High-Dimensional Structured Quantile Regression

Vidyashankar Sivakumar [1] **Arindam Banerjee** [1]

Abstract

Quantile regression aims at modeling the conditional median and quantiles of a response variable given certain predictor variables. In this work we consider the problem of linear quantile regression in high dimensions where the number of predictor variables is much higher than the number of samples available for parameter estimation. We assume the true parameter to have some structure characterized as having a small value according to some atomic norm $R(\cdot)$ and consider the norm regularized quantile regression estimator. We characterize the sample complexity for consistent recovery and give non-asymptotic bounds on the estimation error. While this problem has been previously considered, our analysis reveals geometric and statistical characteristics of the problem not available in prior literature. We perform experiments on synthetic data which support the theoretical results.

1 Introduction

Considerable advances have been made over the past decade on fitting high-dimensional structured linear models when the number of samples n is much smaller than the ambient dimensionality p (Banerjee et al., 2014; Negahban et al., 2012; Chandrasekaran et al., 2012). Most of the advances have been made for linear models: $y_i = \langle \mathbf{x}_i, \theta^* \rangle + \omega_i, i = 1, \ldots, n$, where $\theta^* \in \mathbb{R}^p$ is assumed to be structured, e.g., sparse, group sparse, etc. Estimation of such structured θ is usually done using Lasso-type regularized estimators (Negahban et al., 2012; Banerjee et al., 2014) or Dantzig-type constrained estimators (Chandrasekaran et al., 2012; Chatterjee et al., 2014); other related estimators have also been explored (Hsu & Sabato, 2016; Vershynin, 2015). Such models have been extended to generalized linear models (Banerjee et al., 2014; Negahban et al., 2012), matrix completion (Candès & Recht, 2009), vector auto-regressive models (Melnyk & Banerjee, 2016) among others.

In this paper, we consider the problem of structured quantile regression in high-dimensions, which can be posed as follows: given the response variable y_i and covariates x_i the τth conditional quantile function of y_i given x_i is given by: $F^{-1}_{y_i|x_i}(\tau|x_i) = \langle x_i, \theta^*_\tau \rangle, \tau \in (0,1)$ for some structured θ^*_τ whose structure can be captured by a suitable atomic norm $R(\cdot)$, e.g., l_1-norm for sparsity, l_1/l_2 norm for group sparsity, etc. Here $F^{-1}_{y_i|x_i}(\cdot)$ is the inverse of the conditional distribution function of y_i given x_i. We consider the following regularized estimator for the structured quantile regression problem:

$$\begin{aligned} \hat{\theta}_n &:= \text{argmin}_{\theta \in \mathbb{R}^p} \, \mathcal{L}_\tau(\theta) + \lambda_n R(\theta) \\ &:= \text{argmin}_{\theta \in \mathbb{R}^p} \, \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \theta \rangle) + \lambda_n R(\theta) \, , \end{aligned} \tag{1}$$

where $\rho_\tau(u) = (\tau - \mathbb{I}(u \leq 0))u$ is the asymmetric absolute deviation function (Koenker, 2005), $\mathbb{I}(\cdot)$ is the indicator function. The goal is to get nonasymptotic bounds on the estimation error $\|\hat{\theta} - \theta^*\|_2$.

Many previous papers analyze the asymptotic performance of the estimator in (1) (Li & Zhu, 2008; Kai et al., 2011; Wu & Liu, 2009; Zou & Yuan, 2008; Wang et al., 2012). In the non-asymptotic setting of interest in this paper, special cases of the estimator in (1) have been studied in recent literature (Belloni & Chernozhukov, 2011; Kato, 2011; Fan et al., 2014a), primarily focusing on specific norms like the l_1-norm and nonoverlapping group sparse norm. In contrast, our formulation and analysis is applicable to any atomic norm $R(\cdot)$ giving considerable flexibility in choosing a suitable structure for real world problems, e.g., hierarchical sparsity, k-support norm, OWL norm etc. More recently (Alquier et al., 2017) consider the more general problem of norm regularized regression with Lipschitz loss functions which includes (1) as a special case. They derive similar results to ours for bounds on the estimation error, but their analysis differs significantly and, in our opinion, does not leverage and highlight the key geometric and sta-

[1]Department of Computer Science & Engineering, University of Minnesota, Twin Cities. Correspondence to: Vidyashankar Sivakumar <sivak017@umn.edu>, Arindam Banerjee <banerjee@cs.umn.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

tistical characteristics of the problem.

In the setting of norm regularized regression with square loss, including the widely studied Lasso estimator (Tibshirani, 1996; Negahban et al., 2012; Banerjee et al., 2014), the sample complexity n_0 of the estimator gets determined by a certain restricted strong convexity (RSC) property which simplifies to the restricted eigenvalue (RE) condition on the matrix $X^T X$ (Bickel et al., 2009; Negahban et al., 2012); in the noiseless setting, i.e., when $\omega_i = 0$, the sample complexity determines a phase transition phenomenon so that the probability of recovering the structured θ^* is minimal when $n < n_0$, and one can exactly recover θ^* with high probability when $n > n_0$. Our work gives an equivalent sample complexity characterization for structured quantile regression, which was not available in prior work. The challenge in characterizing RSC in the context of quantile regression stems partly from the non-smoothness of the objective, so one has to work with sub-gradients. However, the unique aspect stems from the geometry of quantile regression, or as the authoritative book on the topic puts it: "How quantile regression works?" (Koenker, 2005)[Section 2.2]. In quantile regression, the n samples get divided into three subsets: ν samples which get exactly interpolated, i.e., $y_i = \langle x_i, \hat{\theta} \rangle$, $(n-\nu)\tau$ samples which lie below the curve, i.e., $y_i < \langle x_i, \hat{\theta} \rangle$, and $(n-\nu)(1-\tau)$ samples which lie above the curve, i.e., $y_i > \langle x_i, \hat{\theta} \rangle$. Note that when $\nu = n$ all samples are interpolated, the loss is zero and the same $\hat{\theta}$ is a solution for all quantiles τ. Quantile regression is clearly not working. The Number of InterPolated Samples (NIPS) ν is an important quantity, inherent to structure in θ^*, and determines the sample complexity of structured quantile regression (1). In fact, we show that when $n > \nu$, the RSC condition associated with the estimator in (1) is satisfied. When there is no structure in θ^*, then $\nu = O(p)$, and quantile regression needs $n > O(p)$ samples to work. However, when θ^* has structure, such as sparsity or group sparsity, ν can be substantially smaller than p. Specifically we show that ν is of the order of square of Gaussian width of the error set (Talagrand, 2014; Chandrasekaran et al., 2012) for a class of atomic norms which includes l_1, l_1/l_2 group sparse, k-support (Argyriou et al., 2012) and the OWL (Bogdan et al., 2013) norms. The Gaussian width as a measure of complexity of a set has been extensively used in prior literature (Chandrasekaran et al., 2012; Tropp, 2015; Banerjee et al., 2014). For example, when θ^* is sparse with s non-zero entries, we show that $\nu = cs \log p$.

When $n > \nu$ and the RSC condition is satisfied, building on recent developments in high-dimensional estimation (Negahban et al., 2012; Banerjee et al., 2014), we show that choosing $\lambda_n \geq 2R^*(\nabla \mathcal{L}_\tau(\theta^*)))$, where $R^*(\cdot)$ is the dual norm of $R(\cdot)$, leads to non-asymptotic bounds on the estimation error $\|\hat{\theta}_n - \theta^*\|_2$. While the specification of λ_n looks complex, with its dependency on the dual norm and its dependency on θ^*, we show it is sufficient to set λ_n based on the Gaussian width (Talagrand, 2014) of the unit norm ball for $R(\cdot)$ (Banerjee et al., 2014; Sivakumar et al., 2015)'. Our analysis and results on the estimation error bound for quantile regression, interestingly, has the same order as that for regularized least squares regression for general norms (Banerjee et al., 2014). In contrast to the least squares loss the quantile loss is more robust as the estimation error is independent of the two norm of the noise. We discuss results for the l_1, l_1/l_2 group sparse and k-support norms as examples, precisely characterizing the sample complexity for recovery and non-asymptotic error bounds. Specifically, our results for the l_1-norm matches those from existing literature on sparse quantile regression (Belloni & Chernozhukov, 2011).

The rest of the paper is organized as follows. In Section 2, we discuss the problem formulation along with assumptions, review the general framework for analyzing regularized estimation problems and discuss the three atomic norms used as examples throughout the paper. In Section 3, we analyze the number of interpolated samples and establish precise sample complexities for a class of atomic norms in terms of the Gaussian widths of sets. In Section 4 we establish key ingredients of the analysis and provide the main bound. We present experimental results in Section 5 and conclude in Section 6.

2 Background and Preliminaries

Problem formulation: We outline the assumptions on the data and estimator. Similar conditions are present in all prior literature on quantile regression and we refer to Section 2.5 in Belloni & Chernozhukov (2011) for examples of data satisfying the conditions.

We consider data is generated as $y = X\theta + \omega$, $X \in \mathbb{R}^{n \times p}$ is the design matrix, $\theta \in \mathbb{R}^p$ and $\omega \in \mathbb{R}^n$ is the noise. We assume subGaussian design matrices $X \in \mathbb{R}^{n \times p}$ which includes the class of all bounded random variables. Note that this is not a restrictive assumption as to avoid the quantile crossing phenomenon the covariate space has to be bounded. Quantile crossing is when the value of the τ_1th quantile is greater than the τ_2th quantile for some $\tau_1 < \tau_2$ (See Section 2.5 in (Koenker, 2005)). We do not make any assumptions on the noise vector $\omega \in \mathbb{R}^n$. More specifically the noise can be sampled from a heavy tailed distribution, can be heteroscedastic as in the location-scale model where $\omega_i = \langle x_i, \eta \rangle \cdot \epsilon_i, \eta \in \mathbb{R}^p$, ϵ_i is noise independent of x_i, bimodal and so on and so forth. Note that this setting is more general than for the least squares loss.

We consider a parametric quantile regression model where the τth conditional quantile function of the response vari-

able y_i given any $x_i \in \mathbb{R}^p$ is given by,

$$F^{-1}_{y_i|x_i}(\tau|x_i) = \langle x_i, \theta^*_\tau \rangle, \theta^*_\tau \in \mathbb{R}^p, \tau \in (0,1) \,, \quad (2)$$

where $F^{-1}_{y_i|x_i}$ is the inverse of the conditional distribution function of y_i given x_i. We will assume the conditional density of y_i evaluated at the conditional quantile $\langle x_i, \theta^*_\tau \rangle$ is bounded away from zero uniformly for all τ, that is, $f_{y_i|x_i}(\langle x_i, \theta^*_\tau \rangle) > \underline{f} > 0$ for all τ and all x_i. The goal is to estimate parameter $\hat{\theta}_\tau$ close to θ^*_τ using n observations of the data when $n < p$. The estimator in this paper belongs to the family of regularized estimators and is of the form:

$$\hat{\theta}_{\lambda_n,\tau} := \text{argmin}_{\theta \in \mathbb{R}^p} \mathcal{L}_\tau(\theta) + \lambda_n R(\theta) \,, \quad (3)$$

where $\mathcal{L}_\tau(\theta) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \langle x_i, \theta \rangle)$, $\rho_\tau(\cdot)$ is the quantile loss function and $R(\cdot)$ is any atomic norm. Examples of atomic norms we consider in this paper are the l_1, l_1/l_2 nonoverlapping group sparse norm and the k-support norm. We present all results assuming any single $\tau \in (0,1)$ and going forward drop the subscripts from θ^*_τ and $\hat{\theta}_{\lambda_n,\tau}$.

Gaussian Width: All results will be in terms of the Gaussian width of suitable sets. For any set $A \in \mathbb{R}^p$, the Gaussian width of the set A is defined as (Gordon, 1985; Chandrasekaran et al., 2012):

$$w(A) = E_g \left[\sup_{u \in A} \langle g, u \rangle \right] \,. \quad (4)$$

where the expectation is over $g \sim N(0, \mathbb{I}_{p\times p})$. Gaussian widths have been widely used in prior literature on structured estimation (Chandrasekaran et al., 2012; Banerjee et al., 2014; Sivakumar et al., 2015; Tropp, 2015).

High-dimensional estimation: Our analysis is built on developments over the past decade for high-dimensional structured regression for linear and generalized linear models using both regularized as well as constrained estimators (Candes & Tao, 2007; Bickel et al., 2009; Chandrasekaran et al., 2012; Negahban et al., 2012; Banerjee et al., 2014). For the regularized formulation considered in this work Banerjee et al. (2014); Negahban et al. (2012) have established a generalized analysis framework when the loss is least squares or more generally the maximum likelihood estimator for generalized linear models. We give a brief overview of the main components of the analysis.

1. Regularization parameter: In Banerjee et al. (2014); Negahban et al. (2012) the regularization parameter is assumed to satisfy the following assumption,

$$\lambda_n \geq 2R^*(\nabla \mathcal{L}_\tau(\theta^*)) \,. \quad (5)$$

With $\Omega_R = \{u | R(u) \leq 1\}$ denoting the unit norm ball, Banerjee et al. (2014) prove that with high probability a value $\lambda_n = O(w(\Omega_R))$ satisfies the above condition for subGaussian design matrices, noise and the least squares loss.

2. Error set: The assumption on the regularization parameter ensures that the error vector $\Delta = \hat{\theta} - \theta^*$ lies in the following error set (Banerjee et al., 2014),

$$\mathcal{C} := \left\{ \Delta \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{2}R(\Delta) \right\} \,. \quad (6)$$

3. The norm compatibility constant: It is defined as follows (Negahban et al., 2012; Banerjee et al., 2014),

$$\Psi_(\mathcal{C}) = \sup_{u \in \mathcal{C}} \frac{R(u)}{\|u\|_2} \,. \quad (7)$$

4. Restricted Strong Convexity (RSC): In Banerjee et al. (2014); Negahban et al. (2012) the loss function is shown to satisfy the following RSC condition with high probability once the number of samples is of the order of the square of the Gaussian width of the error set, that is, $n = O(w^2(\mathcal{C}))$.

$$\begin{aligned} \inf_{u \in \mathcal{C}} \delta\mathcal{L}(\theta^*, u) &= \inf_{u \in \mathcal{C}} (\mathcal{L}(\theta^* + u) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), u \rangle) \\ &\geq \kappa \|u\|^2 \,. \end{aligned} \quad (8)$$

For the squared loss, the RSC condition simplifies to the Restricted Eigenvalue (RE) condition (Bickel et al., 2009)

$$\inf_{u \in \mathcal{C}} \frac{1}{n}\|Xu\|_2^2 \geq \kappa \|u\|_2^2 \,. \quad (9)$$

5. Recovery Bounds: When RSC and bounds on the regularization parameter are satisfied Banerjee et al. (2014) prove the following deterministic error bound,

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq c\frac{\Psi(\mathcal{C}) w(\Omega_R)}{\kappa} \,. \quad (10)$$

where c is any constant.

Atomic Norms: We consider the class of atomic norms for the regularizer. Mathematically consider a set $\mathcal{A} \subseteq \mathbb{R}^p$ the collection of atoms that is compact, centrally symmetric about the origin (that is, $a \in \mathcal{A} \implies -a \in \mathcal{A}$). Let $\|\theta\|_\mathcal{A}$ denote the gauge of $\mathcal{A}$,

$$R(\theta) = \|\theta\|_\mathcal{A} = \inf\{t > 0 : \theta \in t conv(\mathcal{A})\} \quad (11)$$

$$= \inf\{\sum_{a \in \mathcal{A}} c_a : \theta = \sum_{a \in \mathcal{A}} c_a a, c_a \geq 0, \forall a \in \mathcal{A}\} \,. \quad (12)$$

For example when $\mathcal{A} = \{\pm e_i\}_{i=1}^p$ yields $\|\theta\|_\mathcal{A} = \|\theta\|_1$. Although the atomic set $\mathcal{A}$ may contain uncountably many vectors, we assume $\mathcal{A}$ can be decomposed as a union of m sets, $\mathcal{A} = \mathcal{A}_i \cup \mathcal{A}_2 \cup \ldots A_m$ similar to the setting considered

in Chen & Banerjee (2015). Such a decomposition assumption is satisfied by many popular atomic norms like the l_1, l_1/l_2 group sparse norms, k-support norm, OWL norm etc. Throughout the paper we will illustrate our results on the following norms.

1. l_1 **norm:** For the l_1 norm we will consider that θ^* is an s-sparse vector, that is, $\|\theta^*\|_0 = s$.

2. l_1/l_2 **nonoverlapping group sparse norm:** Let $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, G_{N_\mathcal{G}}\}$ denote a collection of groups which are blocks of any vector $\theta \in \mathbb{R}^p$. Let $\theta_{N_\mathcal{G}}$ denote a vector with coordinates $\theta^i_{N_\mathcal{G}} = \theta^i$ if $i \in \mathcal{G}_{N_\mathcal{G}}$, else $\theta^i_{N_\mathcal{G}} = 0$. The maximum size of any group is $l = \max_{i \in [1,\ldots,N_\mathcal{G}]} |\mathcal{G}_i|$. The norm is given as $R(\theta) = \sum_{i=1}^{N_\mathcal{G}} \|\theta_i\|_2$. Let $S_\mathcal{G} \subseteq \{1, 2, \ldots, N_\mathcal{G}\}$ with cardinality $|S_\mathcal{G}| = s_\mathcal{G}$. We consider the true parameter $\theta^* \in \mathbb{R}^p$ is $s_\mathcal{G}$-sparse, that is, $\theta^*_{N_\mathcal{G}} = \vec{0}, \forall N_\mathcal{G} \notin S_\mathcal{G}$.

3. k-support norm: The k-support norm can be expressed as an infimum convolution given by (Argyriou et al., 2012),

$$R(\theta) = \inf_{\sum_i u_i = \theta} \left\{ \sum_i \|u_i\|_2 \quad | \quad \|u_i\|_0 \leq k \right\} . \tag{13}$$

Clearly it is an atomic norm for which $\mathcal{A} = \{a \in \mathbb{R}^p \quad | \quad \|a\|_0 \leq k, \|a\|_2 = 1\}$ and $\mathcal{A}$ is a union of $\binom{p}{k}$ subsets, that is, $\mathcal{A} = \mathcal{A}_1 \cup A_2 \cup \ldots A_{\binom{p}{k}}$. More results on the k-support norm can be looked up in Chatterjee et al. (2014); Chen & Banerjee (2015). We consider the setting where $\|\theta^*\|_0 = s$ and $k < s$. For the results we require the Gaussian widths of the unit norm ball, error set and the norm compatibility constants for the norms. We provide them below for reference. All values are in order notation.

Norm	$w(\Omega_R)$	$w(\mathcal{C})$	$\Psi(\mathcal{C})$
l_1	$c\sqrt{\log p}$	$c\sqrt{s \log p}$	$c\sqrt{s}$
l_1/l_2	$c\sqrt{l + \log N_\mathcal{G}}$	$c\sqrt{l s_\mathcal{G} + s_\mathcal{G} \log N_\mathcal{G}}$	$c\sqrt{s_\mathcal{G}}$
k-sp	$c\sqrt{k + k \log \lceil \frac{p}{k} \rceil}$	$c\sqrt{s + s \log \lceil \frac{p}{k} \rceil}$	$c\sqrt{2s/k}$

3 Number of InterPolated Samples (NIPS)

We begin with intuitions on the geometry of the problem. In the high sample, low dimension setting $n >> p$, when $R(\theta) = 0$, the quantile loss is a linear program and hence the solutions are at the vertices, that is, where any p of the n samples are interpolated. Mathematically we define the quantity $Z = \{i : y_i = \langle x_i, \hat{\theta} \rangle = \langle x_i, \theta^* + u \rangle, u \in \mathbb{R}^p\}$ and note that $\nu = \sup_{u \in \mathbb{R}^p} |Z| = O(p)$. In the high dimensional setting considered in this paper $n < p$ and hence with $R(\theta) = 0$ the number of interpolated samples is $\nu = n$. From an optimization perspective there are multiple such solutions and all solutions are optimal for all quantile parameters τ. But practically quantile regression is not working. Now introducing a regularizer with a suitable choice for the regularization parameter ensures that the error vector lies in a restricted subset of $\mathbb{R}^p$, $\mathcal{C} := \left\{u \,\middle|\, R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2}R(u) \right\} \subseteq \mathbb{R}^p$. We are now interested in characterizing $\nu = \sup_{u \in \mathcal{C}} |Z|, Z = \{i : y_i = \langle x_i, \hat{\theta} \rangle = \langle x_i, \theta^* + u \rangle, u \in \mathcal{C}\}$, that is, the maximum number of interpolated samples with the error restricted to a particular subset of $\mathbb{R}^p$. Again if $\nu = n$, quantile regression is not working. Since there are no restrictions on the number of non-zero elements in the error vector u a first crude estimate will be $\nu \leq \min\{n, p, \|u\|_0\}$, which implies quantile regression will not work unless we have a minimum of p samples. But intuitively ν should depend on properties of the error set $\mathcal{C}$, which the initial crude estimate is failing to take advantage of.

Below we state a result which reinforces the intuition of the relation between ν and the properties of the set $\mathcal{C}$. Specifically we show that for the types of atomic norms considered in this work (which includes all popularly known vector norms) the number of interpolated samples does not exceed the product of the square of the norm compatibility constant and the square of the Gaussian width of the unit norm ball. For the norms considered, this is precisely the square of the Gaussian width of the error set $\mathcal{C}$. For example for the l_1 norm for an s-sparse vector this evaluates to an upper bound of $\nu = O(s \log p)$. While the result statement considers subGaussian design matrices, the result will also hold for design matrices sampled from heavy-tailed distributions using arguments similar to Lecué & Mendelson (2014); Sivakumar et al. (2015).

Theorem 3.1 *Consider X has isotropic subGaussian rows and θ^* is an s-sparse vector that can be written as a linear combination of k atoms from an atomic set of cardinality m,*

$$\theta^* = \sum_{i=1}^{k} c_i a_i, a_i \in \mathcal{A}, c_i \geq 0, |\mathcal{A}| = m \, . \tag{14}$$

For the regularized quantile regression problem penalized with the atomic norm $R(\theta) = \|\theta\|_\mathcal{A}$,

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \mathcal{L}_\tau(\theta) + \lambda R(\theta) = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(\theta) + \lambda R(\theta) \, , \tag{15}$$

let $\mathcal{C} := \left\{u | R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2}R(u)\right\}$ denote the error set, let $\lambda \geq R^*(\nabla \mathcal{L}_\tau(\theta^*))$ and let $n \geq (c_1 s + c_2 \Psi^2(\mathcal{C}) w(\mathcal{A}))$ where $\Psi(\mathcal{C}) = \sup_{u \in \mathcal{C}} \frac{\|u\|_\mathcal{A}}{\|u\|_2}$ is the norm compatibility constant in the error set, $w(\mathcal{A})$ is the Gaussian width of the unit norm ball and c_1 and c_2 are some constants. Then with probability atleast $1 - \exp(-c_2 k_1 \log(em)) - 2\exp(-\eta w^2(\mathcal{C}))$ the number of in-

terpolated samples,

$$\nu = \sup_{u \in \mathcal{C}} |\{i : y_i = \langle x_i, \theta^* + u \rangle, u \in \mathcal{C}\}| \leq c\Psi^2(\mathcal{C})w^2(\mathcal{A}) \,, \tag{16}$$

where c is a constant.

To understand the intuition consider the case of the l_1 norm. For an error vector lying in a particular s-dimensional subspace the maximum number of interpolated samples is $O(s)$ with high probability. Extending the argument to all such s-dimensional subspaces by a union bound argument on the $\binom{p}{s}$ subspaces, the maximum number of interpolated samples when the error vector is any s-sparse vector in p-dimensional space is $O(s \log p)$. Finally the argument is extended to all vectors in the error set using the powerful Maurey's empirical approximation argument previously employed in Sivakumar et al. (2015); Lecué & Mendelson (2014); Rudelson & Zhou (2013).

Suprisingly in prior literature on structured high dimensional quantile regression, the importance of ν has not been explicitly discussed. This intuition about the importance of ν also shows up in an elegant form in the analysis of the RSC condition in Section 4.2.

Below we provide results for the number of interpolated samples for the l_1, l_1/l_2 nonoverlapping group sparse and k-support norms. For the l_1/l_2 nonoverlapping group sparse norm and the k-support norm we first illustrate that they are atomic norms. The results then follow from substituting known values for the norm compatibility constant and Gaussian width of unit norm ball for the different norms. For computation of these quantities for any general norm we refer the interested reader to work in Vershynin (2015); Chen & Banerjee (2015).

Corollary 1 *For the l_1 norm with θ^* being an s-sparse vector, when $n > (c_1 s + c_2 s \log p)$ then with high probability,*

$$\nu = \sup_{u \in \mathcal{C}} |\{i : y_i = \langle x_i, \theta^* + u \rangle\}| \leq cs \log p \,, \tag{17}$$

for some constant c.

Before applying the result to the nonoverlapping group sparse norm note that for any vector $\theta \in \mathbb{R}^p$,

$$\theta = \sum_{i=1}^{N_\mathcal{G}} \sum_j c_{ij} \beta_{ij} \,, \tag{18}$$

where β_{ij} is any unit norm vector in subspace defined by the group i. For any group i, let θ_i denote the vector constructed from θ such that it has component k, $\theta_k = 0$ if $k \notin i$. Now by definition of atomic norm $c_{ij} = \|\theta_i\|_2$ for θ_i in the same direction of β_{ij}, otherwise $c_{ij} = 0$. Therefore the group sparse norm is an atomic norm with a hierarchical set structure with the number of elements $m = N_\mathcal{G}$ in the outer set with each element of the outer set itself being a set of an infinite number of elements with any one element chosen for a particular vector θ, that is, $c_{ij} \neq 0$ for only one j amongst an infinite number of j's.

Corollary 2 *Consider the l_1/l_2 nonoverlapping group-sparse norm with $n > (c_1 s_\mathcal{G} l + c_2 s_\mathcal{G} \log N_\mathcal{G})$. With high probability,*

$$\nu = \sup_{u \in \mathcal{C}} |\{i : y_i = \langle x_i, \theta^* + u \rangle\}| \leq c(l s_\mathcal{G} + s_\mathcal{G} \log N_\mathcal{G}) \,, \tag{19}$$

for some constant c.

For the k-support norm $\|\theta\|_k^{sp} = \inf_{\sum_i u_i = \theta} \{\sum_i \|u_i\|_2 \;|\; \|u_i\|_0 \leq k\}$, can be similarly expressed as,

$$\theta = \sum_{i=1}^{\binom{p}{k}} \sum_j c_{ij} \beta_{ij} \tag{20}$$

where β_{ij} is a unit vector in k-dimensional subspace i. The difference compared to the nonoverlapping group sparse norm is that many of the c_{ij}'s can now be non zero in the inner sum. This is comparably more complex than the group-sparse norm where the inner set becomes a singleton for some θ, but in terms of the analysis nothing changes.

Corollary 3 *Consider the k-support norm with $n > (c_1 s + c_2 s \log \lceil \frac{p}{k} \rceil)$. With high probability,*

$$\nu = \sup_{u \in \mathcal{C}} |\{i : y_i = \langle x_i, \theta^* + u \rangle\}| \leq c(s + s \log \lceil p/k \rceil) \,, \tag{21}$$

for some constant c.

4 Structured Quantile Regression

In this section, we present results for the key components in the general analysis framework of Banerjee et al. (2014) which we briefly described in Section 2 of the paper. We start with results on the regularization parameter by analyzing equation (5) before establishing sample complexity bounds when the restricted strong convexity condition in equation (8) is satisfied. Finally an l_2 bound on the error is obtained using (10). We will consider subGaussian design matrices throughout. All results are in terms of Gaussian widths of sets and the norm compatibility constant. Results for l_1, l_1/l_2-nonoverlapping group sparse and k-support norms are given for illustration purposes.

4.1 Regularization Parameter λ_n

We analyze the bound in equation (5). In prior literature a bound has been established specifically for the l_1 norm in Belloni & Chernozhukov (2011) (See Theorem 1). Below we consider the case of any general atomic norm and obtain a result in terms of the Gaussian width of the unit norm ball. The analysis follows from a similar result for the regularization parameter in Banerjee et al. (2014) for the least squares case.

Theorem 4.1 *Let $X \in \mathbb{R}^{n\times p}$ be a design matrix with independent isotropic subGaussian rows with subGaussian norm $\|\|x_i\|\|_{\psi_2} \leq \kappa$. Define $\Omega_R = \{u : R(u) \leq 1\}$ the unit norm ball and let $\phi = \sup_u \|u\|_2/R(u)$. Then the following holds*

$$E\left[R^*(\nabla\mathcal{L}_\tau(\theta^*))\right] \leq c\frac{\sqrt{\tau(1-\tau)}w(\Omega_R)}{\sqrt{n}}\,, \tag{22}$$

where c is any fixed constant depending only on the subGaussian norm κ. Moreover with probability atleast $1 - c_1\exp\left(-\left(\frac{\tau}{c_2\phi\kappa}\right)^2\right) - 2\exp\left(-2t^2\right)$

$$R^*(\nabla\mathcal{L}_\tau(\theta^*)) \leq c\frac{\sqrt{\tau(1-\tau)}w(\Omega_R)+t}{\sqrt{n}}\,, \tag{23}$$

where c_1, c_2, t are absolute constants.

A major difference to the least squares loss setting, is the independence of the regularization parameter to assumptions on the noise vector (see for example Theorem 3 and Theorem 4 in Banerjee et al. (2014) where the noise is explicitly assumed to be subGaussian and homoscedastic and the noise enters the analysis through properties of $\|\omega\|_2$). This gives the flexibility of considering noise vectors which are heavy tailed or heteroscedastic. Indeed the most interesting applications of quantile regression arise in such settings.

Below we provide bounds for the regularization parameter for different norms by substituting known values of the Gaussian width for the unit norm balls. The result for the l_1 norm matches with Theorem 1 in Belloni & Chernozhukov (2011) for the regularization parameter.

Corollary 4 *If $R(\cdot)$ is the l_1 norm, with high probability*

$$R^*(\nabla\mathcal{L}_\tau(\theta^*)) \leq c\frac{\sqrt{\tau(1-\tau)}\sqrt{\log p}+t}{\sqrt{n}}\,. \tag{24}$$

Corollary 5 *If $R(\cdot)$ is the l_1/l_2 nonoverlapping group sparse norm, with high probability,*

$$R^*(\nabla\mathcal{L}_\tau(\theta^*)) \leq \frac{c\sqrt{\tau(1-\tau)}\sqrt{l+\log N_{\mathcal{G}}}+t}{\sqrt{n}}\,.$$

Corollary 6 *If $R(\cdot)$ is the k-support norm, with high probability,*

$$R^*(\nabla\mathcal{L}_\tau(\theta^*)) \leq \frac{c\sqrt{\tau(1-\tau)}\sqrt{k+k\log\lceil\frac{p}{k}\rceil}+t}{\sqrt{n}}\,.$$

4.2 Restricted Strong Convexity (RSC)

The loss needs to satisfy the RSC condition in equation (8). Prior literature on structured quantile regression has not discussed the RSC condition explicitly, though Fan et al. (2014b) considers it for the quantile huber loss function.

We start by providing an intuition for the RSC formulation for the quantile loss. The RSC condition equation (8) on the error set $\mathcal{C}$ evaluates to the following,

$$\inf_{u\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^n\int_0^{\langle x_i,u\rangle}\left(\mathbb{I}(y_i-\langle x_i,\theta^*\rangle\leq z)-\mathbb{I}(y_i-\langle x_i,\theta^*\rangle\leq 0)\right)dz\,. \tag{25}$$

Let $\nu = \sup_{u\in\mathcal{C}}|Z| = \sup_{u\in\mathcal{C}}|\{i|y_i = \langle x_i,\theta^*+u\rangle\}|$ is the number of interpolated samples. For any $n < p$ if the model can interpolate all points, that is, $\nu = n$ then (25) evaluates to zero. In general, as shown in Section 3, ν gets determined by the structure. For example for the l_1 norm $\nu = O(s\log p)$ rather than the ambient dimensionality p. Thus, the sum over n points in (25) simply reduces to the sum over the $(n-\nu)$ points which are not interpolated, and will ensure the RSC condition when $n > \nu$. The intuition of the NIPS property of Section 3 thus shows up elegantly in the RSC condition.

In equation (25), let $\xi_i = y_i - \langle x_i,\theta^*\rangle, v_i = \int_0^{\langle x_i,u\rangle}(\mathbb{I}(\xi_i\leq z)-\mathbb{I}(\xi_i\leq 0))$ and consider $\frac{1}{n}\sum_{i=1}^n E[v_i]$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n E[v_i] &= \frac{1}{n}\sum_{i=1}^n\int_0^{\langle x_i,u\rangle}(F_i(\xi_i+z)-F_i(\xi))\\ &= \frac{1}{n}\sum_{i=1}^n\int_0^{\langle x_i,u\rangle}f_i(\xi_i)z dz + o(1)\\ &= \frac{1}{2n}\sum_{i=1}^n f_i(\xi_i)\langle x_i,u\rangle^2 + o(1)\\ &\geq \frac{\underline{f}}{2n}\|Xu\|_2^2 \geq \frac{\underline{f}\kappa\|u\|_2^2}{2}\,.\end{aligned}$$

The first line follows from the definition of the cumulative distribution function, the second line by a simple Taylor series expansion, the last line by the assumption that $\underline{f} \leq f_i(\xi_i), \forall i$ and $(1/n)\|Xu\|_2^2 \geq \kappa$, where κ is the restricted eigenvalue (RE) constant. The RE condition is satisfied as the sample complexity bounds for satisfying the NIPS property is same as the RE condition. More generally RSC is a condition on the minimum eigenvalue of the Jacobian matrix $\frac{1}{n}\sum_{i=1}^n f_i(\xi_i)\langle x_i,u\rangle^2$ restricted to the error set

$\mathcal{C}$. This quantity has also been considered in prior literature (see Section 4.2 in Koenker (2005) and the proof in page 121, also see condition D.1 in Belloni & Chernozhukov (2011)). While the above analysis is in expectation of the quantity v_i, we state the following result giving large deviation bounds for the above quantity.

Theorem 4.2 *Consider $X \in \mathbb{R}^{n\times p}$ has subGaussian rows. Let $0 < \underline{f} < f_i(\langle x_i, \theta^*\rangle)$ be a uniform lower bound on the conditional density for all x_i in the support of x. Let κ denote the RE constant satisfying $\frac{1}{n}\|Xu\|_2^2 \geq \kappa\|u\|_2^2$. Let the number of samples $n > cw^2(\mathcal{C})$ where $w(\mathcal{C})$ is the Gaussian width of the error set $\mathcal{C}$ and c is some constant. Then with probability atleast $1 - \exp(-\tau^2/2) - \exp\left(-\frac{\phi_1^2 \underline{f}\xi}{2}n\right)$ where $\phi_1, \xi < 1$ and τ are constants,*

$$\inf_{u\in\mathcal{C}} \delta\mathcal{L}_\tau(\theta^*, u) \geq c_1\kappa\underline{f}\|u\|_2^2 \,. \tag{26}$$

where $c_1 < 0$ is a constant.

Below we provide results for different norms. The sample complexity for the l_1 norm matches the result in Belloni & Chernozhukov (2011) (see equation 2.10).

Corollary 7 *For the l_1 norm with $n > cs\log p$ with high probability the following RSC condition is satisfied,*

$$\inf_{u\in\mathcal{C}} \delta\mathcal{L}_\tau(\theta^*, u) \geq c\kappa\underline{f}\|u\|_2^2 \,. \tag{27}$$

Corollary 8 *For the l_1/l_2 nonoverlapping group sparse norm with $n > c(ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}})$ with high probability the following RSC condition is satisfied,*

$$\inf_{u\in\mathcal{C}} \delta\mathcal{L}_\tau(\theta^*, u) \geq \kappa\underline{f}\|u\|_2^2 \,. \tag{28}$$

Corollary 9 *For the k-support norm with $n > c(s + s\log\lceil\frac{p}{k}\rceil)$ with high probability the following RSC condition is satisfied,*

$$\inf_{u\in\mathcal{C}} \delta\mathcal{L}_\tau(\theta^*, u) \geq c\kappa\underline{f}\|u\|_2^2 \,. \tag{29}$$

4.3 Recovery Bounds

Following the general framework outlined in Banerjee et al. (2014) (see Theorem 2), we state the following high probability bound on the two norm of the error vector $\Delta = \hat{\theta} - \theta^*$.

Theorem 4.3 *For the quantile regression problem, when $\lambda_n \geq \frac{c_1\sqrt{\tau(1-\tau)}w(\Omega_R)}{\sqrt{n}}$, $n > c_2 w^2(\mathcal{C})$ for some constants c_1, c_2 then with high probability,*

$$\|\Delta\|_2 \leq c\frac{\sqrt{\tau(1-\tau)}\Psi(\mathcal{C})w(\Omega_R)}{\underline{f}\kappa} \,. \tag{30}$$

where $\Psi(\mathcal{C})$ is the norm compatibility constant in the error set.

The two norm of the error depends on the two terms $\sqrt{\tau(1-\tau)}$ and $\underline{f}$. The $\sqrt{\tau(1-\tau)}$ term is minimized at the tails and hence has the effect of reducing the estimation error. But typically this is dominated by the lower bound on the density $\underline{f}$ term which makes the estimate less precise in regions of low density. This is to be expected as there are very few samples to make a very precise estimate in low density regions. While similar observations are made in page 72 of Koenker (2005), the results are asymptotic while we show non-asymptotic recovery bounds. Another aspect we reiterate here is the independence of the results from the form of the noise. All results make no assumptions on the noise apart from an assumption on the lower bound of the noise density.

Below we provide recovery bounds for the different norms we consider in the paper.

Corollary 10 *For the l_1 norm when $\lambda_n \geq \frac{c_1\sqrt{\tau(1-\tau)}\sqrt{\log p}}{\sqrt{n}}$ and $n > c_2 s\log p$ with high probability*

$$\|\Delta\|_2 \leq c\frac{\sqrt{s\log p}}{\underline{f}\kappa\sqrt{n}} \tag{31}$$

Corollary 11 *For the l_1/l_2 nonoverlapping group sparse norm when $\lambda_n > \frac{c_1\sqrt{\tau(1-\tau)}\sqrt{l+\log N_{\mathcal{G}}}}{\sqrt{n}}$ and $n > c(ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}})$ with high probability*

$$\|\Delta\|_2 \leq c\frac{\sqrt{ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}}}}{\underline{f}\kappa\sqrt{n}} \tag{32}$$

Corollary 12 *For the k-support norm when $\lambda_n > \frac{c_1\sqrt{\tau(1-\tau)}\sqrt{k+k\log\lceil\frac{p}{k}\rceil}}{\sqrt{n}}$ and $n > cs + s\log\lceil\frac{p}{k}\rceil$ with high probability*

$$\|\Delta\|_2 \leq c\frac{\sqrt{s + s\log\lceil\frac{p}{k}\rceil}}{\underline{f}\kappa\sqrt{n}} \tag{33}$$

5 Experiments

We perform simulations with synthetic data.

5.1 Phase Transition

Data is generated as $y = X\theta^* + \omega$. $\theta^* = [\underbrace{1,1,1,1,1,1}_{6}, \underbrace{0,0,\ldots,0}_{\text{p-6}}] \in \mathbb{R}^p$ for the l_1 norm and $\theta^* = [\underbrace{1,\ldots,1}_{5}, \underbrace{1,\ldots,1}_{5}, \underbrace{1,\ldots,1}_{5}, \underbrace{0,\ldots,0}_{5}, \ldots, \underbrace{0,\ldots,0}_{5}]$ for the l_1/l_2 group sparse norm with $p \in [500, 750, 1000]$. The noise $\omega_i \sim N(0, 0.25), \forall i \in [n]$ is Gaussian with zero mean and 0.25 variance. The design matrix $X \sim$

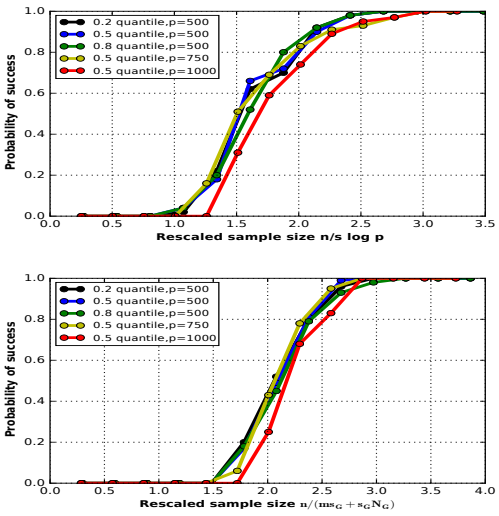

Figure 1: Probability of recovering true parameter versus the rescaled sample size for l_1 norm (top) and l_1/l_2 group sparse norm (bottom). There is a sharp phase transition when the number of samples exceeds NIPS

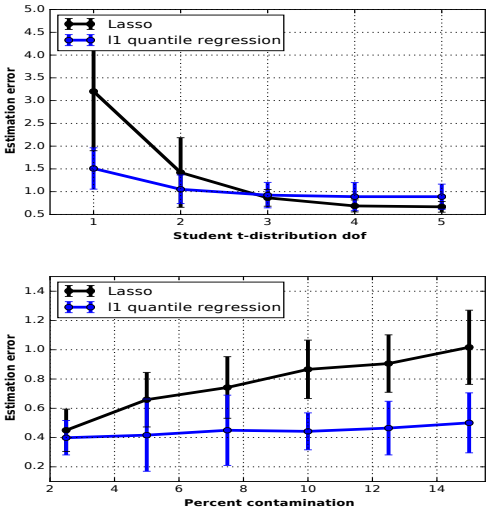

Figure 2: Estimation error of Lasso and l_1-penalized quantile regression against different degrees of freedom of the student t-distribution noise (top) and against percentage contamination (bottom). Quantile regression is robust to heavy-tailed noise and outliers

$N(0, \mathbb{I}_{p\times p})$ is multivariate Gaussian with identity covariance. We vary $n = [10, 20, 30, \ldots, 120, 130]$. For each n we generate 100 datasets with the probability of success defined as the fraction of times we are able to faithfully estimate the true parameter. For $p = 500$ we run simulations for $\tau \in [0.1, 0.5, 0.9]$ and for $p \in [750, 1000]$ we run simulations only for $\tau = 0.5$. For the optimization, we use the Alternating Direction Method of Multipliers (Boyd et al., 2010). The details of the updates can be found in the flare documentation Li et al. (2015). The code was implemented in Python. The plots in Figure 1 clearly show a phase transition for both the l_1 and l_1/l_2 group sparse norms for all quantiles exemplifying the NIPS property described earlier.

5.2 Robustness

We showcase the robustness enjoyed by quantile regression over ordinary least squares estimation against heavy-tailed noise and outliers. We consider the l_1 norm with $y = X\theta^* + \omega$. $\theta^* = [\underbrace{1, 1, 1, 1, 1, 1}_{6}, \underbrace{0, 0, \ldots, 0}_{494}] \in \mathbb{R}^p$. For heavy-tailed noise we consider the student t-distribution with different degrees of freedom, with lower degrees of freedom corresponding to heavier tailed data. To show the robustness to outliers we randomly pick a certain percentage of samples from the dataset and multiply the noise by 10, that is, $\omega_i = 10 * \omega_i$ for a certain proportion of the dataset. We vary the proportion of contamination from 2.5% to 15%. We fix $n = 200$ for this simulation. Again for both exercises, we run 100 simulations and plot the mean and standard deviation of the estimation error $\|\hat{\theta} - \theta^*\|_2$. The plots in Figure 2 show 1. the estimation error against varying degrees of freedom of the student t-distribution and 2. estimation error against the percent contamination. The observations are in agreement with conventional wisdom on robustness of the quantile regression estimator to heavy-tailed noise and outliers.

6 Conclusions

The paper presents a general framework for the analysis of non-asymptotic error and structured recovery for norm regularized quantile regression for any atomic norm. Our results are based on extending the general analysis framework outlined in Banerjee et al. (2014); Negahban et al. (2012) using insights from the geometry of the problem. In particular we introduce the Number of InterPolated Samples (NIPS) as critical for determining the sample complexity for consistent recovery. We prove that once the number of samples crosses the NIPS threshold, we start recovering the true parameter. This phase transition phenomena for norm regularized quantile regression problems has not been discussed in prior literature. We also prove that NIPS is of the order of square of the Gaussian width of the error set for many atomic norms - which is the same order as that for regularized least squares regression and match results from previous work for the l_1 norm (Belloni & Chernozhukov, 2011).

Acknowledgements: We thank reviewers for their valuable comments. This work was supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A.

References

Alquier, P., Cottet, V., and Lecue, G. Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures with Lipschitz Loss Functions. *arXiv:1702.01402*, 2017.

Argyriou, Andreas, Foygel, Rina, and Srebro, Nathan. Sparse Prediction with the k-Support Norm. In *Neural Information Processing Systems (NIPS)*, apr 2012.

Banerjee, Arindam, Chen, Sheng, Fazayeli, Farideh, and Sivakumar, Vidyashankar. Estimation with Norm Regularization. In *Neural Information Processing Systems (NIPS)*, 2014.

Belloni, Alexandre and Chernozhukov, Victor. l1-Penalized Quantile Regression in High-Dimensional Sparse Models. *The Annals of Statistics*, 39(1):82–130, 2011.

Bickel, Peter J., Ritov, Yaacov, and Tsybakov, Alexandre B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. ISSN 0090-5364.

Bogdan, Malgorzata, Berg, Ewout van den, Su, Weijie, and Emmanuel, Candes. Statistical Estimation and Testing via the Sorted L1 Norm. *arXiv:1310.1969*, 2013.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. ISSN 1935-8237.

Candès, Emmanuel J. and Recht, Benjamin. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. ISSN 1615-3375.

Candes, Emmanuel J. and Tao, Terence. The Dantzig selector : statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A., and Willsky, Alan S. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chatterjee, Soumyadeep, Chen, Sheng, and Banerjee, Arindam. Generalized Dantzig Selector: Application to the k-support Norm. In *Advances in Neural Information Processing Systems*, 2014.

Chen, Sheng and Banerjee, Arindam. Structured Estimation with Atomic Norms: General Bounds and Applications. In *Advances in Neural Information Processing Systems*, 2015.

Fan, Jianqing, Fan, Yingying, and Barut, Emre. Adaptive Robust Variable Selection. *Annals of Statistics*, 42(4): 324–351, 2014a.

Fan, Jianqing, Li, Quefeng, and Wang, Yuyan. Robust Estimation of High-Dimensional Mean Regression. *arXiv:1410.2150*, 2014b.

Gordon, Yehoram. Some Inequalitites for Gaussian Processes and Applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Hsu, Daniel and Sabato, Sivan. Loss Minimization and Parameter Estimation with Heavy Tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Kai, B., Li, R., and Zou, H. New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *The Annals of Statistics*, 39:305–332, 2011.

Kato, Kengo. Group Lasso for High Dimensional Sparse Quantile Regression Models. *arXiv:1103.1458*, 2011.

Koenker, Roger. *Quantile Regression*. Cambridge University Press, 2005.

Lecué, Guillaume and Mendelson, Shahar. Sparse recovery under weak moment assumptions. *arXiv:1401.2188*, 2014.

Li, Xingguo, Zhao, Tuo, Yuan, Xiaoming, and Liu, Han. The flare package for high dimensional linear regression and precision matrix estimation in r. *Journal of Machine Learning Research*, 16(1):553–557, 2015.

Li, Y. J. and Zhu, J. L_1-norm Quantile Regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.

Melnyk, Igor and Banerjee, Arindam. Estimating Structured Vector Autoregressive Model. In *International Conference on Machine Learning (ICML)*, 2016.

Negahban, Sahand N., Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012. ISSN 0883-4237.

Rudelson, Mark and Zhou, Shuheng. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, jun 2013.

Sivakumar, Vidyashankar, Banerjee, Arindam, and Ravikumar, Pradeep. Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs. In *Advances in Neural Information Processing Systems*, 2015.

Talagrand, Michel. *Upper and Lower Bounds of Stochastic Processes*. Springer, 2014.

Tibshirani, Robert. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1): 267–288, 1996.

Tropp, Joel A. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory - a Renaissance*. To appear, may 2015.

Vershynin, Roman. Estimation in High Dimensions: A geometric perspective. In *Sampling Theory, a Renaissance*, pp. 3–66. Birkhauser, Basel, 2015.

Wang, Lan, Wu, Yichao, and Li, Runze. Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension. *Journal of the American Statistical Association*, 107:214–222, 2012.

Wu, Y. C. and Liu, Y. F. Variable Selection in Quantile Regression. *Statistica Sinica*, 19:801–817, 2009.

Zou, H. and Yuan, M. Composite Quantile Regression and the Oracle Model Selection Theory. *The Annals of Statistics*, 36:1108–1126, 2008.