# HIGH DIMENSIONAL LINEAR REGRESSION MODELS UNDER LONG MEMORY DEPENDENCE AND MEASUREMENT ERROR

By

Abhishek Kaul

A DISSERTATION

Submitted to<br>
Michigan State University<br>
in partial fulfillment of the requirements<br>
for the degree of

Statistics - Doctor of Philosophy

2015

# ABSTRACT

## HIGH DIMENSIONAL LINEAR REGRESSION MODELS UNDER LONG MEMORY DEPENDENCE AND MEASUREMENT ERROR

By

**Abhishek Kaul**

This dissertation consists of three chapters. The first chapter introduces the models under consideration and motivates problems of interest. A brief literature review is also provided in this chapter.

The second chapter investigates the properties of Lasso under long range dependent model errors. Lasso is a computationally efficient approach to model selection and estimation, and its properties are well studied when the regression errors are independent and identically distributed. We study the case, where the regression errors form a long memory moving average process. We establish a finite sample oracle inequality for the Lasso solution. We then show the asymptotic sign consistency in this setup. These results are established in the high dimensional setup ($p > n$) where $p$ can be increasing exponentially with $n$. Finally, we show the consistency, $n^{\frac{1}{2}-d}$-consistency of Lasso, along with the oracle property of adaptive Lasso, in the case where $p$ is fixed. Here $d$ is the memory parameter of the stationary error sequence. The performance of Lasso is also analysed in the present setup with a simulation study.

The third chapter proposes and investigates the properties of a penalized quantile based estimator for measurement error models. Standard formulations of prediction problems in high dimension regression models assume the availability of fully observed covariates and sub-Gaussian and homogenous model errors. This makes these methods inapplicable to measurement errors models where covariates are unobservable and observations are possi-

bly non sub-Gaussian and heterogeneous. We propose weighted penalized corrected quantile estimators for the regression parameter vector in linear regression models with additive measurement errors, where unobservable covariates are nonrandom. The proposed estimators forgo the need for the above mentioned model assumptions. We study these estimators in both the fixed dimension and high dimensional sparse setups, in the latter setup, the dimensionality can grow exponentially with the sample size. In the fixed dimensional setting we provide the oracle properties associated with the proposed estimators. In the high dimensional setting, 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 this chapter.

# ACKNOWLEDGMENTS

I wish to express my sincere thanks to Professor Hira L. Koul for giving me the opportunity to come here and learn and for guiding me ever since. I thank him for suggesting problems and inspiring and guiding me through them. I also wish to thank him and his family for making this country feel not too alien.

I would also like to thank Professor Ramamoorthi and Professor Maiti for stimulating discussions and for their help during my graduate study. Also, I thank Professor Ping-Shou Zhong and Professor Hao Wang for serving on my dissertation committee.

I thank my parents for their constant love and support which enabled me to complete this work.

Finally I would like to thank my friends Akshita Chawla, Chanakya Kaul, Jiwoong Kim and Ranjit Bawa for making this journey joyous.

This research has been supported in part by the NSF-DMS Grant 1205271: PI. Hira L. Koul, for which I am grateful.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# KEY TO SYMBOLS

In what follows, for any $z = (z_1, \cdots, z_p)^T \in \mathbb{R}^p$, $\|z\|_1 = \sum_{j=1}^p |z_j|$, $\|z\|_2^2 = \sum_{j=1}^p z_j^2$. For any two sequences of positive numbers $\{a_n, b_n\}$, $a_n = O(b_n)$, denotes that for all large $n$, $a_n \leq cb_n$, for some universal constant $c > 0$, which does not depend on any underlying parameters or the sample size $n$. All limits are taken as $n \to \infty$. For any index set $S \subseteq \{1, ..., p\}$, and for any vector $\delta \in \mathbb{R}^p$, denote $\delta_S = \{\delta_j ; j \in S\}$. For any event $A$, denote $I_A$ as the indicator of the event $A$.

# Chapter 1

# Introduction and Literature Review

This dissertation shall analyse and develop estimation and variable selection techniques for linear regression models under two distinct setups. Chapter 2 considers the setup with the model errors being a long memory moving average process and Chapter 3 considers the setup where the design variables are not observed directly but a noisy form of these variables are observable, which is commonly referred to as the measurement error or errors-in-variables setup.

We begin by describing the models under consideration. The $i^{th}$ component of the response vector $y^T = (y_1, ...y_n)$ is assumed to be related to the $i^{th}$ row $x_i^T = (x_{i1}, \cdots, x_{ip})$, $1 \le i \le n$ of the $n \times p$ design matrix $X$ by the relation

$$y_i \;=\; x_i^T\beta + \varepsilon_i, \quad \text{for some } \beta \in \mathbb{R}^p,\ 1 \le i \le n. \tag{1.1}$$

Here for any vector $a$, $a^T$ denotes its transpose. In Chapter 2, the vector $\varepsilon := (\varepsilon_1, ..., \varepsilon_n)^T$ is assumed to be a long memory moving average process, additional assumptions on this vector shall be made precise in Chapter 2. The design variables $\{x_i,\ 1 \le i \le n\}$ shall be assumed to be non random. However as shall become apparent later this condition may be easily relaxed to allow for some common random designs such as sub-Gaussian or sub-Exponential with independent rows.

In the above discussion we assume that the design variables $\{x_i,\ 1 \le i \le n\}$ are completely observed. A classical problem is that of estimation of the parameter vector $\beta$ when the design variables $x_i$'s are not directly observable. This problem shall serve as the content of Chapter 3. In place of the design variables $x_i$'s we observe the surrogate $w_i$'s obeying the model

$$w_i = x_i + u_i, \quad 1 \leq i \leq n. \tag{1.2}$$

Here, $u_i^T = (u_{i1}, \cdots, u_{ip})$ are assumed to be independent of $\{\varepsilon_i\}$ and independent and identically distributed (i.i.d.) $p-$dimensional random vectors. The exact assumptions are made precise in Chapter 3.

The parameter vector of interest throughout this document shall be the $p$-dimensional vector $\beta = (\beta_1, ..., \beta_p)$. We shall consider both the fixed $p$ setup as well as the high dimensional $p$ setup. In the latter case, the dimension $p$ shall be allowed to grow exponentially with the sample size $n$.

A critical assumption made on the parameter vector of interest $\beta$ is "sparsity, " i.e. it is assumed that a large proportion of the columns of the design matrix do not contribute any linear effect to the response vector $y$. In other words, a large proportion of the components of the parameter vector $\beta$ are zero. The problem of interest is to consistently identify and estimate the non-zero components of $\beta$.

The past two decades have contributed extensively to finding solutions to this problem, chief among which has been the $\ell_1$-penalized methods for their desirable finite sample and asymptotic properties and computational efficiency. The estimates for these methods are typically described as

$$(1.3) \qquad \hat{\beta}(\lambda) = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \left\{ l_n(\beta) + \lambda \|d \circ \beta\|_1 \right\}, \quad \lambda > 0,$$

where $l_n(\beta)$ is an appropriately chosen loss function that can be computed using the observed variables. The weights $d = (d_1, ..., d_p)^T$ is a vector of non-negative weights, and '$\circ$' denotes the Hadamard product, i.e., $\|d \circ \beta\|_1 := \sum_{j=1}^p d_j |\beta_j|$. Throughout this thesis, the design variables $x_i$'s may be triangular arrays depending on $n$, but we do not exhibit this dependence for the sake of the transparency of the exposition. Also, all limits are taken as $n \to \infty$, unless mentioned otherwise.

## 1.1 Lasso and long memory

For the estimator described in (1.3) choosing the loss function $l_n(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2$, i.e., the squared loss function and setting $d_j = 1$, $1 \le i \le p$, we obtain the well celebrated least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996). Its statistical properties are well studied when the regression errors are independent and identically distributed (i.i.d.) random variables (r.v.), see, e.g., Knight and Fu (2000), Meinhausen and Bühlmann (2006), Zhao and Yu (2006), Bickel, Ritov and Tsybakov (2009) and Bühlmann and van de Geer (2011). In particular, Knight and Fu (2000) provide the consistency and $\sqrt{n}$-consistency of Lasso estimates under the fixed $p$ setting. In the high dimensional setting, Bickel, Ritov and Tsybakov (2009) and Bühlmann and van de Geer (2011) provide error bounds for the statistical error associated with Lasso that hold with probability tending to 1. Furthermore they provide a detailed discussion of the necessary conditions required to obtain the desired error bounds. Meinhausen and Bühlmann (2006) and Zhao and Yu (2006) made the important contribution of providing model selection consistency results and also detailed the conditions necessary to obtain them.

The above mentioned papers work under the common assumption that the model errors $\{\varepsilon_i,\ 1 \le i \le n\}$ are i.i.d. realizations of a sub-Gaussian r.v., as breifly stated in the introduction the first problem investigated in this dissertation is to analyse the properties of Lasso when the model errors $\{\varepsilon_i,\ 1 \le i \le n\}$ form a long memory moving average process.

The literature in the area of $\ell_1$-penalized estimation with dependence considerations is scant. The first paper dealing with this issue is that of Alquier and Doukhan (2011). They provide finite sample error bounds under weak dependence structures on the model errors $\{\varepsilon_i,\ 1 \le i \le n\}$. Another recent paper addressing dependence concerns is that of Yoon, Park and Lee (2013). Their paper provides asymptotic results in the $n > p$ setup, in a linear regression models with stationary auto-regressive errors. In that paper the error process is assumed to be an AR(q) process, which is known to be a short memory process, see Giraitis, Koul, and Surgailis (2012) (GKS). This thesis investigates the behaviour of Lasso under a stronger dependence structure and less restrictive model assumptions in comparison to the above mentioned papers.

The adaptive Lasso proposed by Zou (2006) differs from Lasso in the way parameters are penalized where the weights $d_j,\ 1 \le i \le p$ are chosen carefully. To be more precise, for any $\eta > 0$, define the weight vector $\mathbf{d} = 1/|(\hat{\beta})|^{\eta}$, with $\hat{\beta}$ being any initial estimate of $\beta$ such that $n^{\frac{1}{2}-d}(\hat{\beta} - \beta) = O_p(1)$ componentwise. To avoid confusion the reader should recall that $d$ denotes the memory parameter of the stationary error sequence whereas $\mathbf{d}$ denotes the vector of weights in the penalty term. The adaptive Lasso estimates $\tilde{\beta}$ are given by

$$\tilde{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} d_j |\beta_j| \right\}. \tag{1.4}$$

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$, $\mathcal{A}_n^\star = \{j : \tilde{\beta}_j \neq 0, 1 \leq j \leq p\}$ and $\beta_\mathcal{A}$, $\tilde{\beta}_\mathcal{A}$ be the corresponding vectors with only those components whose indices are in the set $\mathcal{A}$.

As stated in Zou (2006) for the i.i.d. model error setup when $p$ is fixed, an estimator is said to have **oracle property** if the following hold.

1. Asymptotically, the right model is identified, i.e $\lim_{n \to \infty} P(\mathcal{A}_n^\star = \mathcal{A}) = 1$.
2. The estimator has an optimal estimation rate, $n^{\frac{1}{2}}(\tilde{\beta}_\mathcal{A} - \beta_\mathcal{A}) \to_D \mathcal{N}(0, \Sigma^\star)$, for some covariance matrix $\Sigma^\star$.

Here $\to_D$ denotes convergence in distribution. The adaptive Lasso has an advantage over Lasso, since it possesses the above oracle property under mild assumptions. On the other hand, as proved by Zhao and Yu (2006), for Lasso to be sign consistent a necessary condition is the "strong irrepresentable condition" which is a much stronger assumption.

In Chapter 2 we provide bounds on the statistical error associated with Lasso under long memory dependent model errors, that hold with probability tending to 1. Secondly, we obtain the sign consistency of Lasso under this setup with standard restrictions on the design matrix $X$. Lastly, we provide the consistency and $n^{\frac{1}{2}-d}$-consistency of the Lasso in the case where $p$ is fixed and is less than $n$. This proof is also extended to derive the oracle property for adaptive Lasso. For these results, the price that we pay to tackle the persistent correlation among the error sequence is that the rate of increase of the dimension $p$ in the high dimensional setting, and the rate of convergence in the fixed $p$ setting is slowed down by a factor of $n^d$. Here $d$ is the memory parameter of the stationary error sequence.

## 1.2 Measurement error and penalized quantile regression

Chapter 3 of this dissertation proposes and analyses a quantile based estimation and variable selection technique for the measurement model described in (1.1) and (1.2). In the classical fixed $p$ setting, it is well known that disregarding measurement error in covariates induces an attenuation bias in the estimates of the parameter vector $\beta$, i.e., the estimates are biased towards zero with a non diminishing bias whose magnitude depends on the variance of the covariate noise. This problem for measurement error models when $p$ is fixed has been extensively worked on by several authors including Fuller (1987) and Carroll, Ruppert, Stefanski and Crainiceanu (2006). These authors also provide numerous applications of such models and also provide bias corrected estimators for measurement error models in the context of mean regression, where the objective is to obtain estimates of $\beta$ corresponding to the conditional mean of the response variables, given the covariates.

The problem of estimation and variable selection in high dimensional measurement error models is of recent interest. Authors including Rosenbaum and Tsybakov (2010, 2011) and Loh and Wainwright (2012) study these models and propose penalized estimators which provide consistent estimation and variable selection in the presence of measurement error. In particular, Loh and Wainwright (2012) propose the $\ell_1$-penalised bias corrected least squares approach. They make the important contribution of providing error bounds for the associated statistical error in estimation that hold with probability tending to 1, and they do so with a non-convex loss function. However, both papers assume an underlying sub-Gaussian homoscedastic distribution of the model errors and work under the premise of mean regression.

Another important estimation technique in linear regression models is that of quantile regression where one is interested in estimates of $\beta$ corresponding to a specific conditional quantile of the response $y$, given the covariates, as opposed to mean regression where the problem of interest is to obtain estimates corresponding to the conditional mean of response variables. Quantile regression is robust against outliers and is useful in the presence of heteroscedasticity, see, e.g., Buchinsky (1994). This estimation technique forgoes the need for sub-Gaussian distributional assumptions on the model errors $\{\varepsilon_i,\ 1 \le i \le n\}$ and replaces them with much milder smoothness conditions on the density functions of these errors.

When covariates are completely observed, i.e. without measurement error, 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 (2014), Belloni and Chernozhukov (2011), and Wang, Wu and Li (2012). The convexity of quantile loss function is crucial for the analysis of their inference procedures. More precisely they propose and provide theoretical guarantees for the $\ell_1$-penalised estimator (1.3) with $l_n(\beta) = \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \beta)$ 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. Here $\tau$ is the quantile level of interest, i.e., $P(\varepsilon_i < 0) = \tau$, $1 \le i \le n$.

The problem of correcting for bias induced by measurement error in quantile regression poses a challenging problem. In the case of fixed $p$, this problem has been addressed by Wang, Stefanski and Zhu (2012) (WSZ). They provide a corrected quantile loss function and show that the estimates obtained by minimizing this loss function provides consistent and $\sqrt{n}$-consistent estimates. However since their loss function is un-penalized, it is unable to perform variable selection.

Chapter 3 of this dissertation proposes and analyzes the penalized version of the estimator proposed by WSZ. In the fixed $p$ setup we provide the oracle property of the proposed estimator. In the high dimensional setting we provide bounds on the statistical error of the proposed estimator that hold with probability tending to 1. In this setting, we also establish model selection consistency of this estimator in terms of identifying the correct zero components of the parameter vector.

# Chapter 2

# Lasso with Long Memory Regression Errors

In many problems of practical interest regression models with long memory errors arise naturally in the fields of econometrics and finance, see e.g., Beran (1994), Baillie (1996) and more recent monographs of Giraitis, Koul, and Surgailis (2012) (GKS), and Beran, Feng, Ghosh, Kulik (2013), and the numerous references therein. It is thus of interest to investigate the behavior of Lasso in regression models with long memory errors.

Recall the model (1.1), where $x_i = (x_{i1}, \cdots, x_{ip})^T$, $i = 1, \cdots, n$ are vectors of design variables, $y_i$'s denote the responses. The errors $\varepsilon_i$ are assumed to be long memory moving average with i.i.d. innovations, i.e.,

$$\varepsilon_i = \sum_{k=1}^{\infty} a_k \zeta_{i-k} = \sum_{k=-\infty}^{i} a_{i-k} \zeta_k, \tag{2.1}$$

where, $a_k = c_0 k^{-1+d}$, $k \geq 1$, $0 < d < \frac{1}{2}$ and some constant $c_0 > 0$, and $a_k = 0$ for $k \leq 0$. Also, $\zeta_j, j \in \mathbb{Z} := \{0, \pm 1, \pm 2, \cdots\}$ are i.i.d. r.v.'s with mean zero and variance $\sigma_\zeta^2$. For notational convenience, we shall assume $c_0 = 1$ and $\sigma_\zeta^2 = 1$, without loss of generality. Note that $\{\varepsilon_i, i \in \mathbb{Z}\}$ is a stationary process with autocavariance function

$$(2.2) \quad \gamma_\varepsilon(k) = \sum_{j=1}^{\infty} a_j a_{j+k} = k^{-1+2d} B(d, 1-2d)(1+o(1)), \quad 0 < d < 1/2, \quad k \to \infty,$$

where $B(a,b) := \int_0^1 u^{a-1}(1-u)^{b-1} du$, $a > 0, b > 0$, see, e.g., Proposition 3.2.1(ii) in GKS. This auto-correlation structure induces a long memory structure on the model errors $\varepsilon$, i.e., $\sum_{k=1}^{\infty} |\gamma_\varepsilon(k)| = \infty$. As briefly mentioned in Chapter 1, the Lasso estimate of $\beta$ is defined as,

$$(2.3) \qquad \hat{\beta}(\lambda) = \text{argmin}_\beta \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad \lambda > 0.$$

This chapter is devoted to understanding the properties of Lasso under our long memory setup. The three main contributions of this chapter are as follows. First, we show that the probability bound for a pre defined set controlling the stochastic term $\max_{1 \le j \le p} |X_j^T \varepsilon|$ can be obtained with a long memory moving average probability structure on $\{\varepsilon_i,\ 1 \le i \le n\}$ under appropriate restrictions on the rate of increase of the design variables and with the proper choice of the regularizer $\lambda_n$. Here $X_j$ denotes the $j^{th}$ column of the design matrix $X$. This result can be used to obtain the desired error bound for the statistical error associated with Lasso in this setup. Secondly, we obtain sign consistency of Lasso under the long memory setup with standard restrictions on the design matrix $X$. These results are obtained in the high dimensional setup. Lastly, we provide consistency and $n^{\frac{1}{2}-d}$-consistency of Lasso in the case where $p$ is fixed, under certain assumptions on the design variables $X$. This proof is also extended to derive the oracle property for a modified version of Lasso known as the adaptive Lasso. The price that we pay to tackle the persistent correlation among the error sequence is that the rate of increase of the dimension $p$ in the high dimensional setting, and the rate of convergence in the $n > p$ setting is slowed down by a factor of $n^d$.

All results in the high dimensional setting allow the design variables to grow with the restriction $\sum_{1 \le i \le n} x_{ij}^2 = O(n)$, and hence the results obtained can also easily be extended to case of Gaussian random designs. Furthermore, all results proved in the high dimensional setup in this chapter allow $p$ to grow exponentially with $n$.

This chapter is organized as follows. Section 2.1 below investigates the finite sample properties of Lasso. Section 2.2 investigates the sign consistency of Lasso. Section 2.3 provides the asymptotic properties of Lasso and also the oracle property of adaptive Lasso in the fixe $p$ setup. Section 2.4 presents a simulation study to analyse the finite sample performance of Lasso in the current setup.

## 2.1 Results with Finite Sample

In this section we prove a finite sample oracle inequality for the Lasso solution when the design is non random. This in turn will imply the consistency as well. Based on the assumptions of the design variables it will soon be clear that the results can be easily extended to Gaussian random designs as well. Accordingly, in this subsection we assume $x_i$'s are non random. To proceed further, we shall need the following notation. Let

$$(2.4)\quad W_{nj} \;=\; n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij}\varepsilon_i = n^{-(1/2+d)} \sum_{i=1}^{n} \sum_{k=-\infty}^{i} x_{ij} a_{i-k} \zeta_k = \sum_{k=-\infty}^{n} c_{nk,j} \zeta_k,$$

where

$$
\begin{aligned}
(2.5) \qquad c_{nk,j} &:= n^{-(1/2+d)} \sum_{i=1}^{n} x_{ij} a_{i-k}, \qquad k \in \mathbb{Z},\ j = 1, \cdots, p, \\
c_{n,j} &:= \sup_{-\infty<k\leq n} |c_{nk,j}|, \qquad c_n = \max_{1\leq j\leq p} c_{n,j}.
\end{aligned}
$$

Also, let,

$$
(2.6) \qquad \sigma_{n,j}^2 := Var(W_{nj}), \qquad \sigma_n^2 = \max_{1\leq j\leq p} \sigma_{n,j}^2.
$$

We shall prove that, with an appropriate choice of $\lambda_n$, the Lasso solution obeys the following oracle inequality in the long memory case, with overwhelming probability, i.e., for any $n \geq 1$, with probability approaching 1,

$$
\|X(\hat{\beta} - \beta)\|_2^2/n + \lambda_n \|\hat{\beta} - \beta\|_1 \leq \frac{4\lambda_n^2 s_0}{\phi_0^2}
$$

Here $\lambda_n = (O \log p/n^{1/2-d})$, under some conditions on the design matrix. Also, $s_0$ is the cardinality of the set of nonzero components of $\beta$ and $\phi_0$ is a constant.

The key result required for the proof involves obtaining a probability bound for the set

$$
(2.7) \qquad \Lambda = \left\{ \max_{1\leq j\leq p} 2n^{-1} | \sum_{i=1}^{n} x_{ij} \varepsilon_i | \leq \lambda_{0n} \right\},
$$

for a proper choice of $\lambda_{0n}$. Once this probability bound is obtained, the oracle inequality follows by deterministic arguments (See e.g. Bühlmann and van de Geer (2011)). In fact we have the following

**Proposition 2.1.1** *Let $\varepsilon_i$ be as defined in (2.1) with the innovation distribution satisfying* *the Cramér's condition: For all $k \geq 2$ and some $0 < D < \infty$,*

$$E|\zeta_0|^k \leq D^{k-2} k! E\zeta_0^2. \tag{2.8}$$

*For $t > 0$, define*

$$\lambda_{0n} = \big\{B_n(t^2 + 4\log p) + \sqrt{B_n^2(t^2 + 4\log p)^2 + 16\sigma_n^2(t^2 + 4\log p)}\big\}/2n^{1/2-d}, \tag{2.9}$$

*where $B_n := c_n D$. Then, for all $1 \leq j \leq p$ and for all $n \geq 1$,*

$$P\left(2\big|n^{-1}\sum_{i=1}^{n} x_{ij}\varepsilon_i\big| > \lambda_{0n}\right) \leq 2\exp\{-(t^2 + 4\log p)/4\}. \tag{2.10}$$

*Consequently,*

$$P(\Lambda) \geq 1 - 2\exp(-\frac{t^2}{4}), \qquad n \geq 1. \tag{2.11}$$

The proof of the above proposition will require several lemmas, hence is postponed to Section 2.5 of this chapter. The key to the proof is an application of the Bernstein inequality to finite partial sums and then passing to limit.

We can now proceed to describe the oracle inequality for the Lasso solution. The corresponding results with i.i.d. errors are proved in Bühlmann and van de Geer (2011, chapter 6). In what follows, $S_0$ denotes the collection of indices of the nonzero elements of the true $\beta$ as defined in (1.1) and $s_0$ denotes the cardinality of $S_0$. Also, for any $\delta \in \mathbb{R}^p$, $\delta_{S_0}$ denotes the vector of those components of $\delta$ which have their indices in $S_0$. In order to obtain the following inequality we require the 'compatibility condition' on the design matrix $X$. This condition is as given in Bühlmann and van de Geer (2011), which is restated here for the convenience of the reader.

**Definition 2.1.1** *We say the* **Compatibility condition** *is met for the set* $S_0$, *if for some* $\phi_0$, *and for all* $\beta$ *satisfying* $||\beta_{S_0^c}||_1 \leq 3||\beta_{S_0}||_1$,

$$||\beta_{S_0}||_1^2 \leq \frac{(\beta^T \hat{\Sigma} \beta) s_0}{\phi_0^2},$$

*with* $\hat{\Sigma} = X^T X/n$.

**Theorem 2.1.1** *Assume that the compatibility condition holds for* $S_0$. *For some* $t > 0$ *let the regularization parameter be* $\lambda_n \geq 2\lambda_{0n}$, *where* $\lambda_{0n}$ *is given in (2.9). Then with probability at least* $1 - 2\exp(-t^2/4)$, *we have*

$$\|X(\hat{\beta} - \beta)\|_2^2/n + \lambda\|\hat{\beta} - \beta\|_1 \leq \frac{4\lambda_n^2 s_0}{\phi_0^2}. \tag{2.12}$$

The proof of Theorem 2.1.1 is the same as in (Bühlmann and van de Geer (2011, chapter 6)), with the value of $\lambda_{0n}$ changed to the one given in (2.9). This result holds on the set $\Lambda$ which has the required high probability by Proposition 2.1.1. □

The only assumptions we have made so far are (i) Cramer's Condition in (2.8) on the innovation distribution and (ii) The Compatibility Condition in Definition 3.15 on the design variables. It may be of interest to mention that Gaussianity of the error distribution has not been assumed. The price that we have paid for this generality is that $\lambda_{0n}$ as defined in (2.9) is now itself data driven, i.e. $\lambda_{0n}$ also depends on the design variables $x_i$'s. Thus, keeping in view Theorem 2.1.1, it is of interest to analyse the rate of convergence of $\lambda_{0n}$. The following lemma and remark give additional conditions on the design variables, and the rate of increase of the dimension $p$, under which $\lambda_{0n}$ will converge to 0.

**Lemma 2.1.1** *Let $X = (x_{ij})_{n\times p}$ be the design matrix and suppose the following condition holds $\forall\, 1 \leq j \leq p$,*

$$n^{-1}\sum_{i=1}^{n} x_{ij}^2 \leq C, \quad \textit{for some } C < \infty. \tag{2.13}$$

*Then with $c_n$ and $\sigma_n^2$ as defined in (2.5), (2.6) respectively, we have, $c_n = o(1)$ and $\sigma_n^2 = O(1)$.*

Since $B_n = c_n D$, with $D$ being a fixed constant, the above lemma implies $B_n \to 0$.

**Remark 2.1.1** Now, recall the definition of $\lambda_{0n}$ from (2.9). Assume the design variables satisfy condition (2.13). Further assume, $\log p = o(n^{1/2-d})$, then, $\lambda_{0n} \to 0$.

The following proposition will yield the consistency of the Lasso solution.

**Proposition 2.1.2** *For some $t > 0$, let $\lambda_n \geq 2\lambda_{0n}$, where $\lambda_{0n}$ is defined in (2.9). Then on the set $\Lambda$, with probability at least $1 - 2\exp(-t^2/4)$,*

$$2\|X(\hat{\beta} - \beta)\|_2^2/n \leq 3\lambda\|\beta\|_1. \tag{2.14}$$

As mentioned earlier, the proof of this theorem follows deterministic arguments on the set $\Lambda$, (See Bühlmann and van de Geer (2011, chapter 6)). The probability of the set $\Lambda$ is given in Proposition 2.1.1.

**Remark 2.1.2 Consistency of Lasso**: *Assume the following hold,*

$$\begin{aligned} &(i) \quad \log p/n^{1/2-d} \to 0, \\ &(ii) \quad \textit{Assumption (2.13) holds}, \\ &(iii) \quad \|\beta\|_1 = o(n^{1/2-d}/\log p). \end{aligned} \tag{2.15}$$

Then by Remark 2.1.1 we have $\lambda_{0n} \to 0$. Also, assumption (iii) ensures the right hand side of the inequality (2.14) converges to zero. Hence, Proposition 2.1.2 along with Lemma 2.1.1 yields the consistency of the Lasso solution.

**Remark 2.1.3 Random Design**: There are two assumptions made on the design variables in order to obtain the error bound in Theorem 2.1.1 and the convergence of $\lambda_{0n}$ to zero in Remark 2.1.1. (i) Compatibility condition given in (3.15) and (ii) condition (2.13) which restricts the rate of increase of the design variables. These conditions can be shown to hold in the case of Gaussian random designs with independent rows. Using Theorem 1 of Raskutti (2010), condition (i) can be shown to hold with high probability (increasing to 1 exponentially). If the maximum variance component of the design variables is bounded above by a constant, then (ii) can be shown to hold with high probability using bounds for chi-square distributions given in Johnstone (2001). Hence the above results remain valid with high probability when the design variables are Gaussian with independent rows and independent of the model errors $\varepsilon$.

## 2.2 Sign Consistency of Lasso under Long Memory

In this section we prove the sign consistency of Lasso for the model (1.1), (2.1). The results in this section are similar in spirit to Zhao and Yu (2006) and we shall follow the structure of their proofs. They worked in the i.i.d. setup whereas we will be working in the long memory setup. We begin with a definition and some notations.

**Definition 2.2.1** *Lasso is said to be strongly sign consistent if there exists $\lambda_n = f(n)$, that is, a function of $n$ and independent of $y^n$ or $X^n$ such that,*

$$\lim_{n\to\infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1.$$

Here the equality denotes equality in sign, i.e., $\hat{\beta}^n =_s \beta^n$ if and only if $sign(\hat{\beta}^n) = sign(\beta^n)$, where $sign(\beta_j)$ assigns a value $+1$ to a positive entry, $-1$ to a negative entry and 0 to a zero entry.

Assume $\beta^n = (\beta_1^n, ..., \beta_q^n, \beta_{q+1}^n .. \beta_p^n)^T$, where $\beta_j^n \neq 0$, $j = 1, .., q$, and $\beta_j^n = 0$, $j = q+1, .., p$. Let $\beta_{(1)}^n = (\beta_1^n, ... \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, ... \beta_p^n)^T$. Denote $X(1)$ as the first $q$ columns of $X$, corresponding to the nonzero components of $\beta^n$. Denote $X(2)$ as the last $p-q$ columns of $X$, corresponding to the zero components of $\beta^n$. Let $C^n = n^{-1}X^TX$. Then by setting $C_{11}^n = n^{-1}X(1)^TX(1)$, $C_{22}^n = n^{-1}X(2)^TX(2)$, $C_{12}^n = n^{-1}X(1)^TX(2) = (C_{21}^n)^T$, $C^n$ can then be expressed as

$$C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}.$$

In what follows, we do not exhibit the dependence of $\beta$, $\hat{\beta}$ on $n$ for transparency of the exposition. Assuming $C_{11}^n$ is invertible, the Strong Irrepresentable condition as defined by Zhao and Yu is,

**Strong Irrepresentable Condition**: There exists a vector $\eta$, with constant, positive components, such that,

$$|C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)})| \leq \mathbf{1} - \eta, \tag{2.16}$$

where $\mathbf{1}$ is a $(p-q) \times 1$ vector of ones and the inequality holds element-wise.

The following proposition will serve as a tool to derive the sign consistency in the present setup.

**Proposition 2.2.1** *Assume the strong irrepresentable condition holds with a vector $\eta$, with all components positive. Then*

$$P(\hat{\beta}(\lambda_n) =_s \beta) \geq P(A_n \cap B_n),$$

*for*

$$A_n = \big\{|(C_{11}^n)^{-1} W(1)| < n^{\frac{1}{2}-d}(|\beta_{(1)}| - \frac{\lambda_n}{2}|(C_{11}^n)^{-1} sign(\beta_{(1)})|)\big\}, \tag{2.17}$$

$$B_n = \{|C_{21}^n (C_{11}^n)^{-1} W(1) - W(2)| \leq \frac{\lambda_n}{2} n^{\frac{1}{2}-d} \eta\}, \tag{2.18}$$

*where*

$$W(1) = \frac{X(1)^T \varepsilon}{n^{\frac{1}{2}+d}} \quad and \quad W(2) = \frac{X(2)^T \varepsilon}{n^{\frac{1}{2}+d}}. \tag{2.19}$$

This proposition provides a lower probability bound for the equivalence in sign of the Lasso estimate and the true $\beta$ vector. The proof is deterministic and hence the conclusion holds with any probabilistic structure on $\varepsilon$. It is also worth mentioning that this proposition holds without any restriction on the dimension $p$, hence we shall be able to obtain sign consistency under the case where $p$ is increasing with $n$.

In the following, we shall assume the following conditions on the design matrix and the model parameters. Assume there exists $0 \leq c_1 < c_2 < 1 - 2d$ and $M_1, M_2, M_3 > 0$, so that

$$\frac{1}{n} x_i^T x_i \leq M_1, \quad \forall \quad i \in \{1, ..., n\}, \tag{2.20}$$

$$\alpha' C_{11} \alpha \geq M_2, \quad \forall \quad \alpha \ni ||\alpha||_2^2 = 1, \tag{2.21}$$

$$q_n = O(n^{c_1}), \tag{2.22}$$

$$n^{\frac{1}{2} - d - \frac{c_2}{2}} \min_{1 \leq i \leq q} |\beta_i| \geq M_3. \tag{2.23}$$

Under the above assumptions we obtain the following sign consistency result for Lasso in the long memory case.

**Theorem 2.2.1** *Suppose the long memory regression model (1.1)and (2.1) hold, with the innovation distribution satisfying the Cramér's condition (2.8). Then under the conditions (2.16), (2.20), (2.21), (2.22), (2.23), if for some* $0 < c_3 < c_4 < (c_2 - c_1)/2$, $\lambda_n \propto n^{-(1/2-d-c_4)}$ *and* $p_n = O(e^{n^{c_3}})$, *then*

$$P(\hat{\beta}(\lambda_n) =_s \beta) \to 1. \tag{2.24}$$

The proof is detailed in the Section 2.5.

## 2.3 Asymptotics when $p$ is fixed

### 2.3.1 Asymptotic distribution of $X^T\varepsilon$

When $n > p$ and $p$ is fixed, the asymptotic properties of Lasso rely critically on the asymptotic distribution of suitably normalized $X^T\varepsilon$. This distribution is straightforward to obtain in the case of i.i.d. errors. Here we present the asymptotic distribution of normalized $X^T\varepsilon$. This distribution has essentially been obtained in chapter 4 of GKS, where the authors give CLT's for weighted sums of a long memory moving average process. Define $T_n$ as the non normalized weighted sums $W_n$ as given in (2.4), i.e. $T_n = n^{\frac{1}{2}+d}W_n$. We use $T_n$ instead of $W_n$ to relate the following more closely to GKS. Note that $T_n = X^T\varepsilon$.

Our goal is to establish the asymptotic distribution of suitably normalized $T_n$. This in turn is facilitated by Theorem 4.3.2 of GKS, pp 70. We state a slightly modified version of this theorem which can be proved easily by following the same arguments. In the following denote by $\Sigma_n = \mathrm{Cov}(T_{nj}, T_{nk})_{j,k=1}^p$.

**Theorem 2.3.1** *Let $\{x_{ij}\}_{i=1}^n$, $j = 1, ..., p$, be $p$ arrays of real weights and $\{\varepsilon_i\}$ be the stationary linear process as defined in (2.1). Assume the weights $\{x_{ij}\}_{i=1}^n$ satisfy the following condition $\forall\, j = 1, ..., p$,*

$$(2.25) \qquad (i)\ \max_{1\le i\le n}|x_{ij}| = o(n^{\frac{1}{2}+d}) \quad \textit{and} \quad (ii)\ \sum_{i=1}^{n} x_{ij}^2 \le C_j n^{1+2d},$$

*and for some matrix $\Sigma$,*

$$(2.26) \qquad n^{-(1+2d)}\Sigma_n \to \Sigma.$$

*Then,* $n^{-(\frac{1}{2}+d)}X^T\varepsilon = n^{-(\frac{1}{2}+d)}\Big(T_{n1}, ..., T_{np}\Big) \to_D \mathcal{N}(0, \Sigma)$.

**Corollary 2.3.1** *Suppose the weights* $\{x_{ij}\}$ *satisfy (2.13). Then,* $n^{-(1+2d)}\Sigma_n = O(1)$ *componentwise, for any* $0 < d < 1/2$. *Moreover, if (2.26) holds then,* $n^{-(\frac{1}{2}+d)}X^T\varepsilon \to_D \mathcal{N}(0, \Sigma)$

**Remark 2.3.1** Theorem 2.25 assumes the convergence (2.26), Corollary 3.57 shows that under a further restriction on the design matrix (2.13) we have $n^{-(1+2d)}\Sigma_n = O(1)$, however, we are unable to show convergence or identify the limit $\Sigma$ without further assumptions on the design matrix. On the other hand, if we assume the following structure on the design variables, this limit can then be explicitly computed. Let

$$(2.27)\;\; g_j : [0, 1] \to \mathbb{R},\; j = 1, ..., p, \quad \text{and} \quad x_i = (g_1(i/n), \cdots, g_p(i/n))^T, \quad i = 1, ..., n,$$

where we assume that $g_j$ is a continuous function with $\|g_j\|^2 := \int_0^1 g_j^2(u)du < \infty, \forall 1 \le j \le p$. Under this structure on the design variables, we have $\forall 1 \le j, k \le p$

$$\Sigma_{j,k} := \lim_{n\to\infty} n^{-(1+2d)}\text{Cov}(T_{nj}, T_{nk}) = B(d, 1-2d)\int_0^1\int_0^1 g_j(u)g_k(v)|u-v|^{-1+2d}dudv,$$

where $B(d, 1-2d)$ is defined in (2.2) and $\Sigma_{j,k}$ is the $(j,k)^{th}$ component of $\Sigma$. This structure on the design variables has been used in Dahlhaus (1995) in the context of polynomial regression with long range dependent regression errors. A short proof is given in the Section 2.5.

### 2.3.2 Asymptotic Properties of Lasso

Knight and Fu (2000) proved that in the case of i.i.d. errors, Lasso estimates $\hat{\beta}$ converge in probability to the true coefficient vector $\beta$, with an optimal choice of the regularizer $\lambda_n$.

They also show that Lasso is $\sqrt{n}$-consistent (asymptotic normality). Here we shall present analogous results when the errors are assumed to be long memory moving average. In this section we shall require the following assumption.

(2.28) $\qquad n^{-1}X^TX \to C,$ where C is a positive definite matrix.

**Theorem 2.3.2** *For the long memory regression model (1.1) and (2.1) assume that the design variables satisfy (2.25), (2.26), and (2.28). Further, if $\lambda_n$ is such that $\lambda_n \to \lambda_0 \geq 0$, then $\hat{\beta}^n \to_p \arg\min_\phi(Z(\phi))$, where*

$$Z(\phi) = (\phi - \beta)^T C(\phi - \beta) + \lambda_0 \sum_{j=1}^{p} |\phi_j|, \quad \phi \in \mathbb{R}^p.$$

*Thus, if $\lambda_n = o(1)$ then $argmin_\phi(Z(\phi)) = \beta$ and $\hat{\beta}^n(\lambda_n)$ is consistent for $\beta$*

**Theorem 2.3.3** *For the long memory regression model (1.1) and (2.1) assume that the design variables satisfy (2.25), (2.26), and (2.28). Suppose $n^{\frac{1}{2}-d}\lambda_n \to \lambda_0 \geq 0$ as $n \to \infty$, then*

$$n^{\frac{1}{2}-d}(\hat{\beta}^n - \beta) \to_D \arg\min_u V(u),$$

*where*

$$V(u) = -2u^T W + u^T Cu + \lambda_0 \sum_{j=1}^{p} [u_j sign(\beta_j) I_{[\beta_j \neq 0]} + |u_j| I_{[\beta_j = 0]}|],$$

*and $W$ is a $\mathcal{N}_p(0, \Sigma)$ r.v.*

Note that, when $\lambda_0 = 0$, arg min$V(u) = C^{-1}W$, where $W \sim \mathcal{N}_p(0, \Sigma)$. The above two theorems highlight the desirable asymptotic properties of Lasso in the current setup. In particular, when $\lambda_0 = 0$, Theorem 2.3.2 guarantees estimation consistency, while Theorem 2.3.3 guarantees the $n^{\frac{1}{2}-d}$ – consistency.

The technique used to prove the above theorems is to normalize the dispersion function appropriately in order to use the asymptotic normality of $n^{-(\frac{1}{2}+d)}X^T\varepsilon$, in contrast to $n^{-\frac{1}{2}}X^T\varepsilon$ in the i.i.d. case. The proof is detailed in the Section 2.5.

### 2.3.3 Adaptive Lasso

The adaptive Lasso differs from Lasso in the way parameters are penalized. To be more precise, for any $\eta > 0$, define the weight vector $\hat{w} = 1/|(\hat{\beta}^n)|^\eta$, with $\hat{\beta}^n$ being any estimate of $\beta$ such that $n^{\frac{1}{2}-d}(\hat{\beta}^n - \beta) = O_p(1)$ componentwise. The adaptive Lasso estimates $\tilde{\beta}^n$ are given by,

$$\tilde{\beta}^n = \operatorname{argmin}_\beta \left\{ \frac{1}{n}\|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j| \right\}. \tag{2.29}$$

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$, $\mathcal{A}_n^\star = \{j : \tilde{\beta}_j^n \neq 0, 1 \leq j \leq p\}$ and $\beta_\mathcal{A}, \tilde{\beta}_\mathcal{A}^n$ be the corresponding vectors with only those components whose indices are in the set $\mathcal{A}$.

As stated in Zou (2006), an estimator is said to have **oracle property** if the following hold,

1. Asymptotically, the right model is identified, i.e $\lim_{n\to\infty} P(\mathcal{A}_n^\star = \mathcal{A}) = 1$.
2. The estimator has an optimal estimation rate, $n^{\frac{1}{2}-d}(\tilde{\beta}_\mathcal{A}^n - \beta_\mathcal{A}) \to_D \mathcal{N}(0, \Sigma^\star)$, for some covariance matrix $\Sigma^\star$.

The adaptive Lasso has an advantage over Lasso, since it possesses a desirable variable selection property under mild assumptions. On the other hand, as seen in Section 3, for

Lasso to be sign consistent, we require the strong irrepresentable condition which is a much stronger assumption. The following theorem shows this property of the adaptive Lasso. In other words, the adaptive Lasso enjoys the oracle property in the long memory case. Let $\Sigma_{\mathcal{A}}$ be the limiting covariance matrix in (2.26) with only those components whose indices are in the set $\mathcal{A} \times \mathcal{A}$.

**Theorem 2.3.4** *For the linear model (1.1), assume the design variables satisfy (2.25), (2.26) and (2.28). Let the regularizer $\lambda_n$ be such that $n^{\frac{1}{2}-d}\lambda_n \to 0$, and $n^{\frac{1}{2}+\frac{\eta}{2}-d-d\eta}\lambda_n \to \infty$. Then the adaptive Lasso must satisfy the following.*

1. *Variable selection consistency,* $\lim_{n\to\infty} \mathrm{P}(\mathcal{A}_n^{\star} = \mathcal{A}) = 1$.
2. *Asymptotic normality,* $n^{\frac{1}{2}-d}(\tilde{\beta}_{\mathcal{A}}^n - \beta_{\mathcal{A}}) \to_D (C_{11}^n)^{-1}\mathcal{N}(0, \Sigma_{\mathcal{A}})$

**Remark 2.3.2** For the adaptive weights $\hat{w} = 1/|\hat{\beta}|^{\eta}$, we can choose $\hat{\beta}$ as the ordinary least square estimate. It has already been shown in GKS that $n^{\frac{1}{2}-d}(\hat{\beta}^n - \beta) = O_p(1)$, which is the required condition that the weights must satisfy.

## 2.4 Simulation Study

In this section we numerically analyse the performance of Lasso under long range dependent setup. We also compare its performance to that in the i.i.d. setup. All simulations were done in R, the estimation of Lasso was done using the package 'glmnet' developed by Friedman, Hastie, Tibshirani (2013). The regularizer $\lambda_n$ was chosen by five fold cross-validation.

**Simluation setup**: In this study, $\beta$ was chosen as a $1000 \times 1$ vector, with the first twenty five components chosen independently from a uniform distribution over the interval (-2,5), all other components of $\beta$ were set to zero. The covariates $x_i$ are i.i.d. observations from a 1000 dimensional Gaussian distribution with each component having mean and variance one. We set the pairwise correlation to be $\text{cor}(x_{ij}, x_{ik}) = 0.5^{|j-k|}$. This design matrix has been used by Tibshirani (1996) and many authors since then. The model error vector $\varepsilon$ is generated using the definition (2.1) with $c_0 = 1$ and $d = 0.15, 0.25, 0.35, 0.45$, with the innovations being i.i.d. Gaussian r.v. as given in (2.1) with mean zero and standard deviation $\sigma_\zeta = 3.5$. The simulations were repeated 100 times, i.e. 100 data sets were generated under the above setup with the same parameter vector $\beta$.

Since we have chosen d, the corresponding variance of each component of the stationary error process can be computed as, $\text{Var}(\varepsilon_i) = \sigma_\zeta^2 \sum_{k=1}^{\infty} k^{-2+2d} \quad \forall i$, which turns out to be 25.16, 31.98, 47.64 and 100.94 corresponding to d=0.15, 0.25, 0.35, 0.45 respectively.

We begin by illustrating the significant correlation among the components of the regression error vector $\varepsilon$. Figure 1 and 2 present the sample auto-correlation functions of the error vector $\varepsilon$ of the first model of the 100 simulated data sets. Figure 1 & 2 above exhibit the slow decay of the autocorrelation among the error sequence $\varepsilon$. This slow rate of decay is in coherence with long memory dependence, since $\sum_{k=1}^{\infty} |\gamma_\varepsilon(k)| = \infty$. Also, it is evident from the above two figures that the strength of the dependence is increasing as d increases.

Figure 2.1: Lag vs sample auto-correlation function with d=0.15 and d=0.25.

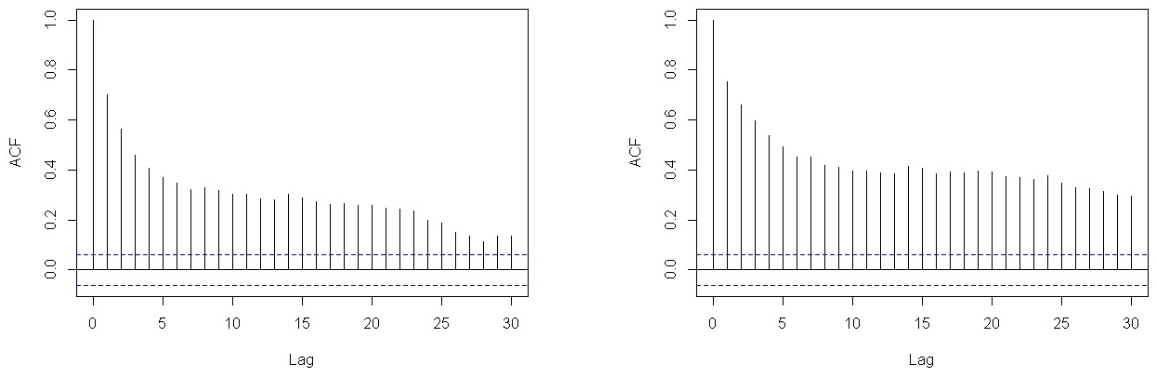
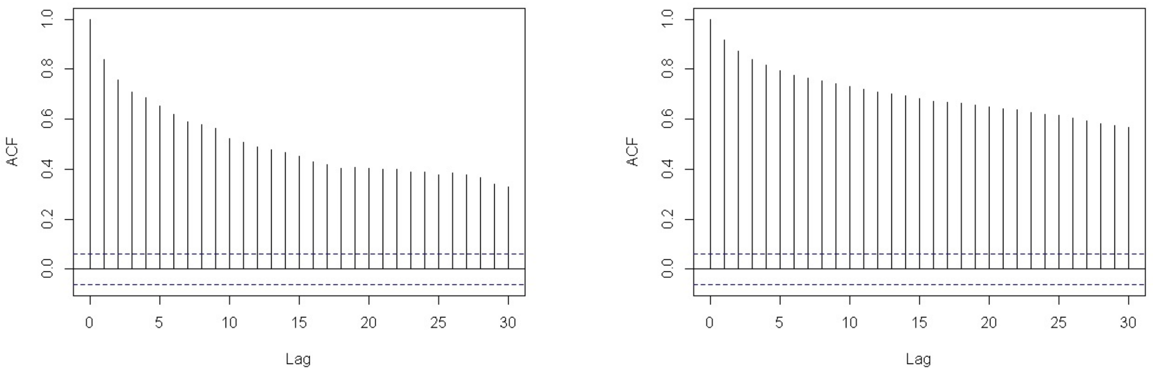


Figure 2.2: Lag vs sample auto-correlation function with d=0.35 and d=0.45.



For comparison purposes, we shall also perform the same simulation study with the errors $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ being i.i.d. Gaussian observations with mean 0 and variance 25.16, 31.98, 47.64, 100.94 which correspond to the variances of the components of the stationary sequence $\varepsilon$ under the long memory setup corresponding to $d = 0.15, 0.25, 0.35, 0.45$. The reason to choose the same variance of $\varepsilon_i$ as in the long memory setup is to maintain the same signal to noise ratio.

Now we proceed to the estimation part. In our study we simulated 100 different realizations of the design matrix $X$ and the error vector $\varepsilon$. Thus leading to 100 data sets with the same parameter vector $\beta$. For performance comparison we shall report the Relative Estimation Error (REE), i.e $\|\hat{\beta} - \beta\|^2/\|\beta\|^2$ and the Relative Prediction Error (RPE) as defined in Zou (2006), i.e. the empirical estimate of $E\|\hat{y} - X^T\beta\|^2/\sigma_\varepsilon^2$. Also, we shall report the number of correctly estimated non-zero parameters (NZ) and the number of incorrectly estimated zero parameters (IZ). Recall that in the true model there are 25 non-zero and 975 zero parameters. Table 1 summarizes the simulation results under the long memory setup and Table 2 summarises the results under the i.i.d. setup.

Table 2.1: Medians of RPE, REE, NZ & IZ with Gaussian design, long mem. errors

| | d=0.15 | | | | d=0.25 | | | | d=0.35 | | | | d=0.45 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (n) | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ |
| 100 | 0.216 | 0.62 | 14 | 33 | 0.23 | 0.62 | 14 | 33.5 | 0.24 | 0.61 | 14 | 34.5 | 0.28 | 0.46 | 14 | 32 |
| 200 | 0.13 | 0.47 | 15 | 30 | 0.13 | 0.44 | 14 | 32.5 | 0.14 | 0.41 | 14 | 35 | 0.18 | 0.38 | 14 | 35 |
| 300 | 0.10 | 0.39 | 15 | 33 | 0.11 | 0.36 | 15 | 35 | 0.11 | 0.38 | 15 | 34 | 0.14 | 0.31 | 15 | 41 |
| 400 | 0.09 | 0.33 | 16 | 39 | 0.09 | 0.31 | 16 | 40 | 0.10 | 0.32 | 15 | 36 | 0.12 | 0.31 | 15 | 41 |
| 700 | 0.05 | 0.23 | 20 | 60 | 0.06 | 0.21 | 19 | 59.5 | 0.07 | 0.23 | 18 | 50 | 0.08 | 0.22 | 17 | 52 |

Table 2.2: Medians of RPE, REE, NZ & IZ with Gaussian design, i.i.d. errors

| | $\text{Var}(\varepsilon_i) = 25.16$ | | | | $\text{Var}(\varepsilon_i) = 31.98$ | | | | $\text{Var}(\varepsilon_i) = 47.64$ | | | | $\text{Var}(\varepsilon_i) = 100.94$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (n) | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ | REE | RPE | NZ | IZ |
| 200 | 0.14 | 0.42 | 14 | 29.5 | 0.15 | 0.38 | 14 | 29 | 0.18 | 0.33 | 14 | 29 | 0.25 | 0.29 | 14 | 30 |
| 400 | 0.09 | 0.28 | 16 | 32 | 0.10 | 0.26 | 15 | 31 | 0.12 | 0.22 | 15 | 28 | 0.15 | 0.18 | 14 | 28 |

Interpretation:

- Lasso is a desirable estimation procedure in our long range dependent setup. It performs accurate estimation at all levels of dependence, from d=0.15 to d=0.45. It is evident from the simulation results that the estimation becomes increasingly accurate in terms of both REE and RPE as the sample size increases. At n=400, the relative error in estimation of $\beta$ is around 10% at all levels of dependence. As the reader might observe, It was expected that at any fixed sample size, RPE should increase as $d$ increases, however this is not the case, the reason for this is, we use cross validation to choose $\lambda_n$ and not the theoretical value of $\lambda_n$ derived earlier.

- In terms of variable selection, Lasso is increasingly successful in choosing the non zero parameters as the sample size increases. By n=700, it identifies around 20 of the non zero parameters for all levels of dependence. The parameters that Lasso is consistently unable to select are the ones that are too small in size, i.e in our model we have four parameters where $|\beta_j| < 0.65, j = 3, 7, 15, 19$, and it is these parameters that Lasso is consistently unable to detect, up to the sample size n=700. The point here being, this is a known drawback of Lasso connected with assumption (2.23), and it is not due to the long memory dependence structure on the errors. This is confirmed by the results for the i.i.d. case, which exhibits the same problem. The above simulation also brings out another familiar drawback, as the reader might observe, although Lasso manages to correctly estimate a significant portion of the zero parameters (around 95% at n=700), however the number of incorrectly estimated zero parameters (IZ) is not decreasing as the sample size increases. This again is not due the long memory errors but is an inherent drawback of cross validation. This can again be confirmed by the results in the i.i.d. case at the variance levels 47.64 and 100.94 where IZ does not decrease as n increases from 200 to 400.

- Comparing RPE in the long memory case and the i.i.d. case, as expected, we observe that the long memory case requires larger number of observations to reach the same level of accuracy, keeping in mind that the variance of components of $\varepsilon$ is similar for both the dependent and independent case.

## 2.5 Proofs for Chapter 2

The proof of **Lemma 2.1.1**, will follow after two key lemma's. To proceed further we require the following notation,

Let $r$ be a finite positive integer and $\forall 1 \leq j \leq r$, let $h_j = (h_{1j}, h_{2j}, ..., h_{nj})^T$ be a vector of weights. Further, let $c_{nk,j} = \sum_{i=1}^{n} h_{ij} a_{i-k}$ and define for $1 \leq j \leq r$,

$$W_{n,j} = h_j^T \varepsilon = \sum_{i=1}^{n} h_{ij} \varepsilon_i = \sum_{i=1}^{n} \sum_{k=-\infty}^{i} h_{ij} a_{i-k} \zeta_k = \sum_{k=-\infty}^{n} c_{nk,j} \zeta_k. \tag{2.30}$$

Further define,

$$c_{n,j} = \sup_{-\infty<k\leq n} |c_{nk,j}|, \quad c_n = \max_{1\leq j\leq r} c_{n,j}. \tag{2.31}$$

Also, denote by

$$\sigma^2_{n,j} = Var(W_{n,j}), \quad \sigma^2_n = \max_{1\leq j\leq r} \sigma^2_{n,j}. \tag{2.32}$$

Observe that,

$$\sigma^2_{n,j} = \sum_{l,m=1}^{n} h_{lj}h_{mj}\gamma_\varepsilon(l-m).$$

Furthermore, if we set $h_{ij} = 0 \ \forall \ i > n$ and $i \leq 0$, then under the assumption $\sum_{i=1}^n h^2_{ij} \leq M/n^{2d}$, $M < \infty$, we obtain using (2.2) that for all $1 \leq j \leq r$,

$$\begin{aligned} \sigma^2_{n,j} &= c_\gamma \sum_{s=-(n-1),s\neq 0}^{n-1} p(s,j)|s|^{-1+2d} + o(1) \\ &= c_\gamma \sum_{m=1}^{n} \sum_{l=1,l\neq m}^{n} h_{mj}h_{lj}|l-m|^{-1+2d} + o(1). \end{aligned} \tag{2.33}$$

Here, $p(s,j) := \sum_{i=1}^n h_{ij}h_{(i+s)j}$, and $c_\gamma = B(d, 1-2d)$ as given by (2.2).

Note that, if we replace $h_{ij}$ by $n^{-(\frac{1}{2}+d)}x_{ij}$ in the above definition of $W_{nj}$ then we obtain (2.4). This more general definition of $W_{nj}$ will be essential later in the proof of sign consistency.

**Lemma 2.5.1** *For any positive integer $r$, and for all $1 \leq j \leq r$, let $h_j = (h_{1j}, ..., h_{nj})^T$ be any vector of weights such that $\|h_j\|_2^2 = \sum_{i=1}^n h^2_{ij} \leq M/n^{2d}$, for some constant $M < \infty$. Let*

*$\sigma_n^2$ be as defined in (2.32). Then $\sigma_n^2 = O(1)$.*

**Proof.** First

$$\begin{aligned} |p(s,j)| \leq \sum_{i=1}^{n} |h_{ij}h_{(i+s)j}| &\leq (\sum_{i=1}^{n} h_{ij}^2)^{1/2}(\sum_{i=1}^{n} h_{ij}^2)^{1/2} \\ &\leq M/n^{2d}, \end{aligned}$$

and hence

$$\begin{aligned} Var(W_{n,j}) &= c_\gamma \sum_{s=-(n-1),s\neq 0}^{n-1} p(s,j)|s|^{-1+2d} + o(1) \\ &\leq c_\gamma \frac{M}{n^{2d}} \sum_{s=-(n-1),s\neq 0}^{n-1} |s|^{-1+2d} + o(1) \\ &\leq c_\gamma \frac{M}{n} \sum_{s=-(n-1),s\neq 0}^{n-1} |s/n|^{-1+2d} + o(1) \\ &\to M' \int_{-1}^{1} |t|^{-1+2d} dt. \end{aligned} \tag{2.34}$$

Observe that the bound in (2.34) is free of $j$, hence the claim follows. □

**Lemma 2.5.2** *For any positive integer $r$, and for all $1 \leq j \leq r$, let $h_j = (h_{1j}, ..., h_{nj})^T$ be any vector of weights such that $\|h_j\|_2^2 = \sum_{i=1}^n h_{ij}^2 \leq M/n^{2d}$, for some constant $M < \infty$. Then for $c_n$ as defined in (2.31) we have, $c_n = o(1)$.*

**Proof** The idea of the proof is borrowed from GKS as part of Proposition 4.3.1, pp 66, where it is used in a different context. First observe, since for all $1 \leq j \leq r$, $\sum_{i=1}^n h_{ij}^2 \leq M/n^{2d}$,

$\Rightarrow \frac{1}{\max_{\substack{1\le i\le n\\ 1\le j\le r}} |h_{ij}|} \ge n^d/\sqrt{M} \to \infty$. Define $K_n := \frac{1}{\max_{\substack{1\le i\le n\\ 1\le j\le p}} |h_{ij}|}$, and consider

$$\begin{aligned}
|c_{nk,j}| &\le \sum_{i=1}^{n} |h_{ij}a_{i-k}| \\
&\le \sum_{i=1}^{n} |h_{ij}a_{i-k}| I(|i-k| \ge K_n) \\
&\quad + \sum_{k=1}^{n} |h_{ij}a_{i-k}| I(|i-k| \le K_n) \\
&=: q_{n,1k,j} + q_{n,2k,j}
\end{aligned}$$

$$\begin{aligned}
q_{n,1k,j} &\le (\sum_{k=1}^{n} h_{ij}^2)^{1/2} (\sum_{k=1}^{n} a_{i-k}^2 I(|i-k| \ge K_n))^{1/2} \\
&\le C/n^d \sum_{l \ge K_n}^{n} a_l^2 \to 0. \qquad (2.35)
\end{aligned}$$

$$\begin{aligned}
q_{n,2k,j} &\le \max_{1\le i\le n, 1\le j\le r} |h_{ij}| \sum_{i=1}^{n} |a_{i-k}| I(|i-k| \le K_n) \\
&\le K_n^{-1} K_n^{1/2} (\sum_{l=0}^{\infty} a_l^2)^{1/2} \\
&\le C K_n^{-1/2} \to 0. \qquad (2.36)
\end{aligned}$$

Since the right hand side of (2.35) and (2.36) are free of $j$, hence we obtain, $c_n = o(1)$. □

***Proof of Lemma 2.1.1***: In the above setup $\forall\ 1 \le j \le p$, let $h_{ij} = n^{-(\frac{1}{2}+d)} x_{ij}, \ \ 1 \le i \le n$. Then, under the assumption (2.13), we have, $\sum_{i=1}^{n} h_{ij}^2 \le C/n^{2d}$. The result now follows from Lemma 2.5.1 and Lemma 2.5.2. □

**Remark 2.5.1** Observe from the proof of Lemma 2.5.2, for $h_{ij} = n^{-(1/2+d)}x_{ij}$, we have,

$$|c_{nk,j}| \leq \frac{(\sum_{i=1}^{n} x_{ij}^2)^{1/2}}{n^{\frac{1}{2}+d}} (\sum_{i=1}^{n} a_{i-k}^2 I(|i-k| \geq K_n))^{1/2} + K_n^{-1/2} (\sum_{l=0}^{\infty} a_l^2)^{1/2},$$

where $K_n = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} |x_{ij}|$. Since $x_{ij} < \infty \; \forall \, 1 \leq i \leq n, \, \forall \, 1 \leq j \leq p$, and the sequence $\{a_l\}$ is square summable, hence each fixed $n$, $c_n < \infty$ without the assumption (2.13).

The following several lemmas are needed to prove **Proposition 2.1.1**. First recall the Bernstein inequality from Doukhan (1994) or Lemma 3.1 from Guo and Koul (2007).

**Lemma 2.5.3** *For each $n \geq 1$, $m \geq 1$, let $Z_{mni}, i = -m, ..., n$, be an array of mean zero finite variance independent random variables. Assume, additionally that they satisfy the Cramérs condition: for some $B_{mn} < \infty$,*

$$(2.37) \qquad E|Z_{mni}|^k \leq B_{mn}^{k-2} k! E Z_{mni}^2, \quad k = 2, 3, ..., \; i = -m, ..., n.$$

*Let $T_{mn} = \sum_{i=-m}^{n} Z_{mni}$, $\sigma_{mn}^2 = \sum_{i=-m}^{n} Var(Z_{mni})$. Then, for any $\eta > 0$ and $n \geq 1$,*

$$(2.38) \qquad P(|T_{mn}| > \eta) \leq 2 \exp\left\{ \frac{-\eta^2}{4\sigma_{mn}^2 + 2B_{mn}\eta} \right\}, \quad \forall \, m \in \mathbb{Z}^+, \, n \geq 1.$$

We need to apply the above Bernstein inequality $p$ times, $j$th time to $Z_{mni,j} := c_{ni,j}\zeta_i$, $-m \leq i \leq n$, $1 \leq j \leq p$. In this case then

$$(2.39) \qquad T_{mnj} = \sum_{i=-m}^{n} c_{ni,j}\zeta_i.$$

For this purpose, we need to verify (2.37) in this case. Let D be as in (2.8) and

$$B_{mn,j} \equiv B_n := c_n D, \quad c_n = \max_{1 \le j \le p} c_{n,j}. \tag{2.40}$$

Then by assumption (2.8),

$$\begin{aligned} |c_{ni,j}|^k E|\zeta_i|^k &\le |c_{ni,j}|^{k-2} D^{k-2} k! c_{ni,j}^2 E\zeta_i^2 \\ &\le B_n^{k-2} k! c_{ni,j}^2 E\zeta_i^2, \qquad -m \le i \le n, \end{aligned} \tag{2.41}$$

thereby verifying the Cramér's condition (2.37) for $Z_{mni,j}$ for each $1 \le j \le p$ with $B_{mn,j} \equiv B_n$, not depending on $m$ and $j$.

To proceed further, we need to obtain an upper bound for $\sigma_{mn,j}^2 := \sum_{i=-m}^n \mathrm{Var}(Z_{mni,j})$. But

$$\begin{aligned} \sigma_{mn,j}^2 &= \sum_{i=-m}^{n} \mathrm{Var}(c_{ni,j}\zeta_i) \le \sum_{i=-\infty}^{n} Var(c_{ni,j}\zeta_i) \\ &= Var(\sum_{i=-\infty}^{n} c_{ni,j}\zeta_i) = Var(\sum_{i=1}^{n} n^{-(1/2+d)} x_{ij}\varepsilon_i) \\ &= n^{-(1+2d)} \sum_{k,\ell=1}^{n} x_{kj} x_{\ell j} \gamma_\varepsilon(k-\ell) = \sigma_{n,j}^2 < \infty, \end{aligned} \tag{2.42}$$

From the above discussion we now readily obtain that for all $\eta > 0$ and $1 \le j \le p$,

$$\begin{aligned} P(|\sum_{i=-m}^{n} c_{ni,j}\zeta_i| > \eta) \quad &\le 2\exp\big[\frac{-\eta^2}{4\sigma_{mn,j}^2 + 2B_n\eta}\big] \\ &\le 2\exp\big[\frac{-\eta^2}{4\sigma_n^2 + 2B_n\eta}\big]. \end{aligned} \tag{2.43}$$

**Remark 2.5.2** By Remark 2.5.1, we see that for each fixed $n \geq 1$, we have, $c_n < \infty$ without assumption (2.13). Hence the Bernstein inequality is applicable for every $n \geq 1$ without assumption (2.13).

We are now almost set to derive the probability bound for $\Lambda$. Before that, we look at the following preliminary lemma, which will help us to obtain this bound from the truncated sums $T_{mnj}$ defined in (2.39) for $W_{nj}$ defined in (2.4) by taking limit as $m \to \infty$.

**Lemma 2.5.4** *For each fixed n, let*

$$\begin{aligned}
&A := \{|\sum_{i=-\infty}^{n} y_{ni}| > r\}, \quad B_m = \{|\sum_{i=-m}^{n} y_{ni}| > r - \delta\},\ r > 0,\ \delta > 0,\ m = 1, 2, ...\\
&B = \liminf_{m\to\infty} B_m.
\end{aligned}$$

*If* $|\sum_{i=-\infty}^{n} y_{ni}| < \infty$, *a.s., then, for each fixed n,* $A \subseteq B$

**Proof.** Let $\omega \in A$. Then $|\sum_{i=-\infty}^{n} y_{ni}(\omega)| > r$. Also, by assumption, $|\sum_{i=-\infty}^{n} y_{ni}(\omega)| < \infty$, which implies $\forall\, \delta > 0\ \exists\, N_{\delta,\omega} \ni |\sum_{i=-\infty}^{-m} y_{ni}(\omega)| < \delta$, $\forall\ m > N_{\delta,\omega}$. Hence $|\sum_{i=-m}^{n} y_{ni}(\omega)| > r - \delta$, $\forall\ m > N_{\delta,\omega}$, which in turn implies

$$\begin{aligned}
\omega &\in \bigcap_{m=N_{\epsilon,\omega}}^{\infty} \{|\sum_{i=-m}^{n} y_{ni}| > r - \delta|\}\\
\Rightarrow \omega &\in \bigcup_{m=1}^{\infty} \bigcap_{l=m}^{\infty} \{|\sum_{i=-l}^{n} y_{ni}| > r - \delta\}\\
\Rightarrow \omega &\in \liminf_{m\to\infty} B_m. \qquad (2.44)
\end{aligned}$$

Since (2.44) is true for any $\delta > 0$, the claim $A \subseteq B$ follows. □

Before proceeding to the next proposition, we see that the assumption in Lemma 2.5.4 is valid for the series in consideration, which is $\sum_{k=-\infty}^{n} c_{nk,j}\zeta_k$. First, consider the series

$\varepsilon_i = \sum_{k=1}^{\infty} a_k\zeta_{i-k}$, since this is an infinite sum of independent zero mean random variables with $\sum_{k=1}^{\infty} \mathrm{Var}(a_k\zeta_{i-k}) < \infty$, hence $\varepsilon_i < \infty$ a.s. (Durrett, Theorem 1.8.3, page 62). Now for each fixed $n$, we have by (2.4), $\sum_{k=-\infty}^{n} c_{nk,j}\zeta_k = n^{-(\frac{1}{2}+d)} \sum_{i=1}^{n} x_{ij}\varepsilon_i$, since this is a finite weighted sum of $\{\varepsilon_i\}$ hence for each fixed $n$, we have, $\sum_{k=-\infty}^{n} c_{nk,j}\zeta_k < \infty$, a.s. $\forall 1 \leq j \leq p$.

***Proof of Proposition 2.1.1.*** Fix a $1 \leq j \leq p$ and an $n \geq 1$. Recall the definition of $c_{nk,j}$ from (2.5). Let $r_{np} := n^{1/2-d}\lambda_{0n}/2$. Then, for any $0 < \delta < r_{np}$, we have the following inequalities.

$$\begin{aligned}
P(|n^{-(1/2+d)}\sum_{i=1}^{n} x_{ij}\varepsilon_i| > r_{np}) &= P(|\sum_{k=-\infty}^{n} c_{nk,j}\zeta_k| > r_{np}) \\
&\leq P(\liminf_{m\to\infty}\{|\sum_{k=-m}^{n} c_{nk,j}\zeta_k| > r_{np}-\delta\}), \quad \text{by Lemma 2.5.4,} \\
&\leq \liminf_{m\to\infty} P(|\sum_{k=-m}^{n} c_{nk,j}\zeta_k| > r_{np}-\delta), \quad \text{Fatou's lemma,} \\
&\leq \liminf_{m\to\infty} 2\exp\Big[\frac{-(r_{np}-\delta)^2}{4\sigma_n^2 + 2B_n(r_{np}-\delta)}\Big],
\end{aligned}$$

where the last inequality follows from (2.43). Upon letting $\delta \to 0$ in this bound we thus obtain

$$P(|n^{-(1/2+d)}\sum_{i=1}^{n} x_{ij}\varepsilon_i| > r_{np}) \leq 2\exp\Big[\frac{-r_{np}^2}{4\sigma_n^2 + 2B_n r_{np}}\Big]. \tag{2.45}$$

Note that $r_{np}$ is a positive solution of the following quadratic equation.

$$\frac{-r_{np}^2}{4\sigma_n^2 + 2B_n r_{np}} = \frac{-(t^2 + 4\log p)}{4}.$$

Hence, (2.45) and the relation

$$2\exp\big[\frac{-r_{np}^2}{4\sigma_n^2+2B_nr_{np}}\big]=2\exp\big[\frac{-(t^2+4\log p)}{4}\big],$$

together imply

$$\begin{aligned}(2.46)\qquad P\big(2\big|n^{-1}\sum_{i=1}^n x_{ij}\varepsilon_i\big|>\lambda_{0n}\big) &= P(|n^{-(1/2+d)}\sum_{i=1}^n x_{ij}\varepsilon_i|>r_{np})\\ &\leq 2\exp\big[\frac{-(t^2+4\log p)}{4}\big].\end{aligned}$$

This completes the proof of (2.10). To prove (2.11), note that

$$\begin{aligned}1-P(\Lambda) &= P(\max_{1\leq j\leq p}2n^{-1}|\sum_{i=1}^n x_{ij}\varepsilon_i|>\lambda_0)\\ &\leq P(\cup_{j=1}^p\{2n^{-1}|\sum_{i=1}^n x_{ij}\varepsilon_i|>\lambda_0\})\\ &\leq \sum_{j=1}^p P(2n^{-1}|\sum_{i=1}^n x_{ij}\varepsilon_i|>\lambda_0).\end{aligned}$$

By (2.46) we get,

$$\sum_{j=1}^p P(2n^{-1}|\sum_{i=1}^n x_{ij}\epsilon_i|>\lambda_0)\ \leq\ 2p\exp\{-(t^2+4\log p)/4\}=2\exp(-\frac{t^2}{4}).$$

This completes the proof of Proposition 2.1.1 □

## Proofs for Section 3

***Proof of Proposition 2.2.1*:** Let $\hat{\beta}$ be as defined in (2.3) and let $\hat{u} = \hat{\beta} - \beta$. Define

$$V_n(u) = \sum_{i=1}^{n} \frac{1}{n}[(\varepsilon_i - X_i^T u)^2 - \varepsilon_i^2] + \lambda_n \|u + \beta\|_1.$$

Then $\hat{u} = \arg\min_u V_n(u)$. Denote the first term in $V_n(u)$ by (I), and the second term by (II). Then (I) can be simplified as

$$\begin{aligned} \sum_{i=1}^{n} \frac{1}{n}[(\varepsilon_i - X_i^T u)^2 - \varepsilon_i^2] &= \Big[ -2\sum_{i=1}^{n} \frac{1}{n} u^T X_i \varepsilon_i + \sum_{i=1}^{n} \frac{1}{n}(u)^T X_i X_i^T u \Big] \\ &= \Big[ \frac{-2u^T W}{n^{\frac{1}{2}-d}} + u^T C^n u \Big], \end{aligned} \tag{2.47}$$

where $W = n^{-1/2-d} X^T \varepsilon$. Differentiate (2.47) with respect to $u$ to obtain

$$2n^{-(\frac{1}{2}-d)}(C^n(n^{\frac{1}{2}-d}u) - W).$$

Let $\hat{u}(1)$, $W(1)$ and $\hat{u}(2)$, $W(2)$ denote the first $q$ and the last $p - q$ entries of $\hat{u}$, $W$, respectively. Now note that (Zhao and Yu, 2006 )

$$\{\text{sign}(\beta_{(1)})\hat{u}(1) > -|\beta_{(1)}|\} \subseteq \{\text{sign}(\hat{\beta}_j) = \text{sign}(\beta_j), j = 1, 2..., q\}. \tag{2.48}$$

Also, by the Karush-Kuhn-Tucker conditions and uniqueness of Lasso, if a solution $\hat{u}$ exists, then the following conditions must hold,

$$(C_{11}^n(n^{\frac{1}{2}-d}\hat{u}(1)) - W(1)) = -\frac{\lambda_n}{2}n^{\frac{1}{2}-d}sign(\beta_{(1)}), \tag{2.49}$$

$$|\hat{u}(1)| < |\beta_{(1)}|, \tag{2.50}$$

$$|(C_{21}^n(n^{\frac{1}{2}-d}\hat{u}(1)) - W(2))| \leq \frac{\lambda_n}{2}n^{\frac{1}{2}-d}\mathbf{1}. \tag{2.51}$$

The set (2.50) is contained in the set on the left of (2.48). Hence (2.49), (2.50), (2.51) together imply $\{sign(\hat{\beta}_{(1)}) = sign(\beta_{(1)})\}$ and $\hat{\beta}_{(2)} = \hat{u}(2) = 0$. The condition $A_n$ implies the existence of $\hat{u}(1)$ which satisfies (2.49) and (2.50) and condition $B_n$ and $A_n$ together imply (2.51). The result follows. □

To maintain clarity of notation in the coming proof, we define the following, for a matrix of weights $h_a = (h_{a1}, ..., h_{aq})$, where $h_{aj} = (h_{a1j}, h_{a2j}, ...h_{anj})$, $\forall 1 \leq j \leq q$, define $W_{n,j}^a, c_n^a, \sigma_{an}^2$ as done in (2.30), (2.31), (2.32) respectively. Also define $B_n^a = c_n^a D$. Repeat similarly for a matrix of weights $h_b = (h_{b1}, ..., h_{b(p-q)})$, with $h_{bj} = (h_{b1j}, h_{b2j}, ...h_{bnj})$, $\forall 1 \leq j \leq (p-q)$.

***Proof of Theorem 2.2.1***: Let $A_n$, $B_n$ be as defined in Proposition 2.2.1.

$$\begin{aligned} 1 - P(A_n \cap B_n) &\leq P(A_n^c) + P(B_n^c) \\ &\leq \sum_{i=1}^{q} P(|z_i| \geq n^{\frac{1}{2}-d}(|\beta_i| - \frac{\lambda_n}{2}b_i)) + \sum_{i=1}^{p-q} P(|\kappa_i| \geq \frac{\lambda_n}{2}n^{\frac{1}{2}-d}\eta_i), \end{aligned}$$

where $z = (z_1, z_2, ..., z_q)^T = (C_{11}^n)^{-1}W(1)$ , $\kappa = (\kappa_1, \kappa_2, ..., \kappa_{p-q})^T = C_{21}^n(C_{11}^n)^{-1}W(1) - W(2)$, $b = (b_1, b_2, ..., b_q) = (C_{11}^n)^{-1}sign(\beta_{(1)})$. Now express $z = h_a^T\varepsilon$, where $h_a^T = (h_{a1}, ..., h_{aq})^T =$

$(C^n_{11})^{-1}(n^{-1/2-d}X(1)^T)$. Then $h^T_a h_a = (C^n_{11})^{-1}n^{-2d}$, and $z_j = h^T_{aj}\varepsilon$ with

$$\|h_{aj}\|^2_2 \leq \frac{1}{n^{2d}M_2} \quad \forall j = 1, ...q, \quad \text{by assumption (2.21)}$$

Similarly write $\kappa = h^T_b\varepsilon$, where $h^T_b = C^n_{21}(C^n_{11})^{-1}(n^{-\frac{1}{2}-d}X(1)^T) - (n^{-\frac{1}{2}-d}X(2)^T)$. Then

$$h^T_b h_b = \frac{1}{n^{1+2d}}X(2)^T\left[I - X(1)(X(1)^TX(1))^{-1}X(1)^T\right]X(2).$$

Since $[I - X(1)(X(1)^TX(1))^{-1}X(1)]$ has eigenvalues between 0 and 1, therefore $\zeta^n_j = h^T_{bj}\varepsilon$, with

$$\|h_{bj}\|^2_2 \leq M_1/n^{2d} \quad \forall j = 1, ...p - q, \quad \text{by assumption (2.20)}.$$

Hence the weight vectors $h_{aj}, 1 \leq j \leq q$ and $h_{bj}, 1 \leq j \leq p - q$ both satisfy Lemma 2.5.1 and Lemma 2.5.2 for $r = q$ and $r = p - q$ respectively. Also,

$$|\lambda_n b| = \lambda_n|(C_{11})^{-1}sign(\beta_{(1)})| \leq \frac{\lambda_n}{M_2}||sign(\beta_{(1)})||_2 = \frac{\lambda_n}{M_2}\sqrt{q}. \tag{2.52}$$

Now, $z_j = h^T_{aj}\varepsilon = \sum^n_{i=1} h_{aij}\varepsilon_i$ Proceed as done earlier in (2.45). Using (2.52), Lemma 2.5.1, Lemma 2.5.2 and the Bernstein's Inequality as applied in (2.45). We get, for some constants $r_1, r_2 > 0$,

$$\begin{aligned} \sum^q_{j=1} P(|z_j| \geq n^{\frac{1}{2}-d}(|\beta_j| - \frac{\lambda_n}{2}b_j)) &\leq \sum^q_{j=1} P(|z_j| \geq r_1 n^{c_2/2}) \\ &\leq 2q\exp\big(\frac{-r^2_1 n^{c_2}}{4\sigma^2_{an} + 2B^a_n r_1 n^{c_2/2}}\big) \to 0. \end{aligned} \tag{2.53}$$

Also

$$(2.54)\quad \begin{aligned}\sum_{j=1}^{p-q} P(|\kappa_j| \geq \frac{\lambda_n}{2} n^{\frac{1}{2}-d}\eta_j) &\leq (p-q)\exp\big(\frac{-r_2^2(\lambda_n n^{1/2-d})^2}{4\sigma_{bn}^2 + 2B_n^b r_2\lambda_n n^{1/2-d}}\big)\\ &\leq (p-q)\exp\big(-r_2\lambda_n n^{1/2-d}\big), \quad \text{for n large,}\\ &\leq \exp\big(n^{c_3} - r_2\lambda_n n^{1/2-d}\big) \to 0.\end{aligned}$$

The result follows from (2.53) and (2.54) together. □

## Proofs for Section 4

***Proof of Corollary 3.57***: Observe that assumption (2.13) implies assumption (2.25)(i) and (2.25)(ii). Hence we only need to show, $n^{-(1+2d)}\Sigma_n = O(1)$ componentwise. For each variance component, this has already been shown in (2.34) in the proof of Lemma 2.1.1, with $h_{ij} = n^{-(\frac{1}{2}+d)}x_{ij}, \forall 1 \leq j \leq p$. The covariance components can be easily dealt with the Cauchy-Schwarz inequality. □

***Proof of Remark 2.3.1***: Using (2.33), we obtain,

$$\begin{aligned}n^{-1-2d}\text{Cov}(T_{nj}, T_{nk}) &= n^{-1-2d}c_\gamma \sum_{l,m=1,l\neq m}^{n} g_j(\frac{l}{n})g_k(\frac{m}{n})|l-m|^{-1+2d} + o(1)\\ &\to c_\gamma \int_0^1\int_0^1 g_j(u)g_k(v)|u-v|^{-1+2d}dudv.\end{aligned}$$

***Proof of Theorem 2.3.2***: Let

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^T\phi)^2 + \lambda_n\sum_{i=1}^{p}|\phi_i|,$$

then $Z_n(\phi)$ is convex. We need to show the pointwise convergence (in probability) of $Z_n(\phi)$ to $Z(\phi)+k^2$ for some constant $k$. Clearly, $\lambda_n \sum_{i=1}^p |\phi_i| \to \lambda_0 \sum_{i=1}^p |\phi_i|$ and consider,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (y_i - X_i^T\phi)^2 &= \frac{1}{n}\sum_{i=1}^n (\varepsilon_i - X_i^T(\phi-\beta))^2 \\
&= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n (\phi-\beta)^T X_i X_i^T(\phi-\beta) - 2n^{-1}(\phi-\beta)^T\sum_{i=1}^n X_i\varepsilon_i \\
&= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n (\phi-\beta)^T X_i X_i^T(\phi-\beta) - 2\frac{1}{n}(\phi-\beta)^T X^T\varepsilon,
\end{aligned}$$

the first term in the above equation converges to $k^2$ by the ergodic theorem (since $\{\varepsilon_i\}$ form a stationary ergodic sequence), the second term converges to $(\phi-\beta)^T C(\phi-\beta)$ and the last term converges to zero in probability (since by Theorem 2.3.1, $n^{-(\frac{1}{2}+d)}X^T\varepsilon$ converges in distribution). This proves the Theorem. □

***Proof of Theorem 2.3.3***: Define

$$(2.55)\ V_n(u) = n^{1-2d}\left[\sum_{i=1}^n \frac{1}{n}[(\varepsilon_i - \frac{X_i^T u}{n^{\frac{1}{2}-d}})^2 - \varepsilon_i^2] + \lambda_n\sum_{j=1}^p[|\beta_j + \frac{u_j}{n^{\frac{1}{2}-d}}| - |\beta_j|]\right].$$

Denote the first term in the above equation by (I), and the second term by (II). Then

$$\begin{aligned}
\textit{(I)} &= n^{1-2d}\left[\sum_{i=1}^n \frac{1}{n}\varepsilon_i^2 + \frac{u^T\sum_{i=1}^n X_i X_i^T u}{n.n^{1-2d}} - 2\frac{\sum_{i=1}^n u^T X_i\varepsilon_i}{n.n^{\frac{1}{2}-d}} - \sum_{i=1}^n \frac{1}{n}\varepsilon_i^2\right] \\
&= [\frac{u^T\sum_{i=1}^n X_i X_i^T u}{n} - 2\frac{\sum_{i=1}^n u^T X_i\varepsilon_i}{n^{\frac{1}{2}+d}}] \\
(2.56) \qquad &\to u^T C u - 2u^T W, \quad \text{as} \quad n\to\infty,
\end{aligned}$$

where $W$ is $\mathcal{N}_p(0,\Sigma)$. Also,

$$\begin{aligned}(II) &= n^{\frac{1}{2}-d}\lambda_n \sum_{j=1}^{p}[|n^{\frac{1}{2}-d}\beta_j + u_j| - n^{\frac{1}{2}-d}|\beta_j|] \\ &\to \lambda_0 \sum_{j=1}^{p}[u_j sign(\beta_j) I_{[\beta_j \neq 0]} + |u_j| I_{[\beta_j = 0]}], \qquad (2.57)\end{aligned}$$

The result follows from (2.56) and (2.57) together. □

***Proof of Theorem 2.3.4***: The structure of the proof is similar to that of Theorem 2 in Zuo (2006). Define,

$$(2.58)\tilde{V_n}(u) = n^{1-2d}\left[\sum_{i=1}^{n}\frac{1}{n}[(\varepsilon_i - \frac{X_i^T u}{n^{\frac{1}{2}-d}})^2 - \varepsilon_i^2] + \lambda_n \sum_{j=1}^{p}\hat{w}_j[|\beta_j + \frac{u_j}{n^{\frac{1}{2}-d}}| - |\beta_j|]\right],$$

then $\tilde{u_j} = n^{\frac{1}{2}-d}(\tilde{\beta}^n - \beta) = \arg\min \tilde{V_n}(u)$. Expanding $\tilde{V_n}(u)$ as done in (2.56) and (2.57) we get,

$$\tilde{V_n}(u) = \frac{u^T \sum_{i=1}^{n} X_i X_i^T u}{n} - 2\frac{\sum_{i=1}^{n} u^T X_i \varepsilon_i}{n^{\frac{1}{2}+d}} + n^{\frac{1}{2}-d}\lambda_n \sum_{j=1}^{p}\hat{w}_j[|n^{\frac{1}{2}-d}\beta_j + u_j| - n^{\frac{1}{2}-d}|\beta_j|.$$

Recall, $n^{-1}X^T X \to C$, and by Theorem 2.3.1 we have $n^{-(\frac{1}{2}+d)}X^T\varepsilon \to_D \mathcal{N}(0, \Sigma)$. Also, since $n^{\frac{1}{2}-d}\lambda_n \to 0$, $n^{\frac{1}{2}+\frac{\eta}{2}-d-d\eta}\lambda_n \to \infty$ and the adaptive weights $\hat{\beta}^n$ are so that $n^{\frac{1}{2}-d}(\hat{\beta}^n - \beta) = O_p(1)$. Hence we obtain $\tilde{V_n}(u) \to \tilde{V}(u)$ where,

$$\tilde{V}(u) = \left\{\begin{array}{cc} u_{\mathcal{A}}^T C_{11} u_{\mathcal{A}} - 2u_{\mathcal{A}}^T W_{\mathcal{A}} & ,\text{if}\, u_j = 0 \forall j \notin \mathcal{A} \\ \infty & , else \end{array}\right\} \qquad (2.59)$$

The unique minimum of $\tilde{V}(u)$ is $(C_{11}^{-1}W_{\mathcal{A}}, 0)^T$. Hence we obtain,

$$\tilde{u}_{\mathcal{A}} = n^{\frac{1}{2}-d}(\tilde{\beta}_{\mathcal{A}}^n - \beta_{\mathcal{A}}) \to_D C_{11}^{-1}W_{\mathcal{A}} \quad \text{and} \quad \tilde{u}_{\mathcal{A}^c} \to_D 0. \qquad (2.60)$$

The variable selection part can be obtained by adjusting normalization in the proof of Zuo (2006). From the asymptotic normality obtained in (2.60), we obtain, $\forall j \in \mathcal{A}$, $P(j \in \mathcal{A}_n^\star) \to 1$. Let $\mathbf{x}_j := (x_{1j}, ...., x_{nj})^T$ be the $j^{th}$ column of the design matrix $X$, $1 \leq j \leq p$. Next we show that if $j \notin \mathcal{A}$, then $P(j \notin \mathcal{A}_n^\star) \to 1$. By the KKT conditions for the Lasso solution, we have, $|2\mathbf{x}_j^T(y - X\tilde{\beta})| \leq n\lambda_n\hat{w}_j$. Consider,

$$\frac{\mathbf{x}_j^T(y - X\tilde{\beta}^n)}{n^{\frac{1}{2}+d}} = \frac{\mathbf{x}_j^T X n^{\frac{1}{2}-d}(\beta - \tilde{\beta}^n)}{n} + \frac{\mathbf{x}_j^T \varepsilon}{n^{\frac{1}{2}+d}}$$

using (2.60), the first term on the right side converges to some normal distribution, and by Theorem 2.3.1 the second term on the right converges to a normal distribution. Also, since $\beta_j = 0$ and $n^{\frac{1}{2}-d}(\beta - \hat{\beta}^n) = O_p(1)$, hence, $n^{\frac{1}{2}-d}\lambda_n\hat{w}_j = n^{\frac{1}{2}-d+\eta-d\eta}\lambda_n\frac{1}{|n^{\frac{1}{2}-d}\hat{\beta}_j|^\eta} \to \infty$. This implies,

$$P(j \notin \mathcal{A}_n^\star) \leq P(|2\mathbf{x}_j^T(y - X\tilde{\beta}^n)| \leq n\lambda_n\hat{w}_j) \to 1.$$

This completes the proof. □

# Chapter 3

# Weighted $\ell_1$-Penalized Corrected Quantile Regression for High Dimensional Measurement Error Models

As described in Chapter 1, we shall now consider the problem of estimation and variable selection in measurement error linear regression models. In this chapter we shall propose a $\ell_1$-penalized corrected quantile estimator that consistently corrects the bias induced by measurement error and also provides consistent variable selection. The problem of bias correction due to measurement error in the context of mean regression is a classical problem in the case of fixed $p$, on the other hand it is of recent interest in the high dimensional setting. Furthermore bias correction in quantile regression has only recently been studied Wang, Stefanski and Zhu (2012) in the fixed $p$ setting and to the best of our knowledge this chapter is the first attempt to do so in the high dimensional setting. The main contributions of this chapter are to provide the oracle property of the proposed estimator in the fixed $p$ setting, and to provide bounds on the statistical error of the proposed estimator in the high dimensional setting. In this setting, we also establish the model selection consistency of this estimator in terms of identifying the correct zero components of the parameter vector. Furthermore we illustrate its empirical success via a simulation study.

## 3.1 Model, Estimator and its Oracle Property

In this section, we describe the model, the proposed estimator and the necessary assumptions on the model parameters. Also we establish the oracle property of the proposed estimator in the fixed $p$ setting. As described in (1.1) and (1.2) we consider a linear regression model with additive error in the design variables. Where $x_i = (x_{i1}, \cdots, x_{ip})^T$, $i = 1, \cdots, n$, are vectors of non-random design variables and $y_i$'s are the responses related to $x_i$'s by the relations

$$y_i \;=\; x_i^T \beta^0 + \varepsilon_i, \quad \text{for some } \beta^0 \in \mathbb{R}^p,\ 1 \le i \le n. \tag{3.1}$$

Here $\beta^0 = (\beta_1^0, ..., \beta_p^0) \in \mathbb{R}^p$ is the parameter vector of interest, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ 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,

$$w_i = x_i + u_i, \quad 1 \le i \le n. \tag{3.2}$$

Here, $u_i^T = (u_{i1}, \cdots, u_{ip})$ 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 3.1.1** 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. (2009) used these distributions in the analysis of global positioning data and Purdom and Holmes (2005) adopted Laplace measurement error model in the analysis of data from some microarray experiments. Stefanski and Carroll (1990) provide an in depth discussion of Laplace measurement errors.

In our setup, we shall consider the model (3.1) and (3.2) in both the fixed and high dimensional settings. In the latter setting, the dimension $p$ of the parameter vector $\beta^0$ is allowed to grow exponentially with $n$, and 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 $S = \big\{j \in \{1, 2, ..., p\}; \beta_j^0 \neq 0.\big\}$, 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 chapter 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 chapter on $x_i$'s can be shown to hold with probability converging to 1 under some random designs, in particular for sub-Gaussian or sub-Exponential designs with independent observations.

When the design variables $x_i$'s are completely observed, several authors including Fan, et al. (2014), Belloni and Chernozhukov (2011) and Wang, Wu and Li (2012) have shown that $\beta^0$ can be estimated consistently by

$$\hat{\beta}_x = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, x_i, \beta) + \lambda_n \|d \circ \beta\|_1 \right\}, \tag{3.3}$$

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 $d = (d_1, ..., d_p)^T$ 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 (2012) (WSZ). To describe their loss function, let $K(\cdot)$ denote a kernel density function and $K'$ be its first derivative. Also, let $h = h_n \to 0$ be sequence of positive window widths, and define $H(x) = \int_{-\infty}^{x} K(u)du$. Let

$$\rho_L^\star(y_i, w_i, \beta, h) = \tilde{\varepsilon}_i(\tau - 1) + \tilde{\varepsilon}_i H\big(\frac{\tilde{\varepsilon}_i}{h}\big) - \frac{\sigma_\beta^2}{2}\Big\{\frac{2}{h}K\big(\frac{\tilde{\varepsilon}_i}{h}\big) + \frac{\tilde{\varepsilon}_i}{h^2}K'\big(\frac{\tilde{\varepsilon}_i}{h}\big)\Big\}, \tag{3.4}$$

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 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

$$\hat{\beta} = \arg\min_{\beta \in \Theta} l_n^{\star}(\beta). \tag{3.5}$$

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

$$\text{(i)} \quad \max_{j \in S} d_j \leq c_{\max}, \quad 0 < c_{\max} < \infty, \qquad \text{(ii)} \quad \min_{j \in S^c} d_j \geq c_{\min}, \quad 0 < c_{\min} < \infty. \tag{3.6}$$

We begin by providing the oracle property of the proposed estimator in the fixed dimension setting, i.e. $p, s$ are constants that do not change with $n$. In the following we provide the necessary notation and assumption required to proceed further.

**(F1)** In the neighborhood of zero, the density function $f_i(\cdot)$ of $\varepsilon_i$, $1 \leq i \leq n$ is bounded away from zero and infinity and has a bounded first derivative.

**(F2)** The kernel function $K(\cdot)$ is a bounded probability density function having finite fourth moment and is symmetric about the origin. In addition, $K(\cdot)$ is twice differentiable and its second derivative $K''(\cdot)$ is bounded and Lipchitz continuous.

**(F3)** Let

$$\Psi_{n1}^*(\beta^0, h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial \rho_L^*(y_i, w_i, \beta^0, h)/\partial\beta \quad \Psi_{n2}^* = \frac{1}{n} \sum_{i=1}^{n} \partial^2 \rho_L^*(y_i, w_i, \beta^0, h)/\partial\beta\partial\beta^T.$$

Assume that there exists positive definite matrices $D$ and $A$ such that $\mathrm{Cov}\big(\Psi_{n1}^*(\beta^0, h)\big) \to D$ and $E\Psi_{n2}^*(\beta^0, h) \to A$.

**(F4)** *Measurement errors:* The measurement errors $\{u_i\}$ are independent of $\{\varepsilon_i\}$, and i.i.d. $L_p(0, \Sigma)$, for all $1 \leq i \leq n$, with a known positive definite matrix $\Sigma$.

**Theorem 3.1.1** *Assume the measurement error model (3.1) and (3.2) along with (F1)-(F4) hold. Furthermore, if* $\lambda_n = O(n^{-1/2})$ *and the weight vector* $\mathbf{d}$ *satisfies (3.6), then there exists a local minimizer* $\hat{\beta}$ *of* $l_n^*(\beta)$ *satisfying* $\|\hat{\beta} - \beta^0\|_2 = O_p(n^{-1/2})$.

The above theorem requires the weight vector to satisfy the condition (3.6). This condition is also satisfied by the ordinary $\ell_1$ penalty, i.e. choosing $d_j = 1$, $1 \le j \le p$. To establish the oracle property we need the weights to be more carefully chosen. Let $d_{\min}^{S^c} = \min\{d_j;\, j \in S^c\}$ then we have the following.

**Theorem 3.1.2** *Assume the measurement error model (3.1) and (3.2) along with (F1)-(F4) hold. Furthermore, if* $n^{1/2}\lambda_n \to 0$ *and the weight vector* $\mathbf{d}$ *satisfies (3.6). In addition assume* $n^{1/2}\lambda_n d_{\min}^{S^c} \to \infty$, *then the* $\sqrt{n}$*-consistent local minimizer* $\hat{\beta} = (\hat{\beta}_S, \hat{\beta}_{S^c})^T$ *in Theorem 3.1.1 satisfies the following*

1. *Sparsity :* $P\big(\hat{\beta}_{S^c} = 0\big) \to 1$,

2. *Asymptotic Normality :* $n^{1/2}(\hat{\beta}_S - \beta_S^0) \to \mathcal{N}(0, A_S^{-1} D_S A_S^{-1})$.

This theorem provides the desired oracle property of the proposed $\ell_1$-penalized corrected quantile estimator.

We now proceed to the high dimensional setup, where the model dimensions $p, s$ are allowed to diverge exponentially with the sample size $n$. Here we shall 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 the paragraph preceding Lemma 3.3.2 below. 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, as observed in part $(a)$ of Theorem 3.1.2 for the fixed $p$ case, setting $d_j = 1$, $1 \leq j \leq p$ may not lead to consistent variable selection. 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 in the high dimensional setup.

**(A1)** *Model errors* ($\varepsilon$)*:* The distribution function (d.f.) $F_i$ of $\varepsilon_i$ has Lebesgue density $f_i$ such that $\sup_{1 \leq i \leq n, x \in \mathbb{R}} f_i(x) < \infty$, 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 $|y| \leq C_1$,

$$\max_{1 \leq i \leq n} |F_i(y) - F_i(0) - y f_i(0)| \leq C_2 y^2.$$

This condition is the same as Condition 1 in Fan et al. (2014). It imposes only mild conditions on the error densities and is slightly stronger than the Lipchitz 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 $1 \leq j \leq p$, $n^{-1} \sum_{i=1}^{n} x_{ij}^2 \leq c_x$, 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 $1 \leq i \leq n$, with a known $\Sigma$. Furthermore, there exists a constant

$0 < \sigma_u^2 < \infty$ such that $\max_{1\leq j\leq p} Var(u_{ij}) \leq \sigma_u^2$.

**(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 Lipchitz continuous, which is detailed in Section 3.5 of this chapter.

## 3.2 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 (3.4) in terms of observed $w_i$'s by the usual convex quantile function $\rho$ defined in (3.3) 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 $(y_i, x_i, \beta, h)$,

$$\rho_L(y_i, x_i, \beta, h) = (y_i - x_i^T\beta)\{\tau - 1 + H(\frac{y_i - x_i^T\beta}{h})\}. \tag{3.7}$$

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

$$\begin{aligned} M_n^*(\beta) \equiv M_n^\star(w, \beta, h) &= n^{-1} \sum_{i=1}^{n} \Big\{ \rho_L^\star(y_i, w_i, \beta, h) - \rho_L^\star(y_i, w_i, \beta^0, h) \Big\}, \\ \tilde{M}_n(\beta) \equiv \tilde{M}_n(x, \beta, h) &= n^{-1} \sum_{i=1}^{n} \Big\{ \rho_L(y_i, x_i, \beta, h) - \rho_L(y_i, x_i, \beta^0, h) \Big\}, \\ M_n(\beta) \equiv M_n(x, \beta) &= n^{-1} \sum_{i=1}^{n} \Big\{ \rho(y_i, x_i, \beta) - \rho(y_i, x_i, \beta^0) \Big\}. \end{aligned} \tag{3.8}$$

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 section 3.5. Throughout, $\gamma_{max}$ denotes the largest eigenvalue of $\Sigma$.

**Theorem 3.2.1** *Assume the measurement error model (3.1) and (3.2) and the assumptions (A1), (A2), (A3) and (A4) hold. Then,*

$$\sup_{\beta \in \Theta} \Big| M_n^\star(\beta) - EM_n^\star(\beta) \Big| = O_p\Big( \gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}} \Big), \tag{3.9}$$

$$\sup_{\beta \in \Theta} \Big| \tilde{M}_n(\beta) - E\tilde{M}_n(\beta) \Big| = O_p\Big( \sqrt{s} \sqrt{\frac{2 \log 2p}{n}} \Big). \tag{3.10}$$

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

$$EM_n^\star(\beta) = E\tilde{M}_n(\beta), \quad \forall\, \beta \in \mathbb{R}^p. \tag{3.11}$$

Secondly, under assumption (A4),

$$\sup_{\beta\in\Theta}|\tilde{M}_n(\beta)-M_n(\beta)| = O(h), \text{ a.s.} \tag{3.12}$$

Claim (3.11) is a direct consequence of Theorem 2 of WSZ while claim (3.12) is proved in WSZ as a part of the proof of their Theorem 3 (page 14). The short proof of this statement is reproduced here for the convenience of a reader.

**Proof of (3.12):** 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 (3.12) is bounded above by 2 times

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

This completes the proof of (3.12).

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 (3.11), (3.12) and Theorem 3.2.1, we obtain

$$\begin{aligned}
&\sup_{\beta\in\Theta}|M_n^\star(\beta)-M_n(\beta)| \\
&\quad\le \sup_{\beta\in\Theta}|M_n^\star(\beta)-EM_n^\star(\beta)| + \sup_{\beta\in\Theta}|\tilde{M}_n(\beta)-E\tilde{M}_n(\beta)| \\
&\qquad + \sup_{\beta\in\Theta}|\tilde{M}_n(\beta)-M_n(\beta)| \\
&\quad= O_p\Big(\gamma_{max}\frac{s^{3/2}}{h^2}\sqrt{\frac{\log 2p}{n}}\Big) + O(h).
\end{aligned} \tag{3.13}$$

The last claim in the above bounds follows since by Theorem 3.2.1, the second term on the right hand side of (3.13) decreases faster than the first term. This approximation plays a pivotal role in the analysis carried out in the sequel.

## 3.3 Results in High Dimensions

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

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

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 3.3.1** *For the measurement error model (3.1) and (3.2), we have,*

$$\begin{aligned} & g_n(\hat\beta) - g_n(\beta^0) + \lambda_n\|d\circ\hat\beta\|_1 \\ & \quad \le \ \lambda_n\|d\circ\beta^0\|_1 + |M^\star_n(\hat\beta) - M_n(\hat\beta)| + |M_n(\hat\beta) - EM_n(\hat\beta)|, \quad \forall\, \beta \in \mathbb{R}^p. \end{aligned} \tag{3.14}$$

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

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

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 neighbourhood of $\beta^0$. Later, using the convexity of $\rho_\tau(\beta)$ and the inequality (3.13), 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

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

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 (3.14). Let $X := (x_1, x_2, \cdots, x_n)^T$ denote the $n \times p$ design matrix, and $\Gamma := diag\{f_1(0), ..., f_n(0)\}$.

**Lemma 3.3.2** *Suppose the model (3.1) 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 Fan et al. (2014) page 341, to obtain the desired result. □

Fan et al. (2014) 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., Bühlmann and Van de Geer (2011) and Raskutti et al. (2010)). The closely related 'Restricted eigenvalue condition' is used by Belloni and Chernozhukov (2011) to provide consistency in estimation for quantile regression, when the covariates are completely observed.

**Definition 3.3.1** *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* $\|\delta_{S^c}\|_1 \leq c_0 \|\delta_S\|_1$*,*

$$\|\delta_S\|_1^2 \leq \frac{bs}{n\phi^2} \delta X' \Gamma X \delta. \tag{3.15}$$

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 $1 \leq j \leq p$, then $c_0 = 3$.

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

$$\begin{aligned} &\text{(i)}\ \kappa_n \to \infty, \quad \gamma_{max} \to \infty, \quad \lambda_n \to 0, \quad \alpha_n \to 0, \\ &\quad \kappa_n \gamma_{max} s_n^{3/2} h^{-2} \sqrt{\frac{\log 2p}{n}} = o(\lambda_n), \quad \alpha_n = o(\kappa_n^{-1}), \\ &\text{(ii)}\ \kappa_n h = o(\lambda_n), \qquad (iii)\ \frac{\lambda_n s_n \kappa_n}{\phi^2} \to 0. \end{aligned} \tag{3.16}$$

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

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

In the above conditions, $\phi$ is the constant defined in (3.15). 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 (3.16)(i) and fast enough to satisfy (3.16)(ii) in the case of unbounded design. Similarly, in the case of bounded design, these rate constraints have to balance between (3.17)(i) and (3.17)(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 3.3.1** *For the measurement error model (3.1) and (3.2), let $\hat{\beta}$ be as in (3.5) and $c_{\min}$, $c_{\max}$ be as in (3.6) with $c_m := c_{\min} + c_{\max}$. Assume (A1), (A2), (A3) and (A4), along with the Compatibility condition (3.15) hold. Also assume that either the rate conditions (3.16) or (3.17) holds. Then the following inequality holds with probability at least 1-o(1).*

$$\begin{aligned} (3.17) \qquad & 3c_a n^{-1}\|\Gamma^{1/2}X(\hat{\beta}-\beta^0)\|_2^2 + 2\lambda_n c_{\min}\|\hat{\beta}-\beta^0\|_1 \\ \leq \; & \frac{4\lambda_n^2 c_m^2 s_n}{\phi^2} + O\big(\gamma_{max} s_n^{3/2} h^{-2}\sqrt{\frac{\log 2p}{n}}\big) + O(h). \end{aligned}$$

The bound (3.17) clearly implies that under the conditions of Theorem 3.3.1, $\|\hat{\beta} - \beta^0\|_1 \rightarrow_p 0$ and $n^{-1}\|\Gamma^{1/2}X(\hat{\beta}-\beta^0)\|_2^2 \rightarrow_p 0$. 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 $d_j$, $1 \leq j \leq p$, do not play a critical role for the consistency of the estimator, i.e. as long as the condition (3.6) 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 (3.17) of Theorem 3.3.1 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 (2011) 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 (3.17) 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.

### 3.3.1 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

$$(3.18) \qquad \alpha_n = \frac{2\lambda_n c_m^2 s}{c_{\min}\phi^2} + \frac{1}{\lambda_n} O\big(\gamma_{max} \frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\big) + \frac{1}{\lambda_n} O(h) \to 0.$$

Then in addition to the conditions of Theorem 3.3.1, we assume there exists a $0 < \delta < 1/2$ such that,

$$(3.19) \qquad (i)\ \kappa_n \leq n^{\delta}, \qquad \text{and} \qquad (ii)\ \log p = o(n^{\delta}).$$

Furthermore along with the conditions (3.6) 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.

$$\begin{aligned} &(i) \quad \max\left\{\alpha_n, \kappa_n^3\alpha_n^2\right\} = o(\lambda_n d_{\min}^{S^c}), \qquad (ii)\, \kappa_n h = o(\lambda_n d_{\min}^{S^c}),\\ &(iii) \quad \max\left\{\gamma_{\max} s n^{\delta} h^{-2}\sqrt{\frac{2\log p}{n}},\ \alpha_n\gamma_{\max} n^{\delta} h^{-3} s\sqrt{\frac{2\log p}{n}}\right\} = o(\lambda_n d_{\min}^{S^c}). \end{aligned} \tag{3.20}$$

**Theorem 3.3.2** *For the measurement error model (3.1) and (3.2), assume the conditions of Theorem 3.3.1 hold. In addition assume that (3.6), (3.19) and (3.20) hold. Then*

$$P\Big(\hat{\beta}_j = 0,\ \forall\, j \in S^c\Big) \to 1. \tag{3.21}$$

This theorem provides the model selection consistency of the proposed $W\ell_1$-$CQ$ estimator $\hat{\beta}$, under suitable choice of the weight vector $d = (d_1, ...d_p)^T$. Note that setting the weights $d_j \equiv 1$, i.e., the un-weighted $\ell_1$-penalty does not satisfy the rate assumptions (3.20) 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 3.3.2 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 (2011). The reason for this being that our result states that zero components are correctly identified, as opposed to Belloni and Chernozhukov (2011), 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 (3.11) and the probability bound (3.47) given below hold.

### 3.3.2 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 (2008) 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 3.3.1 provides the consistency of this estimate. In particular, under conditions of Theorem 3.3.1 we obtain with high probability, $\|\hat{\beta}^{ini} - \beta^0\|_1 \le \alpha_n \to 0$, where $\alpha_n$ is defined in (3.18). 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

$$(3.22) \qquad |\hat{\beta}_j^{ini}| \le b_2 + \alpha_n, \ \ \forall\, j \in S \qquad\qquad |\hat{\beta}_j^{ini}| \le \alpha_n \ \ \forall\, j \in S^c.$$

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

Keeping (3.22) in view, it is easy to verify that when $n$ is large enough, the above weight vector $d$ satisfies the required assumptions (3.6) and (3.20) for some constant $\eta$ chosen appropriately, with probability approaching to 1.

## 3.4 Simulation Study

### 3.4.1 Simulation setup

In this section we numerically analyse 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 (2.3GHz) 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. (2010). To compute $\ell_1$-$CQ$ estimates and the bias corrected Lasso, we use the projected gradient descent algorithm (Agarwal et al. (2012)), which is a tool developed for optimizing penalized smooth loss functions in high dimensions. More precisely, with $\nabla L(\beta)$ denoting the gradient of a loss function $L$, the method of projected descent iterates by the recursions, $\{\beta^r,\, r = 0, 1, 2, ...\}$ as,

$$\beta^{r+1} = \arg\min_{\beta\in\Theta} \left\{ L(\beta^r) + \nabla L(\beta^r)^T(\beta-\beta^r) + \frac{\delta}{2}\|\beta-\beta^r\|_2^2 + \lambda_n\|\beta\|_1 \right\}, \tag{3.23}$$

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. (2012) with the restriction of the parameter space to the $\ell_1$-ball $\Theta$ implemented by the procedure of Duchi et al. (2008). 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 (2008). 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 $w_i^\star = (w_{i1}/d_1, ..., w_{ip}/d_p)^T$, $\forall 1 \leq i \leq n$. Also define $\Sigma^\star = \Big(\sigma_{ij}/d_i d_j\Big)$, $\forall 1 \leq i, j \leq p$ 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 $\hat{\beta}_j = \hat{\beta}_j^\star / d_j$, $\forall 1 \leq j \leq p$.

*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 (2010) provide theoretical justification for using *AIC-BIC* type criteria for several models. The cross validation method is often observed to result in over-fitting (Wang, Li and Tsai, 2007), furthermore it is considerably more time consuming. More recently, Lee, Noh and Park (2014) 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.

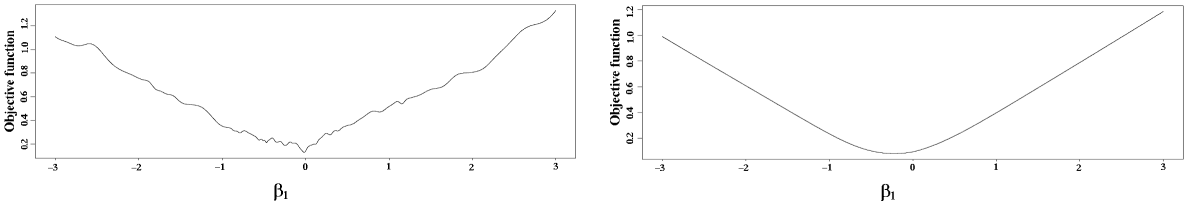
In defining HBIC$(\lambda_n, h)$, we used the corrected quantile loss function instead of the check function as defined by Lee et al. (2013). Although this makes intuitive sense as the corrected quantile loss function is approximated by the check function, however a rigorous theoretical argument justifying its use is missing. This criterion is empirically found to perform well in our setup.

### 3.4.2 Computational Issues

A computational challenge of the proposed estimator is the non-convexity of the loss function $l^*$. 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 behaviour is illustrated in Figure 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.

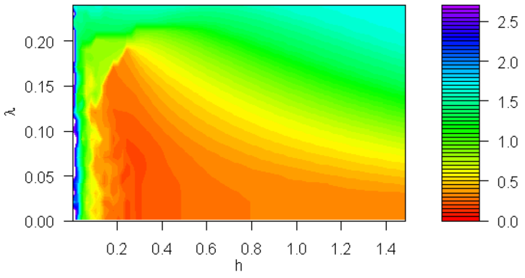
The loss function exhibits several local optimums at smaller values of $h$. On the other hand at relatively higher values of $h$, the loss function appears to be convex shaped around the true parameter and appears to have a unique minimum at the true parameter value. Two

Figure 3.1: $\frac{1}{n}\sum_{i=1}^{n}\rho_L^{\star}(\beta)+\lambda_n\|\beta\|_1$ evaluated around $\beta_1$ at $h=0.01$ (left) and 1.5 (right).



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 Figure 3.2. 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.

Figure 3.2: Colored contours of $\|\hat{\beta}-\beta\|_1$ on $h$ vs. $\lambda_n$ for $\ell_1$-$CQ$.



With regards to computational time, one optimization at $h=0.01$ takes $\approx$ 16seconds as opposed to $\approx 2$ seconds 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.

### 3.4.3 Simulation setup and results

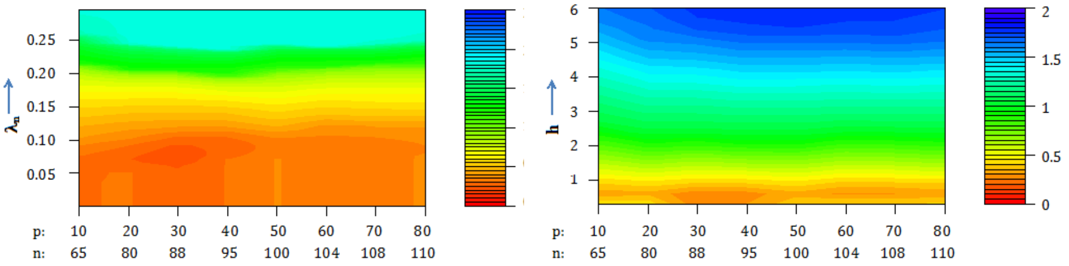
For this simulation study, data is generated from the measurement error model (3.1) and (3.2) under several choices of the underlying parameters and distributional assumptions. The unobserved design variables $\{x_{ij},\ 1 \leq i \leq n, 1 \leq j \leq p\}$ are chosen as i.i.d. r.v.'s from a $\mathcal{N}(0, 1)$ distribution. The measurement errors $\{u_i,\ 1 \leq i \leq n\}$ 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 $\varepsilon_i$, $1 \leq i \leq n$ are independent realizations of Normal, Cauchy or mean centered Pareto r.v.'s.

We begin by numerically verifying the result of Theorem 3.3.1. Observe that this theorem can be viewed as describing the scaling behaviour 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 the following figures 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.

Figure 3.3 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 3.3.1 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 $p$.

Figure 3.3: Colored contours of $\|\hat{\beta} - \beta\|_1$ on $p$ vs. $\lambda$(left) and $h$.(right)



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 $\varepsilon_i$, $1 \leq i \leq n$ 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.

Table 3.1: Simulation results at $p$=40 for Normal, Cauchy and Pareto model errors

| $s = 8$, $u_i \sim L_p(\sigma^2 = 0.2)$, $\varepsilon_i \sim \mathcal{N}(0, 0.2)$, $\tau = 0.5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\ell_1$-$CQ$ | | | $C-Lasso$ | | | $Lasso$ | | |
| n | MEE | MINZ | MIZ | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| 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 |

Tables 3.1 and 3.2 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. (2014), 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 $i = 1, \cdots n$. The dimension $p$ is increased to 500 for this case and the results are provided in Table 3.3.

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 above, where the exponent $\eta$ is chosen by using the selection criteria $\mathrm{e}^{\text{HBIC}}$. 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.

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 3.2: Simulation results at $p$=300 for Normal, Cauchy model errors

| $s = 10,\ u_i \sim L_p(\sigma^2 = 0.1),\ \varepsilon_i \sim \mathcal{N}(0, 0.2),\ \tau = 0.5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\ell_1$-$CQ$ | | | $C - Lasso$ | | | $Lasso$ | | |
| n | MEE | MINZ | MIZ | MEE | MINZ | MIZ | MEE | MINZ | MIZ |
| 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 3.3: Simulation results at $p$=500 under heteroscedasticity

| $s = 10,\ u_i \sim L_p(\sigma^2 = 0.1),\ \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2),\ \sigma_i^2 \sim Uniform(0.1, 9),\ \tau = 0.5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\ell_1$-$CQ$ | | | $C - Lasso$ | | | $Lasso$ | | |
| n | 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 3.4: $W\ell_1$-$CQ$ 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 3.5: $W\ell_1$-$CQ$ 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) |

Table 3.6: $\ell_1$-$CQ$ at $p$=300 with misspecified covariate error distribution.

| $s = 10$, $u_{ij} \sim \mathcal{N}(0, 0.3)$, $\varepsilon_i \sim \mathcal{N}(0, 0.3)$ | | | | | | |
|---|---|---|---|---|---|---|
| | $\ell_1$-$CQ$ | | | $C - Lasso$ | | |
| n | 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 |

## 3.5 Proofs for Chapter 3

We begin by proving Theorem 3.1.1 and Theorem 3.1.2 which provide consistency in estimation and variable selection for the $W\ell_1$-$CQ$ estimator for the case of fixed $p$. For this purpose we shall require the following uniform convergence result stated in WSZ, pp 16 as part of the proof of Theorem 4 of their paper. For any compact set $\mathcal{B}$ in $\mathbb{R}^p$ we have,

$$\sup_{\beta\in\mathcal{B}}\Big|n^{-1}\sum_{i=1}^{n}\big(\frac{\partial^2\rho_L^\star(\beta)}{\partial\beta_j\partial\beta_l}-E\frac{\partial^2\rho_L^\star(\beta)}{\partial\beta_j\partial\beta_l}\big)\Big| = o_p(1). \tag{3.24}$$

The proof of this statement follows from Nolan and Pollard (1987, Lemma 22) and Pollard (1984, Theorem 2.37).

**Proof of Theorem 3.1.1.** It suffices to show that for any $\epsilon > 0$, there exists a sufficiently large constant $C$ such that

$$P\Big(\inf_{\|u\|=C} l_n^*(\beta^0+\alpha_n u) > l_n^*(\beta^0)\Big) > 1-\epsilon \tag{3.25}$$

where $\alpha_n := O(n^{-1/2})$. This implies there exists a local minimum in the ball $(\beta^0+\alpha_n u : \|u\|\le C)$, i.e. there exists a local minimizer such that $\|\hat\beta-\beta^0\| = O_P(\alpha_n)$. Recall $\Psi_{n1}$ and $\Psi_{n2}$ from assumption (F3). Consider,

$$\begin{aligned} D_n(u) &= n^{-1}\sum_{i=1}^{n}\big(\rho_L^*(\beta^0+\alpha_n u)-\rho_L^*(\beta^0)\big)+\lambda_n\sum_{i=1}^{p}d_j(|\beta_j^0+\alpha_n u_j|-|\beta_j^0|) \\ &= T_{n1}+T_{n2}, \quad (say,) \end{aligned} \tag{3.26}$$

here, $T_{n1} = n^{-1}\sum_{i=1}^{n}\big(\rho_L^*(\beta^0+\alpha_n u)-\rho_L^*(\beta^0)\big)$, and $T_{n2} = \lambda_n\sum_{i=1}^{p}d_j(|\beta_j^0+\alpha_n u_j|-|\beta_j^0|)$.

Now by Taylor's expansion we obtain,

$$
\begin{aligned}
T_{n1} &= \alpha_n n^{-1/2}\psi_{n1}^T(w,\beta^0)u + \alpha_n^2\frac{1}{2}u^T\psi_{n2}^*(w,\beta^*)u \\
&= \alpha_n n^{-1/2}\psi_{n1}^T(w,\beta^0)u + \alpha_n^2\frac{1}{2}u^T Au\{1+o_P(1)\}. \qquad (3.27)
\end{aligned}
$$

Where $\beta^*$ is between $\beta^0 + \alpha_n u$ and $\beta^0$ and the second equality follows from (3.24) and assumption (F3) $\sup_{\beta\in\mathcal{B}}|\Psi_{n2}^*(w,\beta) - E\Psi_{n2}^*(w,\beta)| = o_P(1)$ and the assumption (F3).

Now by the Central Limit Theorem $\psi_{n1}^T(w,\beta^0) = O_P(1)$, hence the first term in (3.27) is of the order $O_P(n^{-1/2}\alpha_n)$. Consider the second term on the RHS, since $A$ is positive definite hence the second term is positive and by choosing a sufficiently large $C$ it dominates the first term uniformly in $\|u\| = C$.

Similarly,

$$
T_{n2} = \lambda_n\sum_{i=1}^p d_j(|\beta_j^0 + \alpha_n u_j| - |\beta_j^0|)\lambda_n\alpha_n\sum_{j\in S} d_j|u_j| \le \lambda_n\alpha_n\|u\|(\sum_{j\in S} d_j^2)^{1/2} = O(\lambda_n\alpha_n).
$$

Thus by choosing $C$ large enough, the second term in RHS of (3.26) dominates the other two, thus proving (3.25). □

**Proof of Theorem 3.1.2.** To show the first part of this theorem, observe that since the loss function $\rho_L^*$ is a non convex smooth function, thus the KKT condition for optimatily is necessary but not sufficient. We show that if $(a)$ is not true then with probability tending to 1 the necessity of KKT conditions is violated, i.e., for any $j \in S^c$, let if possible $\hat{\beta}_j \neq 0$ then by the necessity of KKT we have,

$$
n^{-1}\sum_{i=1}^n \frac{\partial\rho_L^*(\hat{\beta})}{\partial\beta_j} = \lambda_n d_j sign(\hat{\beta}_j) \qquad (3.28)
$$

Now by Taylor's expansion we obtain,

$$n^{-1}\sum_{i=1}^{n}\frac{\partial\rho_L^*(\hat{\beta})}{\partial\beta_j} = n^{-1}\sum_{i=1}^{n}\frac{\partial\rho_L^*(\beta^0)}{\partial\beta_j} + n^{-1}\sum_{l=1}^{p}\sum_{i=1}^{n}\frac{\partial^2\rho_L^*(\beta^*)}{\partial\beta_j\partial\beta_l}(\hat{\beta}_l - \beta_l^0)$$

Following standard arguments,

$$n^{-1}\sum_{i=1}^{n}\frac{\partial\rho_L^*(\beta^0)}{\partial\beta_j} = O_P(n^{-1/2})$$

and $n^{-1}\sum_{i=1}^{n}\partial^2\rho_L^*(\beta^*)/\partial\beta_j\partial\beta_l = n^{-1}\sum_{i=1}^{n}E(\partial^2\rho_L^*(\beta^*)/\partial\beta_j\partial\beta_l) + o_P(1)$ by (3.24). Also by assumption $\hat{\beta}-\beta^0 = O_P(n^{-1/2})$. Thus the LHS of (3.28) is of the order $O_P(n^{-1/2})$ hence choosing the condition $\sqrt{n}\lambda_n d_{\min}^{S^c} \to \infty$ contradicts the relation (3.28). Thus proving part $(a)$ of the Theorem.

To prove (b) we begin again with the KKT optimality condition, i.e., in view of Theorem 3.1.1, with probability tending to 1, for any $j \in S$, $\hat{\beta}_j \neq 0$ thus,

$$n^{-1}\sum_{i=1}^{n}\frac{\partial\rho_L^*(\hat{\beta})}{\partial\beta_j} - \lambda_n d_j sign(\hat{\beta}_j) = 0 \tag{3.29}$$

Now,

$$\frac{\partial\rho_L^*(\hat{\beta})}{\partial\beta_j} = \frac{\partial\rho_L^*(\beta^0)}{\partial\beta_j} + \sum_{l=1}^{p}\frac{\partial^2\rho_L^*(\beta^*)}{\partial\beta_j\partial\beta_l}(\hat{\beta}_l - \beta_l^0)$$

Since, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\rho_L^*(\beta^0)}{\partial\beta_S} \Rightarrow \mathcal{N}(0, D_S)$ and $\sqrt{n}\lambda_n d_j \to 0$, since by assumption (3.6) for all $j \in S$, $d_j$'s are bounded above, thus we obtain the required result. □

We now proceed to the proofs of results in the High Dimensional Setting. As briefly stated at the beginning of Section 4, the technique used to prove Theorem 3.3.1 is to use the convexity of the quantile function $\rho(\beta)$ and the weighted $\ell_1$-penalty, along with the approximation of $\rho^\star_L(\beta)$ to $\rho(\beta)$. Some of the steps of the proof are similar to those adopted by Bülmann and Van der Geer (2011, ch. 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,

$$\tilde\beta \in \mathcal{B}(c\alpha_n) \quad \text{implies} \quad \hat\beta \in \mathcal{B}(c\alpha_n/(1-c)), \quad \forall\, 0 < c < 1. \tag{3.30}$$

This fact will be used in the sequel.

Next, by the convexity of $g_n(\beta)$ and $\|d \circ \beta\|_1$, and the inequality (3.14), we obtain

$$\begin{aligned} g_n(\tilde\beta) - g_n(\beta^0) + \lambda_n \|d \circ \tilde\beta\|_1 \ \leq \ & \lambda_n \|d \circ \beta^0\|_1 + \sup_{\beta \in \Theta} |M^\star_n(\beta) - M_n(\beta)| \\ & + \sup_{\beta \in \mathcal{B}(\alpha_n)} |M_n(\beta) - EM_n(\beta)|. \end{aligned}$$

We begin by providing error bounds for $\tilde\beta$, which shall easily extend to $\hat\beta$. By (3.13) and (3.16), the second term in the RHS of the (3.31) is $o_p(1)$. The following lemma provides the rate of decrease of the last term.

**Lemma 3.5.1** *For the measurement error model (3.1), (3.2), assume that (A1) and (A2) hold. Then*

$$\sup_{\beta \in \mathcal{B}(\alpha_n)} |M_n(\beta) - EM_n(\beta)| = O_p\big(\alpha_n \sqrt{\frac{2\log 2p}{n}}\big). \tag{3.31}$$

The proof of Theorem 3.2.1 and Lemma 3.5.1 are provided after the proof of Theorem 3.3.1.

Consider the following events,

$$(i) \quad \Omega_1 = \text{the event that the bounds (3.9) and (3.10) hold,}$$

$$(ii) \quad \Omega_2 = \text{the event that the bound (3.31) holds.}$$

Then by Theorem 3.2.1 and Lemma 3.5.1, $P(\Omega_1 \cap \Omega_2) \geq 1 - o(1)$, and on $\Omega_1 \cap \Omega_2$,

$$\sup_{\beta \in \Theta} |M_n^\star(\beta) - M_n(\beta)| + \sup_{\beta \in \mathcal{B}(\alpha_n)} |M_n(\beta) - EM_n(\beta)| \tag{3.32}$$

$$= O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \tag{3.33}$$

This follows since $\alpha_n \to 0$, and hence the second terms on the LHS of (3.32) 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 (3.31) and (3.32) we now readily obtain that with probability at least $1 - o(1)$,

$$g_n(\tilde{\beta}) - g_n(\beta^0) + \lambda_n \|d \circ \tilde{\beta}\|_1 \leq \lambda_n \|d \circ \beta^0\|_1 + O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \tag{3.34}$$

By Lemma 3.3.2 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 (3.34) yields

$$\begin{aligned} \lambda_n \|(d \circ \tilde{\beta})_{S^c}\|_1 &\leq \lambda_n \|\big(d \circ (\tilde{\beta} - \beta^0)\big)_S\|_1 + O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h) \\ &\leq c_{\max} \lambda_n \|\tilde{\beta}_S - \beta^0_S\|_1 + O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \end{aligned} \tag{3.35}$$

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

$$\frac{\lambda_n}{2} c_{\min} \|\tilde{\beta} - \beta^0\|_1 \geq O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h), \tag{3.36}$$

or **Case (ii)** where,

$$\frac{\lambda_n}{2} c_{\min} \|\tilde{\beta} - \beta^0\|_1 \leq O\Big(\gamma_{max} \frac{s^{3/2}}{h^2} \sqrt{\frac{2 \log 2p}{n}}\Big) + O(h). \tag{3.37}$$

**Proof of Theorem 3.3.1.** First, we prove error bounds for $\tilde{\beta}$, which, in view of (3.30), shall be a precursor to obtaining error bounds for $\hat{\beta}$.

Suppose **Case (i)** (3.36) holds. The fact $\|\tilde{\beta} - \beta^0\|_1 = \|(\tilde{\beta} - \beta^0)_S\|_1 + \|\tilde{\beta}_{S^c}\|_1$ and (3.35) imply

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

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 (3.15) is satisfied for $\delta = \tilde{\beta} - \beta^0$. Now Lemma 3.3.2, the triangle inequality $\|d \circ \tilde{\beta}\|_1 \geq \|d \circ \beta^0\|_1 - \|\big(d \circ (\tilde{\beta} - \beta^0)\big)_S\|_1 + \|(d \circ \tilde{\beta})_{S^c}\|_1$, (3.34), and (3.36) together yield

$$\frac{2c_a}{n} \|\Gamma^{1/2} X(\tilde{\beta} - \beta^0)\|_2^2 + \lambda_n c_{\min} \|\tilde{\beta}_{S^c}\|_1 \leq \lambda_n c_{\min} c_0 \|\tilde{\beta}_S - \beta^0_S\|_1. \tag{3.38}$$

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

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

Here the first inequality follows from (3.38), the second from the Compatibility condition in (3.15), and the third using the identity $4uv \leq u^2 + 4v^2$. Thus

$$3c_a n^{-1}\|\Gamma^{1/2}X(\tilde{\beta}-\beta^0)\|_2^2 + 2\lambda_n c_{\min}\|\tilde{\beta}-\beta^0\|_1 \leq \frac{4\lambda_n^2 c_m^2 s}{\phi^2}. \tag{3.39}$$

Now we consider **Case (ii)**. From (3.34) we obtain,

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

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

$$(3.40)\ 3c_a n^{-1}\|\Gamma^{1/2}X(\tilde{\beta}-\beta^0)\|_2^2 + 2\lambda_n c_{\min}\|\tilde{\beta}-\beta^0\|_1 = O\Big(\gamma_{max}\frac{s^{3/2}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big) + O(h).$$

Hence from (3.39) and (3.40) 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 (3.16), 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,

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

we have by the rate assumptions (3.16), $\kappa_n\alpha_n \to 0$, and hence

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

This along with the construction of $\tilde{\beta}$ and (3.30) 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 (3.17), thereby completing the proof of Theorem 3.3.1. □

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 (3.18). Thus with $\alpha_n$ as in (3.18), with probability $1 - o(1)$,

$$\hat{\beta} \in \mathcal{B}(\alpha_n). \tag{3.41}$$

We now proceed to the proofs of Theorem 3.2.1 and Lemma 3.5.1. For this purpose we first state some facts about the first two summands in the loss function $\rho_L^*$ of (3.4). These facts are consequences of the properties of normal kernel density. Let, for $s$, $y \in \mathbb{R}$,

$$\begin{aligned} l(s,y) &= (y-s)(\tau-1) + (y-s)H\big(\frac{y-s}{h}\big), \\ l'(s,y) &= \tau - 1 + H\big(\frac{y-s}{h}\big) + \frac{y-s}{h}K\big(\frac{y-s}{h}\big), \\ l''(s,y) &= \frac{2}{h}K\big(\frac{y-s}{h}\big) + \frac{y-s}{h^2}K'\big(\frac{y-s}{h}\big), \\ l'''(s,y) &= \frac{3}{h^2}K'\big(\frac{y-s}{h}\big) + \frac{y-s}{h^3}K''\big(\frac{y-s}{h}\big). \end{aligned} \tag{3.42}$$

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 $C > 0$,

$$\begin{aligned} &\text{(i) } |l(s_1,y) - l(s_2,y)| \le C|s_1 - s_2|, && \text{(ii) } |l'(s_1,y) - l'(s_2,y)| \le \frac{C}{h^1}|s_1 - s_2|, \\ &\text{(iii) } |l''(s_1,y) - l''(s_2,y)| \le \frac{C}{h^2}|s_1 - s_2|, && \text{(iv) } |l'''(s_1,y) - l'''(s_2,y)| \le \frac{C}{h^3}|s_1 - s_2|. \end{aligned} \tag{3.43}$$

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 3.2.1.** 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, $\rho^\star_{Li}(\beta) = l_i(\beta) - \frac{\sigma^2_\beta}{2} l''_i(\beta)$. Now,

$$
\begin{aligned}
(3.44) \qquad M^\star_n(\beta) - EM^\star_n(\beta) \;&=\; \frac{1}{n}\sum_{i=1}^n \Big(l_i(\beta) - l_i(\beta^0) - E\big(l_i(\beta) - l_i(\beta^0)\big)\Big) \\
&\quad -\frac{1}{n}\frac{\sigma^2_\beta}{2}\sum_{i=1}^n \Big(l''_i(\beta) - l''_i(\beta^0) - E\big(l''_i(\beta) - l''_i(\beta^0)\big)\Big) \\
&\quad +\frac{1}{n}\frac{\sigma^2_{\beta^0} - \sigma^2_\beta}{2}\sum_{i=1}^n \Big(l''_i(\beta^0) - E\big(l''_i(\beta^0)\big)\Big) \\
&\le\; I - \frac{\sigma^2_\beta}{2} II + \frac{\sigma^2_{\beta^0} - \sigma^2_\beta}{2} III, \quad \text{say.}
\end{aligned}
$$

We shall show that

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

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

Observe that for $\beta \in \Theta$, $\sigma^2_\beta = \beta^T\Sigma\beta \le \gamma_{\max} b_0 s$ and $|\sigma^2_{\beta^0} - \sigma^2_\beta| \le 2b_0\gamma_{\max} s$. 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 $A = \big\{ \max_{1\le j\le p} \frac{1}{n}\sum_{i=1}^n u^2_{ij} \le$

$c_u\}$,

$$\text{(3.45)} \qquad \frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 \le \frac{2}{n}\sum_{i=1}^{n}(x_{ij}^2 + u_{ij}^2) \le 2(c_x + c_u).$$

This bound and the Lipschitz condition (3.43)(i) allow us to apply Lemma 14.20 and Theorem 14.2 as done in Example 14.2 of Bühlmann and Van de Geer (2011) page 503, to yield

$$\begin{aligned} &E\Big(Z_n I_A\Big) \le 32c_1 b_0(c_x + c_u)\sqrt{s}\sqrt{\frac{2\log 2p}{n}}, \\ &P\Big(Z_n I_A \ge 8c_1 b_0(c_x + c_u)\sqrt{s}\Big(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2t}{n}}\Big)\Big) \le \exp\big(-t\big), \end{aligned}$$

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

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

Now to remove the truncation of $Z_n$ on the set $A$, observe that (3.43)(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,

$$\text{(3.46)} \qquad Z_n \le Z_n I_A + c\sqrt{s}(\kappa_n + \max_{ij}|u_{ij}|)I_{A^c} + c\sqrt{s}E\Big((\kappa_n + \max_{ij}|u_{ij}|)I_{A^c}\Big).$$

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 Jhonstone (2001) to obtain

$$
\begin{aligned}
(3.47) \qquad P\Big(A^c\Big) \;&\leq\; \sum_{j=1}^{p} P\Big(\Big\{\frac{1}{n}\sum_{i=1}^{n}|u_{ij}|\Big\}^2 \geq c_u^2\Big) \leq \sum_{j=1}^{p} P\Big(2\frac{1}{n}\sum_{i=1}^{n}\frac{|u_{ij}|}{\sigma_{jj}} \geq 2.8\Big) \\
&=\; \sum_{j=1}^{p} P\Big(\chi^2_{2n} \geq n2.8\Big) \leq \sum_{j=1}^{p} P\Big(|\chi^2_{2n} - 2n| \geq 2n(0.4)\Big) \\
&\leq\; \sum_{j=1}^{p} \exp\Big(\frac{-3n}{100}\Big) \leq \exp\Big(\frac{-3n}{100} + \log p\Big).
\end{aligned}
$$

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

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

Thus, using this bound, (3.47), and the Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
&\text{(a)} \quad P\Big((\max_{ij}|u_{ij}|)I_{A^c} > n^{-k}\Big) \leq n^k E\Big((\max_{ij}|u_{ij}|)I_{A^c}\Big) \leq n^k\sqrt{npc_u \exp\Big(\frac{-3n}{100} + \log p\Big)}, \\
&\text{(b))} \quad E\Big((\max_{ij}|u_{ij}|)I_{A^c}\Big) \leq \sqrt{E\Big((\max_{ij}|u_{ij}|)^2\Big)P(A^c)} \leq \sqrt{npc_u \exp\Big(\frac{-3n}{100} + \log p\Big)}.
\end{aligned}
$$

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 (3.46) 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 (3.46) decreases to zero at an exponential rate. Thus, with probability at least $1 - o(1)$, the remainder two summands in the bound in (3.46) decrease to zero, in probability, faster than $Z_n I_A$. Hence,

$$\sup_{\beta\in\Theta} |I| = Z_n = O_p\Big(\sqrt{s}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{3.48}$$

We can similarly obtain a bound for term $II$ of (3.44). 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, (3.43)(ii) along with the bound (3.45) allow us to apply Lemma 14.20 and Theorem 14.2 of Bühlmann and Van de Geer (2011), page 503, which yields

$$E\Big(\tilde{Z}_n I_A\Big) \le 32c_3b_0(c_x + c_u)\frac{\sqrt{s}}{h^2}\sqrt{\frac{2\log 2p}{n}}, \quad \text{and}$$
$$P\Bigg(Z_n I_A \ge 8c_3b_0(c_x + c_u)\frac{\sqrt{s}}{h^2}\Big(4\sqrt{\frac{2\log 2p}{n}} + \sqrt{\frac{2t}{n}}\Big)\Bigg) \le \exp\big(-t\big), \quad \forall\, t > 0.$$

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

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

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

$$\sup_{\beta\in\Theta} |II| = \tilde{Z}_n = O_p\Big(\frac{\sqrt{s}}{h^2}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{3.49}$$

Lastly, consider the term $III$ in (3.44). Observe that $|l_i''(\beta^0)| \le ch^{-1}$, for $c < \infty$. Then

Lemma 14.11 of Bühlmann and Van de Geer (2011) yields

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

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

$$|III| = \frac{1}{n}\big|\sum_{i=1}^{n}\big(l_i''(\beta^0) - El_i''(\beta^0)\big)\big| = O_p\Big(h^{-1}\sqrt{\frac{\log 2p}{n}}\Big). \tag{3.50}$$

Now use (3.48), (3.49) and (3.50) in (3.44), and the fact that the rate of decrease of (3.49) is the slowest, to conclude (3.9) of Theorem 3.2.1.

The proof of (3.10) similar. This completes the proof of Theorem 3.2.1. □

**Proof of Lemma 3.5.1.** Define $\rho(s, y_i) = \rho_\tau(y_i - s)$. Then observe that it satisfies the following Lipchitz condition,

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

Then proceed as in the proof of (3.9) of Theorem 3.2.1 to obtain the desired bound. □

**Proof of Theorem 3.3.2.** Let $\alpha_n$ be as defined in (3.18). By (3.41), $\hat{\beta} \in \mathcal{B}(\alpha_n)$, 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,

$$\frac{d}{d\beta_j}\Big(n^{-1}\sum_{i=1}^{n}\rho^{\star}_{Li}(\hat{\beta})\Big)=\lambda_n d_j sign(\hat{\beta}_j). \tag{3.51}$$

Recall (3.42). The first derivatives of $\rho^{\star}_{Li}(\beta)$ and $\rho_{Li}(\beta)$ w.r.t $\beta_j$ are

$$\begin{aligned} \rho^{\star'}_{Li,j}(\beta) &:= \frac{d}{d\beta_j}\rho^{\star}_{Li}(\beta)=-w_{ij}l'_i(\beta)+\frac{\sigma^2_{\beta}}{2}\big[l'''_i(\beta)\big]-w_{ij}\sum_{k=1}^{n}\sigma_{kj}\beta_j\big[l''_i(\beta)\big], \\ \rho'_{Li,j}(\beta) &= \frac{d}{d\beta_j}\rho_{Li}(\beta)=-x_{ij}\Big[\tau-1+H\big(\frac{\varepsilon_{i\beta}}{h}\big)+\frac{\varepsilon_{i\beta}}{h}K\big(\frac{\varepsilon_{i\beta}}{h}\big)\Big]. \end{aligned} \tag{3.52}$$

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

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

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

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

We will show the relations (3.53)-(3.53) below hold for all $j\in S^c$ simultaneously with probability at least $1 - o(1)$.

$$
\begin{aligned}
J_1 &:= n^{-1}\Big|\sum_{i=1}^{n}\big(\rho^{\star\prime}_{Li,j}(\beta^0) - \psi^{\star\prime}_{Li,j}(\beta^0)\big)\Big| = o(d_j\lambda_n), && (3.53)\\
J_2 &:= \sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi^{\star\prime}_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big)\Big| = O(\kappa_n h) = o(d_j\lambda_n), && (3.54)\\
J_3 &:= \sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi'_{i,j}(\beta) - \psi'_{i,j}(\beta^0)\big)\Big| = o(d_j\lambda_n), && (3.55)\\
J_4 &:= T^{*}_{\mathcal{B},j} = o(d_j\lambda_n). && (3.56)
\end{aligned}
$$

Lemma 3.5.2 proves (3.54) and (3.55) and Lemma 3.5.3 proves (3.53) and (3.56). Finally, combining (3.53)-(3.56), we obtain that for $n$ large, with probability $1 - o(1)$,

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

This contradicts the optimality condition (3.51), and also completes the proof of Theorem 3.3.2. □

**Lemma 3.5.2** *Under the conditions of Theorem 3.3.2 we have,*

$$
\max_{1\le j\le p,\beta\in\mathbb{R}^p} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi^{\star\prime}_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big)\Big| = O(\kappa_n h) = o(\lambda_n d^{S^c}_{\min}), \tag{3.57}
$$

$$
\max_{j\in S^c}\sup_{\beta\in\mathcal{B}(\alpha_n)} n^{-1}\Big|\sum_{i=1}^{n}\big(\psi'_{i,j}(\beta) - \psi'_{i,j}(\beta^0)\big)\Big| = o(\lambda_n d^{S^c}_{\min}). \tag{3.58}
$$

**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,

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

But

$$(3.59) \qquad \begin{aligned} \psi'_{Li,j}(\beta) - \psi'_{ij}(\beta) &= -x_{ij} E\Big( H\big(\frac{\varepsilon_{i\beta}}{h}\big) - \mathbf{1}\{\varepsilon_{i\beta} > 0\} + \frac{\varepsilon_{i\beta}}{h} K\big(\frac{\varepsilon_{i\beta}}{h}\big)\Big) \\ &= -x_{ij} E\Big( H\big(-|\frac{\varepsilon_{i\beta}}{h}|\big) + \frac{\varepsilon_{i\beta}}{h} K\big(\frac{\varepsilon_{i\beta}}{h}\big)\Big). \end{aligned}$$

Now,

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

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 (3.59) we obtain,

$$(3.60) \qquad \sup_{1\le i\le n, 1\le j\le p, \beta\in\mathbb{R}^p} \big|\psi'_{Li,j}(\beta) - \psi'_{i,j}(\beta)\big| = O(\kappa_n h) = o(\lambda_n d_{min}^{S^c}).$$

The last equality follows from the rate assumptions (3.20). This bound and assumption (A2) completes the proof of (3.57).

Next, we show (3.58). By assumption (A1),

$$(3.61) \qquad \begin{aligned} &\psi'_{i,j}(\beta) - \rho'_{i,j}(\beta^0)) \\ &\qquad = -x_{ij}\Big[F_i\big(x_i'(\beta-\beta^0)\big) - F_i(0)\Big] = -x_{ij} f_i(0) x_i^T(\beta-\beta^0) - x_{ij}\tilde{I}_i, \end{aligned}$$

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 $j \in S^c$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij} f_i(0) x_i'(\beta - \beta^0)\Big| \leq \Big\|\frac{1}{n} x_{ij} f_i(0) x_i'\Big\|_\infty \|\beta - \beta^0\|_1 = O(\alpha_n) = o(\lambda_n d_{\min}^{S^c}), \tag{3.62}$$

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

$$\begin{aligned} \max_{j \in S^c}\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\tilde{I}_i\Big| &\leq \frac{\kappa_n}{n}\sum_{i=1}^{n}\tilde{I}_i \leq C_2 \frac{\kappa_n}{n}\sum_{i=1}^{n}\big(x_i'(\beta - \beta^0)\big)^2. \\ &\leq C\kappa_n^3 \|\beta - \beta^0\|_1^2 = O\big(\kappa_n^3 \alpha_n^2\big) = o(\lambda_n d_{\min}^{S^c}). \end{aligned} \tag{3.63}$$

Now use (3.61)–(3.63) to obtain (3.58), thereby completing the proof of the lemma. □

**Lemma 3.5.3** *Under the conditions of Theorem 3.3.2,*

$$\max_{j \in S^c} \frac{1}{n}\Big|\sum_{i=1}^{n}\big(\rho_{Li,j}^{\star\prime}(\beta^0) - \psi_{Li,j}^{\star\prime}(\beta^0)\big)\Big| = o_p(\lambda_n d_{\min}^{S^c}) \tag{3.64}$$

$$\max_{j \in S^c} T_{\mathcal{B},j}^{*} = o_p(\lambda_n d_{\min}^{S^c}). \tag{3.65}$$

**Proof** The structure of this proof is similar to the proof of Theorem 3.2.1. 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

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

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 (3.47) 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,

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

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

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

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 (2011).

Thus choosing $t = cn^{\delta}\sqrt{2\log 2p/n}$, for some constant $c > 0$, we obtain

$$\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij} l_i'(\beta^0) I_{\mathcal{A}}\Big| = O_p\Big(n^{\delta}\sqrt{\frac{2\log 2p}{n}}\Big). \tag{3.66}$$

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

$$\begin{aligned}\max_{1\le j\le p}|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)| \;\le\; & \max_{1\le j\le p}|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0) I_{\mathcal{A}}| + \max_{1\le j\le p} c(\kappa_n + \max_{i,j}|u_{ij}|)\mathbf{1}_{A^c} \\ & + c\max_{1\le j\le p} E\Big((\kappa_n + \max_{i,j}|u_{ij}|)\mathbf{1}_{A^c}\Big)\end{aligned} \tag{3.67}$$

Proceed as in the proof of Theorem 3.2.1 to show that the last two terms on the RHS converge to zero faster than the first term, in probability. Thus we obtain,

$$\frac{1}{n}\Big|\sum_{i=1}^{n} w_{ij}\gamma_i'(\beta^0)\Big| = O_p\Big(n^{\delta}\sqrt{\frac{2\log 2p}{n}}\Big), \tag{3.68}$$

A similar argument yields that

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

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 (3.20) to obtain (3.64).

To prove claim (3.65), note that

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

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 (3.42) and (3.43) we have,

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

Then as in the above proof of Theorem 3.2.1, apply Lemma 14.2 and Theorem 14.2 of

Bühlmann and Van de Geer (2011 to obtain

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

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 3.2.1, we obtain

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

where the last equality follows by the rate assumption (3.20). A similar argument applied to the terms $I$ and $IV$ yields that

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

An argument similar to the one used for proving (3.64) yields

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

Now claim (3.65) follows from these bounds, the rate condition $(iii)$ of (3.20), 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 3.5.3.

# BIBLIOGRAPHY

# BIBLIOGRAPHY

[1] AGARWAL, A., NEGHBAN, S. AND WAINWRIGHT, M.J. (2012). FAST GLOBAL CONVERGENCE OF GRADIENT METHODS FOR HIGH DIMENSIONAL STATISTICAL RECOVERY. *ANN. STATIST.*, **40**, 2452–2482.

[2] ALQUIER, P. AND DOUKHAN, P. (2011). SPARSITY CONSIDERATIONS FOR DEPENDENT VARIABLES *ELECTRONIC JOURNAL OF STATISTICS*, **5**, 750-774.

[3] BAILLIE, R.T. (1996). LONG MEMORY PROCESSES AND FRACTIONAL INTEGRATION IN ECONOMETRICS *JOURNAL OF ECONOMETRICS*, **73**, 735-59.

[4] BELLONI, A. AND CHERNOZHUKOV, V. (2011) $\ell_1$- PENALIZED QUANTILE REGRESSION IN HIGH DIMENSIONAL SPARSE MODELS, *ANN. OF STATIST.*, **39**, 82–130.

[5] BERAN, J., FENG, Y., GHOSH, S., KULIK, R (2013). LONG MEMORY PROCESSES, *SPRINGER*

[6] BERAN, J. (1992). STATISTICAL METHODS FOR DATA WITH LONG-RANGE DEPENDENCE, *STATISTICAL SCIENCE* **7**, 404-427.

[7] BICKEL, P., RITOV, Y. AND TSYBAKOV, A. (2009). SIMULTANEOUS ANALYSIS OF LASSO AND DANTZIG SELECTOR, *ANN. STATIST.*, **37** 1705–1732.

[8] BUCHINSKY, M. (1994). CHANGES IN THE U.S. WAGE STRUCTURE 1963-1987: APPLICATIONS OF QUANTILE REGRESSION *ECONOMETRICA*, **62** 405–458.

[9] BÜHLMANN, P. AND VAN DE GEER, S. (2011). *STATISTICS FOR HIGH DIMENSIONAL DATA*. SPRINGER, NEW YORK.

[10] CARROLL, R.J., RUPPERT, D., STEFANSKI, L.A. AND CRAINICEANU, C. (2006). *MEASUREMENT ERROR IN NONLINEAR MODELS: A MODERN PERSPECTIVE.* NEW YORK: CHAPMAN AND HALL.

[11] DAHLHAUS, R. (1995). EFFICIENT LOCATION AND REGRESSION ESTIMATION FOR LONG RANGE DEPENDENT REGRESSION MODELS, *ANN. STATIST.*, **23** 1029-1047.

[12] DOUKHAN, P. (1994). MIXING: PROPERTIES AND EXAMPLES, *SPRINGER*

[13] DUCHI, J., SHALEV-SHWARTZ, S., SINGER, Y. AND CHANDRA, T.(2008). EFFICIENT PROJECTIONS ONTO THE $\ell_1$-BALL FOR LEARNING IN HIGH DIMENSIONS. *INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 272-279.* ACM, NEW YORK, NY.

[14] DUREETT, R. (2005). PROBABILITY: THEORY AND EXAMPLES, (THIRD EDITION) *DUXBURY*

[15] FAN, J., FAN, Y., BARUT, E. (2014). ADAPTIVE ROBUST VARIABLE SELECTION *ANN. OF STATIST.*, **42**, 324-351.

[16] FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2010). REGULARIZATION PATHS FOR GENERALIZED LINEAR MODELS VIA COORDINATE DESCENT. *J. STAT. SOFTW.*, **33**, 1–22.

[17] FULLER, W.A. (1987). *MEASUREMENT ERROR MODELS*. WILEY, NEW YORK.

[18] GIRAITIS, L., KOUL, H., SURGAILIS, D. (2012). LARGE SAMPLE INFERENCE FOR LONG MEMORY PROCESSES, *IMPERIAL COLLEGE PRESS*

[19] GUO, H. AND KOUL, H. (2007). NONPARAMETRIC REGRESSION WITH HETEROSCEDASTIC LONG MEMORY ERRORS, *JOURNAL OF STATISTICAL PLANNING AND INFERENCE*, **137** 379-404.

[20] JOHNSTONE, I.M. (2001). CHI-SQUARE ORACLE INEQUALITIES. *STATE OF THE ART IN PROBABILITY AND STATISTICS* (LEIDEN, 1999), 399-418, IMS LECTURE NOTES MONOGR. SER., **36**, IMS, BEACHWOOD, OH.

[21] KNIGHT, K. AND FU, W. (2000). ASYMPTOTICS FOR LASSO-TYPE ESTIMATORS, *ANN. STATIST.*, **28** 1356-1378.

[22] LEE, R., NOH, H. AND PARK, B. (2014). MODEL SELECTION VIA BAYESIAN INFORMATION CRITERION FOR QUANTILE REGRESSION MODELS. *J. AMER. STATIST. ASSOC.*, **109**, 216-229.

[23] LOH, P., AND WAINWRIGHT, M.J. (2012). HIGH-DIMENSIONAL REGRESSION WITH NOISY AND MISSING DATA: PROVABLE GUARANTEES WITH NONCONVEXITY. *ANNALS OF STATIST.*, **40**, 1637–1664.

[24] MCKENZIE, H., JERDE, C. VISSCHER, D., MERRILL, E. AND LEWIS, M. (2009). INFERRING IN THE PRESENCE OF GPS MEASUREMENT ERROR. *ENVIRON. ECOL. STAT.*, **16**, 531–546.

[25] MEINHAUSEN, N. AND BÜHLMANN, P. (2006). HIGH DIMENSIONAL GRAPHS AND VARIABLE SELECTION WITH THE LASSO, *ANN. STATIST.*, **34** 1436-1462.

[26] NOLAN, D. AND POLLARD, D. (1987). U-PROCESSES: RATES OF CONVERGENCE, *ANN. STATIST.*, **15** 780-799.

[27] POLLARD, D. (1984). CONVERGENCE OF STOCHASTIC PROCESSES. SPRINGER, NEW YORK.

[28] PURDOM, E. AND HOLMES, S. (2005). ERROR DISTRIBUTION FOR GENE EXPRESSION DATA. *STATIST. APPL. IN GENETICS AND BIOLOGY,* **4**, ARTICLE 16.

[29] RASKUTTI, G., WAINWRIGHT, M., YU, B. (2010). RESTRICTED EIGENVALUE PROPERTIES FOR CORRELATED GAUSSIAN DESIGNS, *JOURNAL OF MACHINE LEARNING RESEARCH*, **99**, 2241–2259.

[30] ROSENBAUM, M. AND TSYBAKOV, A.B. (2010). SPARSE RECOVERY UNDER MATRIX UNCERTAINTY *ANNALS OF STATIST.*, **38** 2620–2651.

[31] ROSENBAUM, M. AND TSYBAKOV, A.B. (2011). IMPROVED MATRIX UNCERTAINTY SELECTOR *TECHNICAL REPORT.* AVAILABLE AT HTTP://ARXIV.ORG/ABS/1112.4413.

[32] SORENSEN, O., FRIGESSI, A. AND THORESEN, M. (2012). MEASUREMENT ERROR IN LASSO: IMPACT AND LIKELIHOOD BIAS CORRECTION. AVAILABLE AT HTTP://ARXIV.ORG/PDF/1210.5378.PDF.

[33] STEFANSKI, L.A., CARROLL, R.J. (1990). DECONVOLUTING KERNEL DENSITY ESTIMATORS. *STATISTICS*, **21**, 169–184.

[34] TIBSHIRANI, R. (1996). REGRESSION SHRINKAGE AND SELECTION VIA THE LASSO, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, **58** 267288.

[35] WANG, H., LI, R. AND TSAI, C.L. (2007). TUNING PARAMETER SELECTORS FOR SMOOTHLY CLIPPED ABSOLUTE DEVIATION METHOD *BIOMETRIKA*, **3**, 553–668.

[36] WANG, H., STEFANSKI, L.A., AND ZHU, Z. (2012). CORRECTED-LOSS ESTIMATION FOR QUANTILE REGRESSION WITH COVARIATE MEASUREMENT ERRORS. *BIOMETRIKA*, **99**, 405-421.

[37] WANG, L., WU, Y., LI, R. (2012). QUANTILE REGRESSION FOR ANALYZING HETEROGENEITY IN ULTRA-HIGH DIMENSION. *J. AMER. STATIST. ASSOC.*, **107**, 214-222.

[38] YOON, Y., PARK, C., AND LEE, T. (2013), PENALIZED REGRESSION MODELS WITH AUTOREGRESSIVE ERROR TERMS, *JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION*, **83**, 1756-1772.

[39] ZHANG, Y., LI, R. AND TSAI, C.L. (2010). REGULARIZATION PARAMETER SELECTIONS VIA GENERALIZED INFORMATION CRITERION. *J. AMER. STATIST. ASSOC.*, **105**, 312–323.

[40] ZHAO, P. AND YU, B. (2006). ON MODEL SELECTION CONSISTENCY OF LASSO, *JOURNAL OF MACHINE LEARNING RESEARCH*, **7**, 2541–2563.

[41] ZOU, H. (2006). THE ADAPTIVE LASSO AND ITS ORACLE PROPERTIES, *J. AMER. STATIST. ASSOC.*, **101**, 1418-1429.