Journal of Multivariate Analysis 140 (2015) 72–91

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Weighted $\ell_1$-penalized corrected quantile regression for high dimensional measurement error models[☆]

Abhishek Kaul[*], Hira L. Koul

*Michigan State University, United States*

ARTICLE INFO

*Article history:*
Received 17 October 2014
Available online 27 April 2015

*AMS subject classifications:*
62J05
62J07

*Keywords:*
$\ell_1$-consistency
Model selection consistency

ABSTRACT

Standard formulations of prediction problems in high dimension regression models assume the availability of fully observed covariates and sub-Gaussian and homogeneous model errors. This makes these methods inapplicable to measurement errors models where covariates are unobservable and observations are possibly non sub-Gaussian and heterogeneous. We propose a weighted penalized corrected quantile estimator for regression parameters in linear regression models with additive measurement errors, where unobservable covariate is nonrandom. The proposed estimators forgo the need for the above mentioned model assumptions. We study these estimators in a high dimensional sparse setup where the dimensionality can grow exponentially with the sample size. We provide bounds for the statistical error associated with the estimation, that hold with asymptotic probability 1, thereby providing the $\ell_1$-consistency of the proposed estimator. We also establish the model selection consistency in terms of the correctly estimated zero components of the parameter vector. A simulation study that investigates the finite sample accuracy of the proposed estimator is also included in the paper.

## 1. Introduction

In various scientific fields such as econometrics, epidemiology and finance, data is often observed with measurement error in the covariates. It is well known that ignoring this measurement error leads to biased parameter estimates, see, e.g., [10] and Carroll, Ruppert and Stefanski [6]. On the other hand, quantile regression is an important estimation method. It is robust against outliers and is especially useful in the presence of heteroscedasticity, see, e.g., [4]. In the past two decades there has been an abundance of large scale data-sets, where the number of parameters vastly outnumber the number of observations. In standard formulations of prediction problems for these data sets, the covariates are assumed to be fully-observed and the model errors are assumed to be sampled independently from some underlying sub-Gaussian and homoscedastic distribution. It is thus of interest and desirable to develop a quantile based estimator in the presence of measurement error in the covariates, which is capable of dealing with non sub-Gaussian and heterogeneous data.

In the last two decades, a vast amount of literature has been developed on $\ell_1$ penalized estimators, in particular, Lasso and its various cousins, which have been theoretically and computationally very successful in estimation and variables selection, see, e.g., [3] and references therein. See [10,6] for numerous applications of measurement error in linear regression models when the number of parameters $p$ is fixed.

[☆] Reasearch supported in part by NSF-DMS grant 1205271, P.I., H. L. Koul.

[*] Corresponding author.
*E-mail address:* kaulabhi@stt.msu.edu (A. Kaul).

http://dx.doi.org/10.1016/j.jmva.2015.04.009
0047-259X/© 2015 Elsevier Inc. All rights reserved.

In the high dimensional setting where one estimates the mean of a sparse linear model, several authors including Rosenbaum and Tsybakov [17,18] and Loh and Wainwright [13] studied measurement error models and proposed modified versions of the Dantzig selector and the least squares estimators, respectively. Loh and Wainwright [13] make the important contribution of providing statistical error bounds for non convex loss functions. However, both papers assume an underlying sub-Gaussian homoscedastic distribution of the model errors.

The problem of quantile regression with non sub-Gaussianity and heteroscedasticity in sparse high dimensional models has recently been studied by Fan, Fan and Barut [8], Belloni and Chernozhukov [2], and Wang, Wu and Li [23]. These authors investigate regression quantiles in high dimensional set up but when there is no measurement error in the covariate. The convexity of quantile loss function is crucial for the analysis of their inference procedures.

Wang, Stefanski and Zhu [22] (WSZ) introduce a corrected quantile estimator for measurement error models and establish its consistency and asymptotic normality when $p$ is fixed. In this paper, we propose and analyze a weighted penalized version of this estimator ($W\ell_1$-$CQ$) for quantile regression under the high dimensional measurement error setup with possible non sub-Gaussianity and heteroscedasticity.

A major challenge in dealing with noisy covariates is the possible non-convexity of the associated loss function. However, as shown by WSZ in the case of fixed $p$, this loss function is asymptotically approximated by the usual convex quantile loss function. We show that this approximation continues to hold in our high dimensional setting, albeit at a slower rate. We rely on this useful result to establish the validity of the proposed estimator. The main contributions of this paper are to provide bounds on the statistical error of the proposed $W\ell_1$-$CQ$ estimator and also to establish the model selection consistency of this estimator in terms of identifying the correct zero components of the parameter vector. We also illustrate its empirical success via a simulation study. To the best of our knowledge this is the first attempt at providing robust estimates for quantile regression in the high dimensional measurement error setup.

The rest of this paper is organized as follows. Section 2 provides the detailed description of the model and the proposed estimator. Section 3 provides a key approximation result that forms the basis of our results. Section 4 provides the main results regarding the statistical error of the proposed estimator and the model selection consistency. In Section 5, we empirically verify the performance of the proposed estimator. Proofs are postponed to the Appendix.

## 2. Model and estimator

In this section, we describe the model and the proposed estimator. Also, the necessary assumptions on the model parameters are exhibited. We consider a linear regression model with additive error in the design variables. Accordingly, let $x_i = (x_{i1}, \dots, x_{ip})^T,\ i = 1, \dots, n$, be vectors of non-random design variables, where for any vector $a$, $a^T$ denotes its transpose. Let $y_i$'s denote the responses, which are related to $x_i$'s by the relations

$$y_i = x_i^T \beta^0 + \varepsilon_i, \quad \text{for some } \beta^0 \in \mathbb{R}^p,\ 1 \le i \le n. \tag{2.1}$$

Here $\beta^0 = (\beta_1^0, \dots, \beta_p^0) \in \mathbb{R}^p$ is the parameter vector of interest, and $\varepsilon^T = (\varepsilon_1, \dots, \varepsilon_n)$ is an $n$-dimensional vector whose components are independent but not necessarily identically distributed, and satisfy $P(\varepsilon_i \le 0) = \tau$, for every $1 \le i \le n$, where $\tau \in (0, 1)$ is the quantile level of interest.

Furthermore, the design variables $x_i$'s are not observed directly. Instead, we observe the surrogate $w_i$'s obeying the model,

$$w_i = x_i + u_i, \quad 1 \le i \le n. \tag{2.2}$$

Here, $u_i^T = (u_{i1}, \dots, u_{ip})$ are assumed to be independent of $\{\varepsilon_i\}$ and independent and identically distributed (i.i.d.) according to a $p$-dimensional multivariate Laplace distribution which is defined via its characteristic function as follows,

**Definition 2.1.** A random vector $u \in \mathbb{R}^p$ is said to have a multivariate Laplace distribution $L_p(\mu, \Sigma)$, if for some $\mu \in \mathbb{R}^p$ and a nonnegative definite symmetric $p \times p$ matrix $\Sigma$, its characteristic function is $\Big(1 + t^T \Sigma t/2 - i\mu t\Big)^{-1},\ t \in \mathbb{R}^p$.

Note that, if $\mu = 0$, then $\Sigma$ is the covariance matrix of the random vector $u$.

Laplace distributions are often used in practice to model data with tails heavier than normal. McKenzie et al. [14] used these distributions in the analysis of global positioning data and Purdom and Holmes [15] adopted Laplace measurement error model in the analysis of data from some microarray experiments. Stefanski and Carroll [20] provide an in depth discussion of Laplace measurement errors.

In our setup, we shall consider the model (2.1) and (2.2) in the high dimensional setting, i.e., the dimension $p$ of the parameter vector $\beta^0$ is allowed to grow exponentially with $n$, and where the measurement errors $u_i$, $1 \le i \le n$ are i.i.d. $L_p(0, \Sigma)$, with $\Sigma$ known. Furthermore, $\beta^0$ is assumed to be sparse, i.e., only a small proportion of the parameters are assumed to be non zero. The number of non zero components shall be denoted by $s$, where $s$ is allowed to diverge slower than $n$. Let $S = \big\{j \in \{1, 2, \dots, p\};\ \beta_j^0 \ne 0.\big\}$, and $S^c$ denote its compliment set. Note that card$(S) = s$. Also, for any vector $\delta \in \mathbb{R}^p$, let $\delta_S = \{\delta_j; j \in S\}$ and $\delta_{S^c} = \{\delta_j; j \in S^c\}$.

All results presented in the paper shall assume the unobserved design variables $x_i$'s to be non-random. However, it is worth pointing out that the assumptions made in this paper on $x_i$'s can be shown to hold with probability converging to 1 under the usual random designs setup, in particular for sub-Gaussian or sub-Exponential designs with independent observations.

**Notation and convention**: In what follows, parameter values like $p, s$ depend on the sample size $n$ but we do not exhibit this dependence for the sake of brevity. For any $z = (z_1, \ldots, z_p)^T \in \mathbb{R}^p$, $\|z\|_1 = \sum_{j=1}^p |z_j|$, $\|z\|_2^2 = \sum_{j=1}^p z_j^2$. For any two sequences of positive numbers $\{a_n, b_n\}$, $a_n = O(b_n)$, denotes that for all large $n$, $a_n \le cb_n$, for some universal constant $c > 0$, which does not depend on any underlying parameters or the sample size $n$. All limits are taken as $n \to \infty$, unless specified otherwise. For any event $A$, $I_A$ denotes the indicator of the event $A$.

When the design variables $x_i$ are completely observed, several authors including Fan, et al. [8], Belloni and Chernozhukov [2] and Wang, Wu and Li [23] have shown that $\beta^0$ can be estimated consistently by

$$\hat{\beta}_x = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \beta) + \lambda_n \|d \circ \beta\|_1 \right\}, \tag{2.3}$$

where $\rho(y_i, x_i, \beta) = \rho_\tau(y_i - x_i^T\beta)$, $\rho_\tau(v) = v\{\tau - I(v \le 0)\}$ is the quantile loss function, and $d = (d_1, \ldots, d_p)^T$ is a vector of non-negative weights, and '$\circ$' denotes the Hadamard product, i.e. $\|d \circ \beta\|_1 := \sum_{j=1}^p d_j|\beta_j|$.

To overcome the difficulty due to measurement error in the covariates, we begin with the regularized version of corrected quantile estimator $W\ell_1$-$CQ$. The un-penalized version was introduced by Wang, Stefanski and Zhu [22] (WSZ). To describe their loss function, let $K(\cdot)$ denote a kernel density function, $h = h_n \to 0$ be sequence of positive window widths, and define $H(x) = \int_{-\infty}^x K(u)du$. Let

$$\rho_L^\star(y_i, w_i, \beta, h) = \tilde{\varepsilon}_i(\tau - 1) + \tilde{\varepsilon}_i H\left(\frac{\tilde{\varepsilon}_i}{h}\right) - \frac{\sigma_\beta^2}{2}\left\{\frac{2}{h}K\left(\frac{\tilde{\varepsilon}_i}{h}\right) + \frac{\tilde{\varepsilon}_i}{h^2}K'\left(\frac{\tilde{\varepsilon}_i}{h}\right)\right\}, \tag{2.4}$$

where, $\tilde{\varepsilon}_i = y_i - w_i^T\beta$ and $\sigma_\beta^2 = \beta^T\Sigma\beta$. WSZ proposed to approximate the quantile function $\rho_\tau(y_i - x_i^T\beta^0)$ by the smooth function $\rho_L^\star(y_i, w_i, \beta, h)$ and defined their estimator as a minimizer with respect to $\beta$ of the average $n^{-1}\sum_{i=1}^n \rho_L^\star(y_i, w_i, \beta, h)$. Its penalized analog suitable in high dimension is

$$l_n^\star(\beta) := \frac{1}{n}\sum_{i=1}^n \rho_L^\star(y_i, w_i, \beta, h) + \lambda_n\|d \circ \beta\|_1.$$

Observe that $l_n^\star(\beta)$ is non-convex and $l_n^\star(\beta)$ may diverge when $\sigma_\beta^2 = \beta^T\Sigma\beta \to \infty$. Hence, we restrict the parameter space to the expanding $\ell_1$-ball $\Theta = \{\beta \in \mathbb{R}^p; \|\beta\|_1 \le b_0\sqrt{s}\}$, for some $b_0 > 0$. Now, define the $W\ell_1$-$CQ$ estimator as

$$\hat{\beta} = \arg\min_{\beta\in\Theta} l_n^\star(\beta). \tag{2.5}$$

The weights $d_j$, $1 \le j \le p$ are assumed to satisfy

$$\text{(i)}\ \max_{j\in S} d_j \le c_{\max}, \quad 0 < c_{\max} < \infty, \qquad \text{(ii)}\ \min_{j\in S^c} d_j \ge c_{\min}, \quad 0 < c_{\min} < \infty. \tag{2.6}$$

We provide bounds on the statistical error associated with the $W\ell_1$-$CQ$ estimator, namely, bounds on the quantities $\|\hat{\beta} - \beta^0\|_1$ and $n^{-1}\|\Gamma^{1/2}X(\hat{\beta} - \beta^0)\|_2$, where $\Gamma$ is defined in (4.2). The $\ell_1$-consistency of $\hat{\beta}$ will be a direct consequence of these error bounds. Note that the choice $d_j = 1$, for all $1 \le j \le p$, makes $\hat{\beta}$ to be the un-weighted penalized $\ell_1$-$CQ$ estimator. As shall become apparent, the $\ell_1$-$CQ$ estimator is also $\ell_1$-consistent in estimation. This shall also be observed in the simulation study in Section 4. On the other hand, it is now well known that un-weighted $\ell_1$-penalized least squares estimators do not possess variable selection oracle properties, i.e., theoretically, the estimated zero and non zero components may not be the same as the true zero and non zero components, respectively, except under restrictive conditions such as the strong irrepresentable condition, see, e.g., [26,25]. The weights $\{d_j\}$, chosen appropriately, shall serve to improve on this issue, by guaranteeing that the zero components are identified correctly with asymptotic probability 1, thereby making $W\ell_1$-$CQ$ model selection consistent in addition to being $\ell_1$-consistent.

We shall now describe the model more precisely while also providing assumptions necessary to proceed further.

(A1) *Model errors* ($\varepsilon$): The distribution function (d.f.) $F_i$ of $\varepsilon_i$ has Lebesgue density $f_i$ such that $\sup_{1\le i\le n, x\in\mathbb{R}} f_i(x) < \infty$, and $f_i$ is uniformly (in $i$) bounded away from zero, in a neighborhood of zero. Also, there exists universal constants $C_1 > 0$, $C_2 > 0$ such that for any $y$ satisfying $|y| \le C_1$,

$$\max_{1\le i\le n} |F_i(y) - F_i(0) - yf_i(0)| \le C_2y^2.$$

This condition is the same as Condition 1 in [8]. It imposes only mild conditions on the error densities and is slightly stronger than the Lipschitz condition for $f_i$'s around the origin. Gaussianity and homoscedasticity is not imposed. Several distributions, including double exponential and Cauchy, satisfy this condition.

(A2) *Unobserved design matrix* $X$: For all $1 \le j \le p$, $n^{-1}\sum_{i=1}^n x_{ij}^2 \le c_x$, for some constant $c_x < \infty$.

(A3) *Measurement errors*: The measurement errors $\{u_i\}$ are independent of $\{\varepsilon_i\}$, and i.i.d. $L_p(0, \Sigma)$, for all $1 \le i \le n$, with a known $\Sigma$. Furthermore, there exists a constant $0 < \sigma_u^2 < \infty$ such that $\max_{1\le j\le p} \mathrm{Var}(u_{ij}) \le \sigma_u^2$.
(A4) *Kernel function $K$*: $K$ is the probability density function of a standard normal random variable.

This choice of the Kernel function shall play an important role in our analysis. This kernel function is chosen for its many tractable properties, namely, it is symmetric around origin, infinitely differentiable, and more importantly, its derivatives being Lipschitz continuous, which is detailed in the Appendix.

## 3. Relationship between $\rho_L^\star$ and $\rho$

The analysis to follow relies critically on the approximation of the corrected quantile loss function $\rho_L^\star$ defined in (2.4) in terms of observed $w_i$'s by the usual convex quantile function $\rho$ defined in (2.3) involving unobserved $x_i$'s. We begin by establishing this connection.

The approximation result we derive for the current high dimensional set up, where $p$ is increasing exponentially with $n$, is similar to the one used in WSZ in the case of fixed $p$. For that reason we use similar notation as in WSZ. Accordingly, define a smoothed quantile loss function with arguments $(y_i, x_i, \beta, h)$,

$$\rho_L(y_i, x_i, \beta, h) = (y_i - x_i^T\beta)\left\{\tau - 1 + H\left(\frac{y_i - x_i^T\beta}{h}\right)\right\}. \tag{3.1}$$

Note that $\rho_L^\star$ is a function of the observed covariates $w$, whereas the $\rho_L$ and $\rho$ are functions of the unobserved covariates $x$. Now, for $\beta \in \Theta$, define

$$M_n^*(\beta) \equiv M_n^\star(w, \beta, h) = n^{-1}\sum_{i=1}^n\Big\{\rho_L^\star(y_i, w_i, \beta, h) - \rho_L^\star(y_i, w_i, \beta^0, h)\Big\}, \tag{3.2}$$

$$\tilde{M}_n(\beta) \equiv \tilde{M}_n(x, \beta, h) = n^{-1}\sum_{i=1}^n\Big\{\rho_L(y_i, x_i, \beta, h) - \rho_L(y_i, x_i, \beta^0, h)\Big\},$$

$$M_n(\beta) \equiv M_n(x, \beta) = n^{-1}\sum_{i=1}^n\Big\{\rho(y_i, x_i, \beta) - \rho(y_i, x_i, \beta^0)\Big\}.$$

We are now ready to state the following theorem describing the approximation of the processes $M_n^\star(\beta)$ and $M_n(\beta)$ by their respective expectations, uniformly in $\beta \in \Theta$, in probability with rates. Its proof is given in the Appendix section. Throughout, $\gamma_{\max}$ denotes the largest eigenvalue of $\Sigma$.

**Theorem 3.1.** *Assume the measurement error model* (2.1) *and* (2.2) *and the assumptions* (A1), (A2), (A3) *and* (A4) *hold. Then,*

$$\sup_{\beta\in\Theta} |M_n^\star(\beta) - EM_n^\star(\beta)| = O_p\left(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\right), \tag{3.3}$$

$$\sup_{\beta\in\Theta} |\tilde{M}_n(\beta) - E\tilde{M}_n(\beta)| = O_p\left(\sqrt{s}\sqrt{\frac{2\log 2p}{n}}\right). \tag{3.4}$$

To proceed further, we require the following two results of WSZ. First, the twice differentiability of $\rho_L(y, x, \beta, h)$ in the variable $y - x'\beta$ and $u_i \sim L_p(0, \Sigma)$ imply

$$EM_n^\star(\beta) = E\tilde{M}_n(\beta), \quad \forall \beta \in \mathbb{R}^p. \tag{3.5}$$

Secondly, under assumption (A4),

$$\sup_{\beta\in\Theta} |\tilde{M}_n(\beta) - M_n(\beta)| = O(h), \quad \text{a.s.} \tag{3.6}$$

Claim (3.5) is a direct consequence of Theorem 2 of WSZ while claim (3.6) is proved in WSZ as a part of the proof of their Theorem 3 (p. 14). The short proof of this statement is reproduced here for the convenience of a reader.

**Proof of (3.6).** Denote by $\rho_L(e, h) := \rho_L(y, x, \beta, h)$, where $e = y - x'\beta$. Similarly define $\rho(e)$. Let $Z$ denote a r.v. having d.f. $H$. Use the symmetry of $K$, and hence of $H$, the finiteness of its first moment, and the change of variable formula to obtain that the left hand side of (3.6) is bounded above by 2 times

$$\sup_e |\rho_L(e, h) - \rho(e)| \le \sup_e \left|e\Big[H\Big(\frac{e}{h}\Big) - I\{e > 0\}\Big]\right| \le \sup_t \Big|htH\big(-|t|\big)\Big| \le hE|Z|.$$

This completes the proof of (3.6).

It is important to note that both of these results are valid without any restriction on the model dimension $p$, hence applicable in our high dimensional setup. In view of the results (3.5), (3.6) and Theorem 3.1, we obtain

$$\sup_{\beta\in\Theta}|M_n^\star(\beta)-M_n(\beta)| \le \sup_{\beta\in\Theta}|M_n^\star(\beta)-EM_n^\star(\beta)|+\sup_{\beta\in\Theta}|\tilde{M}_n(\beta)-E\tilde{M}_n(\beta)|+\sup_{\beta\in\Theta}|\tilde{M}_n(\beta)-M_n(\beta)|$$
$$= O_p\left(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{\log 2p}{n}}\right)+O(h). \tag{3.7}$$

The last claim in the above bounds follows since by Theorem 3.1, the second term on the right hand side of (3.7) decreases faster than the first term. This approximation plays a pivotal role in the analysis carried out in the sequel.

## 4. Main results and consequences

In this section we shall provide statistical error bounds for the $W\ell_1$-$CQ$ estimator. The following lemma is crucial for obtaining our error bounds. Let

$$v_n(\beta)=n^{-1}\sum_{i=1}^n \rho(y_i,x_i,\beta), \qquad g_n(\beta):=Ev_n(\beta), \quad \beta\in\mathbb{R}^p.$$

We shall some times write $\rho_{Li}^*(\beta)$ for $\rho_L^*(y_i,w_i,\beta,h)$. Similar comment applies to $\rho$ and $\rho_L$. We have

**Lemma 4.1.** *For the measurement error model (2.1) and (2.2), we have,*

$$g_n(\hat\beta)-g_n(\beta^0)+\lambda_n\|d\circ\hat\beta\|_1 \le \lambda_n\|d\circ\beta^0\|_1+|M_n^\star(\hat\beta)-M_n(\hat\beta)|+|M_n(\hat\beta)-EM_n(\hat\beta)|, \quad \forall\,\beta\in\mathbb{R}^p. \tag{4.1}$$

This inequality is obtained by subtracting $M_n(\hat\beta)-EM_n(\hat\beta)$ on both sides of the inequality

$$n^{-1}\sum_{i=1}^n \rho_{Li}^\star(\hat\beta)+\lambda_n\|d\circ\hat\beta\|_1 \le n^{-1}\sum_{i=1}^n \rho_{Li}^\star(\beta^0)+\lambda_n\|d\circ\beta^0\|_1,$$

and then rearranging terms and using the triangle inequality.

The technique adopted to provide the desired error bounds is to first establish results for any $\beta$ chosen in a small neighborhood of $\beta^0$. Later, using the convexity of $\rho_\tau(\beta)$ and the inequality (3.7), we show that the estimator $\hat\beta$ indeed eventually lies in this neighborhood, with probability tending to 1. Let $\kappa_n := \max_{1\le i\le n,1\le j\le p}|x_{ij}|$, and define

$$\mathcal{B}(\alpha_n)=\left\{\beta\in\mathbb{R}^p : \|\beta-\beta^0\|_1\le\alpha_n\right\},$$

where $\alpha_n$ is a sequence of positive numbers decreasing to 0 and satisfying $\alpha_n=o(\kappa_n^{-1})$. The last piece of this jigsaw is the following lemma which shall provide a lower bound for the first term on the left hand side of (4.1). Let $X:=(x_1,x_2,\ldots,x_n)^T$ denote the $n\times p$ design matrix, and

$$\Gamma := \mathrm{diag}\{f_1(0),\ldots,f_n(0)\}, \tag{4.2}$$

**Lemma 4.2.** *Suppose the model (2.1) and assumption* (A1) *hold. Then, there exists a constant* $0<c_a<1$, *such that for any* $\beta\in\mathcal{B}(\alpha_n)$, *and for all large n,*

$$g_n(\beta)-g_n(\beta^0)\ge c_a(\beta-\beta^0)'\frac{X'\Gamma X}{n}(\beta-\beta^0)=c_a n^{-1}\|\Gamma^{1/2}X(\beta-\beta^0)\|_2^2\ge 0.$$

**Proof.** Set $a_i=|x_i'(\beta-\beta^0)|$, $1\le i\le n$. Then for $\beta\in\mathcal{B}(\alpha_n)$, $a_i\le\kappa_n\|\beta-\beta^0\|_1\le\kappa_n\alpha_n\to 0$. Then proceed as in [8, p. 341], to obtain the desired result. □

Fan et al. [8] prove the above result for the oracle estimator, i.e., with the additional information regarding the locations of zero and non zero components of $\beta$ and $\beta^0$. However, as noted above, this result can be obtained without oracle information by defining the set $\mathcal{B}(\alpha_n)$ as above.

To proceed further, we need the following Compatibility condition on the unobserved design matrix $X$. This condition is often used in high dimensional analysis (see, e.g., [5,16]). The closely related 'Restricted eigenvalue condition' is used by Belloni and Chernozhukov [2] to provide consistency in estimation for quantile regression, when the covariates are completely observed.

**Definition 4.1.** We say the *Compatibility condition* is met for the set $S$, if for some $\phi > 0$, and constants $0 < b < 1$, $c_0 > 0$, and for all $\delta \in \mathbb{R}^p$ satisfying $\|\delta_{S^c}\|_1 \le c_0\|\delta_S\|_1$,

$$\|\delta_S\|_1^2 \le \frac{bs}{n\phi^2}\delta X' \Gamma X \delta. \tag{4.3}$$

In our setup the constant $c_0$ can be explicitly computed as $c_0 = (2c_{\max} + c_{\min})/c_{\min}$. Hence, if we are using the $\ell_1$ penalty, where the weights $d_j = 1$, for all $1 \le j \le p$, then $c_0 = 3$.

We also need the following rate conditions on various underlying entities.

$$\begin{aligned} &(\mathrm{i})\, \kappa_n \to \infty, \qquad \gamma_{\max} \to \infty, \qquad \lambda_n \to 0, \qquad \alpha_n \to 0, \\ &\kappa_n \gamma_{\max} s_n^{3/2} h^{-2}\sqrt{\frac{\log 2p}{n}} = o(\lambda_n), \qquad \alpha_n = o(\kappa_n^{-1}), \\ &(\mathrm{ii})\, \kappa_n h = o(\lambda_n), \qquad (\mathrm{iii})\, \frac{\lambda_n s_n \kappa_n}{\phi^2} \to 0. \end{aligned} \tag{4.4}$$

For the bounded designs where $\kappa_n = O(1)$, we shall need the following rate conditions.

$$\begin{aligned} &(\mathrm{i})\, \gamma_{\max} \to \infty, \qquad \lambda_n \to 0, \qquad \alpha_n \to 0, \qquad \gamma_{\max} s_n^{3/2} h^{-2}\sqrt{\frac{\log 2p}{n}} = o(\lambda_n), \\ &(\mathrm{ii})\, h = o(\lambda_n), \qquad (\mathrm{iii})\, \frac{\lambda_n s_n}{\phi^2} \to 0. \end{aligned} \tag{4.5}$$

In the above conditions, $\phi$ is the constant defined in (4.3). As is the case with kernel density estimators, the rate of decrease of the smoothing parameter $h$ has to be appropriately balanced. It has to decrease slowly enough so as to satisfy (4.4)(i) and fast enough to satisfy (4.4)(ii) in the case of unbounded design. Similarly, in the case of bounded design, these rate constraints have to balance between (4.5)(i) and (4.5)(ii). Note that the rate of decrease of $\lambda_n$ is significantly slower than in the case of non measurement error in the covariates. This is mainly due to the presence of the additional noise in the covariates and the smoothing parameter $h$.

We now state the main result providing error bounds for the proposed estimator.

**Theorem 4.1.** *For the measurement error model* (2.1) *and* (2.2), *let* $\hat\beta$ *be as in* (2.5) *and* $c_{\min}$, $c_{\max}$ *be as in* (2.6) *with* $c_m := c_{\min} + c_{\max}$. *Assume* (A1), (A2), (A3) *and* (A4), *along with the Compatibility condition* (4.3) *hold. Also assume that either the rate conditions* (4.4) *or* (4.5) *holds. Then the following inequality holds with probability at least* 1-*o*( 1).

$$3c_a n^{-1}\|\Gamma^{1/2}X(\hat\beta - \beta^0)\|_2^2 + 2\lambda_n c_{\min}\|\hat\beta - \beta^0\|_1 \le \frac{4\lambda_n^2 c_m^2 s_n}{\phi^2} + O\Big(\gamma_{\max} s_n^{3/2} h^{-2}\sqrt{\frac{\log 2p}{n}}\Big) + O(h). \tag{4.6}$$

The bound (4.6) clearly implies that under the conditions of Theorem 4.1, $\|\hat\beta - \beta^0\|_1 \to_p 0$ and $n^{-1}\|\Gamma^{1/2}X(\hat\beta - \beta^0)\|_2^2 \to_p 0$. In other words, the sequence of estimators $W\ell_1$-*CQ* is consistent for $\beta^0$ in $\ell_1$-norm and in the weighted $L_2$-norm $n^{-1}\|\Gamma^{1/2}X(\hat\beta - \beta)\|_2^2$. Secondly, the weights $d_j$, $1 \le j \le p$, do not play a critical role for the consistency of the estimator, i.e. as long as the condition (2.6) is satisfied, the above error bounds will provide the required consistency. Hence, if no prior information is available, one may choose $d_j \equiv 1$, in which case the estimator $\hat\beta$ becomes the $\ell_1$-*CQ* estimator, which is $\ell_1$-consistent. This fact shall also be useful for consistent model selection, which requires carefully chosen weights corresponding to the non-zero and zero indices of the parameter. This shall be further elaborated on after we provide a result on the sparsity properties of the proposed estimator.

The conclusion (4.6) of Theorem 4.1 bounding the $l_1$ and weighted $l_2$ error in estimation resembles in form to that of Theorem 6.2 of Bühlmann and Van de Geer [5] obtained for $\ell_1$-penalized mean regression estimator when there is no measurement error in the covariates and when the errors in regression model are assumed to be sub-Gaussian. In comparison, the above result (4.6) is established here in the presence of heavy tail measurement error in covariates and when the regression model errors are independent heteroscedastic not necessarily sub-Gaussian.

**A sparsity property.** Next, we investigate a model selection property of $W\ell_1$-*CQ*. It is well known that the model selection properties are linked to the first order optimality conditions, also known as the KKT conditions. These conditions are necessary and sufficient when the objective function is convex. However, as noted earlier, the loss function of our estimator is non-convex. In this case, KKT conditions are necessary but not sufficient. We exploit the necessity of KKT conditions to show that the estimator $W\ell_1$-*CQ* identifies all zero components successfully with asymptotic probability 1, provided the weights $d_j$ are chosen appropriately. More precisely, let

$$\alpha_n = \frac{2\lambda_n c_m^2 s}{c_{\min}\phi^2} + \frac{1}{\lambda_n}O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + \frac{1}{\lambda_n}O(h) \to 0. \tag{4.7}$$

Then in addition to the conditions of Theorem 4.1, we assume there exists a $0 < \delta < 1/2$ such that,

$$\text{(i) } \kappa_n \le n^{\delta}, \quad \text{and} \quad \text{(ii) } \log p = o(n^{\delta}). \tag{4.8}$$

Furthermore along with the conditions (2.6) on the weight vector $d$, we assume that $d_j$ for $j \in S^c$, diverge at a fast enough rate, i.e., $d_{\min}^{S^c} = \min\{d_j, j \in S^c\}$ satisfies the following rate conditions.

$$\begin{aligned}&\text{(i) } \max\{\alpha_n, \kappa_n^3\alpha_n^2\} = o(\lambda_n d_{\min}^{S^c}), \qquad \text{(ii) } \kappa_n h = o(\lambda_n d_{\min}^{S^c}),\\ &\text{(iii) } \max\left\{\gamma_{\max} s n^{\delta} h^{-2}\sqrt{\frac{2\log p}{n}},\ \alpha_n \gamma_{\max} n^{\delta} h^{-3} s\sqrt{\frac{2\log p}{n}}\right\} = o(\lambda_n d_{\min}^{S^c}).\end{aligned} \tag{4.9}$$

**Theorem 4.2.** *For the measurement error model (2.1) and (2.2), assume the conditions of Theorem 4.1 hold. In addition assume that (2.6), (4.8) and (4.9) hold. Then*

$$P\left(\hat{\beta}_j = 0, \quad \forall j \in S^c\right) \to 1. \tag{4.10}$$

This theorem provides the model selection consistency of the proposed $W\ell_1$-$CQ$ estimator $\hat{\beta}$, under suitable choice of the weight vector $d = (d_1, \dots, d_p)^T$. Note that setting the weights $d_j \equiv 1$, i.e., the un-weighted $\ell_1$-penalty does not satisfy the rate assumptions (4.9) and hence $\ell_1$-$CQ$ cannot be guaranteed to be model selection consistent as opposed to $W\ell_1$-$CQ$.

As the reader may observe, the conditions required for Theorem 4.2 are only rate conditions on the model parameters, in addition to mild distributional assumptions. These conditions are weaker than those required for model identification in the work of Belloni and Chernozhukov [2]. The reason for this being that our result states that zero components are correctly identified, as opposed to Belloni and Chernozhukov [2], who state a stronger result regarding identifiability of the non-zero components. Thus, we are able to state a weaker result under weaker conditions. We are unable to provide any result regarding the identification of non-zero components due to the non-convexity of the loss function.

We note that the above results will continue to hold for all other measurement error distributions for which the identity (3.5) and the probability bound (A.19) given below hold.

**Adaptive choice of the weight vector** $d$. For model selection we have seen that the choice of the weight vector $d$ plays a critical role for the proposed estimator to guarantee that the zero components are identified correctly. Zou [26] proposed the idea of adaptively choosing these weights by setting $d_j = |\hat{\beta}_j^{ini}|^{-\eta},\ 1 \le j \le p,\ \eta > 0$, where $\hat{\beta}_j^{ini}$ is any initial estimate of $\beta_j^0$ satisfying $\max_{1\le j\le p} |\hat{\beta}_j^{ini} - \beta_j^0| = O_p(\alpha_n)$, with $\alpha_n \to 0$. We use the same approach to select the weight vector $d$ in our setup.

First, we use the $l_1$-$CQ$ estimator, i.e., the proposed estimator with the ordinary $\ell_1$-penalty ($d_j = 1, \forall 1 \le j \le p$), this gives the initial estimate $\hat{\beta}^{ini}$. Theorem 4.1 provides the consistency of this estimate. In particular, under conditions of Theorem 4.1 we obtain with high probability, $\|\hat{\beta}^{ini} - \beta^0\|_1 \le \alpha_n \to 0$, where $\alpha_n$ is defined in (4.7). Here we place an additional assumption on the true parameter vector, namely, we assume that all non-zero components of $\beta^0$ are bounded above and below by a constant, i.e. $b_1 \le |\beta_j^0| \le b_2$. Thus, with high probability

$$|\hat{\beta}_j^{ini}| \le b_2 + \alpha_n, \quad \forall j \in S \qquad |\hat{\beta}_j^{ini}| \le \alpha_n \quad \forall j \in S^c. \tag{4.11}$$

Now, we set $d_j = (|\hat{\beta}_j^{ini}| + c_w)^{-\eta}$, where $c_w = \min_{1\le j\le p}(|\hat{\beta}_j^{ini}|;\ \hat{\beta}_j^{ini} \ne 0)$ is added to the initial estimates to avoid the problem of dividing by zero.

Keeping (4.11) in view, it is easy to verify that when $n$ is large enough, the above weight vector $d$ satisfies the required assumptions (2.6) and (4.9) for some constant $\eta$ chosen appropriately, with probability approaching to 1.

## 5. Simulation study

*Simulation setup*

In this section we numerically analyze the performance of the proposed estimators $\ell_1$-$CQ$ and $W\ell_1$-$CQ$. All computations were done in R, on an ordinary desktop machine with a five core (2.3 GHz) processor. We compare our proposed estimators with least squares based high dimensional procedures including Lasso and the bias corrected Lasso (Loh and Wainwright (2011)), the latter of which is specifically designed to handle sub-Gaussian measurement error in covariates.

While conducting our simulation study, we compute Lasso estimates using the package glmnet developed by Friedman et al. [9]. To compute $\ell_1$-$CQ$ estimates and the bias corrected Lasso, we use the projected gradient descent algorithm [1], which is a tool developed for optimizing penalized smooth loss functions in high dimensions. More precisely, with $\nabla L(\beta)$

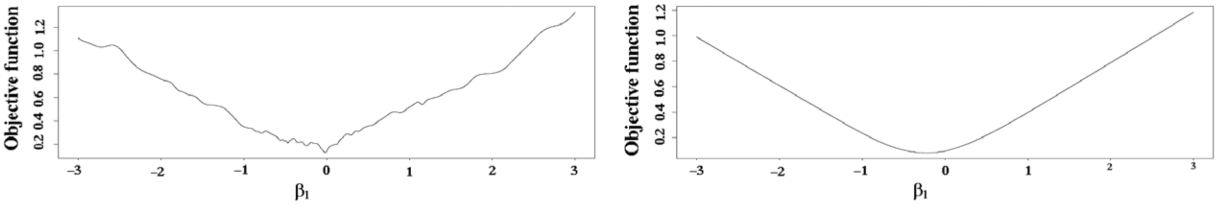

**Fig. 1.** $\frac{1}{n}\sum_{i=1}^{n}\rho_L^\star(\beta)+\lambda_n\|\beta\|_1$ evaluated around $\beta_1$ at $h=0.01$ (left) and 1.5 (right).

denoting the gradient of a loss function $L$, the method of projected descent iterates by the recursions, $\{\beta^r,\ r=0,1,2,\ldots\}$ as,

$$\beta^{r+1}=\arg\min_{\beta\in\Theta}\left\{L(\beta^r)+\nabla L(\beta^r)^T(\beta-\beta^r)+\frac{\delta}{2}\|\beta-\beta^r\|_2^2+\lambda_n\|\beta\|_1\right\}, \tag{5.1}$$

where $\delta>0$ is a stepsize parameter. These recursions can be computed rapidly in $O(p)$ time using the procedure suggested by Agarwal et al. [1] with the restriction of the parameter space to the $\ell_1$-ball $\Theta$ implemented by the procedure of Duchi et al. [7]. This procedure essentially involves two $\ell_2$ projections onto the $\ell_1$ ball $\Theta$.

The weighted version $W\ell_1$-*CQ* can be computed by the procedure described above with the following algorithm similar in spirit to that described by Zou [26]. The proof of this algorithm is straightforward and hence is omitted.

*Algorithm to compute* $W\ell_1$-*CQ by method of projected gradient descent*:

1. Define $w_i^\star=(w_{i1}/d_1,\ldots,w_{ip}/d_p)^T,\ \forall 1\le i\le n$. Also define $\Sigma^\star=\Big(\sigma_{ij}/d_id_j\Big),\ \forall 1\le i,j\le p$ where $\sigma_{ij}$ denote the components of $\Sigma$.
2. Optimize, using the methods of projected gradient descent and Duchi et al.,

$$\hat\beta^\star=\arg\min_{\beta\in\Theta}\Big\{\frac{1}{n}\sum_{i=1}^{n}\rho_L^\star(y_i,w_i^\star,\beta,h)+\lambda_n\|\beta\|_1\Big\}.$$

3. Evaluate $\hat\beta_j=\hat\beta_j^\star/d_j,\ \forall 1\le j\le p$.

*Tuning parameters*: The choice of the tuning parameters $\lambda_n$ and $h$ is still not a completely understood aspect of high dimensional data analysis. Typically in regularized estimation methods, either cross validation or *AIC*-*BIC* type selectors are used to select the tuning parameters. Zhang, Li and Tsai [24] provide theoretical justification for using *AIC*-*BIC* type criteria for several models. The cross validation method is often observed to result in over-fitting [21], furthermore it is considerably more time consuming. More recently, Lee, Noh and Park [12] have suggested a high dimensional BIC type criterion for quantile regression methods. Motivated by their results, one way to proceed is to select $\lambda_n$, $h$ as minimizers of the function

$$\mathrm{HBIC}(\lambda_n,h)=\log\Big(\sum_{i=1}^{n}\rho_L^\star(y_i-w_i^T\hat\beta_{\lambda,h})\Big)+|S_{\lambda,h}|\frac{(\log n)}{2n}C_n,$$

where $|S_\lambda|$ is the number of nonzero coefficients in the estimated parameter vector $\hat\beta_{\lambda,h}$ and $C_n$ is a diverging sequence of positive numbers. However, since $\rho_L^\star$ can take negative values, we shall use

$$\mathrm{e}^{\mathrm{HBIC}}(\lambda_n,h)=\Big(\sum_{i=1}^{n}\rho_L^\star(y_i-w_i^T\hat\beta_{\lambda,h})\Big)e^{|S_{\lambda,h}|\frac{(\log n)}{2n}C_n}$$

to obtain $\lambda_n$ and $h$. The exponential transformation removes the problem of negativity of $\rho_L*$ and also maintains monotonicity. Furthermore, we choose $C_n=O(\log(\log p))$ which is empirically found to work well in this simulation.

In defining $\mathrm{HBIC}(\lambda_n,h)$, we used the corrected quantile loss function instead of the check function as defined by Lee et al. [12]. Although this makes intuitive sense as the corrected quantile loss function is approximated by the check function, however a rigorous theoretical argument justifying its use is missing. This criterion is empirically found to perform well in our setup.

### *Computational issues*

A computational challenge of the proposed estimator is the non-convexity of the loss function $l_n^*$. The objective function $l_n^*$ becomes increasingly volatile around the true parameter as it approaches the check function at values of $h$ very close to zero. This behavior is illustrated in Fig. 1, which plots the loss function against $\beta_1$, keeping all other parameters fixed at the true values. This plot is generated for the first of the 100 simulated models.

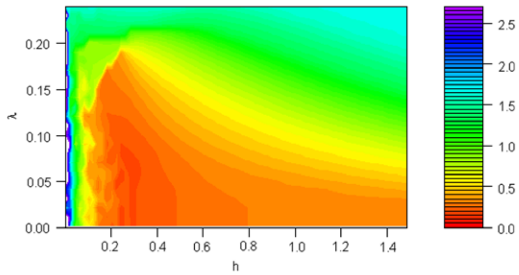

**Fig. 2.** Colored contours of $\|\hat{\beta} - \beta\|_1$ on $h$ vs. $\lambda_n$ for $\ell_1$-*CQ*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

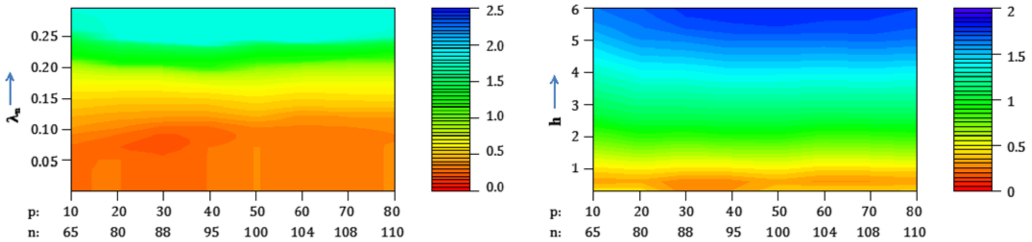

**Fig. 3.** Colored contours of $\|\hat{\beta} - \beta\|_1$ on $p$ vs. $\lambda$(left) and $h$.(right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The loss function exhibits several local optimums at smaller values of $h$. On the other hand at relatively higher values of $h$, the loss function appears to be convex shaped around the true parameter and appears to have a unique minimum at the true parameter value. Two computational consequences of this behavior are that, first, at $h$ close to zero, any optimization procedure becomes excessively time consuming. To avoid this unpleasant feature, we avoid values of $h$ close to zero. It was numerically observed that by doing so, we are able to maintain the accuracy of the estimator along with a reasonable computation time. Second, at values of $h$ outside a neighborhood of zero the optimizations are robust against the starting points used in optimizations. In particular, in all 100 simulation repetitions the starting point for optimization was chosen randomly from a Gaussian distribution. This behavior of the objective function is also represented visually in the contour plot in Fig. 2. Here the $\ell_1$ estimation error $\|\hat{\beta} - \beta\|_1$ is plotted as colored contours with the error increasing from red to blue regions. Values of $h$ are represented on the $x$-axis and values of $\lambda_n$ on the $y$-axis. From this plot it is apparent that the lowest error is given in regions concentrated around the relatively smaller values of $h$ and $\lambda_n$ except when $h$ is in a small neighborhood of zero.

With regards to computational time, one optimization at $h = 0.01$ takes $\approx$ 16 s as opposed to $\approx$ 2 s at $h = 0.5$, at $(p = 40, n = 120)$. This can also be viewed in comparison to corrected Lasso which takes $\approx$ 0.5 second to complete one optimization.

*Simulation setup and results*

For this simulation study, data is generated from the measurement error model (2.1) and (2.2) under several choices of the underlying parameters and distributional assumptions. The unobserved design variables $\{x_{ij},\ 1 \le i \le n, 1 \le j \le p\}$ are chosen as i.i.d. r.v.'s from a $\mathcal{N}(0, 1)$ distribution. The measurement errors $\{u_i,\ 1 \le i \le n\}$ are i.i.d. $L_p(0, \Sigma)$, with $\Sigma = \sigma^2 \times I$, where $I$ is the $p \times p$ identity matrix. The model errors $\varepsilon_i,\ 1 \le i \le n$ are independent realizations of Normal, Cauchy or mean centered Pareto r.v.'s.

We begin by numerically verifying the result of Theorem 4.1. Observe that this theorem can be viewed as describing the scaling behavior of the error $\|\hat{\beta} - \beta\|_1$. In order to visualize this, we perform simulations by varying the dimension of the parameter vector $p$ and the sample size $n$ while holding all other parameters fixed, in particular the number of non zero components $s = 5$, the covariance matrix of the Laplace distribution for the covariate errors is taken to be $0.2I_{p\times p}$ and the model errors are Gaussian with variance 0.2. All of Fig. 3 describe the error of $\ell_1$-*CQ* estimate. The behavior of the error of estimation for the $W\ell_1$-*CQ* is observed to be similar and thus the corresponding plots are omitted for the sake of brevity.

Fig. 3 is a contour plot generated at $h = 0.4$(left) and $\lambda = 0.07$(right). This plot describes the $\ell_1$ error, $\|\hat{\beta} - \beta\|_1$ as a spectrum of colors with red being the least and violet being the maximum. The $y$-axis plots different values of the tuning parameter $\lambda$ and the $x$-axis marks the varying dimension $p$ of the model. Note that for a given model dimension the corresponding sample size is rescaled to maintain the ratio $(n/\log 2p)$ to be constant. This rescaling is done in accordance with the result of Theorem 4.1 and as predicted by the theorem, holding all other parameters fixed for each value of $\lambda$(left) and $h$(right) the error level stays roughly constant (the colors align) across the chosen values of $p$.

**Table 1**
Simulation results at $p = 40$ for Normal, Cauchy and Pareto model errors.

| $n$ | $\ell_1$-*CQ* | | | *C-Lasso* | | | *Lasso* | | |
|---|---|---|---|---|---|---|---|---|---|
| | MEE | MINZ | MIZ | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| $s = 8,\ u_i \sim L_p(\sigma^2 = 0.2),\ \varepsilon_i \sim \mathcal{N}(0, 0.2),\ \tau = 0.5$ | | | | | | | | | |
| 20 | **0.62** | 4 | 5 | 0.63 | 4 | 6 | 0.78 | 4 | 9 |
| | (0.21) | (1.72) | (2.33) | (0.22) | (1.76) | (2.39) | (0.27) | (1.98) | (3.97) |
| 60 | **0.27** | 1 | 9 | 0.28 | 1 | 9 | 0.39 | 1 | 12 |
| | (0.061) | (1.07) | (2.91) | (0.065) | (1.04) | (2.90) | (0.071) | (1.12) | (4.77) |
| 120 | **0.18** | 1 | 12 | 0.20 | 1 | 11 | 0.34 | 1 | 14 |
| | (0.038) | (0.86) | (2.86) | (0.038) | (0.88) | (2.93) | (0.04) | (0.91) | (4.83) |
| $s = 8,\ u_i \sim L_p(\sigma^2 = 0.2),\ \varepsilon_i \sim Cauchy(scale = 0.1),\ \tau = 0.5$ | | | | | | | | | |
| 50 | **0.95** | 5 | 7 | 1.61 | 7 | 9 | 5.21 | 2 | 20.5 |
| | (0.21) | (1.21) | (2.41) | (0.42) | (1.10) | (2.25) | (15.89) | (1.59) | (5.45) |
| 150 | **0.60** | 4 | 9 | 1.54 | 7 | 10 | 5.02 | 2 | 25 |
| | (0.14) | (1.17) | (3.83) | (0.47) | (1.15) | (2.44) | (19.12) | (1.69) | (6.28) |
| 300 | **0.35** | 2 | 11 | 1.56 | 7 | 13 | 4.77 | 2 | 24 |
| | (0.11) | (1.01) | (3.75) | (0.49) | (1.21) | (2.45) | (18.80) | (1.53) | (5.37) |
| $s = 5,\ u_i \sim L_p(\sigma^2 = 0.2),\ \varepsilon_i \sim mean\ centered\ Pareto,\ \tau = 0.75$ | | | | | | | | | |
| 100 | **0.66** | 2 | 8 | 1.01 | 3 | 7 | 1.15 | 3 | 13 |
| | (0.15) | (1.05) | (2.93) | (0.31) | (1.32) | (2.99) | (3.05 | (1.06) | (7.20) |
| 200 | **0.50** | 1 | 8 | 0.84 | 2 | 7.5 | 0.97 | 2 | 15 |
| | (0.10) | (0.94) | (2.56) | (0.32) | 1.31 | (2.64) | (2.49) | (1.78) | (6.70) |
| 300 | **0.38** | 1 | 9 | 0.71 | 2 | 8 | 0.90 | 1 | 13 |
| | (0.07) | (0.84) | (2.89) | (0.27) | (1.16) | (2.86) | (2.41) | (0.94) | 5.64 |

We now proceed to a more detailed numerical comparison of the proposed estimates with Lasso and corrected Lasso estimates. For any given method, we summarize the results obtained by 100 repetitions. For every repetition, each non zero component of the parameter vector $\beta$ is generated from a $\mathcal{N}(0, 1)$ distribution normalized by the $\ell_2$ norm of the generated vector. The dimensions of this vector are chosen to be $p = 40,\ 300,\ 500$. The dimension of the non zero components are set to $s = 5, 8, 10$. As mentioned earlier, the model errors $\varepsilon_i,\ 1 \le i \le n$ are generated from Gaussian, Cauchy or mean centered Pareto r.v.'s. Note that the Pareto distribution can be heavily skewed. For performance comparison we report the following criteria.

MEE : (Median estimation error), median over 100 repetitions of the estimation error $\|\hat{\beta} - \beta^0\|_2$.

MIZ : (Median incorrect number of zeros), median over 100 repetitions of the number of incorrectly identified zero components of the parameter vector.

MINZ : (Median incorrect number of non zeros), median over 100 repetitions of the number of incorrectly identified non zero components of the parameter vector.

In all tables, the standard errors of the corresponding criteria are reported in the parentheses.

Tables 1 and 2 provide results of the simulation study comparing the $\ell_1$-*CQ*, the corrected Lasso (C-Lasso) and Lasso estimators. It is clear from these results that under heavy tailed or skewed model errors (Cauchy and mean centered Pareto), the $\ell_1$-*CQ* estimator significantly outperforms the other two procedures in all three comparison criteria. Furthermore, the standard errors of $\ell_1$-*CQ* are significantly smaller than those of the other two. Under Gaussian model errors, $\ell_1$-*CQ* is comparable (slightly better) in performance to the C-Lasso, while both of these procedures outperform Lasso. Consistency in terms of the estimation error and identifying the correct support of $\beta^0$ is clearly visible as $n \to \infty$. As expected, the $\ell_1$-*CQ* estimator does not provide consistency in identifying the zero components correctly. However, it is still much better in comparison to Lasso. This behavior of Lasso under measurement error has also been observed by Sorensen et al. [19], i.e., measurement error tends to induce over-fitting by naive estimators such as Lasso.

Another instance where the proposed estimators outperform the corrected Lasso and Lasso is the heteroscedastic setup. To illustrate this, we generated independent model errors $\varepsilon_i$ from $\mathcal{N}(0, \sigma_i^2)$, where $\sigma_i^2$ is chosen uniformly from the interval $(0.1, 9)$, for each $i = 1, \ldots, n$. The dimension $p$ is increased to 500 for this case and the results are provided in Table 3.

We next investigate the $W\ell_1$-*CQ* estimator for consistent identification of the zero components, in addition to consistent estimation. Tables 4 and 5 provide simulation results for $W\ell_1$-*CQ* for $p = 40$ and $p = 300$. The weights $d_j$ are chosen as described at the end of Section 4, where the exponent $\eta$ is chosen by using the selection criteria $e^{\text{HBIC}}$. Comparing $W\ell_1$-*CQ*, $\ell_1$-*CQ*, C-Lasso and Lasso, the first and most immediate conclusion is the efficacy of the proposed estimators under heavy tailed or skewed model errors. The $W\ell_1$-*CQ* estimator consistently and significantly outperforms all other procedures in terms of model identification under all chosen distributional and parameter settings.

**Table 2**
Simulation results at $p = 300$ for Normal, Cauchy model errors.

| $n$ | $\ell_1$-*CQ* | | | *C-Lasso* | | | *Lasso* | | |
|---|---|---|---|---|---|---|---|---|---|
| | MEE | MINZ | MIZ | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| $s = 10,\ u_i \sim L_p(\sigma^2 = 0.1),\ \varepsilon_i \sim \mathcal{N}(0, 0.2),\ \tau = 0.5$ | | | | | | | | | |
| 100 | **0.26** | 2 | 21 | 0.27 | 2 | 21.5 | 0.35 | 2 | 37 |
| | (0.04) | (1.20) | (5.20) | (0.05) | (1.19) | (4.70) | (0.07) | (1.09) | (7.43) |
| 200 | **0.17** | 1 | 30 | **0.17** | 1 | 29.5 | 0.23 | 1 | 55 |
| | (0.028) | (0.89) | (6.86) | (0.029) | (0.90) | (6.77) | (0.031) | (0.91) | (11.70) |
| 300 | **0.11** | 1 | 33 | 0.14 | 1 | 32 | 0.18 | 1 | 79 |
| | (0.023) | (0.79) | (5.74) | (0.023) | (0.76) | (5.68) | (0.025) | (0.82) | (16.16) |
| $s = 10,\ u_i \sim L_p(\sigma^2 = 0.1),\ \varepsilon_i \sim Cauchy(scale = 0.1),\ \tau = 0.5$ | | | | | | | | | |
| 100 | **0.44** | 3 | 23 | 0.92 | 7 | 20 | 4.39 | 7 | 27 |
| | (0.10) | (1.57) | (10.36) | (0.41) | (2.35) | (7.21) | (10.70) | (1.96) | (30.16) |
| 200 | **0.30** | 2 | 21.5 | 0.87 | 6 | 22.5 | 3.27 | 6 | 30 |
| | (0.07) | (1.34) | (12.63) | (0.44) | (2.42) | (8.18) | (8.27) | (1.84) | (51.34) |
| 300 | **0.21** | 2 | 13 | 0.85 | 6 | 22 | 3.47 | 6 | 28 |
| | (0.05) | (1.23) | (12.86) | (0.50) | (2.59) | (8.99) | (9.10) | (1.91) | (47.21) |

**Table 3**
Simulation results at $p = 500$ under heteroscedasticity.

| $s = 10,\ u_i \sim L_p(\sigma^2 = 0.1),\ \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2),\ \sigma_i^2 \sim Uniform(0.1, 9),\ \tau = 0.5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $\ell_1$-*CQ* | | | *C-Lasso* | | | *Lasso* | | |
| | MEE | MINZ | MIZ | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| 100 | **0.86** | 7 | 25.5 | 0.99 | 7 | 25 | 1.73 | 7 | 73 |
| | (0.14) | (1.13) | (4.33) | (0.15) | (0.99) | (5.05) | (0.27) | (1.87) | (16.15) |
| 200 | **0.66** | 5 | 30 | 0.78 | 5 | 30 | 1.67 | 5 | 87 |
| | (0.12) | (1.32) | (5.98) | (0.12) | (1.20) | (5.56) | (0.25) | (1.19) | (21.25) |
| 300 | **0.54** | 4 | 34 | 0.69 | 5 | 33 | 1.51 | 3 | 83 |
| | (0.09) | (1.37) | (6.56) | (0.10) | (1.17) | (5.71) | (0.18) | (1.26) | (19.56) |

**Table 4**
$W\ell_1$-*CQ* at $p = 40$ for Normal and Cauchy model errors.

| $\varepsilon_i \sim \mathcal{N}(0, 0.2)$, | | | | $\varepsilon_i \sim Cauchy(scale = 0.1)$, | | | |
|---|---|---|---|---|---|---|---|
| $n$ | MEE | MINZ | MIZ | $n$ | MEE | MINZ | MIZ |
| 20 | 0.70 | 1 | 5 | 50 | 0.98 | 1.5 | 5 |
| | (0.25) | (0.66) | (1.38) | | (0.23) | (0.63) | (2.20) |
| 60 | 0.28 | 0 | 5 | 150 | 0.64 | 1 | **5** |
| | (0.06) | (0.64) | (1.55) | | (0.14) | (0.61) | (2.27) |
| 120 | 0.21 | 0 | 4 | 300 | 0.41 | 1 | 4 |
| | (0.038) | (0.61) | (1.27) | | (0.11) | (0.70) | (2.17) |

**Table 5**
$W\ell_1$-*CQ* at $p = 300$ for Normal and Cauchy model errors.

| $\varepsilon_i \sim \mathcal{N}(0, 0.2)$, | | | | $\varepsilon_i \sim Cauchy(scale = 0.1)$, | | | |
|---|---|---|---|---|---|---|---|
| $n$ | MEE | MINZ | MIZ | $n$ | MEE | MINZ | MIZ |
| 100 | 0.28 | 1 | 7 | 100 | 0.40 | 1 | 8 |
| | (0.04) | (0.69) | (1.75) | | (0.11) | (0.72) | (2.91) |
| 200 | 0.23 | 0 | 6 | 200 | 0.29 | 1 | 6 |
| | (0.02) | (0.62) | (0.98) | | (0.09) | (0.68) | (1.85) |
| 300 | 0.18 | 0 | 5 | 300 | 0.24 | 0 | 5 |
| | (0.02) | (0.41) | (0.97) | | (0.06) | (0.46) | (1.56) |

Finally, to see how robust the $\ell_1$-*CQ* estimator is to the misspecification of the measurement error distribution, we compared its performance with C-Lasso when this error distribution is Gaussian. The results reported in Table 6 show a comparable performance with C-lasso performing only marginally better.

**Table 6**
$\ell_1$-CQ at $p = 300$ with misspecified covariate error distribution.

| $s = 10, u_{ij} \sim \mathcal{N}(0, 0.3),\ \varepsilon_i \sim \mathcal{N}(0, 0.3)$ | | | | | | |
|---|---|---|---|---|---|---|
| $n$ | $\ell_1$-CQ | | | C-Lasso | | |
| | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| 100 | 0.31 | 1 | 18 | **0.28** | 1 | 16 |
| | (0.06) | (0.89) | (3.64) | (0.07) | (0.83) | (3.12) |
| 200 | 0.22 | 1 | 21 | **0.20** | 1 | 18 |
| | (0.04) | (0.71) | (4.10) | (0.03) | (0.71) | (4.44) |
| 300 | 0.20 | 0 | 24 | **0.17** | 0 | 19 |
| | (0.02) | (0.63) | (4.14) | (0.02) | (0.61) | 4.88 |

## Acknowledgments

Authors wish to thank Akshita Chawla and Jiwoong Kim for several useful suggestions regarding computational issues and Po-Ling Loh for providing code to compute the corrected Lasso estimates. The authors also wish to thank a diligent referee for constructive comments that helped to improve the presentation, especially that of the simulation section.

## Appendix

As briefly stated at the beginning of Section 4, the technique used to prove Theorem 4.1 is to use the convexity of the quantile function $\rho(\beta)$ and the weighted $\ell_1$-penalty, along with the approximation of $\rho_L^\star(\beta)$ to $\rho(\beta)$. Some of the steps of the proof are similar to those adopted by Bülmann and Van der Geer [5, Chapter 4].

Let $t = \alpha_n/(\alpha_n + \|\hat{\beta} - \beta^0\|_1)$, and set $\tilde{\beta} = t\hat{\beta} + (1-t)\beta^0$. Note that $\tilde{\beta} \in \mathcal{B}(\alpha_n)$. Moreover,

$$\tilde{\beta} \in \mathcal{B}(c\alpha_n) \quad \text{implies } \hat{\beta} \in \mathcal{B}(c\alpha_n/(1-c)),\ \forall\, 0 < c < 1. \tag{A.1}$$

This fact will be used in the sequel.

Next, by the convexity of $g_n(\beta)$ and $\|d \circ \beta\|_1$, and the inequality (4.1), we obtain

$$g_n(\tilde{\beta}) - g_n(\beta^0) + \lambda_n\|d \circ \tilde{\beta}\|_1 \le \lambda_n\|d \circ \beta^0\|_1 + \sup_{\beta\in\Theta}|M_n^\star(\beta) - M_n(\beta)| + \sup_{\beta\in\mathcal{B}(\alpha_n)}|M_n(\beta) - EM_n(\beta)|. \tag{A.2}$$

We begin by providing error bounds for $\tilde{\beta}$, which shall easily extend to $\hat{\beta}$. By (3.7) and (4.4), the second term in the RHS of (A.2) is $o_p(1)$. The following lemma provides the rate of decrease of the last term.

**Lemma A.1.** *For the measurement error model* (2.1), (2.2), *assume that* (A1) *and* (A2) *hold. Then*

$$\sup_{\beta\in\mathcal{B}(\alpha_n)}|M_n(\beta) - EM_n(\beta)| = O_p\Big(\alpha_n\sqrt{\frac{2\log 2p}{n}}\Big). \tag{A.3}$$

The proof of Theorem 3.1 and Lemma A.1 are provided after the proof of Theorem 4.1. Consider the following events,

(i) $\Omega_1 =$ the event that the bounds (3.3) and (3.4) hold,
(ii) $\Omega_2 =$ the event that the bound (A.3) holds.

Then by Theorem 3.1 and Lemma A.1, $P(\Omega_1 \cap \Omega_2) \ge 1 - o(1)$, and on $\Omega_1 \cap \Omega_2$,

$$\sup_{\beta\in\Theta}|M_n^\star(\beta) - M_n(\beta)| + \sup_{\beta\in\mathcal{B}(\alpha_n)}|M_n(\beta) - EM_n(\beta)| \tag{A.4}$$

$$= O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + O(h). \tag{A.5}$$

This follows since $\alpha_n \to 0$, and hence the second terms on the LHS of (A.4) converges to 0 faster than the first term.

In the sequel, all arguments shall be restricted to the set $\Omega_1 \cap \Omega_2$. Recall that $\tilde{\beta} \in \mathcal{B}(\alpha_n)$. From (A.2) and (A.4) we now readily obtain that with probability at least $1 - o(1)$,

$$g_n(\tilde{\beta}) - g_n(\beta^0) + \lambda_n\|d \circ \tilde{\beta}\|_1 \le \lambda_n\|d \circ \beta^0\|_1 + O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + O(h). \tag{A.6}$$

By Lemma 4.2 we obtain $g_n(\tilde{\beta}) - g_n(\beta^0) \geq 0$. Thus, the triangle inequality $\|d \circ \tilde{\beta}\|_1 \geq \|d \circ \beta^0\|_1 - \|\big(d \circ (\tilde{\beta} - \beta^0)\big)_S\|_1 + \|(d \circ \tilde{\beta})_{S^c}\|_1$ applied to (A.6) yields

$$\begin{aligned} \lambda_n \|(d \circ \tilde{\beta})_{S^c}\|_1 &\leq \lambda_n \|\big(d \circ (\tilde{\beta} - \beta^0)\big)_S\|_1 + O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h) \\ &\leq c_{\max} \lambda_n \|\tilde{\beta}_S - \beta^0_S\|_1 + O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \end{aligned} \tag{A.7}$$

Now we consider two cases, **Case (i)** where,

$$\frac{\lambda_n}{2} c_{\min} \|\tilde{\beta} - \beta^0\|_1 \geq O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h), \tag{A.8}$$

or **Case (ii)** where,

$$\frac{\lambda_n}{2} c_{\min} \|\tilde{\beta} - \beta^0\|_1 \leq O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \tag{A.9}$$

**Proof of Theorem 4.1.** First, we prove error bounds for $\tilde{\beta}$, which, in view of (A.1), shall be a precursor to obtaining error bounds for $\hat{\beta}$.

Suppose **Case (i)** (A.8) holds. The fact $\|\tilde{\beta} - \beta^0\|_1 = \|(\tilde{\beta} - \beta^0)_S\|_1 + \|\tilde{\beta}_{S^c}\|_1$ and (A.7) imply

$$\lambda_n c_{\min} \|\tilde{\beta}_{S^c}\|_1 \leq \lambda_n \|(d \circ \tilde{\beta})_{S^c}\|_1 \leq \lambda_n c_{\max} \|\tilde{\beta}_S - \beta^0_S\|_1 + \frac{\lambda_n}{2} c_{\min} \|\tilde{\beta} - \beta^0\|_1,$$

which implies $\|\tilde{\beta}_{S^c}\|_1 \leq c_0 \|\tilde{\beta}_S - \beta^0_S\|_1$, where $c_0 = (2c_{\max} + c_{\min})/c_{\min}$. Thus the Compatibility condition (4.3) is satisfied for $\delta = \tilde{\beta} - \beta^0$. Now Lemma 4.2, the triangle inequality $\|d \circ \tilde{\beta}\|_1 \geq \|d \circ \beta^0\|_1 - \|\big(d \circ (\tilde{\beta} - \beta^0)\big)_S\|_1 + \|(d \circ \tilde{\beta})_{S^c}\|_1$, (A.6), and (A.8) together yield

$$\frac{2c_a}{n} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + \lambda_n c_{\min} \|\tilde{\beta}_{S^c}\|_1 \leq \lambda_n c_{\min} c_0 \|\tilde{\beta}_S - \beta^0_S\|_1. \tag{A.10}$$

Recall $c_m = c_{\min} + c_{\max}$ and consider

$$\begin{aligned} &4c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + 2\lambda_n c_{\min} \|\tilde{\beta} - \beta^0\|_1 \\ &\quad = 4c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + 2\lambda_n c_{\min} \|\tilde{\beta}_S - \beta^0_S\|_1 + 2\lambda_n c_{\min} \|\tilde{\beta}^c_S\|_1 \\ &\quad \leq 2\lambda_n c_{\min} c_0 \|\tilde{\beta}_S - \beta^0_S\|_1 + 2\lambda_n c_{\min} \|\tilde{\beta}_S - \beta^0_S\|_1 = 4\lambda_n c_m \|\tilde{\beta}_S - \beta^0_S\|_1 \\ &\quad \leq \frac{4\lambda_n c_m \sqrt{s c_a}}{\sqrt{n}\phi} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2 \\ &\quad \leq \frac{c_a}{n} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + \frac{4\lambda_n^2 c_m^2 s}{\phi^2}. \end{aligned}$$

Here the first inequality follows from (A.10), the second from the Compatibility condition in (4.3), and the third using the identity $4uv \leq u^2 + 4v^2$. Thus

$$3c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + 2\lambda_n c_{\min} \|\tilde{\beta} - \beta^0\|_1 \leq \frac{4\lambda_n^2 c_m^2 s}{\phi^2}. \tag{A.11}$$

Now we consider **Case (ii)**. From (A.6) we obtain,

$$\begin{aligned} c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + \lambda_n c_{\min} \|\tilde{\beta}_{S^c}\|_1 &\leq \lambda_n c_{\max} \|\tilde{\beta}_S - \beta^0_S\|_1 + O_p\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O_p(h), \\ &= O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \end{aligned}$$

In particular,

$$c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 = O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h).$$

Thus under **Case (ii)**, we have

$$3c_a n^{-1} \|\Gamma^{1/2} X (\tilde{\beta} - \beta^0)\|_2^2 + 2\lambda_n c_{\min} \|\tilde{\beta} - \beta^0\|_1 = O\Big(\gamma_{\max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \tag{A.12}$$

Hence from (A.11) and (A.12) for any $\tilde{\beta} \in \mathcal{B}(\alpha_n)$ we have, with probability $1 - o(1)$,

$$3c_a n^{-1}\|\Gamma^{1/2}X(\tilde{\beta} - \beta^0)\|_2^2 + 2\lambda_n c_{\min}\|\tilde{\beta} - \beta^0\|_1 \le \frac{4\lambda_n^2 c_m^2 s}{\phi^2} + O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + O(h).$$

Thereby choosing $\lambda_n$ according to the rate assumptions (4.4), with probability $1 - o(1)$,

$$\|\tilde{\beta} - \beta^0\|_1 \le \frac{1}{2}\Big[4\lambda_n c_m^2 s/c_{\min}\phi + \frac{1}{\lambda_n}O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + \frac{1}{\lambda_n}O(h)\Big] \to 0.$$

Thus choosing,

$$\alpha_n \ge \Big(4\lambda_n c_m^2 s/c_{\min}\phi + \frac{1}{\lambda_n}O\Big(\gamma_{\max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + \frac{1}{\lambda_n}O(h)\Big) \to 0,$$

we have by the rate assumptions (4.4), $\kappa_n\alpha_n \to 0$, and hence

$$\|\tilde{\beta} - \beta^0\|_1 \le \frac{\alpha_n}{2}.$$

This along with the construction of $\tilde{\beta}$ and (A.1) applied with $c = 1/2$ implies that $\|\hat{\beta} - \beta^0\|_1 \le \alpha_n$, and thus, $\hat{\beta} \in \mathcal{B}(\alpha_n)$. Repeating the above argument with $\tilde{\beta}$ replaced by $\hat{\beta}$ now gives the desired error bound (4.6), thereby completing the proof of Theorem 4.1. □

For a later use we state the fact about fact $\hat{\beta} \in \mathcal{B}(\alpha_n)$ as follows. Note that the above $\alpha_n$ satisfies (4.7). Thus with $\alpha_n$ as in (4.7), with probability $1 - o(1)$,

$$\hat{\beta} \in \mathcal{B}(\alpha_n). \tag{A.13}$$

We now proceed to the proofs of Theorem 3.1 and Lemma A.1. For this purpose we first state some facts about the first two summands in the loss function $\rho_L^*$ of (2.4). These facts are consequences of the properties of normal kernel density. Let, for $s,\ y \in \mathbb{R}$,

$$\begin{aligned}
l(s,y) &= (y-s)(\tau-1) + (y-s)H\Big(\frac{y-s}{h}\Big), \qquad (A.14)\\
l'(s,y) &= \tau - 1 + H\Big(\frac{y-s}{h}\Big) + \frac{y-s}{h}K\Big(\frac{y-s}{h}\Big),\\
l''(s,y) &= \frac{2}{h}K\Big(\frac{y-s}{h}\Big) + \frac{y-s}{h^2}K'\Big(\frac{y-s}{h}\Big),\\
l'''(s,y) &= \frac{3}{h^2}K'\Big(\frac{y-s}{h}\Big) + \frac{y-s}{h^3}K''\left(\frac{y-s}{h}\right).
\end{aligned}$$

By the MVT and the definition of standard normal density we readily obtain that uniformly in $y \in \mathbb{R}$, the following facts hold for all $s_1, s_2 \in \mathbb{R}$. For some constant $C > 0$,

$$\begin{aligned}
&\text{(i)}\ |l(s_1,y) - l(s_2,y)| \le C|s_1 - s_2|, \qquad \text{(ii)}\ |l'(s_1,y) - l'(s_2,y)| \le \frac{C}{h^1}|s_1 - s_2|, \qquad (A.15)\\
&\text{(iii)}\ |l''(s_1,y) - l''(s_2,y)| \le \frac{C}{h^2}|s_1 - s_2|, \qquad \text{(iv)}\ |l'''(s_1,y) - l'''(s_2,y)| \le \frac{C}{h^3}|s_1 - s_2|.
\end{aligned}$$

The above conditions are the reason for choosing the kernel function $K(\cdot)$ as the p.d.f. of a standard normal r.v. We require that the first three derivatives of $K(\cdot)$ to be bounded uniformly. In the following we denote $l_i(\beta) = l(w_i'\beta, y_i)$ and $l_i'(\beta),\ l_i''(\beta)$ and $l_i'''(\beta)$ are defined similarly.

**Proof of Theorem 3.1.** First note that for $\beta \in \Theta$, we have $\|\beta - \beta^0\|_1 \le 2b_0\sqrt{s}$, by the definition of $\Theta$ and the assumption $\|\beta^0\|_1 \le b_0\sqrt{s}$. Note that, $\rho_{Li}^\star(\beta) = l_i(\beta) - \frac{\sigma_\beta^2}{2}l_i''(\beta)$. Now,

$$\begin{aligned}
M_n^\star(\beta) - EM_n^\star(\beta) &= \frac{1}{n}\sum_{i=1}^n\Big(l_i(\beta) - l_i(\beta^0) - E\big(l_i(\beta) - l_i(\beta^0)\big)\Big)\\
&\quad - \frac{1}{n}\frac{\sigma_\beta^2}{2}\sum_{i=1}^n\Big(l_i''(\beta) - l_i''(\beta^0) - E\big(l_i''(\beta) - l_i''(\beta^0)\big)\Big)\\
&\quad + \frac{1}{n}\frac{\sigma_{\beta^0}^2 - \sigma_\beta^2}{2}\sum_{i=1}^n\Big(l_i''(\beta^0) - E\big(l_i''(\beta^0)\big)\Big)\\
&\le I - \frac{\sigma_\beta^2}{2}\mathrm{II} + \frac{\sigma_{\beta^0}^2 - \sigma_\beta^2}{2}\mathrm{III}, \quad \text{say.}
\end{aligned} \tag{A.16}$$

We shall show that

$$\text{(a)}\ \sup_{\beta\in\Theta}|\mathrm{I}| = O_p\Big(\sqrt{s}\sqrt{\frac{2\log 2p}{n}}\Big), \qquad \text{(b)}\ \sup_{\beta\in\Theta}|\mathrm{II}| = O_p\Big(\frac{s^{1/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big),$$

$$\text{(c)}\ |\mathrm{III}| = O_p\Big(\frac{s}{h}\sqrt{\frac{\log 2p}{n}}\Big).$$

Observe that for $\beta \in \Theta, \sigma_\beta^2 = \beta^T\Sigma\beta \le \gamma_{\max} b_0 s$ and $|\sigma_{\beta^0}^2 - \sigma_\beta^2| \le 2b_0\gamma_{\max}s$. This fact along with bounds for I, II and III shall imply the desired result.

Define the empirical process $\mathcal{G}_n(\beta) := \frac{1}{n}\sum_{i=1}^n\big(l_i(\beta) - El_i(\beta)\big)$ and

$$Z_n := \sup_{\beta\in\Theta}|\mathcal{G}_n(\beta) - \mathcal{G}_n(\beta^0)|.$$

With $\sigma_u$ as in assumption (A3), let $c_u = 1.4\sigma_u$. On the event $A = \big\{\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n u_{ij}^2 \le c_u\big\}$,

$$\frac{1}{n}\sum_{i=1}^n w_{ij}^2 \le \frac{2}{n}\sum_{i=1}^n (x_{ij}^2 + u_{ij}^2) \le 2(c_x + c_u). \tag{A.17}$$

This bound and the Lipschitz condition (A.15))(i) allow us to apply Lemma 14.20 and Theorem 14.2 as done in Example 14.2 of Bühlmann and Van de Geer [5, p. 503], to yield

$$E\Big(Z_nI_A\Big) \le 32c_1b_0(c_x+c_u)\sqrt{s}\sqrt{\frac{2\log 2p}{n}},$$

$$P\Big(Z_nI_A \ge 8c_1b_0(c_x+c_u)\sqrt{s}\Big(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2t}{n}}\Big)\Big) \le \exp\big(-t\big),$$

for any $t > 0$. Choose $t = \log 2p$ in the latter bound to obtain

$$P\Big(Z_nI_A \ge O\Big(\sqrt{s}\sqrt{\frac{2\log 2p}{n}}\Big)\Big) = o(1).$$

Now to remove the truncation of $Z_n$ on the set $A$, observe that (A.15)(i) also implies that,

$$|l_i(\beta) - l_i(\beta^0)| \le C(\kappa_n + \max_{ij}|u_{ij}|)\|\beta - \beta^0\|_1 \le 2Cb_0(\kappa_n + \max_{ij}|u_{ij}|)\sqrt{s},$$

since for any $\beta \in \Theta$, we have $\|\beta - \beta^0\| \le 2b_0\sqrt{s}$. Hence,

$$Z_n \le Z_nI_A + c\sqrt{s}(\kappa_n + \max_{ij}|u_{ij}|)I_{A^c} + c\sqrt{s}E\Big((\kappa_n + \max_{ij}|u_{ij}|)I_{A^c}\Big). \tag{A.18}$$

Now recall that for each $1 \le j \le p$, $\{u_{ij}, 1 \le i \le n\}$ are i.i.d. $L(0, \sigma_{jj}^2)$ r.v.'s. Hence, $2\sum_{i=1}^n|u_{ij}|/\sigma_{jj} \sim \chi_{2n}^2$, where $\chi_{2n}^2$ denotes a chi square r.v. with $2n$ degrees of freedom. Now use the probability bounds for chi-square distributions given by Johnstone [11] to obtain

$$\begin{aligned} P\Big(A^c\Big) &\le \sum_{j=1}^p P\Big(\Big\{\frac{1}{n}\sum_{i=1}^n|u_{ij}|\Big\}^2 \ge c_u^2\Big) \le \sum_{j=1}^p P\Big(2\frac{1}{n}\sum_{i=1}^n\frac{|u_{ij}|}{\sigma_{jj}} \ge 2.8\Big) \\ &= \sum_{j=1}^p P\Big(\chi_{2n}^2 \ge n2.8\Big) \le \sum_{j=1}^p P\Big(|\chi_{2n}^2 - 2n| \ge 2n(0.4)\Big) \\ &\le \sum_{j=1}^p \exp\Big(\frac{-3n}{100}\Big) \le \exp\Big(\frac{-3n}{100} + \log p\Big). \end{aligned} \tag{A.19}$$

Next, use the fact that $|u_{ij}| \sim Exp(\sigma_{jj})$, to obtain

$$E\Big((\max_{ij}|u_{ij}|)^2\Big) \le \sum_{i,j}E(u_{ij}^2) \le npc_u.$$

Thus, using this bound, (A.19), and the Cauchy–Schwarz inequality, we obtain

$$\text{(a)}\ P\Big((\max_{ij}|u_{ij}|)I_{A^c} > n^{-k}\Big) \le n^kE\Big((\max_{ij}|u_{ij}|)I_{A^c}\Big) \le n^k\sqrt{npc_u\exp\Big(\frac{-3n}{100} + \log p\Big)},$$

(b) $E\Big((\max_{ij}|u_{ij}|)I_{A^c}\Big) \le \sqrt{E\Big((\max_{ij}|u_{ij}|)^2\Big)P(A^c)} \le \sqrt{npc_u \exp\Big(\frac{-3n}{100}+\log p\Big)}$.

The exponential bound in (a) implies that the probability of the event in (a) tends to zero, for any $k>0$. This in turn implies that the second summand in (A.18) satisfies

$$\sqrt{s}(\kappa_n + \max_{ij}|u_{ij}|)I_{A^c} = o_p(n^{-k}), \quad \forall\, k>0.$$

Similarly, the bound in (b) implies that the third summand in the bound of (A.18) decreases to zero at an exponential rate. Thus, with probability at least $1-o(1)$, the remainder two summands in the bound in (A.18) decrease to zero, in probability, faster than $Z_nI_A$. Hence,

$$\sup_{\beta\in\Theta}|I| = Z_n = O_p\Big(\sqrt{s}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{A.20}$$

We can similarly obtain a bound for term II of (A.16). An outline is given below. Define the empirical process $\tilde{\mathcal{g}}_n(\beta) := \frac{1}{n}\sum_{i=1}^n\big(l_i''(\beta)-El_i''(\beta)\big)$. Let

$$\tilde{Z}_n := \sup_{\beta\in\Theta}|\tilde{\mathcal{g}}_n(\beta)-\tilde{\mathcal{g}}_n(\beta^0)|.$$

Proceeding as earlier, (A.15)(ii) along with the bound (A.17) allow us to apply Lemma 14.20 and Theorem 14.2 of Bühlmann and Van de Geer [5, p. 503], which yields

$$E\Big(\tilde{Z}_nI_A\Big) \le 32c_3b_0(c_x+c_u)\frac{\sqrt{s}}{h^2}\sqrt{\frac{2\log 2p}{n}}, \quad \text{and}$$

$$P\bigg(Z_nI_A \ge 8c_3b_0(c_x+c_u)\frac{\sqrt{s}}{h^2}\Big(4\sqrt{\frac{2\log 2p}{n}}+\sqrt{\frac{2t}{n}}\Big)\bigg) \le \exp\big(-t\big), \quad \forall\, t>0.$$

Choose $t=\log 2p$ in this bound to obtain

$$P\bigg(\tilde{Z}_nI_A \ge O\Big(\frac{\sqrt{s}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big)\bigg) = o(1).$$

Get rid of the truncation on the set $A$ as done for $I$, to obtain

$$\sup_{\beta\in\Theta}|\text{II}| = \tilde{Z}_n = O_p\Big(\frac{\sqrt{s}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{A.21}$$

Lastly, consider the term III in (A.16). Observe that $|l_i''(\beta^0)| \le ch^{-1}$, for $c<\infty$. Then Lemma 14.11 of Bühlmann and Van de Geer [5] yields

$$P\Big(\frac{1}{n}\Big|\sum_{i=1}^n\big(l_i''(\beta^0)-El_i''(\beta^0)\big)\Big| \ge t\Big) \le 2\exp\Big(-\frac{nt^2h^2}{2c^2}\Big).$$

Choosing $t=h^{-1}\sqrt{\frac{\log 2p}{n}}$, we obtain

$$|\text{III}| = \frac{1}{n}\Big|\sum_{i=1}^n\big(l_i''(\beta^0)-El_i''(\beta^0)\big)\Big| = O_p\Big(h^{-1}\sqrt{\frac{\log 2p}{n}}\Big). \tag{A.22}$$

Now use (A.20)–(A.22) in (A.16), and the fact that the rate of decrease of (A.21) is the slowest, to conclude (3.3) of Theorem 3.1.

The proof of (3.4) similar. This completes the proof of Theorem 3.1. □

**Proof of Lemma A.1.** Define $\rho(s,y_i)=\rho_\tau(y_i-s)$. Then observe that it satisfies the following Lipschitz condition,

$$|\rho(s_1,y_i)-\rho(s_2,y_2)| \le \max\{\tau,1-\tau\}|s_1-s_2|.$$

Then proceed as in the proof of (3.3) of Theorem 3.1 to obtain the desired bound. □

**Proof of Theorem 4.2.** Let $\alpha_n$ be as defined in (4.7). By (A.13), $\hat{\beta}\in\mathcal{B}(\alpha_n)$, with probability $1-o(1)$. Thus, with arbitrarily large probability, for all large $n$, $\hat{\beta}$ is in the interior of $\Theta$ and not on its boundary. Hence, KKT conditions are necessary for this optimum. We prove the desired result via contradiction. For any $j \in S^c$, let if possible $\hat{\beta}_j \neq 0$. Then by the necessity of KKT conditions,

$$\frac{d}{d\beta_j}\Big(n^{-1}\sum_{i=1}^{n}\rho^{\star}_{Li}(\hat{\beta})\Big) = \lambda_n d_j \text{sign}(\hat{\beta}_j). \tag{A.23}$$

Recall (A.14). The first derivatives of $\rho^{\star}_{Li}(\beta)$ and $\rho_{Li}(\beta)$ w.r.t. $\beta_j$ are

$$\rho^{\star'}_{Li,j}(\beta) := \frac{d}{d\beta_j}\rho^{\star}_{Li}(\beta) = -w_{ij}l'_i(\beta) + \frac{\sigma^2_{\beta}}{2}\big[l'''_i(\beta)\big] - w_{ij}\sum_{k=1}^{n}\sigma_{kj}\beta_j\big[l''_i(\beta)\big], \tag{A.24}$$

$$\rho'_{Li,j}(\beta) = \frac{d}{d\beta_j}\rho_{Li}(\beta) = -x_{ij}\Big[\tau - 1 + H\Big(\frac{\varepsilon_{i\beta}}{h}\Big) + \frac{\varepsilon_{i\beta}}{h}K\Big(\frac{\varepsilon_{i\beta}}{h}\Big)\Big].$$

Let $\rho'_{i,j}(\beta) := -x_{ij}\big[\tau - I\{y_i - x'_i\beta \le 0\}\big]$, $\psi^{*'}_{Li,j}(\beta) := E\rho^{\star'}_{Li,j}(\beta)$, $\psi'_{Li,j}(\beta) := E\rho'_{Li,j}(\beta)$, $\psi'_{i,j}(\beta) := E\rho'_{i,j}(\beta)$, and

$$S^{\star}_{n,j}(\beta) = n^{-1}\sum_{i=1}^{n}\big(\rho^{\star'}_{Li,j}(\beta) - \rho^{\star'}_{Li,j}(\beta^0) - \psi^{*'}_{Li,j}(\beta) + \psi^{*'}_{Li,j}(\beta^0)\big), \qquad T^{\star}_{\mathcal{B},j} = \sup_{\beta\in\mathcal{B}(\alpha_n)}\big|S^{\star}_{n,j}(\beta)\big|.$$

The fact that $\psi'_{i,j}(\beta^0) = E(\rho'_{ij}(\beta^0)) = 0$, and triangle inequality yield

$$\begin{aligned}\left|n^{-1}\sum_{i=1}^{n}\rho^{\star'}_{Li,j}(\hat{\beta})\right| &\le n^{-1}\Big|\sum_{i=1}^{n}\big(\rho^{\star'}_{Li,j}(\beta^0) - \psi^{*'}_{Li,j}(\beta^0)\big)\Big| + n^{-1}\Big|\sum_{i=1}^{n}\big(\psi^{\star'}_{Li,j}(\hat{\beta}) - \psi'_{i,j}(\hat{\beta})\big)\Big| \\ &\qquad + n^{-1}\Big|\sum_{i=1}^{n}\big(\psi'_{i,j}(\hat{\beta}) - \psi'_{i,j}(\beta^0)\big)\Big| + T^{*}_{\mathcal{B},j}. \\ &= J_1 + J_2 + J_3 + J_4.\end{aligned}$$

We will show the relations (A.25)–(A.25) hold for all $j \in S^c$ simultaneously with probability at least $1 - o(1)$.

$$J_1 := n^{-1}\Big|\sum_{i=1}^{n}\big(\rho^{\star'}_{Li,j}(\beta^0) - \psi^{\star'}_{Li,j}(\beta^0)\big)\Big| = o(d_j\lambda_n), \tag{A.25}$$

$$J_2 := \sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi^{\star'}_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big)\Big| = O(\kappa_n h) = o(d_j\lambda_n), \tag{A.26}$$

$$J_3 := \sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi'_{i,j}(\beta) - \psi'_{i,j}(\beta^0)\big)\Big| = o(d_j\lambda_n), \tag{A.27}$$

$$J_4 := T^{*}_{\mathcal{B},j} = o(d_j\lambda_n). \tag{A.28}$$

Lemma A.2 proves (A.26) and (A.27) and Lemma A.3 proves (A.25) and (A.28). Finally, combining (A.25)–(A.28), we obtain that for $n$ large, with probability $1 - o(1)$,

$$\left|\frac{d}{d\beta_j}\Big(n^{-1}\sum_{i=1}^{n}\rho^{\star}_{Li}(\hat{\beta})\Big)\right| < d_j\lambda_n, \quad \forall j \in S^c.$$

This contradicts the optimality condition (A.23), and also completes the proof of Theorem 4.2. □

**Lemma A.2.** *Under the conditions of Theorem 4.2 we have,*

$$\max_{1\le j\le p,\beta\in\mathbb{R}^p} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi^{\star'}_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big)\Big| = O(\kappa_n h) = o(\lambda_n d^{S^c}_{\min}), \tag{A.29}$$

$$\max_{j\in S^c}\sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi'_{i,j}(\beta) - \psi'_{i,j}(\beta^0)\big)\Big| = o(\lambda_n d^{S^c}_{\min}). \tag{A.30}$$

**Proof.** Let $a_i = x'_i(\beta - \beta^0)$, and $\varepsilon_{i\beta} := y_i - x'_i\beta = \varepsilon_i - a_i$. By Theorem 2 of WSZ,

$$\sum_{i=1}^{n}\big(\psi^{\star'}_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big) = \sum_{i=1}^{n}\big(\psi'_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big).$$

But

$$\begin{aligned}\psi'_{Li,j}(\beta) - \psi'_{ij}(\beta) &= -x_{ij}E\Big(H\Big(\frac{\varepsilon_{i\beta}}{h}\Big) - \mathbf{1}\{\varepsilon_{i\beta} > 0\} + \frac{\varepsilon_{i\beta}}{h}K\Big(\frac{\varepsilon_{i\beta}}{h}\Big)\Big)\\ &= -x_{ij}E\Big(H\Big(-\Big|\frac{\varepsilon_{i\beta}}{h}\Big|\Big) + \frac{\varepsilon_{i\beta}}{h}K\Big(\frac{\varepsilon_{i\beta}}{h}\Big)\Big).\end{aligned} \tag{A.31}$$

Now,

$$\begin{aligned}E\Big(H\Big(-\Big|\frac{\varepsilon_{i\beta}}{h}\Big|\Big)\Big) &= \int_{x=-\infty}^{\infty} H\Big(-\Big|\frac{x-a_i}{h}\Big|\Big)f_i(x)dx = h\int_{t=-\infty}^{\infty} H(-|t|)f_i(ht+a_i)dt\\ &= h\int_{t=-\infty}^{0} H(t)f_i(ht+a_i)dt + h\int_{t=0}^{\infty} H(-t)f_i(ht+a_i)dt = O(h),\end{aligned}$$

uniformly in $1 \le i \le n, \beta \in \mathbb{R}^p$, because by assumption (A1), $\sup_{1\le i\le n, x\in\mathbb{R}} f_i(x) < \infty$. Similarly, one verifies that $\max_{1\le i\le n,\beta\in\mathbb{R}^p} |h^{-1}E\varepsilon_{i\beta}K(\varepsilon_{i\beta}/h)| = O(h)$. Hence from (A.31) we obtain,

$$\sup_{1\le i\le n, 1\le j\le p, \beta\in\mathbb{R}^p} \big|\psi'_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big| = O(\kappa_n h) = o(\lambda_n d_{\min}^{S^c}). \tag{A.32}$$

The last equality follows from the rate assumptions (4.9). This bound and assumption (A2) completes the proof of (A.29).

Next, we show (A.30). By assumption (A1),

$$\big(\psi'_{i,j}(\beta) - \rho'_{i,j}(\beta^0)\big) = -x_{ij}\Big[F_i\big(x_i'(\beta-\beta^0)\big) - F_i(0)\Big] = -x_{ij}f_i(0)x_i^T(\beta-\beta^0) - x_{ij}\tilde{I}_i, \tag{A.33}$$

where $\tilde{I}_i = F_i(x_i'(\beta-\beta^0)) - F_i(0) - f_i(0)x_i^T(\beta-\beta^0)$. Now for any $j \in S^c$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}f_i(0)x_i'(\beta-\beta^0)\Big| \le \Big\|\frac{1}{n}x_{ij}f_i(0)x_i'\Big\|_\infty \|\beta-\beta^0\|_1 = O(\alpha_n) = o(\lambda_n d_{\min}^{S^c}), \tag{A.34}$$

this follows since $f_i(0)$ and $n^{-1}\sum_{i=1}^n x_{ij}^2$ are bounded by a constant for all $1 \le i \le n$ and $1 \le j \le p$. Also, from assumption (A1) we obtain,

$$\begin{aligned}\max_{j\in S^c}\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\tilde{I}_i\Big| &\le \frac{\kappa_n}{n}\sum_{i=1}^{n}\tilde{I}_i \le C_2\frac{\kappa_n}{n}\sum_{i=1}^{n}\big(x_i'(\beta-\beta^0)\big)^2.\\ &\le C\kappa_n^3\|\beta-\beta^0\|_1^2 = O\big(\kappa_n^3\alpha_n^2\big) = o(\lambda_n d_{\min}^{S^c}).\end{aligned} \tag{A.35}$$

Now use (A.33)–(A.35) to obtain (A.30), thereby completing the proof of the lemma. □

**Lemma A.3.** *Under the conditions of Theorem 4.2,*

$$\max_{j\in S^c}\frac{1}{n}\Big|\sum_{i=1}^{n}\big(\rho^{\star\prime}_{Li,j}(\beta^0) - \psi^{\star\prime}_{Li,j}(\beta^0)\big)\Big| = o_p(\lambda_n d_{\min}^{S^c}) \tag{A.36}$$

$$\max_{j\in S^c} T^*_{\mathcal{B},j} = o_p(\lambda_n d_{\min}^{S^c}). \tag{A.37}$$

**Proof.** The structure of this proof is similar to the proof of Theorem 3.1. In the following proof $c > 0$ shall denote a generic constant that may be different depending on the context. For any $0 < \delta$, define the event

$$\mathcal{A} = \Big\{\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n} u_{ij}^2 \le c_u, \quad \max_{1\le j\le p, 1\le i\le n}|u_{ij}| \le n^\delta\Big\}.$$

Use the fact $|u_{ij}| \sim Exp(\sigma_{jj})$ to obtain

$$P\Big(\max_{1\le i\le n, 1\le j\le p}|u_{ij}| > cn^\delta\Big) \le \sum_{j=1}^{p}\sum_{i=1}^{n} P\Big(|u_{ij}| \ge cn^\delta\Big) \le \frac{1}{\sigma_u}\exp\Big(-\frac{cn^\delta}{\sigma_u} + \log p + \log n\Big).$$

This bound and (A.19) together imply that

$$P(\mathcal{A}^c) \le \frac{1}{\sigma_u}\exp\Big(-\frac{cn^\delta}{\sigma_u} + \log p + \log n\Big) + \exp\Big(\frac{-3n}{100} + \log p\Big).$$

Now,

$$
\begin{aligned}
&n^{-1}\Big|\sum_{i=1}^{n}\big(\rho^{\star'}_{Li,j}(\beta^0)-\psi^{\star'}_{Li,j}(\beta^0)\big)\Big| \\
&\quad\le n^{-1}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)\Big|+\frac{\sigma^2_{\beta^0}}{2}n^{-1}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'''(\beta^0)\Big|+\Big|\sum_{i=1}^{n}\sigma_{ij}\beta_j^0\Big|n^{-1}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i''(\beta^0)\Big|,
\end{aligned}
$$

where $\gamma_i'(\beta^0):=l_i'(\beta^0)-El_i'(\beta^0),\ \gamma_i''(\beta^0):=l''(\beta^0)-El_i''(\beta^0)$ and $\gamma_i'''(\beta^0):=l_i'''(\beta^0)-El_i'''(\beta^0)$. Using $\kappa_n\le n^{\delta}$, we obtain

$$
\text{(i) } |w_{ij}\gamma_i'(\beta^0)I_{\mathcal{A}}|\le cn^{\delta},\qquad \text{(ii) } |w_{ij}\gamma_i''(\beta^0)I_{\mathcal{A}}|\le cn^{\delta}h^{-1},\quad \text{(iii) } |w_{ij}\gamma_i'''(\beta^0)I_{\mathcal{A}}|\le cn^{\delta}h^{-2}.
$$

Hence,

$$
\begin{aligned}
P\Big(\max_{1\le j\le p}\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)I_{\mathcal{A}}\Big|\ge t\Big) &\le \sum_{j=1}^{p} P\Big(\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}\gamma'(\beta^0)I_{\mathcal{A}}\Big|\ge t\Big)\\
&\le 2\exp\big[-cn^{-\delta}nt^2+\log p\big],
\end{aligned}
$$

where the last inequality follows from Lemma 14.11 of Bühlmann and Van de Geer [5]. Thus choosing $t=cn^{\delta}\sqrt{2\log 2p/n}$, for some constant $c>0$, we obtain

$$
\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}l_i'(\beta^0)I_{\mathcal{A}}\Big|=O_p\Big(n^{\delta}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{A.38}
$$

Now to remove the truncation on the set $\mathcal{A}$, observe that,

$$
\max_{1\le j\le p}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)\Big|\le\max_{1\le j\le p}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)I_{\mathcal{A}}\Big|+\max_{1\le j\le p}c(\kappa_n+\max_{i,j}|u_{ij}|)\mathbf{1}_{A^c}+c\max_{1\le j\le p}E\Big((\kappa_n+\max_{i,j}|u_{ij}|)\mathbf{1}_{A^c}\Big). \tag{A.39}
$$

Proceed as in the proof of Theorem 3.1 to show that the last two terms on the RHS converge to zero faster than the first term, in probability. Thus we obtain,

$$
\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)\Big|=O_p\Big(n^{\delta}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{A.40}
$$

A similar argument yields that

$$
\max_{1\le j\le p}\frac{1}{n}\Big|\sum_{i=1}^{n}\gamma_i''(\beta^0)\Big|=O_p\Big(h^{-1}\sqrt{\frac{2\log 2p}{n}}\Big),\qquad \max_{1\le j\le p}\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}\gamma'''(\beta^0)\Big|=O_p\Big(\frac{n^{\delta}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big).
$$

Recall that $\sigma^2_{\beta^0}\le b_0\gamma_{\max}s$, and $|\sum_{k=1}^{n}\sigma_{kj}\beta_j|\le b_0\sigma_u s$. Now combine these results with the rate assumptions (4.9) to obtain (A.36).

To prove claim (A.37), note that

$$
\begin{aligned}
S^{\star}_{n,j}(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} w_{ij}\big(l_i'(\beta)-l_i'(\beta^0)-El_i'(\beta)-El_i'(\beta^0)\big)+\frac{\sigma^2_{\beta}}{2}\frac{1}{n}\sum_{i=1}^{n} w_{ij}\big(l_i'''(\beta)-l_i'''(\beta^0)-El_i''(\beta)-El_i''(\beta^0)\big)\\
&\quad+\frac{\sigma^2_{\beta}-\sigma^2_{\beta_0}}{2}\frac{1}{n}\sum_{i=1}^{n} w_{ij}\big(l_i'''(\beta^0)-El_i'''(\beta^0)\big)-\sum_{k=1}^{n}\sigma_{kj}\beta_j\frac{1}{n}\sum_{i=1}^{n}\big(l_i''(\beta)-l_i''(\beta^0)-El_i''(\beta)-El_i''(\beta^0)\big)\\
&\quad-\sum_{k=1}^{n}\sigma_{kj}(\beta_j-\beta_j^0)\frac{1}{n}\sum_{i=1}^{n}\big(l_i''(\beta^0)-El_i''(\beta^0)\big)\\
&= -I+\frac{\sigma^2_{\beta}}{2}\mathrm{II}+\frac{\sigma^2_{\beta}-\sigma^2_{\beta_0}}{2}\mathrm{III}-\sum_{k=1}^{n}\sigma_{kj}\beta_j\,\mathrm{IV}-\sum_{k=1}^{n}\sigma_{kj}(\beta_j-\beta_j^0)\,\mathrm{V}.
\end{aligned}
$$

We begin with the term II, which turns out to have the slowest rate of convergence. Define the empirical process $\mathcal{G}'''_{n,j}(\beta):=\frac{1}{n}\sum_{i=1}^{n} w_{ij}\big(l_i'''(\beta)-El_i'''(\beta)\big)$ and let,

$$
Z'''_{n,j}=\sup_{\beta\in\mathcal{B}(\alpha_n)}\big|\mathcal{G}'''_{n,j}(\beta)-\mathcal{G}'''_{n,j}(\beta^0)\big|.
$$

Also, observe that from (A.14) and (A.15) we have,

$$\left|w_{ij}l'''(s_1, y_i) - w_{ij}l'''(s_2, y_i)\right| \le c(\kappa_n + \max_{i,j}|u_{ij}|)h^{-3}|s_1 - s_2|.$$

Then as in the above proof of Theorem 3.1, apply Lemma 14.2 and Theorem 14.2 of Bühlmann and Van de Geer [5] to obtain

$$P\left(\max_{1\le j\le p} Z'''_{n,j}I_{\mathcal{A}} \ge 8cb_0(c_x + c_u)n^{\delta}h^{-3}\alpha_n\left(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2t}{n}}\right)\right)$$
$$\le \sum_{j=1}^{p} P\left(Z'''_{n,j}I_{\mathcal{A}} \ge 8cb_0(c_x + c_u)n^{\delta}h^{-3}\alpha_n\left(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2t}{n}}\right)\right) \le \exp\bigl(-t + \log p\bigr).$$

Now choose $t = c\log p$, $c > 0$, so that the last term in the above expression is $o(1)$. Now removing the truncation on the set $\mathcal{A}$ as done in the proof of Theorem 3.1, we obtain

$$\max_{j\in S^c}\sup_{\beta\in\mathcal{B}(\alpha_n)}|\mathrm{II}| = \max_{1\le j\le p} Z'''_{n,j} = O_p\left(n^{\delta}h^{-3}\alpha_n\sqrt{\frac{2\log 2p}{n}}\right),$$

where the last equality follows by the rate assumption (4.9). A similar argument applied to the terms I and IV yields that

$$\max_{j\in S^c}\sup_{\beta\in\mathcal{B}(\alpha_n)}|\mathrm{I}| = O_p\left(n^{\delta}h^{-1}\alpha_n\sqrt{\frac{2\log 2p}{n}}\right), \qquad \max_{j\in S^c}\sup_{\beta\in\mathcal{B}(\alpha_n)}|\mathrm{IV}| = O_p\left(h^{-2}\alpha_n\sqrt{\frac{2\log 2p}{n}}\right).$$

An argument similar to the one used for proving (A.36) yields

$$\max_{j\in S^c}|\mathrm{III}| = O_p\left(\frac{n^{\delta}}{h^2}\sqrt{\frac{2\log 2p}{n}}\right), \qquad \max_{j\in S^c}|\mathrm{V}| = O_p\left(n^{\delta}h^{-1}\sqrt{\frac{2\log 2p}{n}}\right).$$

Now claim (A.37) follows from these bounds, the rate condition (iii) of (4.9), and the facts $|\sum_{k=1}^{n}\sigma_{kj}(\beta_j - \beta^0)| \le 2b_0\sigma_u\sqrt{s}$, and $\sigma_\beta^2 \le b_0\gamma_{\max}s$. This completes the proof of Lemma A.3.

## References

[1] A. Agarwal, S. Neghban, M.J. Wainwright, Fast global convergence of gradient methods for high dimensional statistical recovery, Ann. Statist. 40 (2012) 2452–2482.

[2] A. Belloni, V. Chernozhukov, $\ell_1$- Penalized quantile regression in high dimensional sparse models, Ann. Statist. 39 (2011) 82–130.

[3] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (2009) 1705–1732.

[4] M. Buchinsky, Changes in the US wage structure 1963–1987: Applications of quantile regression, Econometrica 62 (1994) 405–458.

[5] P. Bühlmann, S. van de Geer, Statistics for High Dimensional Data, Springer, New York, 2011.

[6] R.J. Carroll, D. Ruppert, L.A. Stefanski, C. Crainiceanu, Measurement Error in Nonlinear Models: A Modern Perspective, Chapman and Hall, New York, 2006.

[7] J. Duchi, S. Shalev-Shwartz, Y. Singer, T. Chandra, Efficient projections onto the $\ell_1$-ball for learning in high dimensions, in: International Conference on Machine Learning, ACM, New York, NY, 2008, pp. 272–279.

[8] J. Fan, Y. Fan, E. Barut, Adaptive robust variable selection, Ann. Statist. 42 (2014) 324–351.

[9] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.

[10] W.A. Fuller, Measurement Error Models, Wiley, New York, 1987.

[11] I.M. Johnstone, Chi-square oracle inequalities, in: State of the Art in Probability and Statistics (Leiden, 1999), in: IMS Lecture Notes Monogr. Ser., vol. 36, IMS, Beachwood, OH, 2001, pp. 399–418.

[12] R. Lee, H. Noh, B. Park, Model selection via Bayesian information criterion for quantile regression models, J. Amer. Statist. Assoc. 109 (2014) 216–229.

[13] P. Loh, M.J. Wainwright, High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity, Ann. Statist. 40 (2012) 1637–1664.

[14] H. McKenzie, C. Jerde, D. Visscher, E. Merrill, M. Lewis, Inferring in the presence of GPS measurement error, Environ. Ecol. Stat. 16 (2009) 531–546.

[15] E. Purdom, S. Holmes, Error distribution for gene expression data, Stat. Appl. Genet. Biol. 4 (2005) Article 16.

[16] G. Raskutti, M. Wainwright, B. Yu, Restricted eigenvalue properties for correlated Gaussian designs, J. Mach. Learn. Res. 99 (2010) 2241–2259.

[17] M. Rosenbaum, A.B. Tsybakov, Sparse recovery under matrix uncertainty, Ann. Statist. 38 (2010) 2620–2651.

[18] M. Rosenbaum, A.B. Tsybakov, Improved matrix uncertainty selector Technical Report, 2011, Available at: http://arxiv.org/abs/1112.4413.

[19] O. Sorensen, A. Frigessi, M. Thoresen, 2012. Measurement error in Lasso: impact and likelihood bias correction.
Available at: http://arxiv.org/pdf/1210.5378.pdf.

[20] L.A. Stefanski, R.J. Carroll, Deconvoluting kernel density estimators, Statistics 21 (1990) 169–184.

[21] H. Wang, R. Li, C.L. Tsai, Tuning parameter selectors for smoothly clipped absolute deviation method, Biometrika 3 (2007) 553–668.

[22] H. Wang, L.A. Stefanski, Z. Zhu, Corrected-loss estimation for quantile regression with covariate measurement errors, Biometrika 99 (2012) 405–421.

[23] L. Wang, Y. Wu, R. Li, Quantile regression for analyzing heterogeneity in ultra-high dimension, J. Amer. Statist. Assoc. 107 (2012) 214–222.

[24] Y. Zhang, R. Li, C.L. Tsai, Regularization parameter selections via generalized information criterion, J. Amer. Statist. Assoc. 105 (2010) 312–323.

[25] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.

[26] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.