&\quad + \int_{-\infty}^0 \left[(1-\frac{\tau}{F_0(0|\mathbf{x})})(\tau-1)(t-a)+\frac{\tau}{F_0(0|\mathbf{x})}(\tau-1)(y^{-\infty}-a)\right] dF_0(t|\mathbf{x}).
\end{aligned}
$$

$$
\begin{aligned}
\mathscr{P}\rho_{f_0}|\mathbf{x} &= \mathbb{E}[w\rho_\tau(y-\mathbf{x}\boldsymbol{\beta}^o)+(1-w)\rho_\tau(y^{-\infty}-\mathbf{x}\boldsymbol{\beta}^o)|\mathbf{x}] \\
&= \tau\int_0^\infty t dF_0 - \int_0^{\mathbf{x}\boldsymbol{\beta}^o} t dF_0 + (\tau F_0(0|\mathbf{x})-\tau-F_0(0|\mathbf{x}))\,\mathbf{x}\boldsymbol{\beta}^o + \mathbf{x}\boldsymbol{\beta}^o F_0(\mathbf{x}\boldsymbol{\beta}^o|\mathbf{x}) \\
&\quad + \int_{-\infty}^0 \left[(1-\frac{\tau}{F_0(0|\mathbf{x})})(\tau-1)(t-\mathbf{x}\boldsymbol{\beta}^o)+\frac{\tau}{F_0(0|\mathbf{x})}(\tau-1)(y^{-\infty}-\mathbf{x}\boldsymbol{\beta}^o)\right] dF_0(t|\mathbf{x}).
\end{aligned}
$$

$$
\begin{aligned}
\mathscr{P}\rho_f|\mathbf{x} - \mathscr{P}\rho_{f_0}|\mathbf{x} &= -\int_{\mathbf{x}\boldsymbol{\beta}^o}^{a} t dF_0 + (\tau F_0(0|\mathbf{x}) - \tau - F_0(0|\mathbf{x}))\,(a - \mathbf{x}\boldsymbol{\beta}^o) + aF_0(a|\mathbf{x}) - \tau\mathbf{x}\boldsymbol{\beta}^o \\
&\quad + \int_{-\infty}^{0} \left[(1 - \frac{\tau}{F_0(0|\mathbf{x})})(\tau - 1)(\mathbf{x}\boldsymbol{\beta}^o - a) \right. \\
&\qquad \left. + \frac{\tau}{F_0(0|\mathbf{x})}(\tau - 1)(\mathbf{x}\boldsymbol{\beta}^o - a) \right] dF_0(t|\mathbf{x}) \\
&= -\int_{\mathbf{x}\boldsymbol{\beta}^o}^{a} t dF_0 + (\tau F_0(0|\mathbf{x}) - \tau - F_0(0|\mathbf{x}))\,(a - \mathbf{x}\boldsymbol{\beta}^o) + aF_0(a|\mathbf{x}) - \tau\mathbf{x}\boldsymbol{\beta}^o \\
&\quad + (\mathbf{x}\boldsymbol{\beta}^o - a)(\tau - 1) \int_{-\infty}^{0} \left[(1 - \frac{\tau}{F_0(0|\mathbf{x})}) + \frac{\tau}{F_0(0|\mathbf{x})} \right] dF_0(t|\mathbf{x}) \\
&= -\int_{\mathbf{x}\boldsymbol{\beta}^o}^{a} t dF_0 + (\tau F_0(0|\mathbf{x}) - \tau - F_0(0|\mathbf{x}))\,(a - \mathbf{x}\boldsymbol{\beta}^o) + aF_0(a|\mathbf{x}) - \tau\mathbf{x}\boldsymbol{\beta}^o \\
&\quad + (\mathbf{x}\boldsymbol{\beta}^o - a)(\tau - 1)F_0(0|\mathbf{x}) \\
&= -\int_{0}^{a-\mathbf{x}\boldsymbol{\beta}^o} (t + \mathbf{x}\boldsymbol{\beta}^o) d\nu_0(t|\mathbf{x}) \\
&\quad + \left(\tau\nu_0(-\mathbf{x}\boldsymbol{\beta}^o|\mathbf{x}) - \tau - \nu_0(-\mathbf{x}\boldsymbol{\beta}^o|\mathbf{x})\right)(a - \mathbf{x}\boldsymbol{\beta}^o) \\
&\quad + a\nu_0(a - \mathbf{x}\boldsymbol{\beta}^o|\mathbf{x}) - \tau\mathbf{x}\boldsymbol{\beta}^o + (\mathbf{x}\boldsymbol{\beta}^o - a)(\tau - 1)\nu_0(-\mathbf{x}\boldsymbol{\beta}^o|\mathbf{x}) \\
&= -\int_{0}^{a-\mathbf{x}\boldsymbol{\beta}^o} t d\nu_0(t|\mathbf{x}) + (a - \mathbf{x}\boldsymbol{\beta}^o)(\nu_0(a - \mathbf{x}\boldsymbol{\beta}^o|\mathbf{x}) - \tau). \qquad (2.30)
\end{aligned}
$$

Let $z := a - \mathbf{x}\boldsymbol{\beta}^o$, then:

$$
\begin{aligned}
(2.30) \quad &= -\int_0^z t d\nu_0(t|\mathbf{x}) + z(\nu_0(z|\mathbf{x}) - \tau) \\
&= -\int_0^z t d\nu_0(t|\mathbf{x}) + \int_0^z z d\nu_0(t|\mathbf{x}) \\
&= \int_0^z (z-t) d\nu_0(t|\mathbf{x}) \\
&= \int_0^z (z-t)\dot{\nu}_0(t|\mathbf{x}) dt \\
&= \int_0^z (z-t)\dot{\nu}_0(0|\mathbf{x}) dt + \int_0^z (z-t)(\dot{\nu}_0(t|\mathbf{x}) - \dot{\nu}_0(0|\mathbf{x})) dt \\
&\geq \int_0^z (z-t)\dot{\nu}_0(0|\mathbf{x}) dt - \int_0^{|z|} (|z|-t)|\dot{\nu}_0(t|\mathbf{x}) - \dot{\nu}_0(0|\mathbf{x})| dt \\
&\geq^{(i)} \int_0^z (z-t)\dot{\nu}_0(0|\mathbf{x}) dt - L\int_0^{|z|} (|z|-t) t dt \\
&= \frac{1}{2}\dot{\nu}(0|\mathbf{x}) z^2 - \frac{1}{6} L|z|^3. \qquad (2.31)
\end{aligned}
$$

In (i), we use the Lipschitz condition of the density function of error. Because of (2.31) and Condition 4, we can then use the Lemma in Stadler (2010) to conclude that there exists $C_1 > 0$ s.t. $\mathscr{E}(f_{\boldsymbol{\beta}}) \geq C_1^2 ||f_{\boldsymbol{\beta}} - f_0||^2$. □

Proof of Lemma 14.

$$
\begin{aligned}
|\gamma_{\boldsymbol{\beta}}(y,\mathbf{x})| &= |w\rho_\tau(y-\mathbf{x}\boldsymbol{\beta})+(1-w)\rho_\tau(y^{-\infty}-\mathbf{x}\boldsymbol{\beta})-w\rho_\tau(y-\mathbf{x}\boldsymbol{\beta}^o)-(1-w)\rho_\tau(y^{-\infty}-\mathbf{x}\boldsymbol{\beta}^o)| \\
&= |w\rho_\tau(y-\mathbf{x}\boldsymbol{\beta})-w\rho_\tau(y-\mathbf{x}\boldsymbol{\beta}^o)+(1-w)\rho_\tau(y^{-\infty}-\mathbf{x}\boldsymbol{\beta})-(1-w)\rho_\tau(y^{-\infty}-\mathbf{x}\boldsymbol{\beta}^o)| \\
&= |w\left(\rho_\tau(y-\mathbf{x}\boldsymbol{\beta})-\rho_\tau(y-\mathbf{x}\boldsymbol{\beta}^o)\right)+(1-w)(\tau-1)\mathbf{x}(\boldsymbol{\beta}^o-\boldsymbol{\beta})| \\
&\leq_{(i)} w|\max(\tau,1-\tau)\mathbf{x}(\boldsymbol{\beta}-\boldsymbol{\beta}^o)|+(1-w)|(\tau-1)\mathbf{x}(\boldsymbol{\beta}^o-\boldsymbol{\beta})| \\
&= \{w\max(\tau,1-\tau)+(1-w)(1-\tau)\}\,|\mathbf{x}(\boldsymbol{\beta}-\boldsymbol{\beta}^o)| \\
&\leq \max(\tau,1-\tau)|\mathbf{x}(\boldsymbol{\beta}-\boldsymbol{\beta}^o)| \\
&\leq \max(\tau,1-\tau)||\mathbf{x}||_\infty||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \\
&\leq_{(ii)} \max(\tau,1-\tau)K_X||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1.
\end{aligned}
$$

for all x, y, $\boldsymbol{\beta}$ in the range. The inequality (i) is from triangle inequality and property of loss function ρ_τ, and (ii) is because of Condition 5. Therefore, we have

$$
|\gamma_{\boldsymbol{\beta}}(y_i,\mathbf{x}_i)-\mathbb{E}\gamma_{\boldsymbol{\beta}}(y_i,\mathbf{x}_i)| \leq 2\max(\tau,1-\tau)||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1K_X.
$$

Denote $c_{i,\boldsymbol{\beta}} := 2\max(\tau,1-\tau)||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1K_X$, it is easy to show that

$$
\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\leq M}\sum_{i=1}^n c_{i,\boldsymbol{\beta}}^2 \leq \left(4\max(\tau,1-\tau)^2M^2K_X^2\right)n \leq 4M^2K_X^2n.
$$

By the concentration theorem (Massart, 2000), we have

$$
\mathbb{P}(Z_M \geq \mathbb{E}Z_M+t) \leq \exp\left(-\frac{nt^2}{32M^2K_X^2}\right).
$$

Therefore,

$$
\mathbb{P}\left(Z_M \geq \mathbb{E}Z_M+MK_X\sqrt{\frac{32t}{n}}\right) \leq e^{-t}.
$$

By the contraction inequality (Lemma 14.20 in Buhlmann and van de Geer (2011)), we have

$$\mathbb{E}Z_M \leq 4MK_X\sqrt{\frac{2\log(2p)}{n}}.$$

Consequently, for all $t > 0$ and $M > 0$,

$$\mathbb{P}\left(Z_M \geq 4MK_X\sqrt{\frac{2\log(2p)}{n}} + MK_X\sqrt{\frac{32t}{n}}\right) \leq e^{-t}.$$

Let

$$\lambda(t) = 4K_X\sqrt{\frac{2\log(2p)}{n}} + K_X\sqrt{\frac{32t}{n}}, \tag{2.32}$$

we have

$$\mathbb{P}(Z_M \geq M\lambda(t)) \leq e^{-t}.$$

□

2.7 Proofs of Theorems

Proof of Theroem 13.

Lemma 13. *Assuming Conditions 3 and 6, there exists some constant C_1 such that*

$$\mathscr{E}(f_{\boldsymbol{\beta}}) \geq C_1^2||f_{\boldsymbol{\beta}} - f_0||^2.$$

Lemma 14 (Concentration inequality)**.** *Define*

$$\gamma_{\boldsymbol{\beta}}(y,\mathbf{x}) \quad := \quad \rho_{f_{\boldsymbol{\beta}}}(y,\mathbf{x},w) - \rho_{f_{\boldsymbol{\beta}^o}}(y,\mathbf{x},w),$$

$$Z_M := \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \le M} \left| \frac{1}{n}\sum_{i=1}^{n} \gamma_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) - \mathbb{E}\gamma_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) \right|,$$

$$\lambda(t) := 4K_X\sqrt{\frac{2\log(2p)}{n}} + K_X\sqrt{\frac{32t}{n}}.$$

Then we have

$$\mathbb{P}(Z_M \ge M\lambda(t)) \le e^{-t}.$$

The following argument follows Muller and van der Geer (2014). We start with bounding the excess risk for $f_{\widehat{\boldsymbol{\beta}}}$,

$$\begin{aligned}
\mathscr{E}(f_{\widehat{\boldsymbol{\beta}}}) &= \mathscr{P}\rho_{f_{\widehat{\boldsymbol{\beta}}}} - \mathscr{P}\rho_{f_0} \\
&= -(\mathscr{P}_n - \mathscr{P})(\rho_{f_{\widehat{\boldsymbol{\beta}}}} - \rho_{f_0}) && (2.33)\\
&\quad + \widehat{\mathscr{P}_n}(\rho_{f_{\widehat{\boldsymbol{\beta}}}}) + \lambda||\widehat{\boldsymbol{\beta}}||_1 - \left(\widehat{\mathscr{P}_n}(\rho_{f_0}) + \lambda||\boldsymbol{\beta}^o||_1\right) && (2.34)\\
&\quad + \lambda||\boldsymbol{\beta}^o||_1 - \lambda||\widehat{\boldsymbol{\beta}}||_1 && (2.35)\\
&\quad + \mathscr{P}_n(\rho_{f_{\widehat{\boldsymbol{\beta}}}}) - \widehat{\mathscr{P}_n}(\rho_{f_{\widehat{\boldsymbol{\beta}}}}) + \mathscr{P}_n(\rho_{f_0}) - \widehat{\mathscr{P}_n}(\rho_{f_0}). && (2.36)
\end{aligned}$$

The plan is that, for equation (2.33), the empirical process part, we bound the term using concentration inequality. While equation (2.34) is negative by the definition of $\widehat{\boldsymbol{\beta}}$, equation (2.35) can be bounded using triangular inequality. Finally, for equation (2.36), it is negligible because $||w^0 - \widehat{w}||_\infty = o_p(1)$, which is shown in the proof of Lemma 8.

We then bound (2.33), (2.34), (2.35) separately. For (2.35), it is easy to show:

$$\lambda||\boldsymbol{\beta}^o||_1 - \lambda||\widehat{\boldsymbol{\beta}}||_1 \le \lambda \sum_{j\in S(\boldsymbol{\beta}^o)} |\widehat{\beta}_j - \beta_j^o| - \lambda \sum_{j\in S^c(\boldsymbol{\beta}^o)} |\widehat{\beta}_j|.$$

For (2.33), we have

$$-(\mathscr{P}_n - \mathscr{P})(\rho_{f_{\widehat{\boldsymbol{\beta}}}} - \rho_{f_0}) = -(\mathscr{P}_n - \mathscr{P})\gamma_{\widehat{\boldsymbol{\beta}}},$$

and

$$Z_M = \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \leq M} |(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|.$$

Now define

$$Z_M^{\delta} := \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \leq M} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \vee \delta}.$$

We have

$$\begin{aligned}
\mathbb{P}(Z_M^\delta > 2\lambda(t)) &= \mathbb{P}\left(\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \le M} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \vee \delta} > 2\lambda(t)\right) \\
&\le \sum_{j=\lfloor -\log_2 M\rfloor}^{\lceil -\log_2 \delta - 1\rceil} \mathbb{P}\left(\sup_{2^{-j-1}\le||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le 2^{-j}} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \vee \delta} > 2\lambda(t)\right) \\
&\quad + \mathbb{P}\left(\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le\delta} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \vee \delta} > 2\lambda(t)\right) \\
&\le \sum_{j=\lfloor -\log_2 M\rfloor}^{\lceil -\log_2 \delta - 1\rceil} \mathbb{P}\left(\sup_{2^{-j-1}\le||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le 2^{-j}} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{2^{-j-1}} > 2\lambda(t)\right) \\
&\quad + \mathbb{P}\left(\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le\delta} \frac{|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}|}{\delta} > 2\lambda(t)\right) \\
&= \sum_{j=\lfloor -\log_2 M\rfloor}^{\lceil -\log_2 \delta - 1\rceil} \mathbb{P}\left(\sup_{2^{-j-1}\le||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le 2^{-j}} |(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}| > 2^{-j}\lambda(t)\right) \\
&\quad + \mathbb{P}\left(\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1\le\delta} |(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}| > 2\delta\lambda(t)\right) \\
&\le \sum_{j=\lfloor -\log_2 M\rfloor}^{\lceil -\log_2 \delta - 1\rceil} \mathbb{P}\left(Z_{2^{-j}} > 2^{-j}\lambda(t)\right) + e^{-t} \\
&\le \sum_{j=\lfloor -\log_2 M\rfloor}^{\lceil -\log_2 \delta - 1\rceil} e^{-t} + e^{-t} \\
&= \left(\lceil -\log_2\delta - 1\rceil - \lfloor -\log_2 M\rfloor + 2\right)e^{-t} \\
&= \left(\lceil \log_2 M\rceil - \lfloor \log_2\delta + 1\rfloor + 2\right)e^{-t} \\
&\le \left(\lceil \log_2 M\rceil - \lceil \log_2\delta\rceil + 2\right)e^{-t} \\
&\le \log_2\left(\frac{8M}{\delta}\right)e^{-t}.
\end{aligned}$$

Therefore, for any $\boldsymbol{\beta}$ with $||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \le M$, we have

$$|(\mathscr{P}_n - \mathscr{P})\gamma_{\boldsymbol{\beta}}| \le 2\lambda(t)\left(||\boldsymbol{\beta}-\boldsymbol{\beta}^o||_1 \vee \delta\right)$$

with probability at least $1-\log_2\left(\frac{8M}{\delta}\right)e^{-t}$.

It is easy to show that $||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \ll n$. Then let $\delta = p^{-2}$, $t = 2\log(p)$ we have

$$|(\mathscr{P}_n-\mathscr{P})\gamma_{\widehat{\boldsymbol{\beta}}}| \le 2\lambda(t)\left(||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \vee p^{-2}\right)$$

with probability at least $1-\log_2\left(8np^2\right)/p^2$.

If $||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \le p^{-2}$, trivially we have consistency.

If $||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 > p^{-2}$, then because (2.34) is always non-positive by the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\begin{aligned}
\mathscr{E}(f_{\widehat{\boldsymbol{\beta}}}) &\le -(\mathscr{P}_n-\mathscr{P})(\rho_{f_{\widehat{\boldsymbol{\beta}}}}-\rho_{f_0})+\lambda||\boldsymbol{\beta}^o||_1-\lambda||\widehat{\boldsymbol{\beta}}||_1 \\
&\le 2\lambda(t)||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1+\lambda||\boldsymbol{\beta}^o||_1-\lambda||\widehat{\boldsymbol{\beta}}||_1 \\
&= 2\lambda(t)\left(\sum_{j\in S(\boldsymbol{\beta}^o)}|\widehat{\beta}_j-\beta_j^o|+\sum_{j\in S^c(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|\right) \\
&\quad +\lambda\left(\sum_{j\in S(\boldsymbol{\beta}^o)}|\beta_j^o|-\sum_{j\in S(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|-\sum_{j\in S^c(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|\right) \\
&\le 2\lambda(t)\left(\sum_{j\in S(\boldsymbol{\beta}^o)}|\widehat{\beta}_j-\beta_j^o|+\sum_{j\in S^c(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|\right)+\lambda\left(\sum_{j\in S(\boldsymbol{\beta}^o)}|\widehat{\beta}_j-\beta_j^o|-\sum_{j\in S^c(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|\right) \\
&= (2\lambda(t)+\lambda)\sum_{j\in S(\boldsymbol{\beta}^o)}|\widehat{\beta}_j-\beta_j^o|+(2\lambda(t)-\lambda)\sum_{j\in S^c(\boldsymbol{\beta}^o)}|\widehat{\beta}_j|. \qquad (2.37)
\end{aligned}$$

Since $\mathscr{E}(f_{\widehat{\boldsymbol{\beta}}}) \ge 0$ and $\lambda \ge 4\lambda(t)$, from (2.37), we know

$$||\widehat{\boldsymbol{\beta}}_{S_o^c}||_1 \le \frac{\lambda+2\lambda(t)}{\lambda-2\lambda(t)}||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1 \le 3||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1 \qquad (2.38)$$

which allows us to use the compatibility and censoring conditions. And again by (2.37) and

$\lambda \geq 4\lambda(t)$, we have

$$\mathscr{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq (2\lambda(t)+\lambda) \sum_{j \in S(\boldsymbol{\beta}^o)} |\widehat{\beta}_j - \beta_j^o|. \tag{2.39}$$

By Lemma 13, equation (2.39), the censoring condition and the compatibility condition, we have

$$\begin{aligned}
(2\lambda(t)+\lambda)||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1 &\geq C_1^2||f_{\widehat{\boldsymbol{\beta}}} - f_0||_2^2 \\
&= C_1^2(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)^T\mathbb{E}[\mathbf{x}^T\mathbf{x}](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o) \\
&\geq_{(i)} C_1^2 \frac{\phi_0^2}{s_{\boldsymbol{\beta}^o}}||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1^2
\end{aligned} \tag{2.40}$$

where (i) is from the compatibility condition.

By (2.40),

$$||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1 \leq \frac{s_{\boldsymbol{\beta}^o}(2\lambda(t)+\lambda)}{C_1^2\phi_0^2}. \tag{2.41}$$

Equation (2.38) implies that $||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 \leq 4||(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)_{S_o}||_1$, and hence by (2.41),

$$\begin{aligned}
||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1 &\leq \frac{4s_{\boldsymbol{\beta}^o}(2\lambda(t)+\lambda)}{C_1^2\phi_0^2} \\
&\leq \frac{6\lambda s_{\boldsymbol{\beta}^o}}{C_1^2\phi_0^2}.
\end{aligned} \tag{2.42}$$

With $C = 1/C_1^2$, we have Theorem 13. Furthermore, by (2.40), we have

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)^T\mathbb{E}[\mathbf{x}^T\mathbf{x}](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o) \leq \frac{3\lambda C}{2}||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o||_1. \tag{2.43}$$

□

Proof of Theorem 15. Following results from Lemmas 8 - 11, when $\left\|\widehat{\boldsymbol{\Theta}}_{\widehat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^o,j}\right\|_1 = o_p(1)$, the representation (2.17) can be simplified as

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^o\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \psi_i + \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top \phi_i\right) \\ &\quad + O_p\left(K\lambda_j\lambda s_{\boldsymbol{\beta}^o}\sqrt{n} + K\lambda^2 s_{\boldsymbol{\beta}^o}^2\sqrt{n} + \sqrt{\lambda s_{\boldsymbol{\beta}^o} s_j} + \frac{K}{\sqrt{n}} + K\frac{\log^{3/4} n}{\sqrt{n}}\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{\Theta}}_j \mathbf{x}_i^\top (\psi_i + \phi_i)\right) + o_p(1).\end{aligned}$$

The last line follows from assuming both λ and λ_j are of order $O(\sqrt{\log p/n})$, and $Ks_{\boldsymbol{\beta}^o}^2 \log p/n \vee s_{\boldsymbol{\beta}^o}^{1/2} s_j^{1/2} (\log p/n)^{1/4} = o(1)$. Then we have that

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^o\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_j^2\right),$$

where $\sigma_j^2 = \mathbb{E}\widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega} \widehat{\boldsymbol{\Theta}}_j^\top$ and $\boldsymbol{\Omega} := \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i (\psi_i + \phi_i)^2/n$.

The only missing part of the proof is the bound on the estimation error for $\widehat{\sigma}_j^2 := \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Theta}}^\top$ from $\mathbb{E}\widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega} \widehat{\boldsymbol{\Theta}}_j^\top$, where $\widehat{\boldsymbol{\Omega}} = \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \left(\widehat{\psi}_i + \widehat{\phi}_i\right)^2/n$. We start with rewriting the estimation error,

$$\begin{aligned}\left|\widehat{\sigma}_j^2 - \sigma_j^2\right| &= \underbrace{\left|\widehat{\boldsymbol{\Theta}}_j \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{\Theta}}_j^\top - \boldsymbol{\Theta}_j^0 \widehat{\boldsymbol{\Omega}} \boldsymbol{\Theta}_j^{0,\top}\right|}_{T_1} + \underbrace{\left|\boldsymbol{\Theta}_j^0 \widehat{\boldsymbol{\Omega}} \boldsymbol{\Theta}_j^{0,\top} - \boldsymbol{\Theta}_j^0 \boldsymbol{\Omega} \boldsymbol{\Theta}_j^{0,\top}\right|}_{T_2} \\ &\quad + \underbrace{\left|\boldsymbol{\Theta}_j^0 \boldsymbol{\Omega} \boldsymbol{\Theta}_j^{0,\top} - \boldsymbol{\Theta}_j^0 \mathbb{E}\boldsymbol{\Omega} \boldsymbol{\Theta}_j^{0,\top}\right|}_{T_3} + \underbrace{\left|\mathbb{E}\left(\boldsymbol{\Theta}_j^0 \boldsymbol{\Omega} \boldsymbol{\Theta}_j^{0,\top} - \widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega} \widehat{\boldsymbol{\Theta}}_j^\top\right)\right|}_{T_4}\end{aligned}$$

For the term T_1, we can further decompose it as

$$\begin{aligned}T_1 &\leq \left|(\boldsymbol{\Theta}_j^0 - \widehat{\boldsymbol{\Theta}}_j)\boldsymbol{\Omega}\boldsymbol{\Theta}_j^{0,\top}\right| + \left|\widehat{\boldsymbol{\Theta}}_j \boldsymbol{\Omega}(\boldsymbol{\Theta}_j^{0,\top} - \widehat{\boldsymbol{\Theta}}_j^\top)\right| \\ &\leq 2\left|\boldsymbol{\Theta}_j^0 \boldsymbol{\Omega}(\boldsymbol{\Theta}_j^0 - \widehat{\boldsymbol{\Theta}}_j)^\top\right| + \left|(\boldsymbol{\Theta}_j^0 - \widehat{\boldsymbol{\Theta}}_j)\boldsymbol{\Omega}(\boldsymbol{\Theta}_j^0 - \widehat{\boldsymbol{\Theta}}_j)^\top\right| \\ &\leq 2\|\boldsymbol{\Theta}_j^0 \boldsymbol{\Omega}\|_\infty \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^0\|_1 + \|\boldsymbol{\Omega}\|_\infty \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^0\|_1^2.\end{aligned}$$

Because $\left|\mathbf{\Theta}_j^0 \mathbf{x}_i^\top\right| = O(K)$ and $\|\mathbf{x}_i\|_\infty = O(1)$, we know $\|\mathbf{\Theta}_j^0 \mathbf{\Omega}\|_\infty = O(K)$ and $\|\mathbf{\Omega}\|_\infty = O(1)$. Therefore, $T_1 = o_p(1)$ if $K\|\widehat{\mathbf{\Theta}}_j - \mathbf{\Theta}_j^0\|_1 = o_p(1)$. We note that term T_4 can be bounded similarly.

For the term $T_2 + T_3$, denote $\widehat{\xi}_i = \widehat{\psi}_i + \widehat{\phi}_i$, then

$$T_2 + T_3 = \underbrace{\left| \mathbf{\Theta}_j^0 \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \left(\widehat{\xi}_i^2 - \xi_i^2 \right) \right) \mathbf{\Theta}_j^{0,\top} \right|}_{T_2} + \underbrace{\left| \mathbf{\Theta}_j^0 \left(\frac{1}{n} \sum_{i=1}^n \left(\mathbf{x}_i^\top \mathbf{x}_i \xi_i^2 - \mathbb{E} \mathbf{x}_i^\top \mathbf{x}_i \xi_i^2 \right) \right) \mathbf{\Theta}_j^{0,\top} \right|}_{T_3}.$$

For term T_3, since $\|\mathbf{X}\|_\infty = O(1)$ and $|\xi_i| \leq 1$, by Hoeffding's inequality, we have

$$\frac{1}{n} \sum_{i=1}^n \left(\mathbf{\Theta}_j^0 \mathbf{x}_i^\top \mathbf{x}_i \mathbf{\Theta}_j^{0,\top} \xi_i^2 - \mathbb{E} \mathbf{\Theta}_j^0 \mathbf{x}_i^\top \mathbf{x}_i \mathbf{\Theta}_j^{0,\top} \xi_i^2 \right) = O_p\left(\frac{K^2}{\sqrt{n}} \right).$$

Next, note that for T_2

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n \widehat{\xi}_i^2 - \xi_i^2 \right| &= \left| \frac{1}{n} \sum_{i=1}^n \left(\widehat{\xi}_i + \xi_i \right) (\widehat{\xi}_i - \xi_i) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \left(\widehat{\psi}_i + \widehat{\phi}_i + \psi_i + \phi_i \right) \left(\widehat{\psi}_i + \widehat{\phi}_i - \psi_i - \phi_i \right) \right| \\
&\leq 4 \left(\left| \frac{1}{n} \sum_{i=1}^n (\widehat{\psi}_i - \psi_i) \right| + \left| \frac{1}{n} \sum_{i=1}^n \left(\widehat{\phi}_i - \phi_i \right) \right| \right).
\end{aligned}$$

For the first difference, we have

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n (\widehat{\psi}_i - \psi_i) \right| &= \left| \frac{1}{n} \sum_{i=1}^n \left(w_i(\widehat{F}_n) \, \mathbb{1}(Y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}} \geq 0) - w_i(F_0) \, \mathbb{1}(Y_i - \mathbf{x}_i \boldsymbol{\beta}^o \geq 0) \right) \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \left(w_i(\widehat{F}_n) - w_i(F_0) \right) \right| + \left| \frac{1}{n} \sum_{i=1}^n \left(\mathbb{1}(Y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}} \geq 0) - \mathbb{1}(Y_i - \mathbf{x}_i \boldsymbol{\beta}^o \geq 0) \right) \right| \\
&= O_p(1/\sqrt{n}) + O_p(\lambda s_{\boldsymbol{\beta}^o}),
\end{aligned}$$

following results in Lemma 8. In addition,

$$\begin{aligned}
\left|\widehat{\phi}_i - \phi_i\right| &\leq \tau \left| \frac{1}{n} \sum_{\substack{l=1 \\ l\neq i}}^{n} \frac{\mathbb{I}\left(\widehat{F}_n > \tau\right)}{\widehat{F}_n^2} \left(1 - \frac{\mathbb{I}(Y_l = 0)}{\widehat{F}_n}\right) - \frac{1}{n} \sum_{\substack{l=1 \\ l\neq i}}^{n} \frac{\mathbb{I}\left(F_0 > \tau\right)}{F_0^2} \left(1 - \frac{\mathbb{I}(Y_l = 0)}{F_0}\right) \right| \\
&= \frac{\tau}{n} \left| \sum_{\substack{l=1 \\ l\neq i}}^{n} \left(\frac{\mathbb{I}\left(\widehat{F}_n > \tau\right)}{\widehat{F}_n^2} - \frac{\mathbb{I}\left(F_0 > \tau\right)}{F_0^2} \right) \right. \\
&\qquad \left. + \sum_{\substack{l=1 \\ l\neq i}}^{n} \left(\frac{\mathbb{I}\left(F_0 > \tau\right) \mathbb{I}(Y_l = 0)}{F_0^3} - \frac{\mathbb{I}\left(\widehat{F}_n > \tau\right) \mathbb{I}(Y_l = 0)}{\widehat{F}_n^3} \right) \right| \\
&= O_p\left(1/\sqrt{n}\right),
\end{aligned}$$

which then gives that $\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i (\widehat{\xi}_i^2 - \xi_i^2)\right]_{j,k} = O_p(1/\sqrt{n} + \lambda s_{\boldsymbol{\beta}^o})$. Then we conclude that $T_3 = O_p(K^2/\sqrt{n} + K^2 \lambda s_{\boldsymbol{\beta}^o})$.

Finally, when $K s_{\boldsymbol{\beta}^o}^2 \log p/n \vee s_{\boldsymbol{\beta}^o}^{1/2} s_j^{1/2} (\log p/n)^{1/4} \vee K \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^0\|_1 \vee K^2/\sqrt{n} \vee K^2 \lambda s_{\boldsymbol{\beta}^o} = o(1)$, we have that $\widehat{\sigma}_j = \sigma_j + o(1)$, which then completes the proof. □

2.8 Acknowledgement

Chapter 2, in full, is a version of the paper “High-dimensional covariate effects on left-censored quantile event times". The dissertation author is one of the principal investigators of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Chapter 3

Testing Generalized Hypotheses for High-dimensional Precision Matrix

3.1 Introduction

High-dimensional precision matrix arises in many areas of application, such as gene network discovery in [SS05, WZV$^+$04], brain connectivity analysis based on FMRI data in [NVPT13], climate studies in [RPK14], as well as financial data mining and social network analysis. As the precision matrix is often considered as a characterization of the network structure, entailing information regarding the interaction among subjects of the network, it serves as a proxy to a concise network depiction. Thus, it is often the case that an investigation of underlying graph of the network can be transformed into a problem on precision matrix.

It is known that the (i, j) entry in the precision matrix corresponds to the partial correlations between the variables i and j. In addition, the close connection between precision matrix and Gaussian graphical model results in an even stronger property. Under Gaussian graphical model setting, this further indicates conditional independence, see [Lau96]. In other words, if the data follows a multivariate normal distribution, the (i, j) entry of the precision matrix is

zero, if and only if variables i and j are conditionally independent given all other variables. In terms of the graph, this indicates that there is no edge between i and j. Thus, establishing the connection between a sparse graph and a sparse precision matrix. As often only few partial correlations among the large number of variables are significant, sparse precision matrix is a standard assumption high-dimensional setting.

We present here a testing framework for generalized hypotheses in high-dimensional precision matrix, based on the projection pursuit method.

3.1.1 Related Work

Although there is not much results in inference for high-dimensional precision matrix, its estimation problem has been extensively studied. In estimating the sparse precision matrix, one of the major approaches is neighborhood selection, which was introduced in [MB06]. The method estimates the zero entries of the precision matrix, by regressing each variable against the rest with standard Lasso, and thus also the name nodewise regression. [Yua10] uses the Dantzig selector to derive a precision matrix estimator under the nodewise regression framework, whereas [SZ13] proposed an estimator with scaled Lasso. Alternatively, CLIME and its adaptive version ACLIME, presented in [CLL11] and [CLZ$^+$16], offer to solve the problem using a related optimization framework, in place of regression setup.

Another major approach in estimating the precision matrix is through a penalized maximum likelihood estimator for the precision matrix. The method is named graphical Lasso, and is considered more of a global approach than nodewise regression. As opposed to the column-wise nature in the nodewise regression, this approach optimizes for an estimator to the precision matrix in its entirety. [YL07] solved the optimization problem as a max-det problem, and showed convergence result in low-dimensional case. [BGd08] accelerated the optimization process by making use of duality, and solved the problem using semi-definite programming. [FHT08] further improved the computation efficiency by connecting the optimization with Lasso, and hence the

name graphical Lasso. Variants of graphical Lasso have also been proposed. [RBL$^+$08] applied penalty limited to off-diagonal entries, and specified the convergence rates under Frobenius norm loss. [LF09] and [FFW09] explored the graphical Lasso with noncovex penalty functions, such as SCAD and adaptive Lasso. In addition, various pseudo-likelihood based objective functions have been proposed, for example, [FHT10, RZY08, KOR15, PWZZ09]. While these methods preserved symmetry property of the precision matrix, only the CONCORD estimator in [KOR15] is shown to guarantee convergence in optimization and asymptotic consistency.

Despite that precision matrix estimation has been extensively studied, not many works have pursued in high-dimensional precision matrix inference problems. [Liu13] developed a multiple testing procedure for conditional dependence in Gaussian graphical model, capable of asymptotically controlling the false discovery rate. In [WKR$^+$14], Berry-Essen type bounds on the coverage confidence intervals on edge weights are provided, along with bootstrap confidence intervals for certain high dimensional graphs. [RSZ$^+$15] extended the scaled Lasso estimation and nodewise regression. By regressing variables i and j against the remaining ones, a proxy for the covariance matrix of the residuals results in an estimator for the (i, j) entry in the precision matrix, and the inference result for such an estimator has been established. More recently, [JVDG$^+$15] and [JvdG17] have developed confidence intervals for entries in the precision matrix, by de-sparsifying the graphical Lasso and nodewise Lasso estimator respectively, with the help of de-biasing results in [VdGBR$^+$14].

Nevertheless, existing literature has only considered inference for each entry in the precision matrix. In fact, most existing work on inference in linear models, which is closely related to precision matrix estimation, focused on testing hypotheses that specify parameters to be given values [JM14a, VdGBR$^+$14, ZZ14, ZB16]. With an initial Lasso estimator, [VdGBR$^+$14] proposed a bias correction estimator, in order to obtain confidence intervals, while [JM14a] implemented a similar de-biasing procedure, but proposes a different scheme for estimating the inverse covariance matrix required in the bias correcting step. Until recently, [ZB17] and

[JL17] developed frameworks for testing general and complex hypotheses in high-dimensional models. In this paper, inspired by [ZB17], we propose a projection pursuit framework in testing generalized hypotheses for high-dimensional precision matrix.

3.1.2 Contributions

While the problem of testing general hypotheses remains wide open, there is a need in practice for testing general hypotheses. For example, a common assumption in de-biasing framework requires the precision matrix to be row sparse, see [VdGBR$^+$14]. However, no statistical testing procedure has been developed to check such an assumption. Another common assumption, particularly in time series data, is that the precision matrix is banded, which translates to decreasing values as the entries deviate from the diagonal. While bandedness testing in high-dimensional covariance matrix has been studied [CJ$^+$11, QC$^+$12], there has been no testing procedure devised for bandedness testing in high-dimensional precision matrix. Recently, [Bie16] presented graph-guided banding, a more generalized notion of bandedness. Testing the graph-bandedness of precision matrix can be appealing to researchers, who want to apply specific domain knowledge on underlying variable interactions. The following work provides a viable framework for testing a general hypothesis on high-dimensional precision matrix. We demonstrate the framework through three testing hypotheses. In addition, extensive simulation studies and real data analysis have been included.

3.1.3 Content

In Section 3.2, we introduce the projection pursuit approach for testing general hypotheses regarding the precision matrix. In details, we demonstrate the method with concrete testing hypotheses. As examples, we present hypotheses regarding row sparsity, minimum signal strength, bandedness and generalized bandedness. A comprehensive simulation study with numerical

experiment results can be found in Section 3.3. Finally, in Section 3.4, two real data application are demonstrated. As this is still a work in progress, we only present here the methodology and its empirical performance, along with preliminary Lemmas and their proofs in Section 3.5.

3.2 Methodology

Let the vector $\mathbf{Y}^k = (y_1^k, y_2^k, \cdots, y_p^k)^\top$, $k \in [n]$ be n i.i.d. observations, from a multivariate distribution with mean $\mathbf{0}$ and covariance matrix Σ. We also denote $\Omega = \Sigma^{-1} = \left((\omega_{ij})_{(i,j)\in[p]\times[p]}\right)$ as the inverse covariance matrix. We denote the i-th row of a matrix X as $X_{i.}$, and the j-th column as $X_{.j}$.

Often in the literature of high-dimensional statistics, row sparsity of Ω is assumed, namely,

$$\max_i \|\Omega_{i.}\|_0 = \max_i \sum_j \mathbb{I}\left(\Omega_{ij} \neq 0\right) \leq c.$$

Sometimes, we are also interested in testing for the minimum signal strength within the precision matrix, i.e.

$$\min_{(i,j)\in \mathrm{supp}(\Omega)} \left|\Omega_{ij}\right| \geq c,$$

where $\mathrm{supp}(\Omega) = \{(i,j) \in [p]\times[p] | \Omega_{ij} \neq 0\}$. In addition, there are other interesting matrix structures that one may be interested in testing, such as the bandedness of a matrix, or a bandedness that is much more general than the conventional diagonally banded ones.

However, there has been no testing framework for such an assumption. We provide here an extension to the projection pursuit framework in [ZB17] for testing row sparsity assumption of the precision matrix Ω.

3.2.1 Row Sparsity

Row sparsity has become a popular assumption for precision matrix in the literature, especially in de-biasing frameworks for correcting the bias in high dimensional estimates, see [VdGBR$^+$14] for examples. For this reason, we are interested for testing such an assumption. We formally state the hypothesis test for such an assumption. Let $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \max_i \|S_i.\|_0 \leq c\}$, then we are interested in testing

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0.$$

For the initial estimator $\widehat{\Omega}$, we use the CONCORD framework in [KOR15], which is

$$\begin{aligned}\widehat{\Omega} &= \underset{S}{\operatorname{argmin}} L(\mathbf{Y}, S) + P(S) \\ &= \underset{((\omega_{ij}))_{1\leq i,j\leq p}}{\operatorname{argmin}} -\sum_{i=1}^{p} n \log \omega_{ii} + \frac{1}{2}\sum_{i=1}^{p}\left\|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\right\|_2^2 + \lambda \sum_{1\leq i<j\leq p}|\omega_{ij}|. \end{aligned} \tag{3.1}$$

The null $H_0 : \Omega \in \mathscr{S}_0$ is equivalent to $d(\Omega, \mathscr{S}_0) = 0$, where we define the measure of deviation $d(\cdot)$ in this case as the Frobenius norm distance, i.e.

$$d(\Omega, \mathscr{S}_0) = \min_{S\in\mathscr{S}_0} \|\Omega - S\|_F ,$$

or alternatively, the minimization of the infinity norm,

$$d(\Omega, \mathscr{S}_0) = \min_{S\in\mathscr{S}_0} \|\Omega - S\|_\infty .$$

Intuitively, this is a measure on deviation of Ω from the null set $\mathscr{S}_0$. However, as Ω is unknown, we plug in the initial estimator for Ω. In order to obtain the deviation measure, we also need the closest element to $\widehat{\Omega}$ within the null set, which we denote with $\widetilde{\Omega}$. The following optimization

gives us an estimator for $\widetilde{\Omega}$,

$$\widetilde{\Omega} = \underset{S \in \mathscr{S}_0}{\operatorname{argmin}} \left\| \widehat{\Omega} - S \right\|_F \quad \text{or} \quad \widetilde{\Omega} = \underset{S \in \mathscr{S}_0}{\operatorname{argmin}} \left\| \widehat{\Omega} - S \right\|_\infty . \tag{3.2}$$

The optimization results in the solution $\widetilde{\Omega}$, such that $\widetilde{\Omega}_{ij} = \widehat{\Omega}_{ij} \mathbb{I}\left(\left| \widehat{\Omega}_{ij} \right| \geq \left| \widehat{\Omega}_{i\cdot} \right|_{(c)} \right)$, for all $(i, j) \in [p] \times [p]$, where $\left| \widehat{\Omega}_{i\cdot} \right|_{(c)}$ denotes the c-th largest entry of $\left| \widehat{\Omega}_{i\cdot} \right|$. Such a solution is justified through Lemma 15 in Section 3.5.

Finally, for the test statistic, a possible choice is

$$\max_{1 \leq i,j \leq p} \left| \widehat{\Omega}_{ij} - \widetilde{\Omega}_{ij} \right| .$$

However, for high dimensional parameter estimation, we need to correct for the bias introduced by the regularization during initial estimation, so that the test statistic is not driven by the difference in bias in the initial estimator. For the bias correction, we follow the sample splitting approach and de-biasing procedures as in [ZB17]. Assume for simplicity, we have an even number of n samples, which we then split evenly into subsample A and B, each of size $n/2$. Define the combined bias for $\widehat{\Omega}$ and $\widetilde{\Omega}$ with,

$$\widehat{\delta} = (n/2)^{-1} \sum_{k=1}^{n/2} \widehat{\Theta} M\left(\mathbf{Y}^k, \widehat{\Omega} \right) - (n/2)^{-1} \sum_{k=n/2+1}^{n} \widetilde{\Theta} M\left(\mathbf{Y}^k, \widetilde{\Omega} \right),$$

where $M(\mathbf{Y}, S) = \nabla_S L(\mathbf{Y}, S)$ and $\widehat{\Theta}, \widetilde{\Theta} \in \mathbb{R}^{p^2 \times p^2}$ are estimates for the population inverse Hessian matrix $\Theta := \left(\nabla^2_\Omega \mathbb{E} L(\mathbf{Y}, S) \right)^{-1}$ based on respective $\widehat{\Omega}$ and $\widetilde{\Omega}$. The proposed test statistic is then

$$T_n = \sqrt{n} \left\| \operatorname{vec}\left(\widehat{\Omega} - \widetilde{\Omega} \right) - \widehat{\delta} \right\|_\infty , \tag{3.3}$$

where $\operatorname{vec}(\cdot)$ denotes the operation of vectorizing a matrix by stacking the columns on top of one another.

In details, $M(\mathbf{Y}^k, S)$ and $\widehat{\Theta}$ are calculated. For the first derivative, we have

$$\frac{\partial L(\mathbf{Y}^k, S)}{\partial \omega_{il}} = \begin{cases} \left(\omega_{ii} y_i^k + \sum_{j \neq i} \omega_{ij} y_j^k\right) y_l^k, & l \neq i \\ -\frac{1}{\omega_{ii}} + \left(\omega_{ii} y_i^k + \sum_{j \neq i} \omega_{ij} y_j^k\right) y_i^k, & l = i \end{cases}.$$

For the second derivative, we have

$$\frac{\partial^2 \mathbb{E} L(\mathbf{Y}^k, S)}{\partial \omega_{il_1} \partial \omega_{il_2}} = \begin{cases} \mathbb{E} y_{l_1}^k y_{l_2}^k, & l_1 \neq i \text{ or } l_2 \neq i \\ \frac{1}{\omega_{ii}^2} + \mathbb{E}\left(y_i^k\right)^2, & l_1 = l_2 = i \end{cases}.$$

Thus, the population second partial has a block diagonal structure. Define

$$U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p] = \begin{bmatrix} \frac{1}{\omega_{11}} & & & \\ & \frac{1}{\omega_{22}} & & \\ & & \ddots & \\ & & & \frac{1}{\omega_{pp}} \end{bmatrix}.$$

Then the blocks within the population second partial are $\Sigma + \mathbf{u}_i \mathbf{u}_i^\top \in \mathbb{R}^{p \times p}$,

$$\nabla_\Omega^2 \mathbb{E} L(\mathbf{Y}, S) = \begin{bmatrix} \Sigma + \mathbf{u}_1 \mathbf{u}_1^\top & & & \\ & \Sigma + \mathbf{u}_2 \mathbf{u}_2^\top & & \\ & & \ddots & \\ & & & \Sigma + \mathbf{u}_p \mathbf{u}_p^\top \end{bmatrix}.$$

We take a look at the inverse of $\Sigma + \mathbf{u}_i \mathbf{u}_i^\top$ in details. By Sherman-Morrison formula,

$$\left(\Sigma + \mathbf{u}_i \mathbf{u}_i^\top\right)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1} \mathbf{u}_i \mathbf{u}_i^\top \Sigma^{-1}}{1 + \mathbf{u}_i^\top \Sigma^{-1} \mathbf{u}_i} = \Omega - \frac{\Omega \mathbf{u}_i \mathbf{u}_i^\top \Omega}{1 + \mathbf{u}_i^\top \Omega \mathbf{u}_i} = \Omega - \frac{\frac{1}{\omega_{ii}^2} \Omega_{\cdot i} \Omega_{i \cdot}}{1 + \frac{1}{\omega_{ii}}} = \Omega - \frac{\Omega_{\cdot i} \Omega_{i \cdot}}{\omega_{ii}^2 + \omega_{ii}}.$$

Thus, a good estimator $\widehat{\Theta}$ takes the form of

$$\widehat{\Theta} = \begin{bmatrix} \widehat{\Omega} - \frac{\widehat{\Omega}_{\cdot 1}\widehat{\Omega}_{1\cdot}}{\widehat{\omega}_{11}^2+\widehat{\omega}_{11}} & & & \\ & \widehat{\Omega} - \frac{\widehat{\Omega}_{\cdot 2}\widehat{\Omega}_{2\cdot}}{\widehat{\omega}_{22}^2+\widehat{\omega}_{22}} & & \\ & & \ddots & \\ & & & \widehat{\Omega} - \frac{\widehat{\Omega}_{\cdot p}\widehat{\Omega}_{p\cdot}}{\widehat{\omega}_{pp}^2+\widehat{\omega}_{pp}} \end{bmatrix}.$$

Finally, considering the difficulty to derive the actual distribution of the test statistic T_n, we apply multiplier bootstrap to derive the critical value for the test statistic. Notice that under the null hypothesis, T_n is set to approximate the quantity

$$\sqrt{n}\left\| - \left(\frac{2}{n}\sum_{k=1}^{n/2} R_k - \frac{2}{n}\sum_{k=n/2+1}^{n} R_k \right) \right\|_\infty, \text{ where } R_k = \begin{cases} \Theta M(\mathbf{Y}^k, \Omega), & 1 \le k \le n/2 \\ \Theta M(\mathbf{Y}^k, \Omega), & n/2+1 \le k \le n. \end{cases}$$

Then given a set of Gaussian multipliers $\{\xi_k\}_{k=1}^n$, where ξ_k follows a p^2-variate multivariate standard normal distribution $\mathcal{N}(0, I)$, for $1 \le k \le n$. The bootstrap statistic is defined as

$$T_n^* = \sqrt{n}\left\| - \left(\frac{2}{n}\sum_{k=1}^{n/2} \left(\widehat{R}_k - \bar{R}_A\right)\xi_k - \frac{2}{n}\sum_{k=n/2+1}^{n} \left(\widehat{R}_k - \bar{R}_B\right)\xi_k \right) \right\|_\infty, \tag{3.4}$$

where $\bar{R}_A = (n/2)^{-1}\sum_{k=1}^{n/2}\widehat{R}_k$ and $\bar{R}_B = (n/2)^{-1}\sum_{k=n/2+1}^{n}\widehat{R}_k$, and

$$\widehat{R}_k = \begin{cases} \widehat{\Theta} M\left(\mathbf{Y}^k, \widehat{\Omega}\right), & 1 \le k \le n/2 \\ \widetilde{\Theta} M\left(\mathbf{Y}^k, \widetilde{\Omega}\right), & n/2+1 \le k \le n \end{cases}.$$

The α-level critical value is taken to be $(1-\alpha)$ quantile of $\{T_n^*\}$, denoted as $T_{n,1-\alpha}^*$.

3.2.2 Minimum Signal Strength

Minimum signal assumption is common for parameter estimation in high-dimensional linear models. Here, we expand the idea, and test whether partial correlations among variables have a minimum signal strength. For the hypothesis testing on minimum signal strength in a precision matrix, we formally state the hypothesis. Let $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \min_{(i,j)\in \text{supp}(S)} \left|S_{ij}\right| \geq c\}$, then we are interested in testing

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0.$$

While we still use the same initial estimator as in (3.1) and the same projection optimization to acquire the closest estimator to the initial $\widehat{\Omega}$ under the null, the solution to $\widetilde{\Omega}$, however, is changed accordingly. Specifically, the optimization results in $\widehat{\Omega}$ such that $\widetilde{\Omega}_{ij} = \widehat{\Omega}_{ij}\, \mathbb{I}\left(\left|\widehat{\Omega}_{ij}\right| \geq c\right) + c\, \mathbb{I}\left(\left|\widehat{\Omega}_{ij}\right| \in (c/2, c)\right)$. The solution is justified with Lemma 16 in Section 3.5. The rest of the testing procedure follows as in the test for precision matrix row sparsity.

3.2.3 Bandedness

For sparse large matrices, often the bandedness assumption is imposed. The nonzero entries of a banded matrix are confined to a diagonal band with certain bandwidth. It is a well studied type of matrix structure for high-dimensional covariance matrix, see [BL08] and [CZZ$^+$10] for bandable covariance matrix. Defining such a banded structure is not only for theoretic convenience, but it also has intrinsic meanings attached. Often in financial time series and genomics data variables interact only with the ones in vicinity.

In this section, we present a testing framework for high-dimensional precision matrix bandedness. Such hypothesis can be set up as the following. Let $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | S_{ij} =$

0, for $|i-j|>c\}$, we are interested in testing

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0. \tag{3.5}$$

After obtaining the same initial estimator as in (3.1), following the Frobenius projection as in (3.2) gives us $\widetilde{\Omega}$, such that $\widetilde{\Omega}_{ij} = \widehat{\Omega}_{ij}\, \mathbb{I}(|i-j| \leq c)$. Such a projection is justified by Lemma 17 in Appendix. The testing procedure then follows as in the test for row sparsity.

3.2.4 Generalized Bandedness

In addition to the conventional definition of bandedness as $\mathscr{S}_0$ defined in (3.5), we also consider a more generalized banded structure. Recently, graph-guided banding presented in [Bie16] expanded the traditional diagonal band, and redefined bandedness under the context of graphs. In practice, this enables researchers to incorporate background information into the testing problem, and test the progress in development from the original graph. We begin with the definitions of generalized bandedness.

We denote a known graph $G = ([p], E)$, where $[p]$ denotes the p nodes and E denotes the edges. The B-th power of a graph G, denoted as G^B, connects nodes that are B hops of each other in the original graph G. In other words, using an adjacency matrix A to describe G^B, we have $A_{ij} \neq 0$ for $d_G(i,j) \leq B$, where $d_G(i,j)$ denotes the distance between node i and j. Thus, G is also referred as the seed graph. We formally define graph-guided bandedness with the following two definitions.

Definition 1. *A matrix Ω is b-banded with respect to a graph G, if* $\text{supp}(\Omega) = E\left(G^b\right)$*, that is* $\Omega_{ij} \neq 0 \iff d_G(i,j) \leq b$.

Definition 2. *A matrix Ω is $(b_1, \cdots, b_p)$-banded with respect to a graph G, if* $\Omega_{ij} \neq 0 \iff d_G(i,j) \leq \max\{b_i, b_j\}$.

We also denote $\mathscr{G}(B_1,\cdots,B_p)$ as the set of $(b_1,\cdots,b_p)$-banded matrices with respect to G, where $b_1 \leq B_1,\cdots,b_p \leq B_p$. The two definitions can be regarded as graphs with a global bandwidth and a local bandwidth respectively. The latter one is more general, as one with a global bandwidth can be seen as a special case to one with a local bandwidth. For our testing purposes, we are interested in testing the bandedness based on a seed graph, i.e. given a seed graph G,

$$H_0 : \Omega \in \mathscr{G}(B_1,\cdots,B_p) \text{ vs. } H_1 : \Omega \notin \mathscr{G}(B_1,\cdots,B_p). \tag{3.6}$$

Intuitively, the test utilizes the difference in sparsity pattern among graphs with different bandwidths. However, the change in sparsity patterns with B_j stops when $B_j > \text{diam}_j(G)$, where $\text{diam}_j(G)$ denotes the diameter of the j-th node in graph G. Thus, given a seed graph G, the generalized bandwidth test is only effective for testing hypothesis j-th node bandwidth less than or equal to the j-th node diameter in the graph.

With the initial estimation as described in (3.1), we derive the Frobenius projection as in (3.2), which gives us $\widetilde{\Omega}_{ij}$ such that $\widetilde{\Omega}_{ij} = \widehat{\Omega}_{ij}\, \mathbb{I}\left[d_G(i,j) \leq \max\{B_i,B_j\}\right]$. The testing procedure then follows as in previous sections.

3.3 Simulations

We evaluate the empirical performance of the projection pursuit high-dimensional precision matrix testings with an extensive simulation study, which include numerical experiments of all the precision matrix testing hypotheses mentioned in Section 3.2.

Three scenarios are considered, with dimensionality settings (n,p) as $(100,200)$, $(200,300)$ and $(300,400)$. Under each setting, we provide simulation results for precision matrix test for row sparsity, minimum signal strength, bandedness and generalized bandedness. While the significance levels for all tests are held at 0.05, we vary the alternative hypothesis and examine the power performance of the testing method.

The testing procedure is straightforward. Given generated dataset $\mathbf{Y}$ and the prespecified significance level $\alpha = 0.05$, our testing procedure is as following. We apply CONCORD algorithm as in (3.1) to obtain an initial precision matrix estimator $\widehat{\Omega}$. Then we derive the projection estimator $\widetilde{\Omega}$ under the null hypothesis as in (3.2). The test statistic is calculated as (3.3), with bootstrap iterations chosen to be 200. Finally, the α-level critical value results from the multiplier bootstrap as in (3.4). The testing conclusion follows accordingly.

3.3.1 Row Sparsity

The underlying true precision matrix is chosen to be modified Toeplitz matrix, i.e.

$$\Omega_{ij} = \frac{1}{2}\left(\rho^{|i-j|}\, \mathbb{I}(|i-j| < t) + \mathbb{I}(i=j)\right), \tag{3.7}$$

where the Toeplitz parameter $\rho = 0.9$, and $t = 4$. The underlying true precision matrix has row sparsity $s_0 = 2t - 1 = 7$. We test for the hypothesis that

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.8}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \max_i \|S_i.\|_0 \leq c\}$, for $c \in \{1,3,5,7,9,11\}$. In total, we generate data $\mathbf{Y}$ according to multivariate normal distribution $\mathscr{N}(\mathbf{0}, \Omega^{-1})$, and perform 100 iterations of the test. Within each iteration, the test statistic using multiplier bootstrap is carried out with 200 bootstrap iterations. The results under various dimensionality settings are summarized in Figure 3.1. As the plot indicates, for the test in (3.8) with $c = 7$, the true underlying sparsity, the proportion of rejecting the null hypothesis is close to the nominal level 0.05 for each of the three dimensionality settings. Once the null deviates from the true sparsity, the power of our row sparsity test gains power quickly.

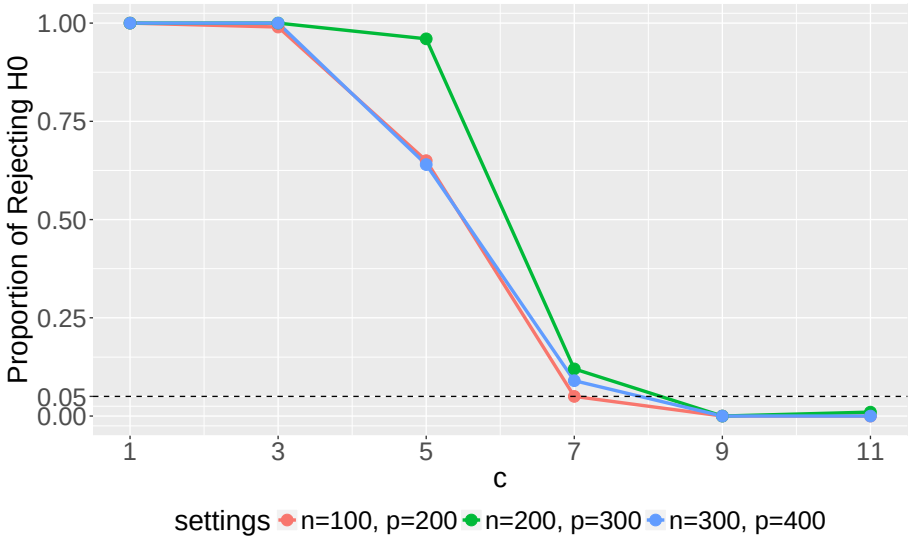

Figure 3.1: Power curves for precision matrix row sparsity test as in (3.8) under various dimensionality settings. The underlying true precision matrix takes form of (3.7), with $\rho = 0.9$ and $t = 4$. The tested sparsity level $c \in \{1,3,5,7,9,11\}$, and the true sparsity is 7.

3.3.2 Minimum Signal Strength

For the minimum signal strength, we use the modified Toeplitz matrix structure as in (3.7) for the underlying true precision matrix. The Toeplitz parameter $\rho = 0.9$. However, $t = 17$, in order to provide a sufficient range for analysis in statistical power. The underlying true precision matrix has minimum signal strength of $0.5 \times 0.9^{16} \approx 0.0927$. We test for the following hypothesis

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.9}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \min_{(i,j)\in \text{supp}(S)} |S_{ij}| \geq c\}$ for $c \in \{0.075 \times l + 0.5 \times 0.9^{16}\}$, where l is a integer such that $-1 \leq l \leq 4$. The results under various dimensionality settings are summarized in Figure 3.2. As the minimum signal of interest increases in the null hypothesis, we observe that the proportion of rejecting the null increases with the increase of testing minimum signal.

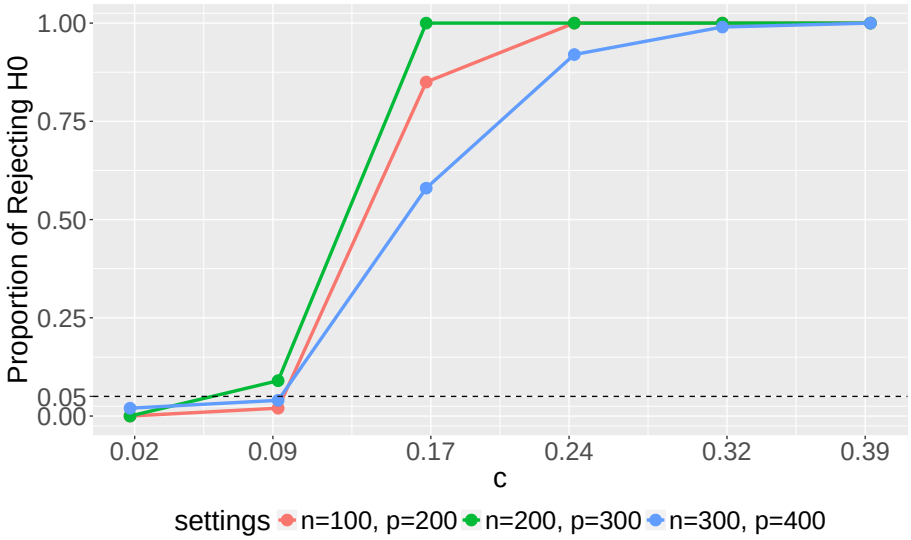

Figure 3.2: Power curves for precision matrix minimum signal test as in (3.9) under various dimensionality settings. The underlying true precision matrix takes form of (3.7), with $\rho = 0.9$ and $t = 17$. The tested sparsity level $c \in \{0.075 \times l + 0.5 \times 0.9^{16}, l \in \mathcal{N}, -1 \leq l \leq 4\}$, and the true minimum signal is $0.5 \times 0.9^{16} \approx 0.0927$.

3.3.3 Bandedness

We demonstrate two examples for the conventionally defined banded matrices. In addition, we compare our precision matrix test performance with the covariance matrix bandedness test of [QC+12]. In order for the covariance matrix test to be comparable to the precision matrix test, the underlying matrix structure of choice and the test hypothesis are set for the covariance matrix and the precision matrix respectively.

We begin with an example matrix structure used in [QC+12]. In details, we generate the precision matrix with a vector $\gamma = (\gamma_0, \gamma_1, \cdots, \gamma_t)$. The precision matrix Ω is then generated as the following,

$$\Omega_{ij} = \begin{cases} \sum_{k=0}^{t-|j-i|} \gamma_k \times \gamma_{|j-i|+k}, & \text{for } |j-i| \leq t \\ 0, & \text{otherwise} \end{cases}. \tag{3.10}$$

Specifically, we let $\gamma = (1, 0.4, 0.4, 0.4, 0.4, 0.4)$, resulting in an underlying precision matrix with

bandwidth 5. This matrix structure corresponds to one used in the test of $H_0 : \Sigma = B_5(\Sigma)$ with $\gamma_1 = \cdots = \gamma_5 = 0.4$ in [QC+12]. To make the two tests comparable, we specify (3.10) as the precision matrix for precision matrix test, and (3.10) as the covariance matrix for covariance matrix test. In other words, we take the inverse of (3.10) to generate data for the precision matrix test, in contrast to using (3.10) directly for data generation.

We apply testing procedure as described in [QC+12]. We perform hypothesis test T_1,

$$H_0 : \Sigma \in \mathscr{S}_0 \text{ vs. } H_1 : \Sigma \notin \mathscr{S}_0, \tag{3.11}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | S_{ij} = 0, \text{ for } |i-j| > c\}$, for $c \in \{0,1,2,3,4,5,6\}$. In comparison, we also apply our precision matrix bandedness test, using (3.10) as our underlying precision matrix Ω and apply our precision matrix bandedness test, which results in the test T_2,

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.12}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | S_{ij} = 0, \text{ for } |i-j| > c\}$, for $c \in \{0,1,2,3,4,5,6\}$. We summarize the power curves of the two tests in Figure 3.3. As n increases, our precision matrix test gains more and more statistical power in rejecting the null hypothesis when the alternative is true, and becomes comparable to the performance of covariance matrix testing.

The second example is under the modified Toeplitz matrix setting, with parameter $\rho = 0.9$ and $t = 4$. The underlying true precision matrix thus has bandwidth 3. We perform the hypothesis test T_1 for covariance matrix bandedness,

$$H_0 : \Sigma \in \mathscr{S}_0 \text{ vs. } H_1 : \Sigma \notin \mathscr{S}_0, \tag{3.13}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | S_{ij} = 0, \text{ for } |i-j| > c\}$, for $c \in \{0,1,2,3,4\}$. We also perform the

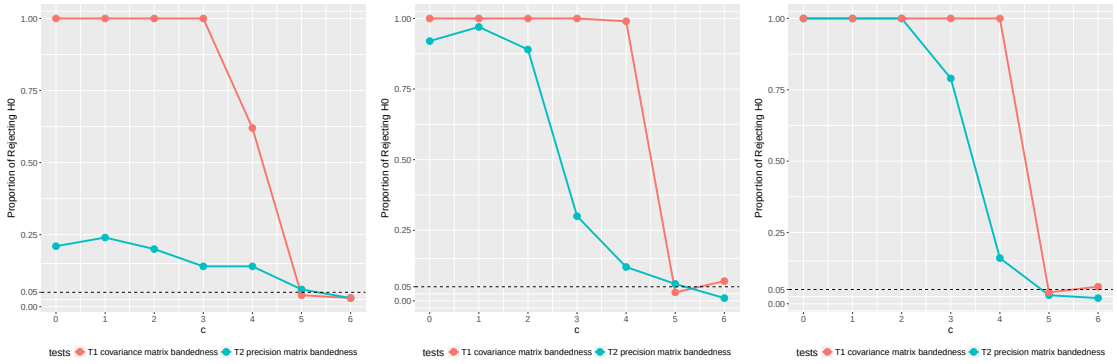

Figure 3.3: Power curves for covariance matrix and precision matrix bandedness tests as in (3.11) and (3.12) under various dimensionality settings. The underlying true covariance matrix for T_1 and precision matrix for T_2 take form of (3.10), with $\gamma_1 = \cdots = \gamma_5 = 0.4$ and $t = 5$. The tested bandwidth level $c \in \{0,1,2,3,4,5,6\}$, and the true bandwidth is 5. Left: $n = 100, p = 200$; Center: $n = 200, p = 300$; Right: $n = 300, p = 400$.

hypothesis test T_2 for precision matrix bandedness,

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.14}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | S_{ij} = 0, \text{ for } |i-j| > c\}$, for $c \in \{0,1,2,3,4\}$. The results are summarized in Figure 3.4. In contrast to covariance matrix testing T_1 as in (3.13), our testing procedure T_2 for precision matrix is more powerful in detecting small deviations from the null hypothesis.

3.3.4 Generalized Bandedness

For generalized bandedness, we generate a seed graph G by connecting two nodes with 0.0015 probability. To generate the underlying precision matrix, we let

$$A_{ij} = \begin{cases} \frac{1}{d_G(i,j)} \mathbb{1}\left[d_G(i,j) \leq \max\{B_i, B_j\}\right], & j \neq k \\ a, & j = k \end{cases}, \tag{3.15}$$

where a is chosen to ensure the minimum eigenvalue of A is at least σ^2 ($\sigma = 0.01$ throughout). Finally, we standardize the diagonal of A, and $\Omega = (\text{diag}(A))^{-1/2} A (\text{diag}(A))^{-1/2}$. For graph-

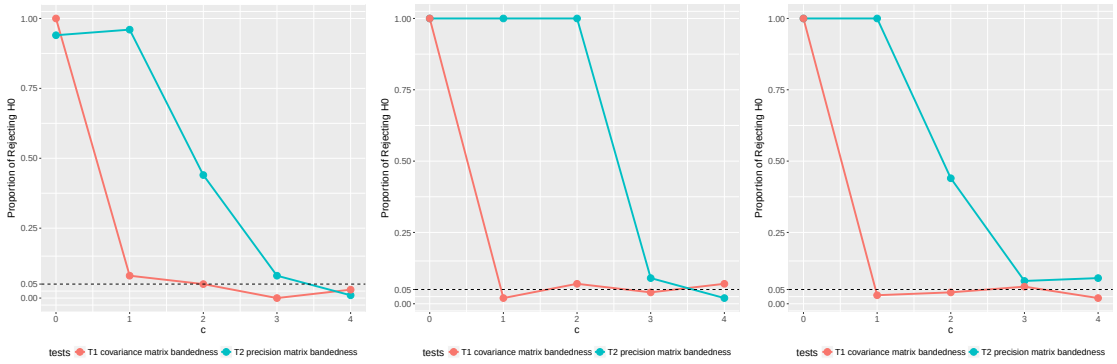

Figure 3.4: Power curves for covariance matrix and precision matrix bandedness tests as in (3.11) and (3.12) under various dimensionality settings. The underlying true covariance matrix for T_1 and precision matrix for T_2 take form of (3.7), with $\rho = 0.9$ and $t = 4$. The tested bandwidth level $c \in \{0,1,2,3,4\}$, and the true bandwidth is 3. Left: $n = 100, p = 200$; Center: $n = 200, p = 300$; Right: $n = 300, p = 400$.

guided banding with global bandwidths, we let the true bandwidth to be $b = 3$. A visualization of the seed graph G and its global bandwidth 3 graph under various dimensionality settings can be found in Figure 3.5.

To investigate the power performance, we vary the hypothesis,

$$H_0 : \Omega \in \mathscr{G}(B_1, \cdots, B_p) \text{ vs. } H_1 : \Omega \notin \mathscr{G}(B_1, \cdots, B_p), \tag{3.16}$$

where $B_j = c$, for all $j \in [p]$, and we vary $c \in \{1,2,3,4\}$. The result is summarized in Figure 3.6

For graph-guided banding with local bandwidths, we generate random local bandwidths according to the following distribution,

$$b_j = \begin{cases} 1, & \text{with probability } 0.1 \\ 2, & \text{with probability } 0.1 \cdot \\ 3, & \text{with probability } 0.8 \end{cases} \tag{3.17}$$

Figure 3.7 offers a visualization of the locally banded graphs, in contrast to their seed graphs, under various dimensionality settings.

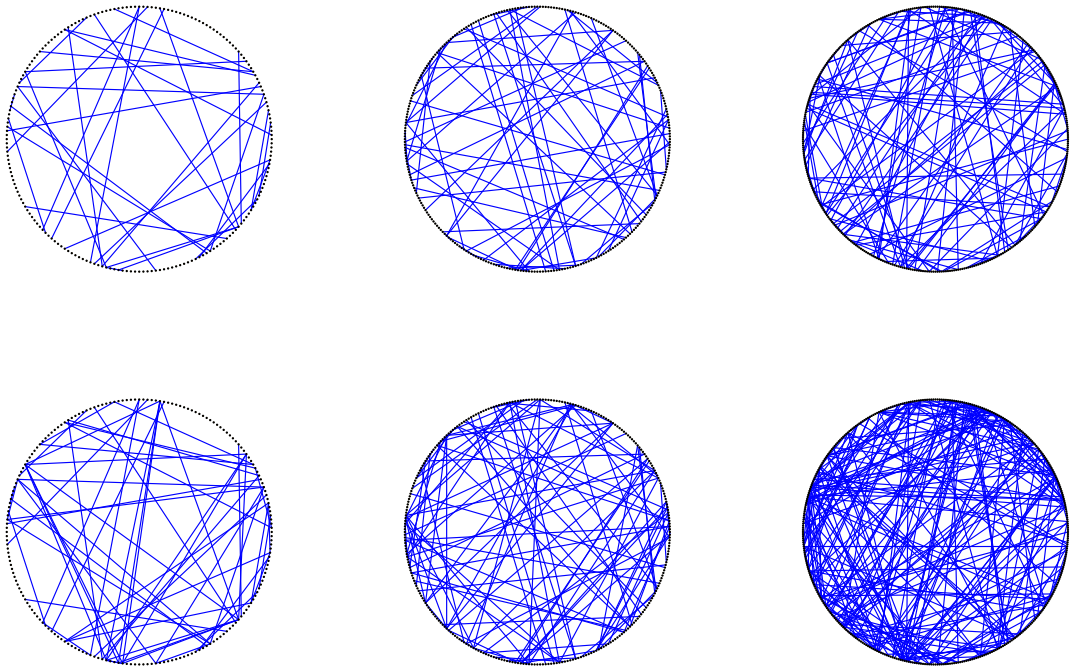

Figure 3.5: Visualizations of seed graphs and their globally banded graphs under various dimensionality settings. Top: seed graph under various dimensionality settings; Bottom: global bandwidth of 3 for respective seed graphs above. Left: $p = 200$; Center: $p = 300$; Right: $p = 400$.

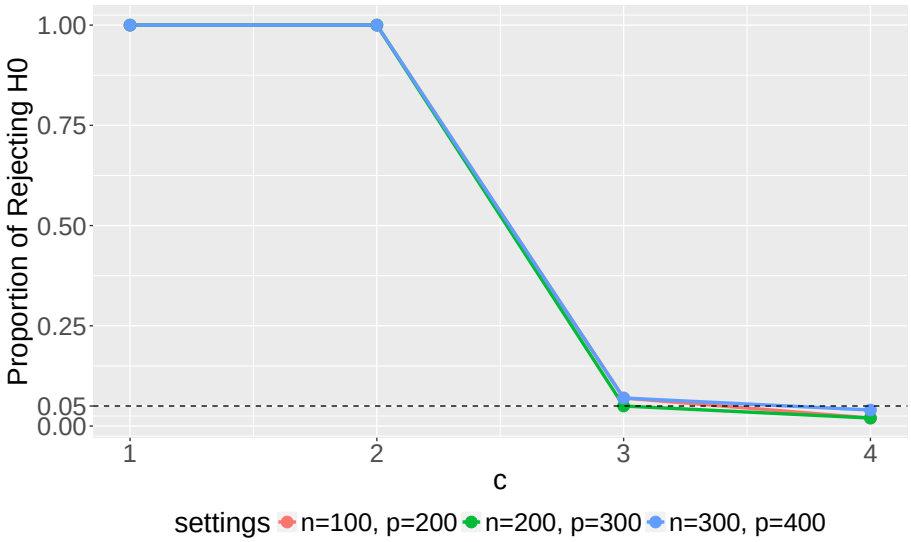

Figure 3.6: Power curves for precision matrix graph-guided globally bandedness test as in (3.16). The underlying true precision matrix of the seed graphs and their globally banded graphs are visually presented in Figure 3.5. The tested bandwidth $c \in \{1,2,3,4\}$, and the true global bandwidth is 3.

We investigate the power performance with the following hypothesis,

$$H_0 : \Omega \in \mathscr{G}(B_1, \cdots, B_p) \text{ vs. } H_1 : \Omega \notin \mathscr{G}(B_1, \cdots, B_p), \tag{3.18}$$

where $B_j = b_j - c \times \mathbb{I}(b_j = 3)$, and we vary $c \in \{2,1,0,-1\}$. The result is summarized in Figure 3.8.

3.4 Real Data

We also apply the methodology for two real datasets. Specifically, we compare the findings of row sparsity test with the literature, and demonstrate that an estimation of the underlying precision matrix row sparsity can be achieved through multiple tests by varying the alternative hypothesis.

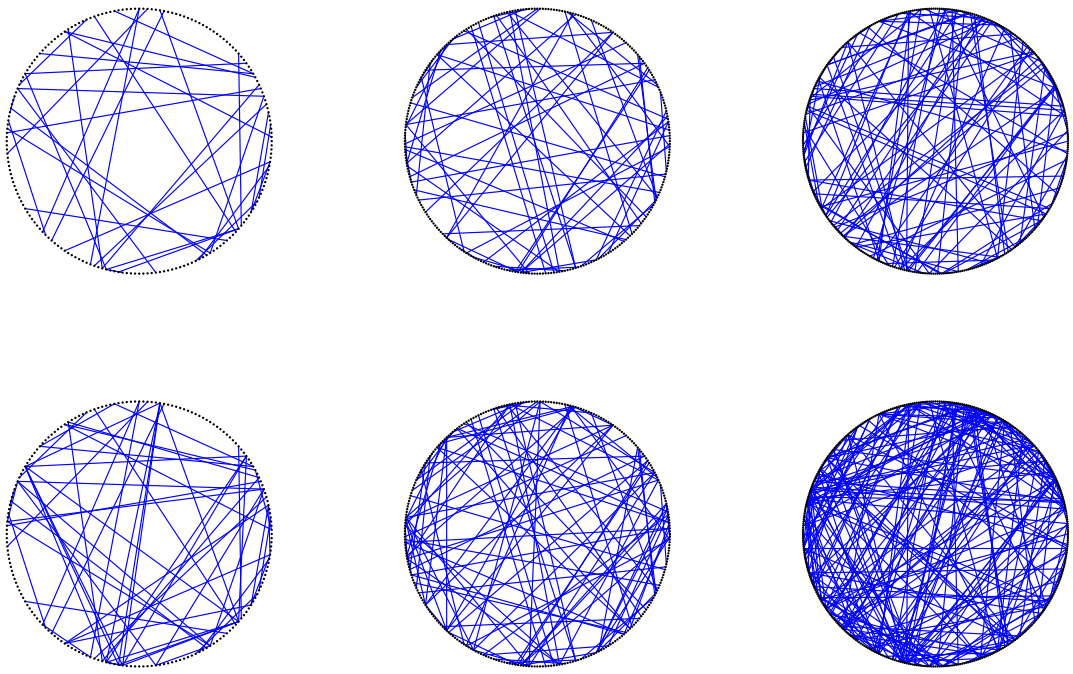

Figure 3.7: Visualizations of seed graphs and their locally banded graphs under various dimensionality settings. Top: seed graph under various dimensionality settings; Bottom: local bandwidths generated according to (3.17) for respective seed graphs above. Left: $p = 200$; Center: $p = 300$; Right: $p = 400$.

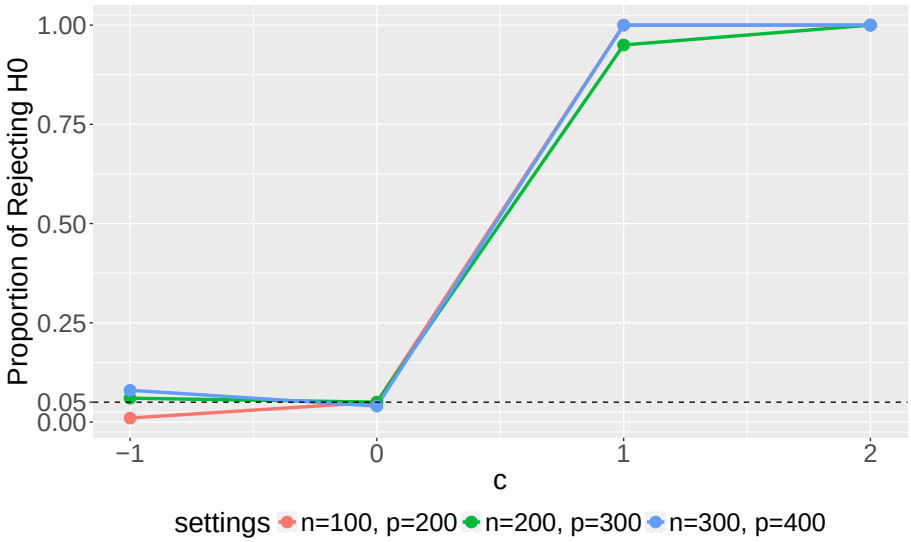

Figure 3.8: Power curves for precision matrix graph-guided locally bandedness test as in (3.18). The underlying true precision matrix of the seed graphs and their locally banded graphs are visually presented in Figure 3.7. The tested bandwidth perturbation parameter $c \in \{2,1,0,-1\}$, where the true bandwidth perturbation parameter is 0.

3.4.1 Riboflavin Data

The first dataset considered is riboflavin production by bacillus subtilis, and is readily available from the R package hdi. The dataset contains $n = 71$ observations of genetically engineered mutants of bacillus subtilis, while each observation is comprised of a record $p = 4088$ logarithms of gene expression levels. Instead of investigating the conditional independence structure among the covariates, we are interested in determining the sparsity of such a structure. Thus, we consider the top 500 covariates with the highest variances, and test the precision matrix Ω as the following,

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.19}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \max_i \|S_{i\cdot}\|_0 \leq c\}$ and $1 \leq c \leq 10$.

As we perform the row sparsity test by varying the parameter c, theoretically the observed p-values stay below the nominal level 0.05 for c less than the true underlying row sparsity.

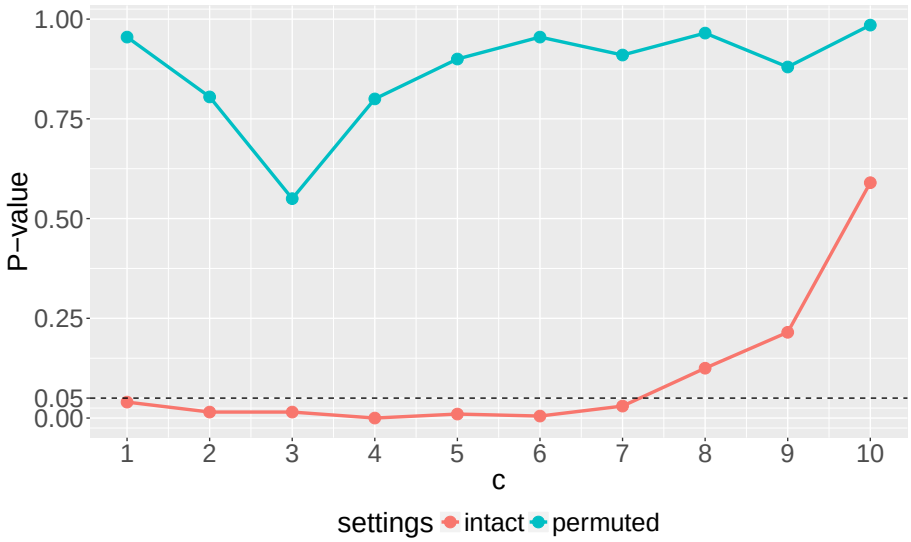

Figure 3.9: P-values for precision matrix row sparsity test with riboflavin dataset as in (3.19). The curves correspond to intact and permuted dataset respectively.

However, once the parameter c is larger than the true row sparsity, p-values of the tests start increasing. Thus, we take the parameter c, after which the p-value begins increasing above nominal level 0.05, as our estimate for the row sparsity. In order to achieve a stable p-value, we set the number of bootstrap iterations to 1000. We estimate that the row sparsity of the precision matrix of the riboflavin dataset to be 7.

In addition, we also independently permuted variables, in order to remove existing conditional dependencies among the variables. As we performed the same testing procedures, it is observed that the row sparsity of the precision matrix is 1, which indeed confirms the fact that variables are conditionally independent after the permutation. For the same dataset, [JVDG$^+$15] identifies 5 edges as significant. Our finding is inline with [JVDG$^+$15], if the 5 edges happen to be related through a gene expression acting as a hub. Otherwise, our results may indicate there are more significant edges. We summarize the test results in Figure 3.9.

3.4.2 Breast Cancer Data

Our second real dataset originates from breast cancer research. After [HAS+06] first analyzed the dataset, it was made available online at `http://bioinformatics.mdanderson.org`. Due to its accessibility, the dataset has been examined in [FFW09], [CLL11], [ZL13] and [WRG16] under the context of precision matrix estimation. The dataset consists of 133 subjects, each with an observation of $22,283$ gene expression levels. Among the 133 subjects, 34 of them have obtained pathological complete response (pCR), which is associated with excellent long-term cancer-free survival. On the contrary, the other 99 subjects have residual disease (RD). In previous works, the problem has been regarded as a classification task, and the performance depends on the estimation of the precision matrix. The common assumption for the problem is that the gene expression data follows a multivariate normal distribution with different mean, but the same covariance matrix, for the two groups.

As the methodology focuses on precision matrix testing, we decide to apply the methodology and estimate the row sparsity of the precision matrix. We select 110 gene expressions that are most statistically distinctive between the pCR and RD group of subjects, following procedures in [FFW09]. We test the precision matrix Ω as the following,

$$H_0 : \Omega \in \mathscr{S}_0 \text{ vs. } H_1 : \Omega \notin \mathscr{S}_0, \tag{3.20}$$

where $\mathscr{S}_0 = \{S \in \mathbb{R}^{p\times p} | \max_i \|S_i.\|_0 \leq c\}$ and $1 \leq c \leq 5$.

Our estimate for the row sparsity of the precision matrix for the 110 gene expression levels is at 2. We note that the gene networks constructed in [FFW09] exhibit a similar pattern, where the edges among genes are very scarce. In addition, we also independently permuted variables to remove existing conditional dependencies among the variables. Our method remains valid, as the p-values stay at a level close to 1 throughout the choices of the sparsity parameter c. We summarize the test results in Figure 3.10.

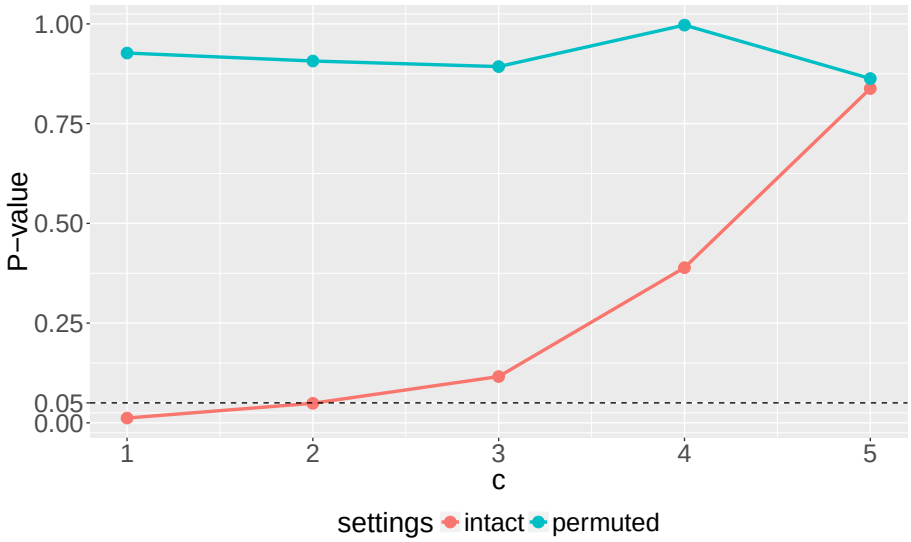

Figure 3.10: P-values for precision matrix row sparsity test with riboflavin dataset as in (3.19). The curves correspond to intact and permuted dataset respectively.

3.5 Proofs of Preliminary Lemmas

Lemma 15. *Let $X \in \mathbb{R}^{p\times p}$ and s_0 be a nonnegative integer. Suppose that $\pi_i : \{1,\cdots,p\} \to \{1,\cdots,p\}$ are permutations such that $\left|X_{i,\pi_i(1)}\right| \geq \left|X_{i,\pi_i(2)}\right| \geq \cdots \geq \left|X_{i,\pi_i(p)}\right|$, for $i = 1,\cdots,p$. Let $\widetilde{X} \in \mathbb{R}^{p\times p}$, where*

$$\widetilde{X}_{i,j} = \begin{cases} X_{i,j}, & j \in \{\pi_i(1),\cdots \pi_i(s_0)\} \\ 0, & \textit{otherwise} \end{cases} \tag{3.21}$$

for $i \in [p]$. Then $\widetilde{X}$ solves $\min_{S\in\mathbb{R}^{p\times p}} \|X-S\|_F$ s.t. $\max_i \|S_i.\|_0 \leq s_0$, and $\min_{S\in\mathbb{R}^{p\times p}} \|X-S\|_\infty$ s.t. $\max_i \|S_i.\|_0 \leq s_0$.

Proof of Lemma 15. We fix an arbitrary $S \in \mathbb{R}^{p\times p}$, with $\max_i \|S_i.\|_0 \leq s_0$. We have that

$$\|X-S\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\left|X_{ij}-S_{ij}\right|^2 \tag{3.22}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\left|X_{i,\pi_i(j)}-S_{i,\pi_i(j)}\right|^2 \tag{3.23}$$

$$= \sum_{i=1}^{p}\left(\sum_{j\in \mathrm{supp}(S_i.)}\left|X_{i,\pi_i(j)}-S_{i,\pi_i(j)}\right|^2 + \sum_{j\notin \mathrm{supp}(S_i.)}\left|X_{i,\pi_i(j)}\right|^2\right) \tag{3.24}$$

$$\geq \sum_{i=1}^{p}\left(\sum_{j\notin \mathrm{supp}(S_i.)}\left|X_{i,\pi_i(j)}\right|^2\right) \geq \sum_{i=1}^{p}\left(\min_{J\subseteq[p],|J|\geq p-s_0}\sum_{j\in J}\left|X_{i,\pi_i(j)}\right|^2\right) \tag{3.25}$$

$$= \sum_{i=1}^{p}\sum_{j=s_0+1}^{p}\left|X_{i,\pi_i(j)}\right|^2 = \left\|X-\widetilde{X}\right\|_F^2, \tag{3.26}$$

for an arbitrary S. In addition, since $\max_i \|\widetilde{X}_i.\|_0 \leq s_0$, $\widetilde{X}$ is the minimizer. The proof for $\widetilde{X}$ as the minimizer for the optimization with infinity norm follows similarly. □

Lemma 16. *Let $X \in \mathbb{R}^{p\times p}$ and r_0 be a nonnegative integer. In addition, let $\widetilde{X} \in \mathbb{R}^{p\times p}$, where*

$$\widetilde{X}_{ij} = X_{ij}\,\mathbb{I}\left(\left|X_{ij}\right| \geq r_0\right) + r_0\,\mathbb{I}\left(\left|X_{ij}\right| \in (r_0/2, r_0)\right) \tag{3.27}$$

for $1 \leq i,j \leq p$. Then $\widetilde{X}$ solves

$$\min_{S\in\mathbb{R}^{p\times p}} \|X-S\|_F \;\; s.t. \;\; \min_{(i,j)\in\mathrm{supp}(S)}\left|S_{ij}\right| \geq r_0,$$

and

$$\min_{S\in\mathbb{R}^{p\times p}} \|X-S\|_\infty \;\; s.t. \;\; \min_{(i,j)\in\mathrm{supp}(S)}\left|S_{ij}\right| \geq r_0.$$

Proof of Lemma 16. We fix an arbitrary $S \in \mathbb{R}^{p\times p}$, with $\min_{(i,j)\in\mathrm{supp}(S)}\left|S_{ij}\right| \geq r_0$. Thus, $\left|S_{ij}\right| \in$

$\{0\} \cup [r_0, \infty)$, for $1 \leq i, j \leq p$. We can see that

$$\left|X_{ij} - S_{ij}\right|^2 \geq \min_{|t| \in \{0\} \cup [r_0, \infty)} \left|X_{ij} - t\right|^2 = \left|X_{ij} - \widetilde{X}_{ij}\right|^2. \tag{3.28}$$

Rewriting the Frobenius norm and summing up all the entries,

$$\begin{aligned} \|X - S\|_F^2 &= \sum_{i=1}^{p} \sum_{j=1}^{p} \left|X_{ij} - S_{ij}\right|^2 && (3.29) \\ &\geq \sum_{i=1}^{p} \sum_{j=1}^{p} \min_{|t_{ij}| \in \{0\} \cup [r_0, \infty)} \left|X_{ij} - t_{ij}\right|^2 && (3.30) \\ &\geq \sum_{i=1}^{p} \sum_{j=1}^{p} \left|X_{ij} - \widetilde{X}_{ij}\right|^2 = \left\|X - \widetilde{X}\right\|_F^2, && (3.31) \end{aligned}$$

for an arbitrary S. In addition, since $\min_{(i,j) \in \text{supp}(S)} \left|\widetilde{X}_{ij}\right| \geq r_0$, $\widetilde{X}$ is the minimizer. The proof for $\widetilde{X}$ as the minimizer for the optimization with infinity norm follows similarly. □

Lemma 17. *Let $X \in \mathbb{R}^{p \times p}$ and t_0 be a nonnegative integer. In addition, let $\widetilde{X} \in \mathbb{R}^{p \times p}$, where*

$$\widetilde{X}_{ij} = X_{ij} \, \mathbb{1}(|i - j| \leq t_0) \tag{3.32}$$

for $1 \leq i, j \leq p$. Then $\widetilde{X}$ solves

$$\min_{S \in \mathbb{R}^{p \times p}} \|X - S\|_F \ \text{ s.t. } S_{ij} = 0, \text{ for } |i - j| > t_0,$$

and

$$\min_{S \in \mathbb{R}^{p \times p}} \|X - S\|_\infty \ \text{ s.t. } S_{ij} = 0, \text{ for } |i - j| > t_0.$$

Proof of Lemma 17. We fix an arbitrary $S \in \mathbb{R}^{p \times p}$, with $S_{ij} = 0$ for $|i - j| > t_0$. Thus, we have

that

$$\|X-S\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\left|X_{ij}-S_{ij}\right|^2 \tag{3.33}$$

$$= \sum_{|i-j|\le t_0}\left|X_{ij}-S_{ij}\right|^2 + \sum_{|i-j|>t_0}\left|X_{ij}-S_{ij}\right|^2 \tag{3.34}$$

$$= \sum_{|i-j|\le t_0}\left|X_{ij}-S_{ij}\right|^2 + \sum_{|i-j|>t_0}\left|X_{ij}\right|^2 \tag{3.35}$$

$$\ge \sum_{|i-j|>t_0}\left|X_{ij}\right|^2 = \left\|X-\widetilde{X}\right\|_F^2, \tag{3.36}$$

for an arbitrary S. In addition, since $\widetilde{X}_{ij}=0$ for $|i-j|>t_0$, $\widetilde{X}$ is the minimizer. The proof for $\widetilde{X}$ as the minimizer for the optimization with infinity norm follows similarly. □

3.6 Acknowledgement

Chapter 3, in full, is a version of the paper “Testing generalized hypotheses for high-dimensional precision matrix". The dissertation author is the principal investigator of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Bibliography

[Ame73] Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016, 1973.

[BCCW17] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *to appear in the Annals of Statistics*, 2017.

[BCK13] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Robust inference in high-dimensional approximately sparse quantile regression models. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2013.

[BCK14] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, page asu056, 2014.

[BFW11] Jelena Bradic, Jianqing Fan, and Weiwei Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

[BG16] Jelena Bradic and Jiaqi Guo. Robust confidence intervals in high-dimensional left-censored regression. *arXiv preprint arXiv:1609.07165*, 2016.

[BGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[BH98] Moshe Buchinsky and Jinyong Hahn. An alternative estimator for the censored quantile regression model. *Econometrica*, pages 653–671, 1998.

[Bic75] Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

[Bie16] Jacob Bien. Graph-guided banding of the covariance matrix. *arXiv preprint arXiv:1606.00451*, 2016.

[BL08] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[CH93] Clint W Coakley and Thomas P Hettmansperger. A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, 88(423):872–880, 1993.

[Chi92] Siddhartha Chib. Bayes inference in the tobit censored regression model. *Journal of Econometrics*, 51(1-2):79–99, 1992.

[CJ+11] T Tony Cai, Tiefeng Jiang, et al. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525, 2011.

[CLL11] Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[CLZ+16] T Tony Cai, Weidong Liu, Harrison H Zhou, et al. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

[CZZ+10] T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[DKES06] Sorin Draghici, Purvesh Khatri, Aron C Eklund, and Zoltan Szallasi. Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics*, 22(2):101–109, 2006.

[Efr67] Bradley Efron. The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853, 1967.

[Efr07] Sam Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, pages 2504–2535, 2007.

[FFW09] Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521, 2009.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical report, Stanford University, 2010.

[Fit97] Bernd Fitzenberger. Computational aspects of censored quantile regression. *Lecture Notes-Monograph Series*, pages 171–186, 1997.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[FW07] Bernd Fitzenberger and Peter Winker. Improving the computation of censored quantile regressions. *Computational Statistics & Data Analysis*, 52(1):88–108, 2007.

[GJP97] Amos Golan, George Judge, and Jeffrey Perloff. Estimation and inference with censored and ordered multinomial response data. *Journal of Econometrics*, 79(1):23–51, 1997.

[GMCS94] W Gonzalez-Manteiga and C Cadarso-Suarez. Asymptotic properties of a generalized kaplan-meier estimator with some applications. *Communications in Statistics-Theory and Methods*, 4(1):65–78, 1994.

[HAS+06] Kenneth R Hess, Keith Anderson, W Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Jaime A Mejia, Daniel Booser, Richard L Theriault, Aman U Buzdar, Peter J Dempsey, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology*, 24(26):4236–4244, 2006.

[Hil77] Richard Walter Hill. *Robust regression when there are outliers in the carriers*. PhD thesis, Harvard University, 1977.

[HY05] Peter Hall and Qiwei Yao. Approximating conditional distribution functions using dimension reduction. *Annals of statistics*, pages 1404–1421, 2005.

[IL15] Rafael Izbicki and Ann B Lee. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, (just-accepted), 2015.

[JL17] Adel Javanmard and Jason D Lee. A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971*, 2017.

[JM14a] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[JM14b] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[JVDG+15] Jana Jankova, Sara Van De Geer, et al. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[JvdG17] Jana Janková and Sara van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

[KBJ78] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[KBJ82] Roger Koenker and Gilbert Bassett Jr. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 43–61, 1982.

[KG01] Roger Koenker and Olga Geling. Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468, 2001.

[KOR15] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

[KP96] Roger Koenker and Beum J Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283, 1996.

[KP16] Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimension? *arXiv preprint arXiv:1608.00696*, 2016.

[Lau96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[LF09] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[Liu13] Weidong Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, pages 2948–2978, 2013.

[LS86] Shaw-Hwa Lo and Kesar Singh. The product-limit estimator and the bootstrap: some asymptotic representations. *Probability Theory and Related Fields*, 71(3):455–465, 1986.

[LZ08] Youjuan Li and Ji Zhu. L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[MvdG16] Patric Müller and Sara van de Geer. Censored linear model in high dimensions. *Test*, 25(1):75–92, 2016.

[MY09] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

[NL17] Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, 45(1):158–195, 02 2017.

[NNLL15] Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv preprint arXiv:1510.08986*, 2015.

[NP90] Whitney K Newey and James L Powell. Efficient estimation of linear and type i censored regression models under conditional quantile restrictions. *Econometric Theory*, 6(03):295–317, 1990.

[NVPT13] Bernard Ng, Gaël Varoquaux, Jean Baptiste Poline, and Bertrand Thirion. A novel sparse group gaussian graphical model for functional connectivity estimation. In *International Conference on Information Processing in Medical Imaging*, pages 256–267. Springer, 2013.

[NYWR09] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[Obe82] Walter Oberhofer. The consistency of nonlinear regression minimizing the l1-norm. *The Annals of Statistics*, pages 316–319, 1982.

[Por03] Stephen Portnoy. Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012, 2003.

[Pow84] James L Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984.

[Pow86a] James L Powell. Censored regression quantiles. *Journal of econometrics*, 32(1):143–155, 1986.

[Pow86b] James L Powell. Symmetrically trimmed least squares estimation for tobit models. *Econometrica: journal of the Econometric Society*, pages 1435–1460, 1986.

[PSA+10] Maya C Poffenberger, Iryna Shanina, Connie Aw, Nahida El Wharry, Nadine Straka, Dianne Fang, Annie E Baskin-Hill, Sabrina H Spiezio, Joseph H Nadeau, and Marc S Horwitz. Novel nonmajor histocompatibility complex–linked loci from mouse chromosome 17 confer susceptibility to viral-mediated chronic autoimmune myocarditisclinical perspective. *Circulation: Cardiovascular Genetics*, 3(5):399–408, 2010.

[PWZZ09] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[QC+12] Yumou Qiu, Song Xi Chen, et al. Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *The Annals of Statistics*, 40(3):1285–1314, 2012.

[RBL+08] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[RDK+00] Charles H Redfern, Michael Y Degtyarev, Andrew T Kwa, Nathan Salomonis, Nathalie Cotte, Tania Nanevicz, Nick Fidelman, Kavin Desai, Karen Vranizan, Elena K Lee, et al. Conditional expression of a gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proceedings of the National Academy of Sciences*, 97(9):4826–4831, 2000.

[RPK14] Jakob Runge, Vladimir Petoukhov, and Jürgen Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27(2):720–739, 2014.

[RSZ+15] Zhao Ren, Tingni Sun, Cun-Hui Zhang, Harrison H Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[RWG+16] Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

[RZY08] Guilherme V Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*, 2008.

[SCG+14] Luke C Swenson, Bryan Cobb, Anna Maria Geretti, P Richard Harrigan, Mario Poljak, Carole Seguin-Devaux, Chris Verhofstede, Marc Wirden, Alessandra Amendola, Jurg Boni, et al. Comparative performances of hiv-1 rna load assays at low viral load levels: results of an international collaboration. *Journal of clinical microbiology*, 52(2):517–523, 2014.

[SDC03] Mark R Segal, Kam D Dahlquist, and Bruce R Conklin. Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980, 2003.

[Son11] Weixing Song. Distribution-free test in tobit mean regression model. *Journal of Statistical Planning and Inference*, 141(8):2891–2901, 2011.

[SS05] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[SZ13] Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

[Tob58] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

[VDGB+09] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[VdGBR+14] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

[WKR+14] Larry Wasserman, Mladen Kolar, Alessandro Rinaldo, et al. Berry-esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics*, 8(1):1188–1224, 2014.

[WRG16] Lingxiao Wang, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In *Artificial Intelligence and Statistics*, pages 177–185, 2016.

[WW12] Huixia Judy Wang and Lan Wang. Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 2012.

[WZV+04] Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):R92, 2004.

[YL07] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[Yua10] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

[ZB16] Yinchu Zhu and Jelena Bradic. Significance testing in non-sparse high-dimensional linear models. *arXiv preprint arXiv:1610.02122*, 2016.

[ZB17] Yinchu Zhu and Jelena Bradic. A projection pursuit framework for testing general high-dimensional hypothesis. *arXiv preprint arXiv:1705.01024*, 2017.

[ZBW+14] Yudong Zhao, Bruce M Brown, You-Gan Wang, et al. Smoothed rank-based procedure for censored data. *Electronic Journal of Statistics*, 8(2):2953–2974, 2014.

[ZGR16] Mikhail Zhelonkin, Marc G Genton, and Elvezio Ronchetti. Robust inference in sample selection models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):805–827, 2016.

[ZKL14] Tianqi Zhao, Mladen Kolar, and Han Liu. A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*, 2014.

[ZL13] Tuo Zhao and Han Liu. Sparse inverse covariance estimation with calibration. In *Advances in Neural Information Processing Systems*, pages 2274–2282, 2013.

[ZP96] Kenneth Q. Zhou and Stephen L. Portnoy. Direct use of regression quantiles to construct confidence sets in linear models. *Ann. Statist.*, 24(1):287–306, 02 1996.

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.