Electronic Journal of Statistics
Vol. 13 (2019) 582–645
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1463

Generalized M-estimators for high-dimensional Tobit I models

Jelena Bradic and Jiaqi Guo

Department of Mathematics
University of California, San Diego
e-mail: jbradic@ucsd.edu; jig026@ucsd.edu

Abstract: This paper develops robust confidence intervals in high-dimensional and left-censored regression. Type-I censored regression models, where a competing event makes the variable of interest unobservable, are extremely common in practice. In this paper, we develop smoothed estimating equations that are adaptive to censoring level and are more robust to the misspecification of the error distribution. We propose a unified class of robust estimators, including one-step Mallow's, Schweppe's, and Hill-Ryan's estimator that are adaptive to the left-censored observations. In the ultra-high-dimensional setting, where the dimensionality can grow exponentially with the sample size, we show that as long as the preliminary estimator converges faster than $n^{-1/4}$, the one-step estimators inherit asymptotic distribution of fully iterated version. Moreover, we show that the size of the residuals of the Bahadur representation matches those of the pure linear models – that is, the effects of censoring disappear asymptotically. Simulation studies demonstrate that our method is adaptive to the censoring level and asymmetry in the error distribution, and does not lose efficiency when the errors are from symmetric distributions.

Keywords and phrases: Left-censoring, p-values, inference after model selection, empirical processes, smoothing.

Received May 2017.

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 (Swenson et al., 2014). 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,

[*] This work has been supported by the grant from the NSF, DMS-1712481.

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 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 (Zhao and Yu, 2006; Fan and Li, 2001), 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 we 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 Bradic et al. (2011). Such results heavily rely on assumptions named "irrepresentative condition" and variants thereof including but not limited to the minimal signal strength (Van De Geer et al., 2009). 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. In this paper, we seek to address this challenge: we showcase 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 (Tobin, 1958) include maximum likelihood approaches (Amemiya, 1973), consistent estimators of the asymptotic covariance matrix (Powell, 1984), bayesian methods (Chib, 1992), maximum entropy principles (Golan et al., 1997), etc. 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. In this paper, we explore the use of ideas from the high-dimensional literature to improve the performance of these classical methods with many covariates.

We focus on the family of de-biased estimators introduced by (Zhang and Zhang, 2014), 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 (Bickel, 1975) in that they improve on an initial estimator by following a Newton-Raphson updating rule; 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.

One-step M-estimators even in low-dimensions, however, have not been defined for censored regression models, where measurements are censored by fixed constants. Moreover, knowledge of the underlying data generating mechanism is seldom available, and thus models with fixed-censoring are more prone to the distributional misspecification. To overcome that, we aim at developing a semi-parametric one-step estimator that makes no distributional assumptions.

Despite their widespread success in estimation problems, there are important hurdles that need to be cleared before one-step M-estimators are directly useful. Ideally, an estimator should be consistent with a well-understood asymptotic sampling distribution regardless of the error distribution, so that a researcher can use it to test hypotheses and establish confidence intervals. Yet, the asymptotics of censored one-step estimators have been largely left open, even in the standard regression contexts. This paper addresses these limitations, developing a regularization-based method for the high-dimensional Tobit model that allows for a tractable asymptotic theory and valid statistical inference.

We begin our theoretical analysis by developing the consistency and asymptotic normality results in the context of least absolute deviation regression. We prove these results for a carefully developed estimator that uses one-step corrections to remove regularization bias, while relying on a new technique, named smoothing estimating equations, which allows for efficient semi-parametric inference.

We also show that the generalized M-estimators that are robust to the outliers in the feature distribution can be effectively constructed for the Tobit model. Our methodology builds upon classical ideas from Hampel (1974), as well as Huber (1973). Given these general constructions, we show that our consistency and asymptotic normality result holds when the number of features is larger than the sample size.

1.1. Related work

From a technical point of view, the main contribution of this paper is an asymptotic normality theory enabling statistical inference in high-dimensional Tobit I models. Results by Powell (1986a), Powell (1986b) and Newey and Powell (1990) have established asymptotic properties in low-dimensional setting where the number of features is fixed, while Song (2011) and Zhao et al. (2014b) developed distribution free and rank-based tests. Müller and van de Geer (2016) offered a penalized version of Powell's estimator (penalized CLAD). Robustness properties of sample-selection models in low-dimensions were recently studied in Zhelonkin et al. (2016). To the best of our knowledge, however, we provide the first set of conditions under which semi-parametric estimators are both asymptotically unbiased and Gaussian in high-dimensional settings, thus allowing for classical statistical inference. The extension to the robust high-dimensional estimates robust to both feature and model outliers in this paper is also new.

A growing literature, including Van de Geer et al. (2014), Zhang and Zhang (2014), Ren et al. (2015) and Rinaldo et al. (2016), 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, Belloni et al. (2014, 2013), Zhao et al. (2014a) and Javanmard and Montanari (2014) use robust approaches to estimate the asymptotic variance, and then use related bias correction step to remove the effect of regularization. A limitation of this line of work is that, until now, it has lacked formal statistical inference results in the presence of measurements with fixed censoring.

We view our contribution as complementary to this literature, by showing that bias correction methods may be applied to partially observed data. We believe that the new methodological tools developed here will be useful beyond the specific class of models studied in the paper. In particular, tools of utilizing unknown error distribution as a kernel smoother allow for direct analysis of many estimators with non-smooth loss functions.

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 Van de Geer et al. (2014), which proposes bias correction estimator for least squares losses and obtain valid confidence intervals. Other related approaches include those of Javanmard and Montanari (2014) and Ning and Liu (2017), 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). Belloni et al. (2014) and Zhao et al. (2014a) discuss one-step approaches for quantile inference; however, the tools and techniques heavily depend on the convexity of the quantile loss (observe that our loss is non-convex due to the fixed or constant left-censoring mechanism; random censoring typically does not suffer from this problem). It is worth mentioning that the double-robust approach of Belloni et al. (2017), which proposes a powerful inference method for quantile regression, is based on leveraging principles of doubly-robust scores and their estimating equations. Intriguingly, even in low dimensions, doubly robust methods are not necessary for achieving semiparametric efficiency. We showcase that the newly proposed method achieves efficiency on its own.

1.2. Organization of the paper

In Section 2, we propose the smoothed estimating equations (SEE) for left-censored linear models. In Section 3, we present our main result on confidence regions. In Section 4, we develop robust and left-censored Mallow's, Schweppe's and Hill-Ryan's estimators, and present their theoretical analysis. Section 5 provides numerical results on simulated data sets. In Section 6, we include discussions and conclusions for our work. We defer more general results for confidence regions, as well as the Bahadur representation of the SEE estimator, to Section 7 in the Appendix Section. Section 8 and 9 in the Appendix consist of technical details and proofs.

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 debiasing, score, Wald, and etc.) 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.

2.1. Left-censored linear model

We consider the problem of confidence interval construction where we observe a vector of responses $Y = (y_1, \dots, y_n)$ and their censoring level $c = (c_1, \dots, c_n)$ together with covariates $X_1, \dots X_p$. 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. 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 our 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,

$$y_i = \max\left\{0, x_i\boldsymbol{\beta}^* + \varepsilon_i\right\}. \tag{2.1}$$

2.2. Smoothed estimating equations (SEE)

In this paper, we take a general approach to the problem of designing robust and semi-parametric inference for left-censored linear models. Our estimator 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

$$\mathbb{E}\Big[\Psi(\boldsymbol{\beta}^*)\Big] = 0. \tag{2.2}$$

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 our interest, that allows error misspecifications, is

$$\psi_i(\boldsymbol{\beta}) = \operatorname{sign}\left(y_i - \max\{0, x_i\boldsymbol{\beta}\}\right) w_i^\top(\boldsymbol{\beta}) \tag{2.3}$$

where $w_i(\boldsymbol{\beta}) = x_i \,\mathbb{I}\{x_i\boldsymbol{\beta} > 0\}$. The motivation comes from famous least absolute deviation l_1 loss. The advantage of the function ψ_i above is then 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, our 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 (2.1), 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 (2.2) 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

$$\mathbb{E}[\Psi(\boldsymbol{\beta}^*)] + \boldsymbol{\Upsilon}[\boldsymbol{\beta}^* - \boldsymbol{\beta}] = 0. \tag{2.4}$$

For certain choices of the matrix $\boldsymbol{\Upsilon}$ we aim to avoid both non-convexity and huge dimensionality of the system of equations (2.2). 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 (2.1). 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 (2.3), the respective smoothed score function that we will be working with is

$$S(\boldsymbol{\beta}^*) = n^{-1}\sum_{i=1}^n \left[1 - 2P_\varepsilon\left(y_i - x_i\boldsymbol{\beta}^* \le 0\right)\right]\left(w_i(\boldsymbol{\beta}^*)\right)^\top, \tag{2.5}$$

where P_ϵ denotes the probability measure generated by the errors ε (2.1). Smoothed score will typically depend 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 (2.5) – for homoscedastic errors ε_i, the unknown cdf 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

$$\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*) = 2f_\varepsilon(0)\mathbb{E}_X\left[n^{-1}\sum_{i=1}^n w_i(\boldsymbol{\beta}^*)^\top w_i(\boldsymbol{\beta}^*)\right] := 2f_\varepsilon(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*). \tag{2.6}$$

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 $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) := n^{-1}\sum_{i=1}^n w_i(\boldsymbol{\beta})^\top w_i(\boldsymbol{\beta})$. To infer the parameter $\boldsymbol{\beta}^*$, we need to efficiently solve the SEE equation (2.4). We can observe that solving SEE equations (2.4) 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 $\hat{\boldsymbol{\beta}}$ and a sample plug-in estimate $\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})$ of $\boldsymbol{\Upsilon}(\boldsymbol{\beta}^*)$,

$$\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}}) = 2n^{-1}\hat{f}_\varepsilon(0)\sum_{i=1}^n w_i(\hat{\boldsymbol{\beta}})^\top w_i(\hat{\boldsymbol{\beta}}) \tag{2.7}$$

and a sample estimate of $\mathbb{E}\Psi(\boldsymbol{\beta}^*)$, denoted with $\Psi(\hat{\boldsymbol{\beta}})$ and a suitable density estimate $\hat{f}_\varepsilon(0)$. However, when $p \gg n$, this is highly inefficient. Instead, it is more efficient to directly estimate $\boldsymbol{\Upsilon}^{-1}(\boldsymbol{\beta}^*) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)/2f_\varepsilon(0)$. Let $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ be an estimate of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)$ (see Section 2.3 for discussion). Then, we proceed to solve SEE equations approximately, by defining the SEE estimator as

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\Psi(\hat{\boldsymbol{\beta}})/\hat{f}_\varepsilon(0).$$

Remark 1. *The proposed SEE can be viewed as a high-dimensional extension of inference from estimating equations. Although we consider a left-censored linear model, the proposed SEE methodology applies more broadly. For example, our framework includes loss functions based on ranks or non-convex loss functions for the fully observed data. For instance, the method in Van de Geer et al. (2014) 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 Neykov et al. (2015) do not apply to non-differentiable estimating equations (see Section 2.1 where a twice-differentiable assumption is imposed).*

2.3. Estimation of the scale in left-censored models

We will introduce the methodology for estimating each row of the matrix $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)$. 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 $\mathbf{\Gamma}_{(j)}(\boldsymbol{\beta}^*) = [-\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_1, \ \cdots, \ -\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_{j-1}, \ 1, \ -\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_{j+1}, \ \cdots, \ -\boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)_p]$ and

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

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 $\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}})$ and $\hat{\tau}_j^2$ 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,

$$\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}\left\{n^{-1}\left\|W_j(\hat{\boldsymbol{\beta}}) - W_{-j}(\hat{\boldsymbol{\beta}})\boldsymbol{\gamma}\right\|_2^2 + 2\lambda_j\|\boldsymbol{\gamma}\|_1\right\}. \tag{2.8}$$

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 $\hat{\boldsymbol{\zeta}}_j = W_j(\hat{\boldsymbol{\beta}}) - W_{-j}(\hat{\boldsymbol{\beta}})\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}})$, and a bias corrected estimate of τ_j^2 defined as

$$\hat{\tau}_j^2(\lambda_j) = n^{-1}\hat{\boldsymbol{\zeta}}_j^\top\hat{\boldsymbol{\zeta}}_j + \lambda_j\left\|\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}})\right\|_1. \tag{2.9}$$

Observe that the naive estimate $n^{-1}\hat{\boldsymbol{\zeta}}_j^\top\hat{\boldsymbol{\zeta}}_j$ does not suffice due to the bias carried over by the penalized estimate $\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}})$. Lastly, the matrix estimate of $\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)$, much in the same spirit as Zhang and Zhang (2014) is defined with

$$\mathbf{\Omega}_{jj}(\hat{\boldsymbol{\beta}}) = \hat{\tau}_j^{-2}, \qquad \mathbf{\Omega}_{j,-j}(\hat{\boldsymbol{\beta}}) = -\hat{\tau}_j^{-2}\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}}), \qquad j = 1, \ldots, p. \tag{2.10}$$

The proposed scale estimate can be considered as the censoring adaptive extension of the graphical lasso estimate of Van de Geer et al. (2014).

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), we propose a novel density estimator designed to be adaptive to the left-censoring in the observations. For a positive bandwidth sequence $\hat{h}_n$, we define the density estimator of $f(0)$ as

$$\hat{f}(0) = \hat{h}_n^{-1} \sum_{i=1}^{n} \frac{\mathbb{I}(x_i\hat{\boldsymbol{\beta}} > 0)\, \mathbb{I}(0 \le y_i - x_i\hat{\boldsymbol{\beta}} \le \hat{h}_n)}{\sum_{i=1}^{n} \mathbb{I}(x_i\hat{\boldsymbol{\beta}} > 0)}. \tag{2.11}$$

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 $h_n^{-1} = \mathcal{O}(\sqrt{n/(s\log p)})$ suffices. However, we also propose an adaptive choice of the bandwidth sequence and consider $\hat{h}_n = o(1)$ such that let $u_i := y_i - x_i\hat{\boldsymbol{\beta}}$,*

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

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

2.5. Confidence intervals

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

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\Psi(\hat{\boldsymbol{\beta}})/2\hat{f}(0). \tag{2.12}$$

For the presentation of our coverage rates of the confidence interval (2.15) and (2.16), we start with the Bahadur representation. Lemmas 1–6 (presented in the Appendix) enable us to establish the following decomposition for the introduced one-step estimator $\tilde{\boldsymbol{\beta}}$,

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) + \Delta, \tag{2.13}$$

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 $\hat{\boldsymbol{\beta}}$ is presented below.

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

One particular choice of such estimator can be l_1 penalized CLAD estimator studied in Müller and van de Geer (2016)

$$\hat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{B}} \left\{ \frac{1}{n} \|Y - \max\{0, X\boldsymbol{\beta}\}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{2.14}$$

which satisfies the Condition (I) with $r_n^2 = s_{\boldsymbol{\beta}^*} \log p/n$ and $\|\hat{\boldsymbol{\beta}}\|_0 = \mathcal{O}_P(s_{\boldsymbol{\beta}^*} \times \lambda_{\max}(X^\top X)/n)$, 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 (2.14); any local minima suffices as long as the prediction error is bounded accordingly.

With the normality result of the proposed estimator $\tilde{\boldsymbol{\beta}}$ (as shown in Theorem 8, Section 7 in the Appendix), 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 7 in the Appendix, the standard studentized approach leads to a $(1-2\alpha)100\%$ confidence interval for $\mathbf{c}^\top\boldsymbol{\beta}^*$ of the form

$$\mathrm{I}_n = \left(\mathbf{c}^\top\tilde{\boldsymbol{\beta}} - a_n, \mathbf{c}^\top\tilde{\boldsymbol{\beta}} + a_n\right), \tag{2.15}$$

where $\tilde{\boldsymbol{\beta}}$ is defined in (2.12) and

$$a_n = z_\alpha \sqrt{\mathbf{c}^\top \boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\mathbf{c}} \Big/ 2\sqrt{n}\hat{f}(0) \tag{2.16}$$

with $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ defined in (2.10), $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})$ defined in (2.7) and $\hat{f}(0)$ as defined in (2.11). 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.

3. High-dimensional asymptotics

Within this section, we will present the theoretical results using a specific initial estimator. However, our 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 7 in the Appendix. We begin with a set of very mild model error assumptions.

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, Bickel et al. (2009), Meinshausen and Yu (2009) and Negahban et al. (2009). Meanwhile, their sampling variability has been analyzed by Van de Geer et al. (2014). However, to the best of our knowledge, our 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 Belloni et al. (2014) and Zhao et al. (2014a), 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 Belloni et al. (2014) relies on the score vector being a convex function of unknown parameters, and the Hessian matrix in Zhao et al. (2014a) 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 optimal bias-variance decomposition can be achieved. Although the methods of Belloni et al. (2014) and Zhao et al. (2014a) 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 Belloni et al. (2013) as well as Belloni et al. (2017) 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 our 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 Karoui and Purdom (2016), 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.

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 $G_i(z, \boldsymbol{\beta}, r) = \mathbb{E}[\mathbb{1}(|x_i\boldsymbol{\beta}| \leq \|x_i\|_2 \cdot z)\|x_i\|_2^r]$ and assume that $G_i(z, \boldsymbol{\beta}, r) \leq K_1 \cdot z$, 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 as well as the design matrix of the model (2.1) for further analysis.

Condition (C). *There exist constants $C_2 > 0$ and $\phi_0 > 0$, such that for all $\boldsymbol{\beta}$ satisfying $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^C_{\boldsymbol{\beta}^*}}\|_1 \leq 3\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S_{\boldsymbol{\beta}^*}}\|_1$, $\|\max\{0, X\boldsymbol{\beta}^*\} - \max\{0, X\boldsymbol{\beta}\}\|_2^2 \geq C_2\|X(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2$, 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 $1/v_n = \mathcal{O}(1)$ and assume $\max_j \boldsymbol{\Sigma}_{jj}(\boldsymbol{\beta}^*) = \mathcal{O}(1)$. Moreover, assume that $\max_j \|X\mathbf{v}\|_\infty = \mathcal{O}(K)$ where $\mathbf{v} \in \mathbb{R}^p$.*

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 5) 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 (Bickel et al., 2009). 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.

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

The preceding condition is not uncommon, and can also be found in Van de Geer et al. (2014); Belloni et al. (2014).

With the conditions above, we present our main result. More generalized results for initial estimators satisfying Condition (I) are presented in Theorem 8 and 9 in the Appendix.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}$ be defined as in (2.14) 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 $j = 1, \dots, p$

$$\left\| \hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*) \right\|_1 = \mathcal{O}_P \left(\frac{1}{\phi_0^2 C_2} s_j \sqrt{\log p/n} \right). \tag{3.1}$$

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

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

(iii) Let $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ defined in (2.10). Then, for $\hat{\tau}_j^2$ as in (2.9), we have $\hat{\tau}_j^{-2} = \mathcal{O}_P(1)$. Moreover,

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

(iv) Let $\tilde{\boldsymbol{\beta}}$ be defined as in (2.12) with $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ defined in (2.10), $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})$ defined in (2.7) and $\hat{f}(0)$ as defined in (2.11). Then, for $\bar{s} = s_{\boldsymbol{\beta}^*} \vee s_\Omega$ with $s_\Omega = \max_j s_j$, the size of the residual term in (2.13) is

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

(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 (2.15) and (2.16). Then, for all vectors $\mathbf{c} = \mathbf{e}_j$ and any $j \in \{1, \dots, p\}$, 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 $\|\hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\gamma}^*_{(j)}(\boldsymbol{\beta}^*)\|_1$ 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, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 < n^{-1/4}$ 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 semiparametric estimators of the precision matrix for censored regression models. Even in low-dimensional setting, this result appears to be new and 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 $\hat{\boldsymbol{\beta}}$. In comparison to linear models the established rate is slower for a factor of $\sqrt{s_j}$, whereas in comparison to the results of section 3 of (Van de Geer et al., 2014) (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 Appendix. There we show that a large class of initial estimates suffices.

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

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = U + \Delta. \tag{3.2}$$

with

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

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

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

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

whenever $\bar{s}(\log p)^{1/4}/n^{1/4} = \mathcal{O}(1)$, 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 $o_P(1)$; 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*

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

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. Zhou and Portnoy (1996)), 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 Belloni et al. (2013) 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 (2.1), 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 9 in the Appendix).

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 (Bickel, 1975). There have been considerable challenges in developing good robust methods for more general problems. To the best of our knowledge, there is no prior work that discusses robust one-step estimators for the case of left-censored models (for either high- or low-dimensions).

We propose 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 (2.1) rarely follows distribution with light tail. Namely, model (2.1) can be reparametrized as $y_i = z_i(\boldsymbol{\beta}^*)\boldsymbol{\beta}^* + \xi_i$, where $z_i(\boldsymbol{\beta}^*) = x_i\, \mathbb{1}\{x_i\boldsymbol{\beta}^* + \varepsilon_i \geq 0\}$ and $\xi_i = \varepsilon_i\, \mathbb{1}\{x_i\boldsymbol{\beta}^* + \varepsilon_i \geq 0\}$. Hence ξ_i will often have skewed distribution with heavier tails, and it is in this regard very important to design estimators that are robust. We introduce Mallow's, Schweppe's and Hill-Ryan's estimators for left-censored models.

4.1. Smoothed robust estimating equations (SREE)

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

$$\mathbb{E}[\Psi^r(\boldsymbol{\beta})] = 0 \tag{4.1}$$

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

$$\psi_i^r(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^n q_i w_i^\top(\boldsymbol{\beta})\ \psi\Big(v_i\big(y_i - \max\{0, x_i\boldsymbol{\beta}\}\big)\Big), \tag{4.2}$$

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 (4.1) similarly as before and consider the system of equations

$$\mathbb{E}[\Psi^r(\boldsymbol{\beta}^*)] + \boldsymbol{\Upsilon}^r[\boldsymbol{\beta}^* - \boldsymbol{\beta}] = 0, \tag{4.3}$$

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 (4.1).

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 (2.1). To infer the parameter $\boldsymbol{\beta}^*$, we adapt a one-step approach in solving the empirical counterpart of the population equations above. We name the empirical equations 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

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

The particular form of the matrix $\mathbf{\Upsilon}^r(\boldsymbol{\beta}^*)$ depends on the choice of the weight functions q and v and the function ψ. In particular, for the left-censored model, (2.1)

$$\nabla_{\boldsymbol{\beta}^*}\mathbb{E}_\varepsilon[\Psi^r(\boldsymbol{\beta}^*)] = n^{-1}\sum_{i=1}^n q_i \nabla_{\boldsymbol{\beta}^*}\mathbb{E}_\varepsilon\left[\psi\left(v_i(y_i - \max\{0, x_i\boldsymbol{\beta}^*\})\right)\right] \tag{4.4}$$

leads to the following form

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

whenever the function ψ is differentiable. We denote $\psi'\left(v_i r_i(\boldsymbol{\beta})\right) := \partial\psi\left(v_i r_i(\boldsymbol{\beta})\right)/\partial\boldsymbol{\beta}$, 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 $\psi(\cdot) = \mathrm{sign}(\cdot)$, then $g(\boldsymbol{\beta})$ is equal to $1 - 2P(r_i(\boldsymbol{\beta}) \le 0)$ and $g'(\boldsymbol{\beta}^*) = 2f_{\varepsilon_i}(0)\,\mathbb{I}(x_i\boldsymbol{\beta}^* > 0)$.

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., $\{\mathbf{\Upsilon}^r\}^{-1}(\boldsymbol{\beta}^*)$. 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 $\tilde{W}(\boldsymbol{\beta}) = Q^{1/2}W(\boldsymbol{\beta})$ and

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

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

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

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

$$\tilde{\boldsymbol{\theta}}_{(j)}(\boldsymbol{\beta}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}{\mathrm{argmin}}\ \mathbb{E}\left\|\tilde{W}_j(\boldsymbol{\beta}) - \tilde{W}_{-j}(\boldsymbol{\beta})\boldsymbol{\theta}\right\|_2^2/n. \tag{4.5}$$

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

$$\tilde{\boldsymbol{\Omega}}_{jj}(\hat{\boldsymbol{\beta}}) = \tilde{\mathcal{T}}_j^{-2}, \qquad \tilde{\boldsymbol{\Omega}}_{j,-j}(\hat{\boldsymbol{\beta}}) = -\tilde{\mathcal{T}}_j^{-2}\tilde{\boldsymbol{\theta}}_{(j)}(\hat{\boldsymbol{\beta}}), \tag{4.6}$$

with

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

and the normalizing factor

$$\tilde{\mathcal{T}}_j^2 = n^{-1} \left\| \tilde{W}_j(\hat{\boldsymbol{\beta}}) - \tilde{W}_{-j}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{\theta}}_{(j)}(\hat{\boldsymbol{\beta}}) \right\|_2^2 + \lambda_j \|\tilde{\boldsymbol{\theta}}_{(j)}(\hat{\boldsymbol{\beta}})\|_1.$$

Remark 4. *Estimator* (4.6) *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. Hill (1977) 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,

$$\check{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}}) \left(n^{-1} \sum_{i=1}^n q_i w_i^\top(\hat{\boldsymbol{\beta}})\ \psi\Big(v_i \big(y_i - \max\{0, x_i\hat{\boldsymbol{\beta}}\} \big) \Big) \right). \tag{4.7}$$

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

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

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

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

and $\hat{S} = \{j : \hat{\boldsymbol{\beta}}_j \neq 0\}$. Extending the work of Coakley and Hettmansperger (1993), it is easy to see that Mallow's one-step estimator with $\alpha = 1$ and $b = \chi^2_{\hat{s},0.95}$ quantile of chi-squared distribution with $\hat{s} = |\hat{S}|$ 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

$$v_i = q_i = 1 / \left\| \boldsymbol{\Omega}_{\hat{S},\hat{S}}(\hat{\boldsymbol{\beta}})(w_{i,\hat{S}}(\hat{\boldsymbol{\beta}}) - \bar{w}_{\hat{S}}(\hat{\boldsymbol{\beta}})) \right\|_2, \tag{4.8}$$

and the one-step left-censored Schweppe's estimator with the same q_i as the left hand side of (4.8), 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.

4.3. Theoretical results

Similar to the concise version of Bahadur representation presented in (2.13) for the standard one-step estimator with $q_i = 1$ and $v_i = 1$, we also have the expression for robust generalized M-estimator,

$$\sqrt{n}\left(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = U^{\mathrm{r}} + \Delta^{\mathrm{r}}, \tag{4.9}$$

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 (2.14) 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 such that $|\{k : \boldsymbol{\theta}^*_{(j),k}(\boldsymbol{\beta}^*) \neq 0\}| \leq \tilde{s}_j$ 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 (4.9), we obtain the following result.

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

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

For the residual term of the decomposition (4.9) we obtain the following statement.

Theorem 3. *Let Conditions (C), (r*Γ*) 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* $\bar{s} = s_{\boldsymbol{\beta}^*} \vee \tilde{s}_\Omega$ *with* $\tilde{s}_\Omega = \max_j \tilde{s}_j$*. 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 $\hat{\boldsymbol{\beta}}$ 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 $\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \gamma_n$, and let for simplicity $s_{\boldsymbol{\beta}^*} = s$. Then results of Theorem 3 extend to hold for the confidence interval defined as $\bar{I}_n = (\mathbf{c}^\top\tilde{\boldsymbol{\beta}} - a_n, \mathbf{c}^\top\tilde{\boldsymbol{\beta}} + a_n)$ with a_n as in (4.11). In particular, the error rates are of the order of

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

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 $\scriptstyle\mathcal{O}(n^{-\epsilon})$ for a small $\epsilon > 0$.

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

$$\mathrm{I}_n^{\mathrm{r}} = \left(\mathbf{c}^\top\breve{\boldsymbol{\beta}} - \breve{a}_n, \mathbf{c}^\top\breve{\boldsymbol{\beta}} + \breve{a}_n\right), \tag{4.10}$$

where $\breve{\boldsymbol{\beta}}$ is defined in (4.7), $\mathbf{c} = \mathbf{e}_j$ for some $j \in \{1,2,\ldots,p\}$,

$$\breve{a}_n = z_\alpha\sqrt{\mathbf{c}^\top\tilde{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Upsilon}}^r(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\mathbf{c}}\Big/\sqrt{n}, \tag{4.11}$$

with the robust covariance estimate that we define as

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

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 the Appendix,

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

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 the Appendix imply

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

Together with Theorem 5 in the Appendix, 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 $\mathcal{O}\left(\bar{s}^2(\log(p \vee n))^{3/4} n^{-1/4}\right)$.

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

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

whereas for the Schweppe's estimator

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

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 1. *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 $1 - \alpha$.*

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 (2.14) as linear program within the compact set $\mathcal{B}$, and solve accordingly (Powell, 1984),

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

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, $\hat{\lambda} = \arg\min_{\lambda\in\{\lambda^1,\ldots,\lambda^m\}} \mathrm{CV}(\lambda)$. We move λ in the direction of decreasing regularization until it ceases to be true that $\mathrm{CV}(\lambda) \leq \mathrm{CV}(\hat{\lambda}) + \mathrm{SE}(\hat{\lambda})$. Standard error for the cross-validation curve, $\mathrm{SE}(\hat{\lambda})$, 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

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

with $\hat{\boldsymbol{\beta}}^{-k}(\lambda)$ 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 $\mathrm{CV}^j(\lambda_j) \leq \mathrm{CV}^j(\hat{\lambda}_j) + \mathrm{SE}^j(\hat{\lambda}_j)$ for $\hat{\lambda}_j$ as the cross-validation parameter value. The cross-validation statistic is here defined as

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

   with $\hat{\gamma}_j^{-k}(\lambda_j)$ denoting estimators (2.8) 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 (2.11), with a constant $c = 10$. We compute the above estimator by splitting the sample into two parts: the first sample is used for computing $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ and the other sample is to compute the estimate $\hat{f}(0)$. Optimal value of h is of special independent interest; however, it is not the main objective of this work.
4. Obtain $\tilde{\boldsymbol{\beta}}$ by plugging $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ and $\hat{f}(0)$ into (2.12) 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 and Figures 1 and 2. The *High-Dimensional Regime* are summarized in Table 2 and Figures 3 and 4. 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 $\tilde{X}$. 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 twice the quantity of the maximum entry in X, i.e. $\tilde{X}_{ij} = X_{ij} + 2 \times \max_{ij} X_{ij}$. Such perturbations create a considerate proportion of outliers in the design. The results are summarized in Table 3 and Figures 5 and 6. As coverages under various scenarios are close to the nominal level, the results show that the SREE estimator is robust to high leverage points.

TABLE 1
Coverage Probability for Low-Dimensional Regime with 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 2
Coverage Probability for High-Dimensional Regime with 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 3
Coverage Probability for High-Dimensional Regime with 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

6. Discussion and conclusion

In this article, we 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

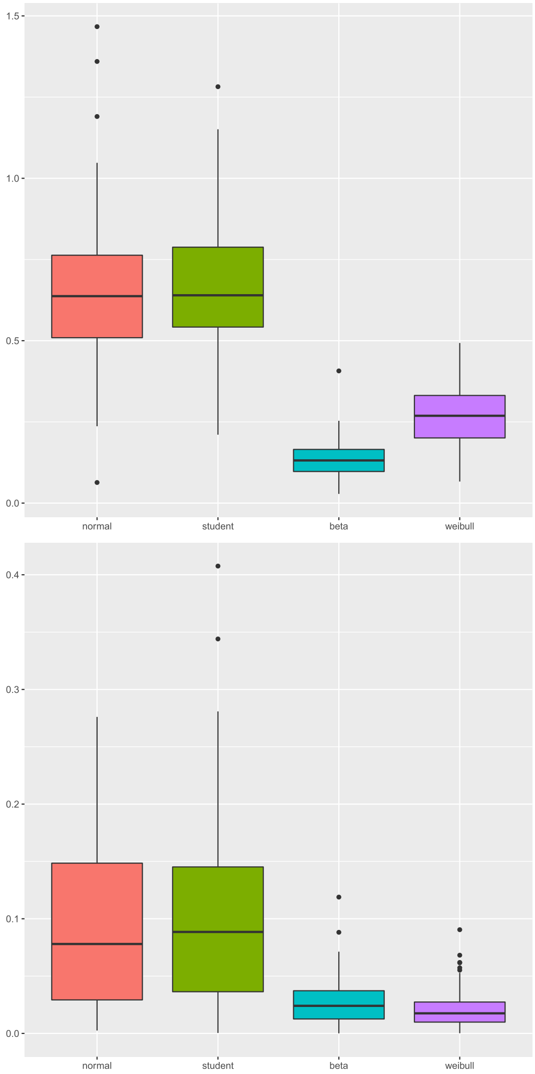

Fig 1. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SEE estimator $p \ll n$ and Toeplitz Design with $\rho = 0.4$.*

methods have been aiming to create efficient methods with asymptotically exact or pivotal properties in a class of specific models. However, 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

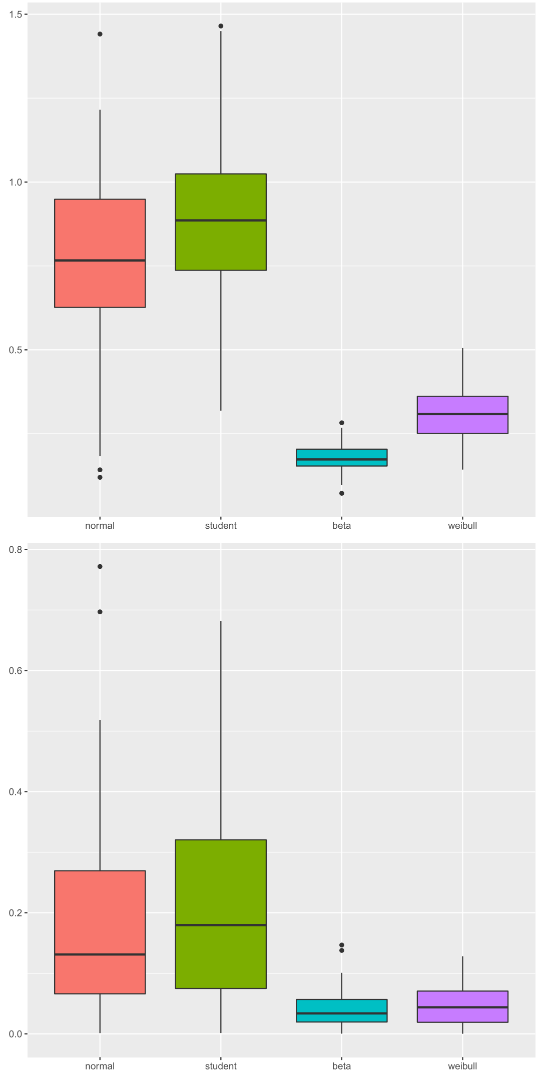

Fig 2. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SEE estimator $p \ll n$ and Identity Design.*

iterate the estimate, and identify misleading observations. Therefore, limiting the effect of poor data quality.

The aim of this article is to establish a new framework for high-dimensional robust inference. 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. We provide a novel theory, with emphasis on diverging dimensions and left-censoring. Future work will be devoted to how to better utilize longitudinal and heterogeneous observations.

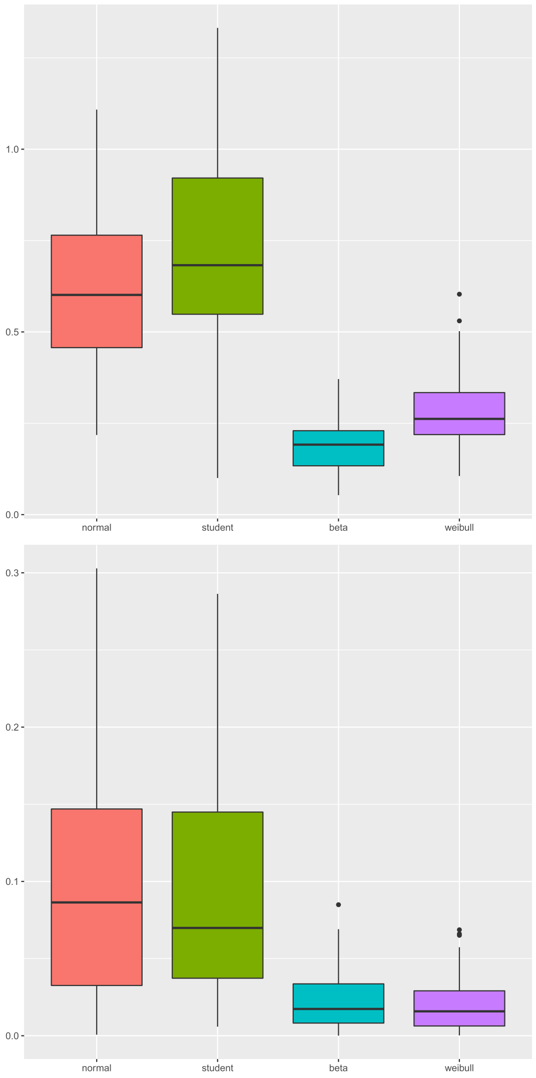

Fig 3. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SEE estimator $p \gg n$ and Toeplitz Design with $\rho = 0.4$.*

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. Our proposed method allows for left-censoring, non-smooth, non-convex losses and/or non-monotone equations, and complements the existing methods in these domains. Our method achieves rates comparable the ones of efficient methods (with full observations), and our analysis provides tight control over both Type I and Type II error rates, which makes it a practically useful and efficient alternative.

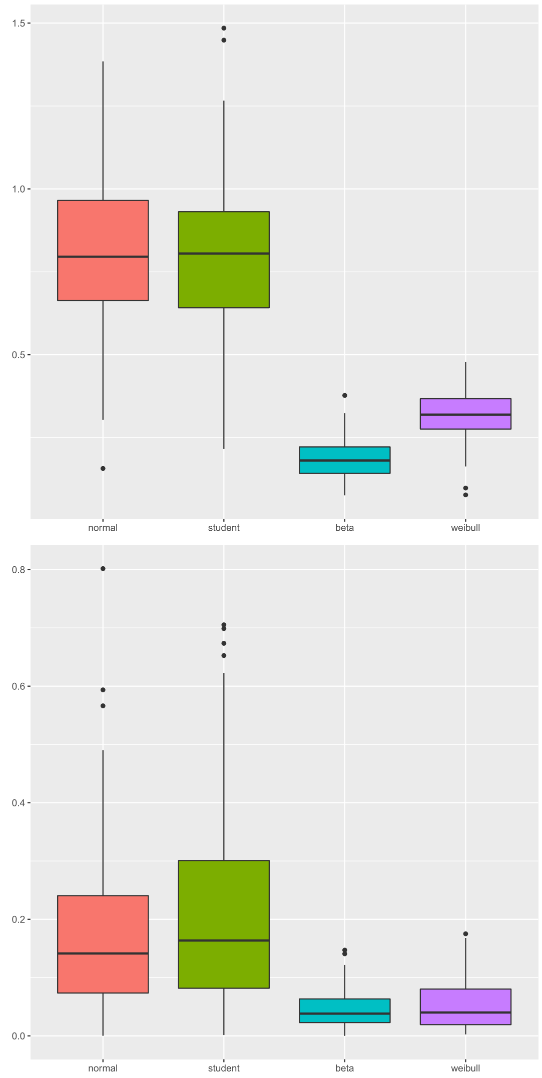

FIG 4. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SEE estimator $p \gg n$ and Identity Design.*

Appendix

General results along with theoretical considerations are presented. In addition, statements and proofs of Lemmas 1–6 and Theorems 5–9, as well as proofs of main Theorems 1–4, are included.

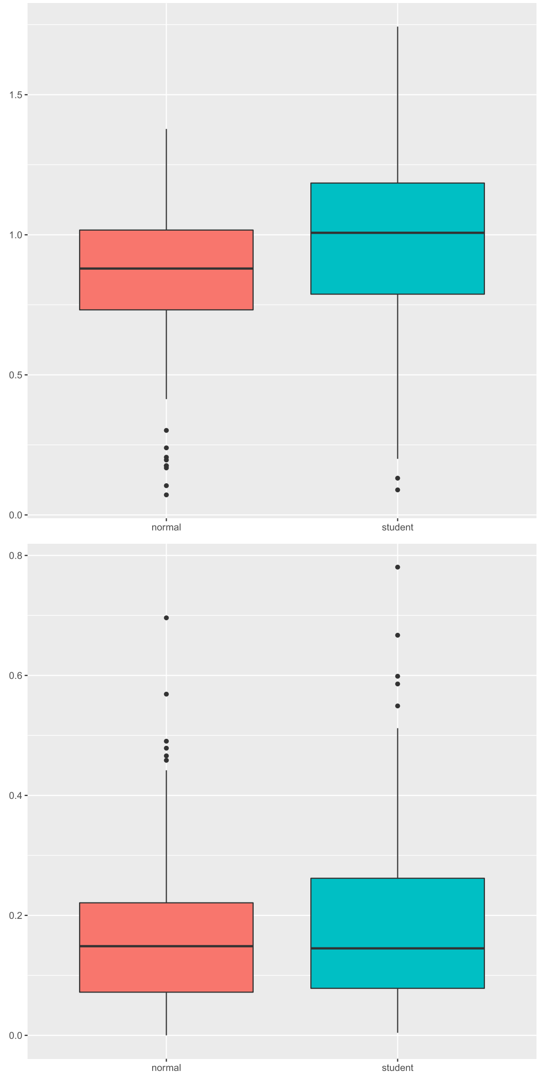

Fig 5. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SREE estimator $p \gg n$ and Identity Design.*

7. General results

We begin theoretical analysis with the following decomposition of (2.12)

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) + \frac{1}{2f(0)}\left(\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)$$

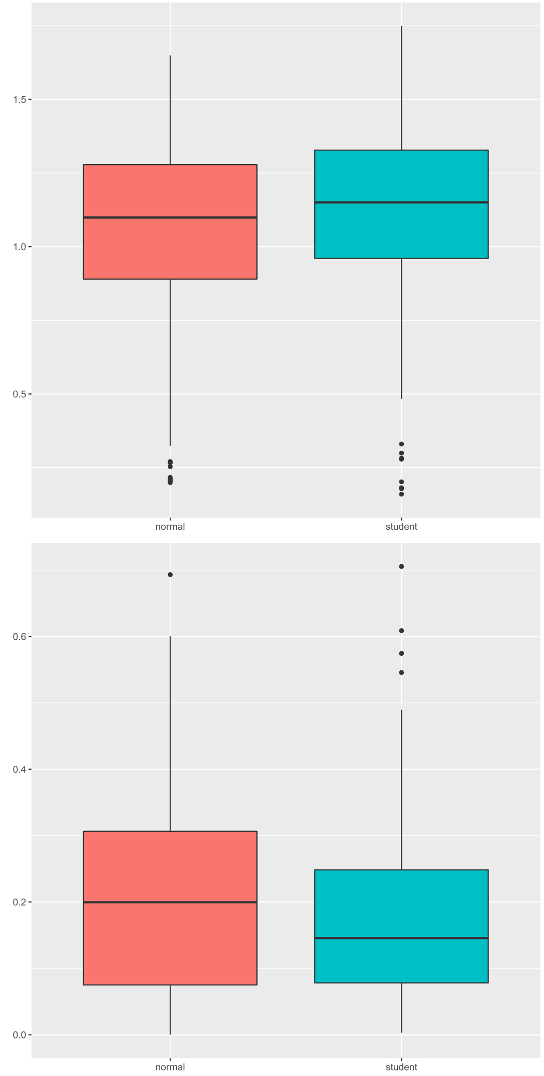

Fig 6. *Comparative boxplots of the average Interval length of Signal (left) and Noise (right) variables. Case of SREE estimator $p \gg n$ and Toeplitz Design.*

$$+\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + \frac{1}{2f(0)}\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\psi_i(\hat{\boldsymbol{\beta}}) - n^{-1}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*)\right). \tag{7.1}$$

We can further decompose the last factor of the last term in (7.1) as $n^{-1}\sum_{i=1}^{n}\psi_i(\hat{\boldsymbol{\beta}}) - n^{-1}\sum_{i=1}^{n}\psi_i(\boldsymbol{\beta}^*) = \mathbb{G}_n(\hat{\boldsymbol{\beta}}) - \mathbb{G}_n(\boldsymbol{\beta}^*) + n^{-1}\sum_{i=1}^{n}\mathbb{E}\Big[\psi_i(\hat{\boldsymbol{\beta}}) -$

$\psi_i(\boldsymbol{\beta}^*)\Big]$, where

$$\mathbb{G}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \left[\psi_i(\boldsymbol{\beta}) - \mathbb{E}\psi_i(\boldsymbol{\beta})\right]. \tag{7.2}$$

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 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, $\mathcal{C}(r,t) = \{\mathbf{w} \in \mathbb{R}^p \mid \|\mathbf{w}\|_2 \le r_n, \sum_{j=1}^p \mathbb{I}\{w_j \neq 0\} \le t\}$ where r_n is a sequence of positive numbers. Then, there exists a fixed constant C (independent of p and n), such that the process $\mu_i(\boldsymbol{\delta}) = \mathbb{I}\{x_i\boldsymbol{\delta} \ge x_i\boldsymbol{\beta}^*\} - \mathbb{I}\{0 \ge x_i\boldsymbol{\beta}^*\}$ satisfies with probability $1-\delta$.*

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

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 (7.1).

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

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

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$ is defined in (2.6).

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 $\psi_i(\hat{\boldsymbol{\beta}})$ 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 (7.1), the Bahadur representation of $\tilde{\boldsymbol{\beta}}$ becomes

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}^*) + I_1 + I_2 + I_3 + I_4 \tag{7.3}$$

where

$$I_1 = \sqrt{n}\left(I - \boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right),$$

$$I_2 = -\frac{1}{2f(0)}\mathbf{\Omega}(\hat{\boldsymbol\beta})\sqrt{n}\cdot\mathcal{O}_P(\|\hat{\boldsymbol\beta}-\boldsymbol\beta^*\|_1)(\hat{\boldsymbol\beta}-\boldsymbol\beta^*)$$

$$I_3 = \frac{1}{2f(0)}\left(\mathbf{\Omega}(\hat{\boldsymbol\beta})-\mathbf{\Sigma}^{-1}(\boldsymbol\beta^*)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\boldsymbol\beta^*),$$

$$I_4 = \frac{1}{2f(0)}\mathbf{\Omega}(\hat{\boldsymbol\beta})\sqrt{n}\left[\mathbb{G}_n(\hat{\boldsymbol\beta})-\mathbb{G}_n(\boldsymbol\beta^*)\right].$$

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 (7.3).

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 $\hat{\boldsymbol\gamma}_{(j)}(\hat{\boldsymbol\beta})$ as defined in (2.8). 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 $\lambda_j = C((\log p/n)^{1/2}\bigvee(r_n^{1/2}\bigvee t^{1/4}(\log p/n)^{1/2})t^{3/4}(\log p/n)^{1/2})$ for a constant $C>1$ and let Conditions (I), (E), (C) and (Γ) hold. Then,*

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

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 $r_n \le t^{-3/4}$ 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 $t \le n^{1/4}$ and $s_j \le n^{1/4}$.*

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

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

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

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

Lemma 5. *Let the setup of Lemma 4 hold. Let $\mathbf{\Omega}(\hat{\boldsymbol\beta})$ be the estimator as in (2.10). Then, for $\hat\tau_j^2$ as in (2.9), we have $\hat\tau_j^{-2} = \mathcal{O}_P(1)$. Moreover,*

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

The one-step estimator $\tilde{\boldsymbol{\beta}}$ 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 $\mathbb{V}_n(\boldsymbol{\delta}) = \boldsymbol{\Omega}(\boldsymbol{\delta}+\boldsymbol{\beta}^*)[\mathbb{G}_n(\boldsymbol{\delta}+\boldsymbol{\beta}^*) - \mathbb{G}_n(\boldsymbol{\beta}^*)]$ satisfies*

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

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 $\|I_4\|_\infty = \mathcal{O}_P\left(r_n^{1/2}t^{3/4}(\log p)^{1/2} \bigvee r_n t(\log p)^{1/2} \bigvee t\log p/n^{1/2}\right)$ in (7.3).

Size of the remainder term in (2.13) is controlled by the results of Lemmas 1–6 and we provide details below.

Theorem 5. *Let $\lambda_j = C((\log p/n)^{1/2} \bigvee (r_n^{1/2} \bigvee t^{1/4}(\log p/n)^{1/2})t^{3/4}(\log p/n)^{1/2})$ for a constant $C>1$ and let Conditions (I), (E), (C) and (Γ) hold. With $s_\Omega = \max_j s_j$,*

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

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 Zhang and Zhang (2014) for example). Then, in the case of low-dimensional problems with $s=\mathcal{O}(1)$ and $p=\mathcal{O}(1)$, we observe that whenever the initial estimator of rate r_n, is in the order of $n^{-\epsilon}$, for a small constant $\epsilon>0$, then $\|\Delta\|_\infty = \mathcal{O}_P(n^{-\epsilon/2})$. In particular, for a consistent initial estimator, i.e. $r_n = \mathcal{O}(n^{-1/2})$ we obtain that $\|\Delta\|_\infty = \mathcal{O}_P(n^{-1/4})$. For high-dimensional problems with s and p growing with n, for all initial estimators of the order r_n such that $r_n = \mathcal{O}(s_{\boldsymbol{\beta}^*}^a(\log p)^b/n^c)$ and $t=\mathcal{O}(s_{\boldsymbol{\beta}^*})$ we obtain that

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

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

Next, we present the result on the asymptotic normality of the leading term of the Bahadur representation (2.13).

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

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

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

Remark 9. A few remarks are in order. Theorem 6 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 (2.11).

Theorem 7. *There exists a sequence* h_n *such that* $h_n = \mathcal{O}(1)$ *and* $\lim_{n\to\infty} \hat{h}_n/h_n = 1$ *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 6 we can provide the next result.

Corollary 2. *With the choice of density estimator as in* (2.11), *under conditions of Theorem 6 and 7, the results of Theorem 6 continue to hold unchanged, i.e.,*

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

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* (2.1). *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 8. *Let* $\lambda_j = C((\log p/n)^{1/2} \bigvee (r_n^{1/2} \bigvee t^{1/4}(\log p/n)^{1/2})t^{3/4}(\log p/n)^{1/2})$ *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 (2.15) and (2.16). Then, for all vectors $\mathbf{c} = \mathbf{e}_j$ and any $j \in \{1, \dots, p\}$, 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 (2.1). Then the following holds.

Theorem 9. *Under the setup and assumptions of Theorem 8 when $n, p, \bar{s} \to \infty$*

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

8. Proofs of main theorems

Proof of Theorem 1. The proof for the result with initial estimator chosen as the penalized CLAD estimator of Müller and van de Geer (2016) follows directly from Lemma 1-6 and Theorem 5-8 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 6 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

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

Moreover, observe that R_i^r are necessarily bounded random variables (see Condition (rΓ)). Following similar steps as in Theorem 6 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

$$n^{-1} \sum_{i=1}^n \mathbb{E}_\varepsilon[\psi_i^r(\boldsymbol{\beta})] = n^{-1} \sum_{i=1}^n x_i^\top \, 1\!\!1\{x_i \boldsymbol{\beta} > 0\} q_i \mathbb{E}_\varepsilon\left[\psi\Big(-v_i x_i(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - v_i \varepsilon_i\Big)\right]. \tag{8.1}$$

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 (8.1) with $\Lambda_n^r(\boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})$, i.e.

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

Next, we observe that by Condition (rΓ),

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

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 (rΓ), and with W replaced with $\tilde{W}$. 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

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

with $\tilde{g}_i(\boldsymbol{\delta}) = q_i\psi(v_i(x_i\boldsymbol{\delta} + \varepsilon_i))$. Moreover, $\mathbb{E}_\varepsilon[f_i(\boldsymbol{\delta})\tilde{g}_i(\boldsymbol{\delta})] = f_i(\boldsymbol{\delta})\mathbb{E}_\varepsilon[q_i\psi(v_i(x_i\boldsymbol{\delta} + \varepsilon_i))] := \tilde{w}_i(\boldsymbol{\delta})$. 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

$$f_i(\boldsymbol{\delta})\tilde{g}_i(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \ge -x_i\boldsymbol{\beta}^*\} q_i\psi(v(\varepsilon_i)) + \mathbb{1}\{x_i\boldsymbol{\delta} \ge -x_i\boldsymbol{\beta}^*\} q_i v_i x_i\boldsymbol{\delta}\psi'(\xi_{\boldsymbol{\delta}})$$

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

$$f_i(\boldsymbol{\delta})\tilde{g}_i(\boldsymbol{\delta}) - \mathbb{E}\left[f_i(\boldsymbol{\delta})\tilde{g}_i(\boldsymbol{\delta})\right] = T_{211}^r(\boldsymbol{\delta}) + T_{212}^r(\boldsymbol{\delta})$$

where

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

and

$$T_{212}^r(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \ge -x_i\boldsymbol{\beta}^*\} q_i v_i x_i\boldsymbol{\delta}\psi'(\xi_{\boldsymbol{\delta}}) - \mathbb{E}\left[\mathbb{1}\{x_i\boldsymbol{\delta} \ge -x_i\boldsymbol{\beta}^*\} q_i v_i x_i\boldsymbol{\delta}\psi'(\xi_{\boldsymbol{\delta}})\right]$$

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

$$T_{2121}^r(\boldsymbol{\delta}) = \mathbb{1}\{x_i\boldsymbol{\delta} \ge -x_i\boldsymbol{\beta}^*\} q_i v_i x_i\boldsymbol{\delta}\left(\psi'(\xi_{\boldsymbol{\delta}}) - \mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}}))\right)$$

and

$$
\begin{aligned}
T_{2122}^{r}(\boldsymbol{\delta}) = & \mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_iv_ix_i\boldsymbol{\delta}\mathbb{E}_\varepsilon(\psi'(\xi_{\boldsymbol{\delta}})) \\
& - \mathbb{E}\left[\mathbb{1}\{x_i\boldsymbol{\delta} \geq -x_i\boldsymbol{\beta}^*\}q_iv_ix_i\boldsymbol{\delta}\mathbb{E}_\varepsilon\psi'(\xi_{\boldsymbol{\delta}})\right]
\end{aligned}
$$

Let us focus on the last expression as it is the most difficult one to analyze. Observe that we are interested in the difference $T_{2122}^{r}(\boldsymbol{\delta}) - T_{2122}^{r}(\tilde{\boldsymbol{\delta}}_k)$. 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

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

First due to monotonicity of indicators and (9.16) we have

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

with

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

As $\sup\psi' < \infty$, I_{11} can be handled in the same manner as T_{21} of the proof of Lemma 6, whereas $I_{12} = \mathcal{O}_P(\tilde{L}_n)$. 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

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

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

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

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 9 applies, but only replace the matrix $\boldsymbol{\Sigma}$ with the matrix $\boldsymbol{\Sigma}^{\mathrm{r}}$. 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 5. The proof of the theorem follows from the bounding residual terms in the Bahadur representation (7.3) with the help of Lemma 3 - 6.

Recall in Lemma 6, we showed that

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

For the term I_3, we have that

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

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

For the term I_2, we have

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

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 (7.3), I_1, we apply Hölder's inequality and Lemma 5,

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

□

Proof of Theorem 6. We begin the proof by noticing that

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

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 $R_i = \mathrm{sign}(-\varepsilon_i)$. 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

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}|V_i|^{2+\delta} &= \left(\frac{1}{\sqrt{n}}\right)^{2+\delta}\mathbb{E}\sum_{i=1}^n |X_{ij}\,\mathbb{1}(x_i\boldsymbol{\beta}^* > 0)|^{2+\delta}\\ &\le n^{-1-\delta/2}\mathbb{E}\sum_{i=1}^n |X_{ij}|^{2+\delta} \le n^{-\delta/2}K. \qquad (8.2)\end{aligned}$$

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,

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

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

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

Thus,

$$\begin{aligned}\text{Var}T_n &\ge \frac{1}{n}\left(t_n^2F(t_n)\,\big|\,_{-n}^{n} - 2\int_{-n}^n t_nF(t_n)dt_n\right) \qquad (8.3)\\ &\ge \frac{1}{n}\left(n^2F(n) - n^2F(-n) - 2\int_{-n}^n t_n dt_n\right)\end{aligned}$$

$$= \frac{1}{n}\left(2n^2F(n) - n^2\right) = n\left(2F(n) - 1\right)$$

Now combining (8.2) and (8.3), 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

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

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

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

Therefore, we have the following conclusion,

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

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

$$\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)_{jj}\right]^{-\frac{1}{2}}\left[\frac{1}{2f(0)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}^*)\right]_j \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f(0)^2}\right) \quad (8.4)$$

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 8, that

$$\left\|\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_{\max} = o_P(1) \quad (8.5)$$

Let $a = \left[\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\right]_{jj}$ 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 (8.5), and so is $\sqrt{ab}(\sqrt{b}+\sqrt{a})$, since we have

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

The rate above follows from (8.9) in the proof of Theorem 8. Notice the rate is of order smaller than the rate assumption in Theorem 5.

Thus, we can deduce that

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

for some finite constant C. Applying Slutsky theorem on (8.4) with the inequality above, the desired result is obtained. □

Proof of Theorem 7. We can rewrite the expression $\hat{f}(0)$ in (2.11) as

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

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
&\mathbb{1}\{0 \le \hat{\eta}_i \le \hat{h}_n\} - \mathbb{1}\{0 \le \eta_i \le h_n\} \\
&\quad = \mathbb{1}\{0 \le \hat{\eta}_i\}\, \mathbb{1}(\hat{\eta}_i \le \hat{h}_n\} - \mathbb{1}\{0 \le \eta_i\}\, \mathbb{1}\{\eta_i \le h_n\} \\
&\quad = \mathbb{1}\{0 \le \hat{\eta}_i\} \Big(\mathbb{1}(\hat{\eta}_i \le \hat{h}_n\} - \mathbb{1}(\eta_i \le h_n\} \Big) \\
&\qquad + (\mathbb{1}\{0 \le \hat{\eta}_i\} - \mathbb{1}\{0 \le \eta_i\})\, \mathbb{1}\{\eta_i \le h_n\} \\
&\quad \le \mathbb{1}\{\hat{\eta}_i \le \hat{h}_n\} - \mathbb{1}\{\eta_i \le h_n\} + \mathbb{1}\{0 \le \hat{\eta}_i\} - \mathbb{1}\{0 \le \eta_i\}.
\end{aligned}
$$

Next, we notice that for real numbers a and b, we have $\mathbb{1}(a > 0) - \mathbb{1}(b > 0) \le \mathbb{1}(|b| \le |a - b|)$. Thus, we have

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

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

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

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 $\mathcal{O}_P(1)$. Then by Hoeffding's inequality, we have that with probability approaching 1 that T_{21} is of $\mathcal{O}_P(1)$. T_{22} can be bounded in exactly the same steps.

Finally, we are ready to put everything together that

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

By applying Slutsky theorem, the result follows directly,

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

Proof of Corollary 2. By multiplying and dividing the term $f(0)$, we can rewrite the term on the left hand side as

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

Also, as a result of theorem 7, we have

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

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 6, we have

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

Proof of Theorem 8. The result of Theorem 8 is a simple consequence of Wald's device and results of Corollary 2. The only missing link is an upper bound on

$$\left\|\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_{\max}. \tag{8.6}$$

First, observe that

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

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

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

We notice that

$$
\begin{aligned}
T_{21} = {} & \Bigg\|\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\Bigg(n^{-1}\sum_{i=1}^{n} w_i^\top(\hat{\boldsymbol{\beta}})w_i(\hat{\boldsymbol{\beta}}) - n^{-1}\sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*) \\
& + n^{-1}\sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*) - n^{-1}\mathbb{E}\sum_{i=1}^{n} w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*)\Bigg)\Bigg\|_{\max} \\
\leq {} & \left\|\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\left(n^{-1}\sum_{i=1}^{n}\left(w_i(\hat{\boldsymbol{\beta}}) + w_i(\boldsymbol{\beta}^*)\right)^\top\left(w_i(\hat{\boldsymbol{\beta}}) - w_i(\boldsymbol{\beta}^*)\right)\right)\right\|_{\max} && (8.7) \\
& + \left\|\mathbf{\Sigma}^{-1}(\boldsymbol{\beta}^*)\left(n^{-1}\sum_{i=1}^{n}\left(w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*) - \mathbb{E}w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*)\right)\right)\right\|_{\max}. && (8.8)
\end{aligned}
$$

For (8.7), we have the following bound

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

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 $\mathcal{O}_P\left(K^2 s_\Omega^{1/2}(r_n^{1/2}t^{3/4}(\log p/n)^{1/2} \vee t\log p/n)\right)$. For (8.8), 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 $\mathcal{O}_P(n^{-1/2})$. Then we bound term (8.8) as following, for some constant C,

$$\begin{aligned}(8.8) &\leq \left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_\infty \left\| n^{-1}\sum_{i=1}^n \left(w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*) - \mathbb{E}w_i^\top(\boldsymbol{\beta}^*)w_i(\boldsymbol{\beta}^*)\right)\right\|_{\max} \\ &\leq C s_\Omega^{1/2} \max_{j,k}\left\{n^{-1}\sum_{i=1}^n (W_{ij}(\boldsymbol{\beta}^*)W_{ik}(\boldsymbol{\beta}^*) - \mathbb{E}W_{ij}(\boldsymbol{\beta}^*)W_{ik}(\boldsymbol{\beta}^*))\right\} = {\scriptstyle\mathcal{O}}_P(1)\end{aligned}$$

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

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

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

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

Thus, $\|\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})\|_{\max}$ is $\mathcal{O}_P(1)$, and so can T_2 be shown similarly. The expression (8.6) is then bounded as,

$$\begin{aligned}&\left\|\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)\right\|_{\max} \\ &\quad = \mathcal{O}_P\left((s_\Omega^{1/2} \vee s_\Omega^{3/2}\lambda_j)(r_n^{1/2}t^{3/4}(\log p/n)^{1/2} \vee t\log p/n) \bigvee s_\Omega^{3/2}\lambda_j\right)\end{aligned} \tag{8.9}$$

which then completes the proof. □

Proof of Theorem 9. The result of Theorem 9 holds by observing that Bahadur representations (7.3) remain accurate uniformly in the sparse vectors $\boldsymbol{\beta} \in \mathcal{B}$; hence, all the steps of Theorem 5 apply in this case as well. □

9. Proofs of lemmas

Proof of Lemma 1. Let $\{\tilde{\boldsymbol{\delta}}_k\}_{k\in[N_\delta]}$ be the centers of the balls of radius $r_n\xi_n$ that cover the set $\mathcal{C}(r_n,t)$. Such a cover can be constructed with $N_\delta \leq \binom{p}{t}(3/\xi_n)^t$ (see, for example Van der Vaart, 2000). 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

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

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

$$\begin{aligned}\sup_{\boldsymbol{\delta}\in\mathcal{C}(r_n,t)}|\mathbb{D}_n(\boldsymbol{\delta})| &= \max_{k\in[N_\delta]}\sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k,r_n\xi_n)}|\mathbb{D}_n(\boldsymbol{\delta})| \\ &\leq \underbrace{\max_{k\in[N_\delta]}|\mathbb{D}_n(\tilde{\boldsymbol{\delta}}_k)|}_{T_1} + \underbrace{\max_{k\in[N_\delta]}\sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k,r_n\xi_n)}\left|\mathbb{D}_n(\boldsymbol{\delta}) - \mathbb{D}_n(\tilde{\boldsymbol{\delta}}_k)\right|}_{T_2}.\end{aligned} \tag{9.1}$$

We first bound the term T_1 in (9.1). To that end, let $Z_{ik} = (\mu_i(\tilde{\boldsymbol{\delta}}_k) - \mathbb{E}[\mu_i(\tilde{\boldsymbol{\delta}}_k)])$. With a little abuse of notation we use l to denote the density of $x_i\boldsymbol{\beta}^*$ for all i. Observe,

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

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

$$\begin{aligned}
&\mathrm{Var}[Z_{ik}] \\
&\quad = \mathbb{E}\Big[\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big) + \mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big) - 2\,\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big)\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big)\Big] \\
&\qquad - \Big[\mathbb{E}\,\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big) - \mathbb{E}\,\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big)\Big]^2 \\
&\quad \overset{(i)}{\leq} \mathbb{E}\Big[\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big) + \mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big) - 2\,\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big)\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big)\Big] \\
&\qquad + 2\mathbb{E}\Big[\Big(\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big) - \mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq x_i\tilde{\boldsymbol{\delta}}_k\Big)\Big)\mathbb{1}\Big(x_i\boldsymbol{\beta}^* \leq 0\Big)\Big] \\
&\quad \overset{(ii)}{\leq} w_i(\tilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0}) + 2\left|w_i(\tilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})\right| \leq 3\left|w_i(\tilde{\boldsymbol{\delta}}_k) - w_i(\mathbf{0})\right|,
\end{aligned} \tag{9.2}$$

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

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

where (iii) follows by the mean value theorem and (iv) from the Condition (E). Hence, we have that almost surely, $|Z_{ik}| \leq C\max_i\left|x_i\tilde{\boldsymbol{\delta}}_k\right|$ for a constant $C < \infty$. For a fixed k, Bernstein's inequality (see, for example, Section 2.2.2 of Van Der Vaart and Wellner, 1996) gives us

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

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

$$\sum_{i\in[n]}\left|x_i\tilde{\boldsymbol{\delta}}_k\right| \leq C^2 n\sqrt{\tilde{\boldsymbol{\delta}}_k^\top X^\top X\tilde{\boldsymbol{\delta}}_k} \leq C^2 n r_n t^{1/2} \tag{9.3}$$

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

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

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

$$\begin{aligned} Z'_{ik} := Q_i(\boldsymbol{\delta}) - Q_i(\tilde{\boldsymbol{\delta}}_k) = &\left[\mathbb{1}(x_i\boldsymbol{\delta} \geq x_i\boldsymbol{\beta}^*) - \mathbb{1}(x_i\tilde{\boldsymbol{\delta}}_k \geq x_i\boldsymbol{\beta}^*) \right] \\ &- \left[\mathbb{E}\,\mathbb{1}(x_i\boldsymbol{\delta} \geq x_i\boldsymbol{\beta}^*) + \mathbb{E}\,\mathbb{1}(x_i\tilde{\boldsymbol{\delta}}_k \geq x_i\boldsymbol{\beta}^*) \right]. \end{aligned}$$

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

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

Moreover,

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

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

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

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 $1-2\delta$,

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

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

$$n^{-1}\sum_{i=1}^n \psi_i(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n x_i^\top \mathbb{1}(x_i\boldsymbol{\beta} > 0)[1 - 2\cdot \mathbb{1}(y_i - x_i\boldsymbol{\beta} < 0)]. \tag{9.4}$$

Using the fundamental theorem of calculus, we notice that if $x_i\boldsymbol{\beta}^* > 0$, $\int_{x_i(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^0 f(\epsilon_i)d\varepsilon_i = F(0) - F(x_i(\boldsymbol{\beta}-\boldsymbol{\beta}^*)) = \frac{1}{2} - P(y_i < x_i\boldsymbol{\beta})$, 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 $\Delta := \left|[\mathbb{E}_X\Lambda_n(\boldsymbol{\beta}) - 2f(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)]_{jk}\right|$, 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,

$$\Delta \le \left|n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)x_{ij}x_{ik}\cdot 2(f(u^*) - f(0))\right| \tag{9.5}$$

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

$$\left. - n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta}^* > 0)x_{ij}x_{ik}\cdot 2f(0)\right|. \tag{9.6}$$

Notice that $\mathbb{1}(x_i\boldsymbol{\beta} > 0) - \mathbb{1}(x_i\boldsymbol{\beta}^* > 0) \le \mathbb{1}(x_i\boldsymbol{\beta} \ge 2x_i\boldsymbol{\beta}^*) = \mathbb{1}[x_i\boldsymbol{\beta}^* \le x_i(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]$. Moreover, the original expresion is also smaller than or equal to $\mathbb{1}(|x_i\boldsymbol{\beta}^*| \le |x_i(\boldsymbol{\beta} - \boldsymbol{\beta}^*)|)$. The term (9.6) can be bounded by the design matrix setup and Condition (E),

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

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

With the help of Hölder's inequality, $|(9.5)| \leq n^{-1}\sum_{i=1}^n \mathbb{E}_X\, \mathbb{1}(x_i\boldsymbol{\beta} > 0)\|x_i\|_\infty^2 \cdot 2\left|f(u^*) - f(0)\right|$. 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 $n \to \infty$

$$\begin{aligned} n^{-1}\sum_{i=1}^n \mathbb{E}\psi_i(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbb{E}_X\mathbb{E}_\varepsilon\psi_i(\boldsymbol{\beta}) &= 2f(0)\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \\ &+ \mathcal{O}(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1)(\boldsymbol{\beta}^* - \boldsymbol{\beta}). \end{aligned} \tag{9.7}$$

□

Proof of Lemma 3. For the simplicity in notation we fix $j = 1$ and denote $\hat{\boldsymbol{\gamma}}_{(1)}(\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}})$. 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 $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ belongs to the appropriate cone set with high probability. We introduce the loss function $l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^n (W_{i,1}(\boldsymbol{\beta}) - W_{i,-1}(\boldsymbol{\beta})\boldsymbol{\gamma})^2$. The loss function above is convex in $\boldsymbol{\gamma}$ hence

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

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

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

$$
\begin{aligned}
&= \sum_{j \in S_1^c} -\lambda_1 |\boldsymbol{\delta}_j| + \sum_{j \in S_1} \lambda_1 |\boldsymbol{\delta}_j| + h^* \|\boldsymbol{\delta}_{S_1}\|_1 + h^* \|\boldsymbol{\delta}_{S_1^c}\|_1 \\
&= (h^* - \lambda_1) \|\boldsymbol{\delta}_{S_1^c}\|_1 + (\lambda_1 + h^*) \|\boldsymbol{\delta}_{S_1}\|_1.
\end{aligned}
$$

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

$$
\mathcal{C}(a, S_1) = \{\mathbf{x} \in \mathbb{R}^{p-1} : \|\mathbf{x}_{S_1^c}\|_1 \leq a \|\mathbf{x}_{S_1}\|_1\} \tag{9.8}
$$

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

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

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

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

Next, by Lemma 1 we observe

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

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

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

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

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

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

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

where $\|u_i\|_{\psi_1} \leq K\|\boldsymbol{\zeta}^*_{1,i}\|_{\psi_1} := \tilde{K} < \infty$. 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 (9.8) to be true holds with high probability.

Step 2. We begin by a basic inequality

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

guaranteed as $\hat{\boldsymbol{\gamma}}$ minimizes the penalized loss (2.8). Here and below in the rest of the proof we suppress the subscript 1 and $\boldsymbol{\beta}$ in the notation of $W_1(\hat{\boldsymbol{\beta}})$ and $W_{-1}(\hat{\boldsymbol{\beta}})$ and use $\hat{W}$ and $\hat{W}^-$ instead and similarly $W^* := W_1(\boldsymbol{\beta}^*)$ and $W^{-*} = W_{-1}(\boldsymbol{\beta}^*)$. Rewriting the inequality above we obtain

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

Observe that $W_{ij}(\hat{\boldsymbol{\beta}}) = W_{ij}(\boldsymbol{\beta}^*) + X_{ij}[\mu_i(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \mu_i(0)]$. Let $\alpha_{ij} = X_{ij}[\mu_i(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \mu_i(0)]$. 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 $\hat{W}_i = W_i^{-*}\boldsymbol{\gamma}^* + \zeta^*_i + \mathbf{A}_i := W_i^{-*}\boldsymbol{\gamma}^* + \varepsilon^*_i$. Hence, the basic inequality above becomes,

$$- 2n^{-1}\left(W^{-*}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}^*\right)^\top (W^{-*} + \mathbf{A}^-)\hat{\boldsymbol{\gamma}} + n^{-1}\hat{\boldsymbol{\gamma}}^\top (W^{-*} + \mathbf{A}^-)^\top (W^{-*} + \mathbf{A}^-)\hat{\boldsymbol{\gamma}}$$

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

With reordering the terms in the inequality above, we obtain

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

for

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

Next, we observe that A_i are bounded, mean zero random variables and hence $n^{-1}|\sum_{i=1}^n A_i| = \mathcal{O}_P(n^{-1/2})$. 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

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

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

Lastly, observe that

$$\delta_3 \leq n^{-1}\boldsymbol{\gamma}^{*\top}\left(\mathbf{A}^{-\top}\mathbf{A}^- + 2W^{-*\top}A^-\right)\boldsymbol{\gamma}^* + n^{-1}\hat{\boldsymbol{\gamma}}^\top\left(\mathbf{A}^{-\top}\mathbf{A}^- + 2W^{-*\top}A^-\right)\hat{\boldsymbol{\gamma}} \tag{9.9}$$

Moreover, as $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ belongs to the cone $C(a, S_1)$ (9.8) by Step 1, by convexity arguments it is easy to see that $\hat{\boldsymbol{\gamma}}$ belongs to the same cone. Together with Hölder's inequality we obtain

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

Utilizing Lemma 1 now provides

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

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

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

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

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

$$n^{-1} C_2 \left\| X(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right\|_2^2 \leq 3\lambda_1 \|\hat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}_{S_1}^*\|_1 + 2\lambda_1 \|\boldsymbol{\gamma}_{S_1}^*\|_2^2 + \lambda_1 \|\hat{\boldsymbol{\gamma}}_{S_1} - \boldsymbol{\gamma}_{S_1}^*\|_2^2 \tag{9.10}$$

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

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

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

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

The result then follows from a simple norm inequality

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

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

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

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

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

$$\left\| \hat{\boldsymbol{\zeta}}_j + \boldsymbol{\zeta}_j^* \right\|_\infty \leq \left\| W_j(\boldsymbol{\beta}^*) - W_{-j}(\boldsymbol{\beta}^*) \boldsymbol{\gamma}_{(j)}^*(\boldsymbol{\beta}^*) \right\|_\infty + \left\| W_j(\hat{\boldsymbol{\beta}}) - W_{-j}(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}}) \right\|_\infty . \tag{9.11}$$

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

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

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

Its multiplying term can be decomposed as following

$$n^{-1}\left\|\hat{\boldsymbol{\zeta}}_j - \boldsymbol{\zeta}_j^*\right\|_1 \leq \underbrace{n^{-1}\left\|X_j \circ \left(\mathbb{1}(X\hat{\boldsymbol{\beta}} > 0) - \mathbb{1}(X\boldsymbol{\beta}^* > 0)\right)\right\|_1}_{i} + \underbrace{n^{-1}\left\|W_{-j}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}}) - W_{-j}(\boldsymbol{\beta}^*)\boldsymbol{\gamma}_{(j)}^*(\boldsymbol{\beta}^*)\right\|_1}_{ii}, \qquad (9.12)$$

where ∘ 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

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

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

For the term ii, we have

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

Observe, that the right hand side is upper bounded with

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

by the design matrix condition. Utilizing Lemma 1, Lemma 3 and Condition (Γ) together we obtain

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

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

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

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 $|\zeta^*_{ij}| = \mathcal{O}_P(K + K s_j^{1/2})$. We then apply Hoeffding's inequality for bounded random variables, to obtain $\left| n^{-1}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^* - n^{-1}\mathbb{E}\boldsymbol{\zeta}_j^{*\top}\boldsymbol{\zeta}_j^* \right| = O_P(K^2 s_j n^{-1/2})$. □

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

$$0 = W_j(\hat{\boldsymbol{\beta}}) - W_{-j}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}}) \tag{9.13}$$

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

In order to show results about the matrices $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\Omega}(\boldsymbol{\beta}^*)$, we first provide a bound on the $\hat{\tau}$ 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 $\hat{\boldsymbol{\zeta}}_j$ and $\boldsymbol{\zeta}_j^*$

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

Moreover, by the Karush-Kuhn-Tucker conditions of problem (2.8) we have $\lambda_j \|\hat{\gamma}_{(j)}(\hat{\boldsymbol{\beta}})\|_1 = n^{-1}\hat{\boldsymbol{\zeta}}_j^\top W_{-j}(\hat{\boldsymbol{\beta}})\hat{\gamma}(\hat{\boldsymbol{\beta}})$, which in turn enables a representation

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

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

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

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 $\mathcal{O}_P\left(K^2 s_j \lambda_j\right)$. Regarding term II, we first point out one result due to the Karush-Kuhn-Tucker conditions of (6),

$$\begin{aligned}\lambda_j \cdot 1^\top \ge \lambda_j \mathrm{sign}\left(\hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}})\right)^\top &= n^{-1}\left(W_j(\hat{\boldsymbol{\beta}}) - W_{-j}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{(j)}(\hat{\boldsymbol{\beta}})\right)^\top W_{-j}(\hat{\boldsymbol{\beta}}) \\ &= n^{-1}\hat{\boldsymbol{\zeta}}_j^\top W_{-j}(\hat{\boldsymbol{\beta}}).\end{aligned}$$

For the term II, we then have

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

since by Lemma 3 we have

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

Putting all the pieces together, we have shown that rate

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

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

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

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

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

Observe that $\mathbb{I}\left\{y_i - x_i\hat{\boldsymbol{\beta}} \leq 0\right\} = \mathbb{I}\left\{x_i\boldsymbol{\delta} \geq \varepsilon_i\right\}$ and hence $1 - 2\,\mathbb{I}\{y_i - x_i\hat{\boldsymbol{\beta}} > 0\} = 2\,\mathbb{I}\left\{y_i - x_i\hat{\boldsymbol{\beta}} \leq 0\right\} - 1$. 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 $\{\tilde{\boldsymbol{\delta}}_k\}_{k\in[N_\delta]}$ be centers of the balls of radius $r_n\xi_n$ that cover the set $\mathcal{C}(r_n, t)$. Such a cover can be constructed with $N_\delta \leq \binom{p}{t}(3/\xi_n)^t$ (see, for example Van der Vaart, 2000). Furthermore, let

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

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

$$\begin{aligned}\sup_{\boldsymbol{\delta}\in\mathcal{C}(r_n,t)} ||\mathbb{V}_n(\boldsymbol{\delta})||_\infty &= \max_{k\in[N_\delta]} \sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k,r_n\xi_n)} ||\mathbb{V}_n(\boldsymbol{\delta})||_\infty \\ &\le \underbrace{\max_{k\in[N_\delta]} ||\mathbb{V}_n(\tilde{\boldsymbol{\delta}}_k)||_\infty}_{T_1} + \underbrace{\max_{k\in[N_\delta]} \sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k,r_n\xi_n)} ||\mathbb{V}_n(\boldsymbol{\delta}) - \mathbb{V}_n(\tilde{\boldsymbol{\delta}}_k)||_\infty}_{T_2} . \qquad (9.14)\end{aligned}$$

Observe that the term T_1 arises from discretization of the sets $\mathcal{C}(r_n,t)$. 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 $\tilde{\boldsymbol{\delta}}_k$. 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 (9.14). 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

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

and

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

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

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

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.

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

$$\overset{(iii)}{\leq} 3a_{ij}^2(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}}_k)\left|x_i\tilde{\boldsymbol{\delta}}_k\right| K_1$$

where (i) follows similarly as in equation (9.2) 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,

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

We will work on the event

$$\mathcal{A} = \left\{\max_{j\in[p]}\left\|\boldsymbol{\Omega}_j(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}}_k) - \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\beta}^*)\right\|_1 \leq C_n\right\} \tag{9.15}$$

which holds with probability at $1-\delta$ using Lemma 5. For a fixed j and k Bernstein's inequality (see, for example, Section 2.2.2 of Van Der Vaart and Wellner, 1996) gives us

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

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

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

where the line follows using the Cauchy-Schwartz inequality, equation (9.3) in proof of Lemma 1, and results of Lemma 5. Combining all of the results above, with probability $1-2\delta$ we have that

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

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

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

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

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

with probability $1-\delta$, since on the event $\mathcal{A}$ in (9.15) we have that $\left|\tilde{Z}_{ijk}\right| \le C_n\Lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\beta}^*))$ and

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

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

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

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

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

since $r_n = \mathcal{O}_P(1)$. 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 $\sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k, r_n\xi_n)}\left|\mathbf{e}_j^\top\left(\mathbb{V}_n(\boldsymbol{\delta}) - \mathbb{V}_n(\tilde{\boldsymbol{\delta}}_k)\right)\right|$ is upper bounded with

$$\underbrace{\sup_{\boldsymbol{\delta}\in\mathcal{B}(\tilde{\boldsymbol{\delta}}_k, r_n\xi_n)}\left| n^{-1}\sum_{i\in[n]}\mathbb{Q}_{ij}(\boldsymbol{\delta}) - \mathbb{Q}_{ij}(\tilde{\boldsymbol{\delta}}_k)\right|}_{T_{21}}$$

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

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,

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

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

$$\max_{k \in [N_\delta]} \max_{i \in [n]} \sup_{\boldsymbol{\delta} \in \mathcal{B}(\tilde{\boldsymbol{\delta}}_k, r_n \xi_n)} \left| x_i\boldsymbol{\delta} - x_i\tilde{\boldsymbol{\delta}}_k \right| \leq r_n\xi_n\sqrt{t} \max_{i,j} |x_{ij}| \leq C r_n \xi_n \sqrt{t} =: \tilde{L}_n. \tag{9.16}$$

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

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

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

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

and the second

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

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

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

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 $\bar{\boldsymbol{\beta}}_k = \alpha\boldsymbol{\beta}^* + (1-\alpha)(\boldsymbol{\beta}^* + \tilde{\boldsymbol{\delta}}_k)$ and $\alpha \in (0,1)$, instead of $|a_{ij}(\boldsymbol{\beta}^*)|$. Observe that $a_{ij}(\boldsymbol{\beta})\hat{\tau}_j^2 = -\hat{\gamma}_{(j),i}$. Moreover, by construction $\hat{\tau}_j$ is a continuous, differentiable and convex function of $\boldsymbol{\beta}$ and is bounded away from zero by Lemma 5. Additionally, $\hat{\boldsymbol{\gamma}}_{(j)}$ 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, $\hat{\gamma}_j$ is a bounded random variable according to Lemma 5. Hence, $|a'_{ij}(\boldsymbol{\beta}^*)| \le K'$, for a large enough constant K'. Therefore, for a large enough constant C we have

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

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 $N_\delta \lesssim (pn^2)^t$. With these choices, the term T_2 is negligible compared to T_1 and we obtain

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

which completes the proof. □

References

Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016, 1973. MR0440773

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.

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. MR3335097

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. MR3852664

Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975. MR0386168

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. MR2533469

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. MR2815779

Siddhartha Chib. Bayes inference in the tobit censored regression model. *Journal of Econometrics*, 51(1–2):79–99, 1992. MR1151954

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. MR1242938

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. MR1946581

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. MR1457696

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974. MR0362657

Richard Walter Hill. *Robust regression when there are outliers in the carriers*. PhD thesis, Harvard University, 1977. MR2940734

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973. MR0356373

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. MR3265038

Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimension? *arXiv preprint arXiv:1608.00696*, 2016.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009. MR2488351

Patric Müller and Sara van de Geer. Censored linear model in high dimensions. *Test*, 25(1):75–92, 2016. MR3463803

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.

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. MR1085576

M. Neykov, Y. Ning, J. S. Liu, and H. Liu. A Unified Theory of Confidence Regions and Testing for High Dimensional Estimating Equations. *ArXiv e-prints*, October 2015. MR3843384

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. URL `https://doi.org/10.1214/16-AOS1448`. MR3611489

James L Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984. MR0752444

James L Powell. Censored regression quantiles. *Journal of econometrics*, 32(1): 143–155, 1986a. MR0853049

James L Powell. Symmetrically trimmed least squares estimation for tobit models. *Econometrica: journal of the Econometric Society*, pages 1435–1460, 1986b. MR0868151

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. MR3346695

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.

Weixing Song. Distribution-free test in tobit mean regression model. *Journal of Statistical Planning and Inference*, 141(8):2891–2901, 2011. MR2787753

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.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958. MR0090462

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. MR3224285

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. MR2576316

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. MR1652247

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996. MR1385671

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. MR3153940

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006. MR2274449

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*, 2014a.

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, 2014b. MR3299129

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. MR3534351

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. URL `http://dx.doi.org/10.1214/aos/1033066210`. MR1389891