NON-PARAMETRIC QUANTILE REGRESSION VIA THE K-NN FUSED LASSO

Steven Siwei Ye
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
stevenysw@g.ucla.edu

Oscar Hernan Madrid Padilla
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
oscar.madrid@stat.ucla.edu

December 4, 2020

ABSTRACT

Quantile regression is a statistical method for estimating conditional quantiles of a response variable. In addition, for mean estimation, it is well known that quantile regression is more robust to outliers than l_2-based methods. By using the fused lasso penalty over a K-nearest neighbors graph, we propose an adaptive quantile estimator in a non-parametric setup. We show that the estimator attains optimal rate of $n^{-1/d}$ up to a logarithmic factor, under mild assumptions on the data generation mechanism of the d-dimensional data. We develop algorithms to compute the estimator and discuss methodology for model selection. Numerical experiments on simulated and real data demonstrate clear advantages of the proposed estimator over state of the art methods. All codes that implement the algorithms and the datasets used in the experiments are publicly available on the author's Github page (`https://github.com/stevenysw/qt_knnfl`).

***K*eywords** quantile regression · non-parametric · fused lasso · bounded variation

1 Introduction

Assume that we have n observations, $(x_1, y_1), ..., (x_n, y_n)$, of the pair of random variables (X, Y). The response variable Y is a real-valued vector and X is a multivariate covariate or predictor variable in a metric space $\mathcal{X}$ with metric $d_{\mathcal{X}}$. A standard goal of non-parametric regression is to infer, in some way, the underlying relationship between Y and X. The generative model behind this can be expressed as

$$y_i = f_0(x_i) + \epsilon_i, \text{ for } i = 1, ..., n, \tag{1}$$

where f_0 is an unknown function that we want to estimate. While usual regression considers estimation of the conditional mean of the response variable, quantile regression estimates the conditional median (or other desired quantiles) of the response variable. Specifically, given a quantile level $\tau \in (0, 1)$, we can rewrite (1) as

$$y_i = \theta_i^* + \epsilon_i, \text{ for } i = 1, ..., n, \tag{2}$$

where

$$\theta_i^* = F^{-1}_{y_i|x_i}(\tau). \tag{3}$$

Here, θ^* is the vector of τ-quantiles of y, $F_{y_i|x_i}$ represents the cumulative distribution function of y_i given x_i, and $\mathbb{P}(\epsilon_i \leq 0) = \tau$.

The goal of quantile regression is to estimate θ^* as accurately as possible, and it usually involves an optimization problem in the form

$$\hat{\theta} \in \underset{\theta \in \zeta \subset \mathbb{R}^n}{\arg\min}\, L(\theta), \tag{4}$$

where $L(\theta)$ is the loss function is defined by

$$L(\theta) = \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i), \tag{5}$$

with $\rho_\tau(t) = (\tau - 1\{t \leq 0\})t$, the asymmetric absolute deviation function (Koenker and Bassett Jr, 1978).

In this paper, we utilize total variation denoising for non-parametric quantile regression in a multivariate setup and combine it with the K-NN procedure. We leverage insights gained from recent results for quantile trend filtering (Padilla and Chatterjee, 2020) and K-NN fused lasso (Padilla et al., 2020a). Our proposed estimator, quantile K-NN fused lasso, can address to the piecewise linear or piecewise polynomial structure in the true vector properly.

It takes simply two steps to compute our proposed quantile K-NN fused lasso estimator. We first construct a K-nearest-neighbor graph corresponding to the given observations. The second step involves a constrained optimization problem with a lasso-type penalty along the $K-$NN graph as follow

$$\hat{\theta} \in \underset{\theta \in \zeta \subset \mathbb{R}^n}{\arg\min}\, L(\theta) + \lambda \|\nabla_G \theta\|_1, \tag{6}$$

where $\lambda > 0$ is a tuning parameter. The notation ∇_G in the penalty term represents the oriented incidence matrix of the K-NN graph, and we will provide the detailed definition in Section 2.

We study the rate of convergence of the estimator defined in (6). Towards that end, we denote a loss function $D_n^2 : \mathbb{R}^n \to \mathbb{R}$ by

$$D_n^2(\delta) := \frac{1}{n} \sum_{i=1}^{n} \min\{|\delta_i|, \delta_i^2\}.$$

Thus, D_n^2 is similar to the Huber cost function (see page 471 in Wainwright, 2018), which offers a compromise between the least-squares cost and the l_1-norm cost function and is thus less sensitive to outliers in data. We show that under mild conditions, (6) attains a convergence rate of $n^{-1/d}$ for d-dimensional data in terms of D_n^2, ignoring the logarithmic factor. The rate is nearly minimax and it matches the mean squared error rates from Padilla et al. (2020a). However, unlike Padilla et al. (2020a), our result holds under general errors allowing for heavy-tailed distributions. Another notable point in our theoretical analysis, different from previous quantile regression work, is that we only require a bounded variation class of signals and hence guarantee our result to hold under very general models.

1.1 Previous Work

Quantile regression, since introduced by Koenker and Bassett Jr (1978), has become a commonly used class of methods in many applications thanks to its flexibility and robustness for modelling conditional distributions. The study of quantile regression in non-parametric setups can be dated back to Utreras (1981), Cox (1983) and Eubank (1988), whose works were mainly developed for median regression on one-dimensional data. Later, Koenker et al. (1994) proposed a more general estimator for any desired quantile τ, quantile smoothing spline, and He et al. (1998) provided a bivariate version of the estimator. Quantile smoothing spline problems are of a l_1 penalty structure, which led to a more general study on l_1-norm regularized quantile regression by Li and Zhu (2008). Other methods for non-parametric quantile regression have also been proposed in the literature. Yu and Jones (1998), Cai and Xu (2008), and Spokoiny et al. (2013) explored local polynomial quantile regression. Belloni et al. (2019) studied non-parametric series quantile regression, and Meinshausen (2006) introduced quantile random forests. Quantile regression with rectified linear unit (ReLU) neural networks is another edge-cutting approach that utilizes some knowledge from the field of deep learning, and the theory is studied in Padilla et al. (2020b).

Our approach exploits the local adaptivity of total variation denoising. Rudin et al. (1992) first proposed total variation denoising for the application of image processing, and Tibshirani and Saunders (2005) studied fused lasso thoroughly. Later, Kim et al. (2009) extended the discussion to the so-called trend filtering on one-dimensional data, and Wang et al. (2016) provided a generalization of trend filtering to the setting of estimation on graphs. These problems can be formulated similarly into an optimization problem of the form

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^n}{\arg\min}\, \frac{1}{2} \sum_{i=1}^{n} (\theta_i - y_i)^2 + \lambda \|D\theta\|_1, \tag{7}$$

where $\lambda > 0$ is a regularization parameter to be chosen carefully and D is a matrix. For instance, trend filtering order k (fused lasso $k = 0$) consists of matrices D that capture the $(k+1)$th order total variation of a given signal, see Tibshirani (2014). A substantial amount of literature focused on theoretical guarantees of these estimators. Mammen and Geer

(1997) and Tibshirani (2014) showed that trend filtering attains nearly minimax rates in mean squared error (MSE) for estimating functions of bounded variation. Hütter and Rigollet (2016) showed a sharp oracle rate of total variation denoising along grid graphs. More recently, Padilla et al. (2020a) incorporated fused lasso with the K-NN procedure and proved that the K-NN fused lasso also achieves nearly minimax convergence rate. In the quantile regression literature, Belloni and Chernozhukov (2011), Kato (2011), Fan et al. (2014) studied quantile model selection via l_1 regularization, but most of these works required strict linear assumptions. Padilla and Chatterjee (2020) provided proofs for theoretical properties of quantile trend filtering estimator in one dimension. They showed that under minimal assumptions of the data generation mechanism, quantile trend filtering estimator attains a minimax rate for one-dimensional piecewise polynomial regression.

On the computational side, quantile regression is different from l_2 based methods because it requires a non-trivial reformulation of the optimization problem due to non-differentiability of the loss as (5). The most well-known algorithm for computing a quantile estimator is due to Koenker (2005) and it uses an interior point (IP) approach. Pietrosanu et al. (2020) studied high-dimensional quantile regression problems and obtained estimators by applying the alternating direction method of multipliers (ADMM; Boyd et al., 2011), majorize-minimize (MM; Hunter and Lange, 2000), and coordinate descent (CD; Wu and Lange, 2008) algorithms for variable selection. For computing trend filtering estimates, Hochbaum and Lu (2017) developed a fast algorithm for quantile fused lasso in $O(n \log n)$ operations. Recently, Brantley et al. (2020) proposed an ADMM based algorithm for computing kth order quantile trend filtering estimators for one-dimensional data.

1.2 Outline of the Paper

In Section 2, we provide the definition of quantile K-nearest-neighbors fused lasso estimator and the constrained version of the problem. Two algorithms to compute the proposed estimators numerically – alternating directions method of multipliers (ADMM), and majorize-minimize (MM), are introduced in Section 3, and the discussion on how to select an appropriate penalty parameter in practice is also included. Section 4 presents the two theoretical developments regarding the constrained and penalized estimators. The theorems demonstrate that under general assumptions, both estimates converge at a rate of $n^{-1/d}$, up to a logarithmic factor, for estimating d-dimensional data under the loss function D_n^2 defined above. Section 5 lists the results of numerical experiments on multiple simulated datasets and two real datasets, California housing data and Chicago crime data. The experiments show that the proposed estimator outperform state-of-the-art methods on both simulated and real datasets. Moreover, the comparison on accuracy and computational time among the algorithms introduced in Section 3 provides the audience with some insights on choosing a suitable algorithm for specific problems. The proofs of theorems in the paper are provided in the Appendix.

2 Quantile K-NN Fused lasso

The first step to construct the quantile quantile K-NN fused lasso estimator is build a K-NN graph G. Specifically, given the observations, G has vertex set $V = \{1, ..., n\}$, and its edge set E_K contains the pair (i, j), for $i \in V$, $j \in V$, and $i \neq j$, if and only if x_i is among the K-nearest neighbors of x_j, with respect to the metric $d_{\mathcal{X}}$, or vice versa. After constructing the K-NN graph, we can formalize an optimization problem for quantile K-NN fused lasso as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\nabla_G \theta\|_1, \tag{8}$$

where $\lambda > 0$ is a tuning parameter, and ∇_G is an oriented incidence matrix of the K-NN graph G. Thus, we define ∇_G as follow: each row of the matrix corresponds to one edge in G; for instance, if the p-th edge in G connects the i-th and j-th observations, then

$$(\nabla_G)_{p,q} = \begin{cases} 1 & \text{if } q = i, \\ -1 & \text{if } q = j, \\ 0 & \text{otherwise.} \end{cases}$$

In this way, the p-th element in $\nabla_G \theta$, $(\nabla_G \theta)_p = \theta_i - \theta_j$. Notice that we choose the ordering of the nodes and edges in ∇_G arbitrarily without loss of generality.

Once $\hat{\theta}$ in (8) has been computed, we can predict the value of response corresponding to a new observation $x \in \mathcal{X} \backslash \{x_1, ..., x_n\}$ by the averaged estimated response of the K-nearest neighbors of x in $\{x_1, ..., x_n\}$. Mathematically, we write

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{n} \hat{\theta}_i \cdot \mathbf{1}\{x_i \in \mathcal{N}_K(x)\}, \tag{9}$$

where $\mathcal{N}_K(x)$ is the set of K-nearest neighbors of x in the training data. A similar prediction rule was used in Padilla et al. (2020a).

A related estimator to the penalized estimator $\hat{\theta}$ defined in (8) is the the constrained estimator $\hat{\theta}_C$, of which the corresponding optimization problem can be written as

$$\begin{aligned} \hat{\theta}_C &= \underset{\theta \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \\ &\quad \text{subject to } \|\nabla_G \theta\|_1 \leq C, \end{aligned} \tag{10}$$

for some positive constant C.

3 Algorithms and Model Selection

To compute the quantile K-NN fused lasso estimator, the first step is to construct the K-NN graph from the data. The computational complexity of constructing the K-NN graph is of $O(n^2)$, although it is possible to accelerate the procedure to $O(n^t)$ for some $t \in (1, 2)$ using divide and conquer methods (Bentley, 1980; Chen et al., 2009).

The second step of computation is to solve a constrained optimization problem as (8). Here, we introduce three algorithms to solve the problem numerically. Before presenting our algorithms, we stress that both the problems (8) and (10) are linear programs, therefore we can use any linear programming software to derive an optimal solution. Noticeably, we can take the advantage of sparsity in the penalty matrix for faster computation. However, a shortcoming of linear programming is that the algorithm can become very time-consuming for large sized problems, especially when n is greater than 5000.

3.1 Alternating Directions Method of Multipliers (ADMM)

The alternating directions method of multipliers (ADMM) algorithm (Boyd et al., 2011) is a powerful tool for solving constrained optimization problems.

We first reformulate the optimization problem (8) as

$$\begin{aligned} &\underset{\theta \in \mathbb{R}^n, z \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\nabla_G z\|_1 \\ &\text{s.t.} \qquad z = \theta, \end{aligned} \tag{11}$$

and the augmented Lagrangian can then be written as

$$L_R(\theta, z, u) = \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\nabla_G z\|_1 + \frac{R}{2} \|\theta - z + u\|^2,$$

where R is the penalty parameter that controls step size in the update. Thus we can solve (11) by iteratively updating the primal and dual

$$\theta \leftarrow \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \frac{R}{2} \|\theta - z + u\|^2 \right\}, \tag{12}$$

$$z \leftarrow \underset{z \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \|\theta + u - z\|^2 + \frac{\lambda}{R} \|\nabla_G z\|_1 \right\}. \tag{13}$$

The primal problem (12) can be solved coordinate-wise in closed form as

$$\theta_i = \begin{cases} z_i - u_i + \frac{\tau}{R} & \text{if } y_i - z_i + u_i > \frac{\tau}{R}, \\ z_i - u_i + \frac{\tau - 1}{R} & \text{if } y_i - z_i + u_i < \frac{\tau - 1}{R}, \\ y_i & \text{otherwise;} \end{cases}$$

See Appendix A for the steps to derive the solution. The dual problem (13) is a generalized lasso problem that can be solved with the parametric max-flow algorithm from Chambolle and Darbo (2009).

The entire procedure is presented in Algorithm 1. In practice, we can simply choose the penalty parameter R to be $\frac{1}{2}$. We require the procedure to stop if it reaches the maximum iteration or the primal residual $\|\theta^{(k)} - \theta^{(k-1)}\|_2$ is within a threshold (e.g., 10^{-4}). Actually, the ADMM algorithm converges very quickly with only tens of iterations and hence we find it to be faster than linear programming.

Algorithm 1: Alternating Directions Method of Multipliers for quantile K-NN fused lasso

Input: Number of nearest neighbor: K, quantile: τ, penalty parameter: λ, maximum iteration: N_{iter}
Data: $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$
Output: $\hat{\theta} \in \mathbb{R}^n$
1. Compute K-NN graph incidence matrix ∇_G from X.
2. Initialize $\theta^{(0)} = y, z^{(0)} = y, u^{(0)} = 0$.
3. For $k = 1, 2, ...,$ until $\|\theta^{(k)} - \theta^{(k-1)}\|_2$ converges or the procedure reaches N_{iter}:
(a) For $i = 1, ..., n$, update

$$\theta_i^{(k)} \leftarrow \arg\min \left\{ \rho_\tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i^{(k-1)} + u_i^{(k-1)})^2 \right\}.$$

(b) Update

$$z^{(k)} \leftarrow \arg\min \left\{ \frac{1}{2}\|\theta^{(k)} + u^{(k-1)} - z\|^2 + \frac{\lambda}{R}\|\nabla_G z\|_1 \right\}.$$

(c) Update

$$u^{(k)} \leftarrow u^{(k-1)} + \theta^{(k)} - z^{(k)}.$$

3.2 Majorize-Minimize (MM)

We now exploit the majorize-minimize (MM) approach from Hunter and Lange (2000) for estimating the conditional median. The main advantage of the MM algorithm versus ADMM is that it is conceptually simple and it is a descent algorithm.

Recall that for $\tau = 0.5$, quantile regression becomes least absolute deviation, or L_1-regression, and the loss function $L(\theta)$ of our problem becomes

$$L(\theta) = \sum_{i=1}^{n} |y_i - \theta_i| + \lambda\|\nabla_G\theta\|_1. \tag{14}$$

Next, notice that $L(\theta)$ can be majorized at $\theta^{(k)}$ by $Q(\theta \mid \theta^{(k)})$ given as

$$Q(\theta \mid \theta^{(k)}) = \sum_{i=1}^{n} \frac{(y_i - \theta_i)^2}{|y_i - \theta_i^{(k)}|} + \lambda \sum_{(i,j)\in E_K} \frac{(\theta_i - \theta_j)^2}{|\theta_i^{(k)} - \theta_j^{(k)}|}, \tag{15}$$

since it holds that

$$\begin{aligned} Q(\theta^{(k)} \mid \theta^{(k)}) &= L(\theta^{(k)}), \\ Q(\theta \mid \theta^{(k)}) &\geq L(\theta^{(k)}) \text{ for all } \theta. \end{aligned} \tag{16}$$

To avoid possible occurrences of zero, we add a perturbation ϵ to the denominator each time. Then, the iterative algorithm optimize $Q(\theta \mid \theta^{(k)})$ at each iteration. The stopping criterion for MM algorithm remains the same as for ADMM. Because the optimization problem here has a closed-form solution, we can compute the solution directly by solving a linear system (see Step 3c in Algorithm 2). We find the MM algorithm to be faster in running time than linear programming and ADMM for large-size problems and it produces reasonably stable solutions as the others; see the experiments and discussion in Section 5.1. A major drawback of our fast algorithm is that it can only handle median regression at this moment. We leave for future work studying an extension of an MM-based algorithm for estimating general quantiles in the future.

3.3 Model Selection

The choice of the tuning parameter λ in (8) is an important practical issue in estimation because it controls the degree of smoothness in the estimator. The value of λ can be chosen through K-fold cross-validation. Alternatively, we can select the regularization parameter based on Bayesian Information Criteria (BIC; Schwarz, 1978). The BIC for quantile regression (Yu and Moyeed, 2001) can be computed as

$$\text{BIC}(\tau) = \frac{2}{\sigma}\sum_{i=1}^{n} \rho_\tau(y_i - \hat{\theta}_i) + \nu \log n,$$

Algorithm 2: Majorize-Minimize for quantile K-NN fused lasso, $\tau = 0.5$

Input: Number of nearest neighbor: K, penalty parameter: λ, maximum iteration: N_{iter}
Data: $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$
Output: $\hat{\theta} \in \mathbb{R}^n$
1. Compute K-NN graph incidence matrix ∇_G from X.
2. Initialize $\theta_i^{(0)} = \text{median}(y)$ for $i = 1, ..., n$.
3. For $k = 1, 2, ...,$ until $\|\theta^{(k)} - \theta^{(k-1)}\|_2$ converges or the procedure reaches N_{iter}:
(a) Compute weight matrix $W \in \mathbb{R}^{n\times n}$:

$$W = \text{diag}(1/[|y - \theta^{(k-1)}| + \epsilon]).$$

(b) Compute weight matrix

$$\tilde{W} = \text{diag}(1/[|\theta_i^{(k-1)} - \theta_j^{(k-1)}| + \epsilon]) \text{ if } (i,j) \in E_K.$$

(c) Update

$$\theta^{(k)} \leftarrow [W + \lambda \nabla_G^\top \tilde{W}^2 \nabla_G]^{-1} W y.$$

where ν denotes the degree of freedom of the estimator and $\sigma > 0$ can be empirically chosen as $\sigma = \frac{1-|1-2\tau|}{2}$. It is also possible to use Schwarz Information Criteria (SIC; Koenker et al., 1994) given by

$$\text{SIC}(\tau) = \log\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \hat{\theta}_i)\right] + \frac{1}{2n}\nu \log n.$$

However, BIC is more stable than SIC in practice because when λ is small, SIC may become ill-conditioned as the term inside the logarithm is close to 0.

Tibshirani and Taylor (2012) demonstrates that a lasso-type problem has degrees of freedom ν equal to the expected nullity of the penalty matrix after removing the rows that are indexed by the boundary set of a dual solution at y. In our case, we define ν as the number of connected components in the graph ∇_G after removing all nodes j's where $|(\nabla_G \hat{\theta})_j|$ is beyond a threshold κ, typically very small (e.g., 10^{-2}).

4 Theoretical Analysis

Before arriving at our main results, we introduce some notation. For a set $A \subset \mathcal{A}$ with $(\mathcal{A}, d_{\mathcal{A}})$ a metric space, we write $B_\epsilon(A) = \{a : \text{exists } a' \in A, \text{ with } d_{\mathcal{A}} \leq \epsilon\}$. The Euclidean norm of a vector $x \in \mathbb{R}$ is denoted by $\|x\|_2 = (x_1^2 + ... + x_d^2)^{1/2}$. The l_1 norm of x is denoted by $\|x\|_1 = |x_1| + ... + |x_d|$. The infinity norm of x is denoted by $\|x\|_\infty = \max_i |x_i|$. In the covariate space $\mathcal{X}$, we consider the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, induced by the metric $d_{\mathcal{X}}$, and we let μ be a measure on $\mathcal{B}(\mathcal{X})$. We assume that the covariates in the model (1) satisfy $x_i \overset{\text{ind}}{\sim} p(x)$. In other word, p is the probability density function associated with the distribution of x_i, with respect to the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. Let $\{a_n\}$ and $\{b_n\} \subset \mathbb{R}$ be two sequences. We write $a_n = O_{\mathbb{P}}(b_n)$ if for every $\epsilon > 0$ there exists $C > 0$ such that $\mathbb{P}(a_n \geq C b_n) < \epsilon$ for all n. We also write poly(x) as a polynomial function of x, but the functions vary from case to case in the content below. Throughout, we assume the dimension of $\mathcal{X}$ to be greater than 1, since the study of the one-dimensional model, quantile trend filtering, can be found in Padilla and Chatterjee (2020).

We first make some necessary assumptions for analyzing the theoretical guarantees of both the constrained and penalized estimators.

Assumption 1. (Bounded Variation) We write $\theta_i^* = F_{y_i|x_i}^{-1}(\tau)$ for $i = 1, ..., n$ and require that $V^* := \|\nabla_G \theta^*\|_1 / [n^{1-1/d}\text{poly}(\log n)]$ satisfies $V^* = O_{\mathbb{P}}(1)$. Here $F_{y_i|x_i}$ is the cumulative distribution function of y_i given x_i for $i = 1, ..., n$.

The first assumption simply requires that θ^*, the vector of τ-quantiles of y, has bounded total variation along the K-NN graph. The scaling $n^{1-1/d}\text{poly}(\log n)$ comes from Padilla et al. (2020a), and we will discuss the details after we present Assumptions 3-5 below.

Assumption 2. There exists a constant $L > 0$ such that for $\delta \in \mathbb{R}^n$ satisfying $\|\delta\|_\infty \leq L$ we have that

$$\min_{i=1,...,n} f_{y_i|x_i}(\theta_i^* + \delta_i) \geq \underline{f} \quad \text{a.s.},$$

for some $\underline{f} > 0$, and where $f_{y_i|x_i}$ is the the conditional probability density of y_i given x_i.

The assumption that the density of the response variable is bounded below is standard in quantile regression analysis. Related conditions appears as D.1 in Belloni and Chernozhukov (2011) and Condition 2 in He and Shi (1994).

The next three assumptions inherit from the study of K-NN graph in Von Luxburg et al. (2014) and Padilla et al. (2020a).

Assumption 3. The density of the covariates, p, satisfies $0 < p_{\min} < p(x) < p_{\max}$, for all $x \in \mathcal{X}$, where $p_{\min}, p_{\max} > 0$ are fixed constants.

We only require the distribution of covariates to be bounded above and below by positive constants. In Gyorfi et al. (2006) and Meinshausen (2006), p is assumed to be the probability density function of the uniform distribution in $[0,1]^d$.

Assumption 4. The base measure μ in the metric space $\mathcal{X}$, in which X is defined, satisfies

$$c_{1,d} r^d \leq \mu\{B_r(x)\} \leq c_{2,d} r^d, \text{ for all } x \in \mathcal{X},$$

for all $0 < r < r_0$, where $r_0, c_{1,d}, c_{2,d}$ are positive constants, and $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

Although $\mathcal{X}$ is not necessarily a Euclidean space, we require in this condition that balls in $\mathcal{X}$ have volume, with respect to some measure μ on the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, that behaves similarly to the Lebesgue measure of balls in $\mathbb{R}^d$.

Assumption 5. There exists a homeomorphism $h : \mathcal{X} \to [0,1]^d$, such that

$$L_{\min} d_{\mathcal{X}}(x, x') \leq \|h(x) - h(x')\|_2 \leq L_{\max} d_{\mathcal{X}}(x, x'),$$

for all $x, x' \in \mathcal{X}$ and for some positive constants $L_{\min}, L_{\max}$, where $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

The existence of a continuous bijection between $\mathcal{X}$ and $[0,1]^d$ ensures that the space has no holes and is topologically equivalent to $[0,1]^d$.

On another note, we point that Padilla et al. (2020a) showed that under Assumptions 3-5, $\|\nabla_G \theta^*\|_1 \asymp n^{1-1/d}$ for a K-NN graph G up to a polynomial of $\log n$, with an extra condition on the function $f_0 \circ h^{-1}$ being piecewise Lipschitz. Hence, under such conditions Assumption 1 holds. For completeness, the definition of the class of piecewise Lipschitz functions is provided in the Appendix. It is rather remarkable that Padilla et al. (2020a) also present an alternative condition on $f_0 \circ h^{-1}$ than piecewise Lipschitz to guarantee the results hold; see Assumption 5 in the same paper.

Now, we are ready to present our first theoretical result on quantile K-NN fused lasso estimates.

Theorem 1. *Under Assumptions 1-5, by setting $C = \frac{V}{n^{1-1/d}}$ in (10) for a tuning parameter V, where $V \asymp V^*$ and $V \geq V^*$, we have*

$$D_n^2\left(\theta^* - \hat{\theta}_C\right) = O_{\mathbb{P}}\left\{n^{-1/d}\text{poly}(\log n)\right\}.$$

The first theorem shows that quantile K-NN fused lasso attains the optimal rate of $n^{-1/d}$ under the loss $D_n^2(\cdot)$ defined in Section 1 for estimating signals in a constrained set.

Theorem 2. *Under Assumptions 1-5, there exists a choice of λ for (8) satisfying*

$$\lambda = \begin{cases} \Theta\left\{\log n\right\} & \text{for } d = 2, \\ \Theta\left\{(\log n)^{1/2}\right\} & \text{for } d > 2, \end{cases}$$

such that

$$D_n^2\left(\theta^* - \hat{\theta}\right) = O_{\mathbb{P}}\left\{n^{-1/d}\text{poly}(\log n)\right\}.$$

The second theorem states that, under certain choice of the tuning parameter, the penalized estimator achieves the convergence rate of $n^{-1/d}$, similar to the constrained estimator, ignoring the logarithmic factor. It is noteworthy that for all d, the rate of $n^{-1/d}$ is actually minimax, up to a logarithmic factor, and it matches the rate from Willett et al. (2006).

5 Experiments

In this section, we will examine the performance of quantile K-NN fused lasso (QKNN) on various simulated and real datasets. The two benchmark estimators we compare against are K-NN fused lasso (KNN; Padilla et al., 2020a) and

quantile random forest (QRF; Meinshausen, 2006). The performance of an estimator is measured by its mean squared error, defined by

$$\text{MSE}(\hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta} - \theta_i^*)^2,$$

where θ^* is the vector of τ-quantiles of the true signal.

For quantile K-NN fused lasso, we use the ADMM algorithm and select the tuning parameter λ based on the BIC criteria described in Section 3.3; for K-NN fused lasso, we use the algorithm from Chambolle and Darbo (2009) and the corresponding penalty parameter is chosen to minimize the average mean squared error over 100 Monte Carlo replicates. For quantile random forest, we directly use the R package "quantregForest" with defaulted choice of tree structure and tuning parameters.

Throughout, for both K-NN fused lasso and quantile K-NN fused lasso, we set K to be 5 for sufficient information and efficient computation.

5.1 Simulation Study

We generate 100 data sets from models under each scenario described below with sample size between 10^2 and 10^4 and then report the mean squared errors of the three estimators with respect to different quantiles. For each scenario the data are generated as

$$y_i = \theta_i^* + \epsilon_i, \text{ and } \theta_i^* = f_0(x_i),\ i = 1, ..., n,$$

where θ_i^* comes from some underlying functions f_0, and the errors $\{\epsilon_i\}_{i=1}^n$ are independent with $\epsilon_i \sim F_i$ for some distributions F_i, where we select from Gaussian, Cauchy, and t-distributions.

Scenario 1
We generate x_i uniformly from $[0,1]^2$, and define $f_0 : [0,1]^2 \to \{0,1\}$ by

$$f_0(x) = \begin{cases} 1 & \text{if } \frac{5}{4}x_{i1} + \frac{3}{4}x_{i2} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 2
In this case, we generate $X \in \mathbb{R}^2$ according to the probability density function

$$p(x) = \frac{1}{5}\mathbf{1}_{\{[0,1]^2 \setminus [0.4,0.6]^2\}}(x) + \frac{16}{25}\mathbf{1}_{\{[0.45,0.55]^2\}}(x) + \frac{4}{25}\mathbf{1}_{\{[0.4,0.6]^2 \setminus [0.45,0.55]^2\}}(x).$$

The function $f_0 :\ [0,1]^2 \to \mathbb{R}$ is defined as

$$f_0(x) = \mathbf{1}_{\{\|x - \frac{1}{2}(1,1)^\top\|_2^2 \le \frac{2}{1000}\}}(x).$$

Scenario 3
Again, x_i are from uniform $[0,1]^2$. The smooth function $f_0 :\ [0,1]^2 \to \mathbb{R}$ is defined as

$$f_0(x_i) = 0.4x_{i1}^2 + 0.6x_{i2}^2.$$

Scenario 4
The function $f_0 :\ [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 1 & \text{if } \|x - \frac{1}{4}1_d\|_2 < \|x - \frac{3}{4}1_d\|_2, \\ -1 & \text{otherwise,} \end{cases}$$

and the density p is uniform in $[0,1]^d$. The errors are chosen as $(x^\top\beta)\epsilon$, where $\beta = (\frac{1}{d}, ..., \frac{1}{d})^\top$. Here we simulate with $d = 5$.

The results presented in Table 1 indicate that overall, quantile K-NN fused lasso outperforms the competitors in most scenario. As expected, for estimating functions with Gaussian errors like Scenario 1, regular K-NN fused lasso is the best method. For Scenarios 2 and 3, when estimating the conditional median of piecewise continuous or continuous functions with heavy-tail errors (such as Cauchy and t-distributions), quantile K-NN fused lasso achieve the smallest mean square errors over the other two methods. Figure 1 displays the true data and the estimates from both quantile K-NN fused lasso and quantile random forest under Scenario 2 and 3 with t-distribution errors. Clearly, quantile K-NN

n	Scenario	ϵ	τ	QKNN	QRF	KNN
100	1	N(0,1)	0.5	0.2558	0.2173	**0.1606**
1000	1	N(0,1)	0.5	0.1320	0.1558	**0.0811**
5000	1	N(0,1)	0.5	0.1086	0.1345	**0.0532**
10000	1	N(0,1)	0.5	0.0639	0.1289	**0.0463**
100	1	Cauchy(0,1)	0.5	**0.2722**	2652.702	836.5628
1000	1	Cauchy(0,1)	0.5	**0.1578**	43154.31	3126.1
5000	1	Cauchy(0,1)	0.5	**0.1127**	41390.26	3344.1
10000	1	Cauchy(0,1)	0.5	**0.0991**	13450.92	3563.1
100	2	t_3	0.5	**0.2322**	0.5570	0.2433
1000	2	t_3	0.5	**0.0898**	0.3784	0.1401
5000	2	t_3	0.5	**0.0359**	0.3445	0.1260
10000	2	t_3	0.5	**0.0299**	0.3253	0.0526
100	3	t_2	0.5	**0.0403**	2.7943	0.1919
1000	3	t_2	0.5	**0.0146**	2.0091	0.0406
5000	3	t_2	0.5	**0.0085**	0.9847	0.0465
10000	3	t_2	0.5	**0.0187**	0.9935	0.0469
100	4	t_3	0.9	0.8037	**0.6511**	*
1000	4	t_3	0.9	**0.3643**	0.5389	*
5000	4	t_3	0.9	**0.2859**	0.4658	*
10000	4	t_3	0.9	**0.2764**	0.4163	*
100	4	t_3	0.1	1.2871	**0.8751**	*
1000	4	t_3	0.1	**0.4354**	0.6329	*
5000	4	t_3	0.1	**0.3860**	0.4954	*
10000	4	t_3	0.1	**0.2974**	0.4597	*

Table 1: Mean squared error $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2$, averaging over 100 Monte Carlo simulations for the different methods, sample sizes, errors, quantiles considered.

fused lasso is able to capture local patterns in piecewise continuous signals, but quantile random forest completely fails to provide an accurate estimate to the data in these cases. For estimating quantiles other than median of piecewise continuous signals as Scenario 4, quantile K-NN fused lasso can still perform better than quantile random forest when the sample size is large enough.

We also compare the performance of linear programming the two algorithms discussed in Section 3, with simulated data from Scenario 3. We obtain almost identical estimators from the three algorithms under the same choice of λ. Regarding computational time, we record the averaged time consumed over 100 simulations for each algorithm. Figure 2 demonstrates that majorize-minimize (MM) is the most efficient one among the three algorithms, and linear programming (LP) can be very expensive in operational time for large-size problems.

5.2 Real Data

5.2.1 California Housing Data

In this section, we conduct an experiment of predicting house value in California based on median income and average occupancy, similar to the experiment in Petersen et al. (2016). The data set, consisting of 20,640 measurements, was originally used in Pace and Barry (1997) is publicly available from the Carnegie Mellon StatLib data repository (lib.stat.cmu.edu).

We perform a train-test split on the data set, with training sizes 1000, 5000, and 10000 separately. For median estimation, we compare mean squared prediction errors (after taking the log of housing price) from quantile K-NN fused lasso and quantile random forest; besides, we construct 90% and 95% confidence intervals from both methods and report the proportion of true observations in testing sets that locate in the intervals. Both evaluations are averaged over 100 repetitions for each method. The tuning parameter λ for quantile K-NN fused lasso are chosen based on the BIC criteria for each training and the parameters for quantile random forest are selected as default.

From the results in Table 2, quantile K-NN fused lasso has better performance than quantile random forest in all cases. The result agrees with the nature of piecewise continuity in housing price, that guarantee the advantage of our proposed method over the competitor. When we illustrate the predicted values for the experiment with a training size of 10,000

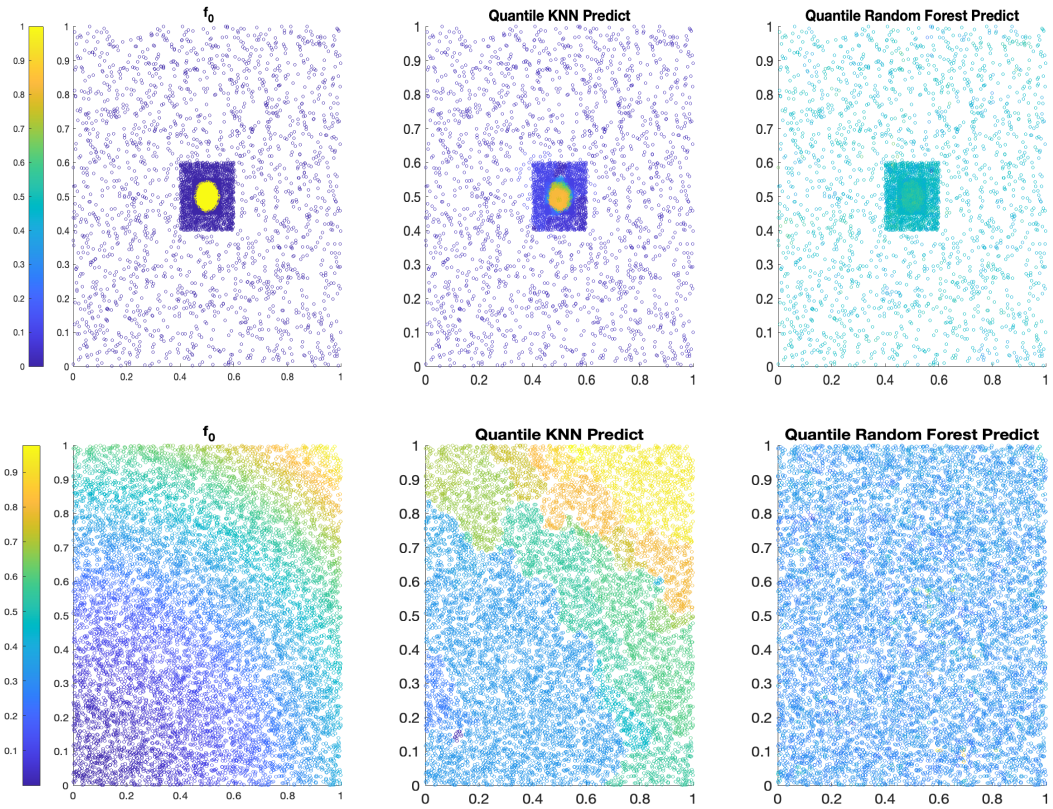

Figure 1: Comparison among observations and estimates from Scenario 2 (*Above*) and Scenario 3 (*Below*). *Left*: the function f_0, with $n = 10000$. *Middle*: the estimate of f_0 obtained via quantile K-NN fused lasso. *Right*: the estimate of f_0 obtained via quantile random forest.

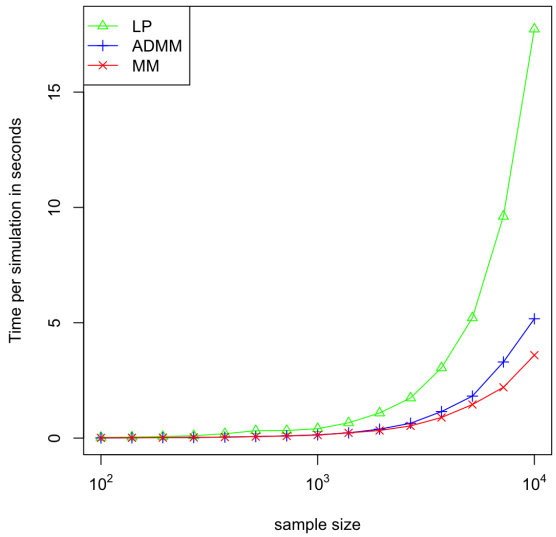

Figure 2: A log-scaled plot of time per simulation of LP, ADMM and MM algorithm against problem size n (15 values from 10^2 up to 10^4). For each algorithm, the time to compute the estimate for one simulated data is averaged over 100 Monte Carlo simulations.

visually in Figure 3, we also observe a piecewise continuous structure in quantile K-NN fused lasso estimates, while estimates from quantile random forest contain more noise, especially for lower and higher quantiles.

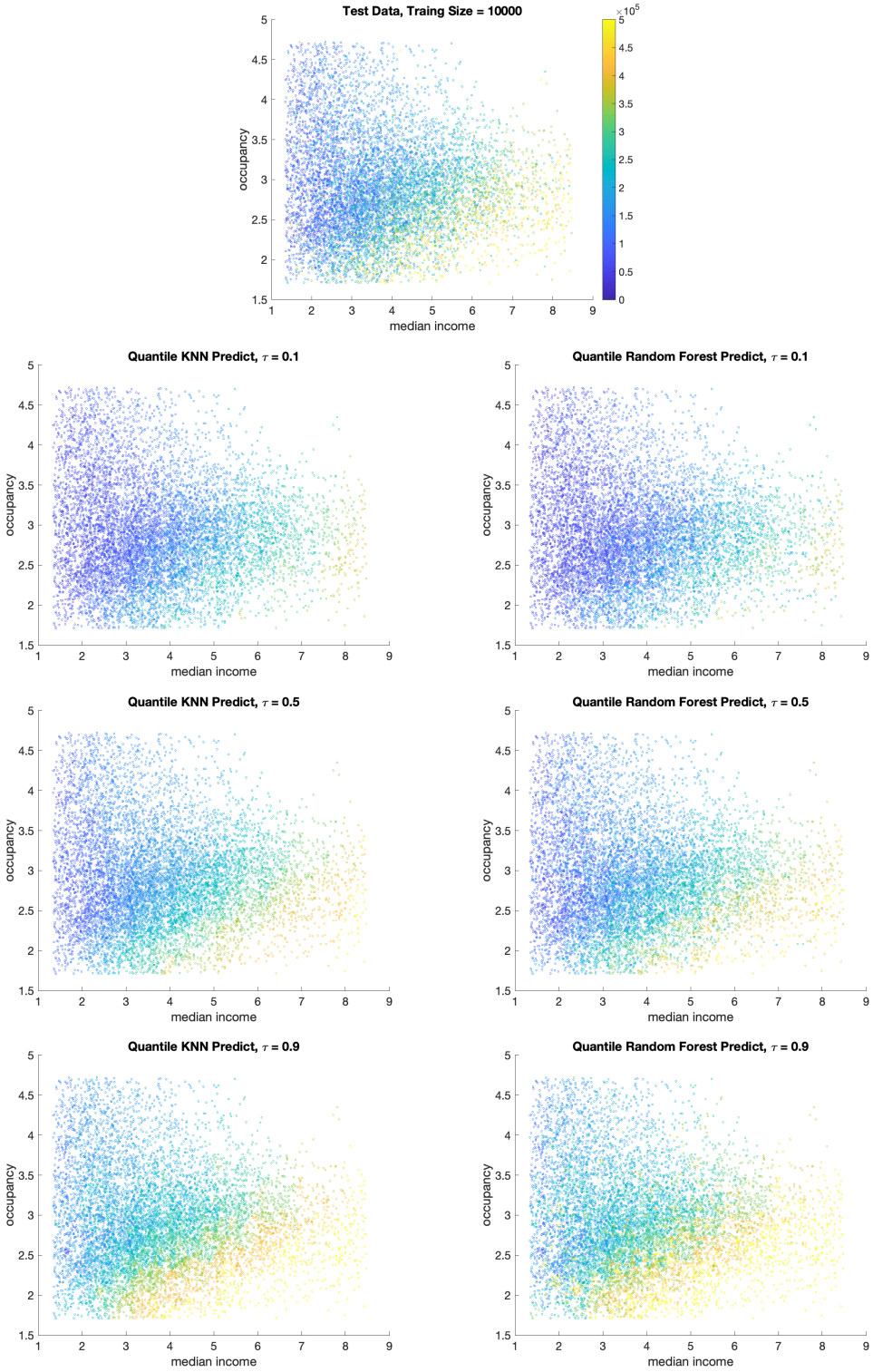

Figure 3: Comparison between predictions from two methods for California housing data, with respect to $\tau = 0.1, 0.5, 0.9$. The plot at the top presents the true testing data value, and the color scale is the same among all plots.

Training Size	τ	QKNN	QRF
1000	0.5	**0.1469**	0.1689
5000	0.5	**0.1417**	0.1629
10000	0.5	**0.1396**	0.1599

Training Size	CI	QKNN	QRF
1000	90%	**0.8361**	0.8335
5000	90%	**0.8477**	0.8360
10000	90%	**0.8479**	0.8395
1000	95%	**0.8900**	0.8895
5000	95%	**0.8997**	0.8896
10000	95%	**0.9037**	0.8902

Table 2: Average test set prediction error (across the 50 test sets) on California housing data. For median, we report the mean squared errors; for other quantiles, we report the averaged proportion of the true data located in the predicted confidence interval. The number of nearest neighbors, K is chosen as 5 for quantile K-NN fused lasso.

5.2.2 Chicago Crime Data

We apply quantile K-NN fused lasso and quantile random forest to a dataset of publicly-available crime report counts in Chicago, Illinois in 2015. We preprocess the dataset in the same way as Tansey et al. (2018) by merging all observations into a fine-grained 100×100 grid based on latitude and longitude, taking the log of the total counts in each cell, and omitting all cells with zero count. The resulting preprocessed data contains a total number of 3756 data points along the grid. Similar to the experiment on California housing data, we perform a train-test split with training size 500, 1000, 1500, and 2000, and model on the median counts according to the position on the X and Y axes of the grid. We then predict the value on test set and report the averaged square errors over 50 test sets. The parameters for both methods are selected in the same way as in the previous experiment.

Training Size	τ	QKNN	QRF
500	0.5	1.2051	**1.1453**
1000	0.5	0.9907	**0.9823**
1500	0.5	**0.9156**	0.9220
2000	0.5	**0.8149**	0.8329

Table 3: Average test set prediction error (across the 50 test sets) on Chicago Crime data.

From the results of Table 3, we see that quantile K-NN fused lasso still outperforms quantile random forest in MSE if the training size is at least 1500. We notice that for sample size, our method suffers. This is presumably due to the fact that the raw data is not smooth enough. Yet, Figure 4 showss that quantile K-NN fused lasso capture local patterns more successfully than quantile random forest in most regions.

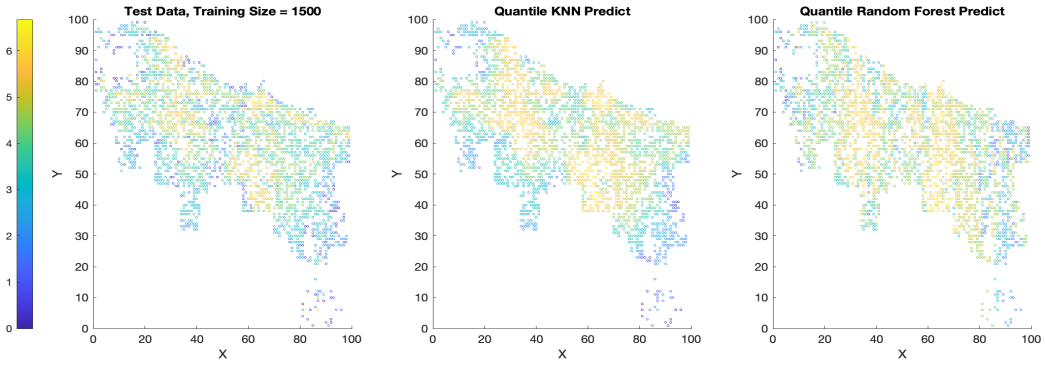

Figure 4: *Left*: one test data from Chicago Crime Data, under training size 1500. *Middle*: the estimate obtained via quantile K-NN fused lasso. *Right*: the estimate obtained via quantile random forest.

A Closed-form Solution to the Primal in ADMM Algorithm

In Section 3.1, we introduce an ADMM algorithm to compute quantile K-NN estimates. The algorithm requires to solve the primal problem (12)

$$\theta \leftarrow \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \frac{R}{2}\|\theta - z + u\|^2 \right\}.$$

We can solve the problem coordinate-wisely: for $i = 1, ..., n$, we find the minimizer

$$\theta_i \leftarrow \underset{\theta_i \in \mathbb{R}}{\arg\min} \left\{ \rho_\tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2 \right\}.$$

By definition,

$$\rho_\tau(y_i - \theta_i) = \begin{cases} \tau(y_i - \theta_i) & \text{if } y_i - \theta_i > 0, \\ (\tau - 1)(y_i - \theta_i) & \text{if } y_i - \theta_i < 0, \\ 0 & \text{if } y_i - \theta_i = 0. \end{cases}$$

We discuss the three cases separately.

(1) When $y_i - \theta_i > 0$, $\theta_i \leftarrow \arg\min \left\{ \tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2 \right\}$. Take the derivative and set to 0 to obtain $\theta_i = z_i - u_i + \tau/R$. The condition $y_i - \theta_i > 0$ then becomes $y_i - z_i + u_i > \tau/R$.

(2) When $y_i - \theta_i < 0$, $\theta_i \leftarrow \arg\min \left\{ (\tau - 1)(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2 \right\}$. Take the derivative and set to 0 to obtain $\theta_i = z_i - u_i + (\tau - 1)/R$. The condition $y_i - \theta_i < 0$ then becomes $y_i - z_i + u_i < (\tau - 1)/R$.

(3) When $y_i - \theta_i = 0$, it is simple to get $\theta_i = y_i$.

To summarize, the closed-form solution to the primal (12) is

$$\theta_i = \begin{cases} z_i - u_i + \frac{\tau}{R} & \text{if } y_i - z_i + u_i > \frac{\tau}{R}, \\ z_i - u_i + \frac{\tau - 1}{R} & \text{if } y_i - z_i + u_i < \frac{\tau - 1}{R}, \\ y_i & \text{otherwise;} \end{cases}$$

for $i = 1, ..., n$.

B General Lemmas

Definition 1. The function Δ^2 is defined as

$$\Delta^2(\delta) := \sum_{i=1}^{n} \min\{|\delta_i|, \delta_i^2\},$$

where $\delta_i \in \mathbb{R}^n$. We also write $\Delta(\delta) = \{\Delta^2(\delta)\}^{1/2}$.

Definition 2. For a set $S \subset \mathbb{R}^n$, the sub-Gaussian width of S is defined as

$$SGW(S) = \mathbb{E}\left(\sup_{v \in S} \sum_{i=1}^{n} s_i v_i \right),$$

where $s_1, ..., s_n$ are independent 1-subgaussian random variables.

The notation of sub-Gaussian width is not used very often in literature compared to a similar definition of Gaussian width,

$$GW(S) = \mathbb{E}\left(\sup_{v \in K} \sum_{i=1}^{n} z_i v_i \right),$$

where $z_1, ..., z_n$ are independent standard normal random variables. In fact, the sub-Gaussian width shares many common properties with the Gaussian width, as we can upper bound the sub-Gaussian width by a constant times the Gaussian width using generic chaining; see Chapter 4 in Talagrand (2005) and also Banerjee et al. (2014) for precise explanations.

Definition 3. We define the empirical loss function

$$\hat{M}(\theta) = \sum_{i=1}^{n} \hat{M}_i(\theta_i),$$

where

$$\hat{M}_i(\theta_i) = \rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*).$$

Setting $M_i(\theta_i) = \mathbb{E}(\rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*))$, the population version of $\hat{M}$ becomes

$$M(\theta) = \sum_{i=1}^{n} M_i(\theta_i).$$

Now, we consider the M-estimator

$$\begin{aligned} \hat{\theta} \;&=\; \underset{\theta \in \mathbb{R}^n}{\arg\min} \quad \hat{M}(\theta) \\ &\qquad \text{subject to } \; \theta \in S, \end{aligned} \tag{17}$$

and $\theta^* \in \arg\min_{\theta \in \mathbb{R}^n} M(\theta)$. Throughout, we assume that $\theta^* \in S \subset \mathbb{R}^n$.

Lemma 3. *With the notation from before,*

$$M(\hat{\theta}) \leq \sup_{v \in S} \left\{ M(v) - \hat{M}(v) \right\}.$$

Proof. See the proof of Lemma 6.2 in Padilla and Chatterjee (2020).

Lemma 4. *Suppose that Assumption 2 holds. Then there exists a constant $c_\tau > 0$ such that for all $\delta \in \mathbb{R}^n$, we have*

$$M(\theta + \delta) \geq c_\tau \Delta^2(\delta).$$

Proof. See the proof of Lemma 6.7 in Padilla and Chatterjee (2020).

Corollary 5. *Under Assumption 2, if $\theta^* \in S$, we have that*

$$\mathbb{E}\left\{ \Delta^2(\hat{\theta} - \theta^*) \right\} \leq \mathbb{E}\left[\sup_{v \in S} \left\{ M(v) - \hat{M}(v) \right\} \right]. \tag{18}$$

Next, we proceed to bound the right hand side of (18).

Lemma 6. *(Symmetrization) Under Assumption 2, we have that*

$$\mathbb{E}\left[\sup_{v \in S} \left\{ M(v) - \hat{M}(v) \right\} \right] \leq 2\mathbb{E}\left\{ \sup_{v \in S} \sum_{i=1}^{n} \xi_i \hat{M}_i(v_i) \right\},$$

where $\xi_1, ..., \xi_n$ are independent Rademacher variables independent of $y_1, ..., y_n$.

Proof. See the proof of Lemma 6.3 in Padilla and Chatterjee (2020).

Lemma 7. *(Contraction principle) Under Assumption 2, we have that*

$$\mathbb{E}\left\{ \sup_{v \in S} \sum_{i=1}^{n} s_i \hat{M}_i(v_i) \right\} \leq 2SGW(S - \theta^*) = 2SGW(S),$$

where $s_1, ..., s_n$ are independent 1-subgaussian random variables independent of $y_1, ..., y_n$.

Proof. Recall that $\hat{M}_i(v_i) = \rho_\tau(y_i - v_i) - \rho_\tau(y_i - \theta_i^*)$. Clearly, these are 1-Lipschitz continuous functions. Therefore,

$$\begin{aligned}
\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} s_i\hat{M}_i(v_i)\right\} &= \mathbb{E}\left(\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} s_i\hat{M}_i(v_i)\Big| y\right\}\right) \\
&\leq \mathbb{E}\left(\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} s_i v_i\Big| y\right\}\right) \\
&= \mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} s_i(v_i-\theta^*)\right\} + \mathbb{E}\left(\sum_{i=1}^{n} s_i\theta^*\right) \\
&= \mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} s_i(v_i-\theta^*)\right\}
\end{aligned}$$

where the inequality follows from the Gaussian version of Talagrand's contraction principle, see corallary 3.17 in Ledoux and Talagrand (2013) and 7.2.13 in Vershynin (2018). □

C K-NN Embeddings

This section follows from Section D and E in Padilla et al. (2020a) and it originally appeared in the flow-based proof of Theorem 4 in Von Luxburg et al. (2014). The main idea is to embed a mesh on a K-NN graph corresponding to the given observations $X = (x_1, ..., x_n)$ under Assumptions 3-5 so that we are able to bound the total variation along the grid graph from below. In this way, we can further derive an upper bound on the loss of the optimization problem (8) and (10) with respect to the loss function defined in Definition 1.

First, we need to construct a valid grid discussed in Definition 17 of Von Luxburg et al. (2014) with a minor modification. With high probability, a valid grid graph G satisfies the following: (i) the grid width is not too small: each cell of the grid contains at least one of the design points; (ii) the grid width is not too large: points in the same or neighboring cells of the grid are always connected in the K-NN graph.

Given $N \in \mathbb{N}$, a d-dimensional grid graph $G_{\text{lat}} \in [0,1]^d$ has vertex set V_{lat} and edge set E_{lat}. The grid graph has equal side lengths, and the total number of nodes $|V_{\text{lat}}| = N^d$. Without loss of generality, we assume that the nodes of the grid correspond to the points

$$P_{\text{lat}}(N) = \left\{\left(\frac{i_1}{N} - \frac{1}{2N}, ..., \frac{i_d}{N} - \frac{1}{2N}\right) : i_1, ..., i_d \in \{1, ..., N\}\right\}. \tag{19}$$

Moreover, for $z, z' \in P_{\text{lat}}(N)$, $(z, z') \in E_{\text{lat}}(N)$ if and only if $\|z - z'\|_2 = \frac{1}{N}$.

Now, we define $I(N) = h^{-1}\{P_{\text{lat}}(N)\}$ as the mesh in the covariate space $\mathcal{X}$ corresponding to the grid graph $G_{\text{lat}}(N) \in [0,1]^d$ through the homeomorphism h from Assumption 5. In general, $I(N)$ performs as a quantifization in the domain $\mathcal{X}$; see Alamgir et al. (2014) for more details. We denote the elements in $I(N)$ by $u_1, ..., u_{N^d}$, and define a collection of cells $\{C(x)\} \in \mathcal{X}$ for $x \in I(N)$ as

$$C(x) = h^{-1}\left(\left\{z \in [0,1]^d : h(x) = \underset{x' \in P_{\text{lat}}(N)}{\arg\min} \|z - x'\|_\infty\right\}\right). \tag{20}$$

In order to analyze the behavior of the proposed estimator $\hat{\theta}$ through the grid embedding, we construct two vectors denoted by $\theta_I \in \mathbb{R}^n$ and $\theta^I \in \mathbb{R}^{N^d}$ for any signal $\theta \in \mathbb{R}^n$.

The first vector, $\theta_I \in \mathbb{R}^n$ incorporates information about the samples $X = (x_1, ..., x_n)$ and the cells $\{C(x)\}$. The idea is to force covariates x_i fallen in the same cell to take the same signal value after mapping with the homeomorphism. Formally, we define

$$(\theta_I)_i = \theta_j \quad \text{where } j = \underset{l=1,...,n}{\arg\min} \|h(P_I(x_i)) - h(x_l)\|_\infty, \tag{21}$$

where $P_I(x)$ is the point in $I(N)$ such that $x \in C(P_I(x))$ and if there exists multiple points satisfying the condition, we arbitrarily select one.

The second vector, $\theta^I \in \mathbb{R}^{N^d}$ records coordinates corresponding to the different nodes of the mesh (centers of cells), and is associated with θ_I. We first induce a signal in $\mathbb{R}^{N^d}$ corresponding to the elements in $I(N)$ as

$$I_j = \{i = 1, ..., n : P_I(x_i) = u_j\}, \text{ for } j = 1, ..., N^d.$$

If $I_j \neq \emptyset$, then there exists $i_j \in I_j$ such that $(\theta_I)_i = \theta_{i_j}$ for all $i \in I_j$; if I_j is empty, we require $\theta_{i_j} = 0$. We can thus define

$$\theta^I = \left(\theta_{i_1}, ..., \theta_{i_{N^d}}\right).$$

Note that in the proof of Lemma 11 in Padilla et al. (2020a), the authors showed that under Assumptions 3-5, with probability close to 1,

$$\max_{x \in I(N)} |C(x)| \leq \text{poly}(\log n). \tag{22}$$

We will use this inequality in our proofs later, but we will not make the polynomial function of $\log n$ explicit.

D Definition of Piecewise Lipschitz

To make sure Assumption 1 is valid, we require a piecewise Lipschitz condition on the regression function f_0. In this section, we provide the detailed definition of the class of piecewise Lipschitz functions, followed from Definition 1 in Padilla et al. (2020a). All notations follow the same from the main context, besides an extra notation on the boundary of a set A, denoted by ∂A.

Definition 4. Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. We say that a bounded function $g : [0,1]^d \to \mathbb{R}$ is *piecewise Lipschitz* if there exists a set $\mathcal{S} \subset (0,1)^d$ that has the following properties:

- The set $\mathcal{S}$ has Lebesgue measure zero.
- For some constants $C_{\mathcal{S}}, \epsilon_0 > 0$, we have that $\mu(h^{-1}\{B_\epsilon(\mathcal{S}) \cap ([0,1]^d \backslash \Omega_\epsilon)\}) \leq C_{\mathcal{S}}\epsilon$ for all $0 < \epsilon < \epsilon_0$.
- There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, then $|g(z) - g(z')| \leq L_0 \|z - z'\|_2$.

Roughly speaking, a bounded function g is piecewise Lipschitz if there exists a small set $\mathcal{S}$ that partitions $[0,1]^d$ in such a way that g is Lipschitz within each connected component of the partition. Theorem 2.2.1 in Ziemer (2012) implies that if g is piecewise Lipschitz, then g has bounded variation on any open set within a connected component.

E Theorem 1

E.1 Notations

For a matrix D, we denote its kernel by $\text{Ker}(D)$ and its Moore–Penrose inverse by $D^\dagger$. We also write Π as the projection matrix onto $\text{Ker}(D)$, and denote $\text{Ker}(D)^\perp$ as the orthogonal complement to the kernel space of D.

Our goal is to upper bound the expectation of $M(\cdot) - \hat{M}(\cdot)$ in the constrained set S defined by

$$S = \left\{\theta \in \mathbb{R}^n \ : \ \|\nabla_G \theta\|_1 \leq V n^{1-1/d} \text{poly}(\log n)\right\}, \tag{23}$$

where $V = \|\nabla_G \theta^*\|_1 / [n^{1-1/d} \text{poly}(\log n)]$ and $V \geq V^*$. Through the K-NN embedding, we can instead bound the expected loss in the embedded set defined by

$$\tilde{S} = \left\{\theta \in \mathbb{R}^n \ : \ \|D\theta^I\|_1 \leq V n^{1-1/d} \text{poly}(\log n)\right\}, \tag{24}$$

where θ^I follows the definition in Appendix C.

E.2 Auxiliary lemmas for Proof of Theorem 1

Lemma 8. *For $v \in \mathbb{R}^n$, if $\|\nabla_G v\|_1 \leq \tilde{V}$, and $\Delta^2(v - \theta^*) \leq t^2$, then $\|Dv^I\|_1 \leq \tilde{V}$, and $\Delta^2(v^I - \theta^{*,I}) \leq c_1 t^2$ for some constant c_1.*

Proof. Lemma 4 in Padilla et al. (2020a) obtains the inequality

$$\|Dv^I\|_1 \leq \|\nabla_G v\|_1, \ \forall v \in \mathbb{R}^\cdot$$

Hence, the first claim follows.
Next, we observe that for any vector $u \in \mathbb{R}^n$,

$$\Delta^2(u^I) = \frac{1}{N^d}\left(\sum_{j=1}^{N^d} \min\left\{|u_{i_j}|, u_{i_j}^2\right\}\right) \leq \frac{1}{N^d}\left(\sum_{i=1}^{n} \min\left\{|u_i|, u_i^2\right\}\right) \leq c_1 \Delta^2(u),$$

for some positive constant c_1. The second claim follows then. □

Lemma 8 gives us the fact that

$$\left\{v \in S : \Delta^2(v-\theta^*) \le t^2\right\} \subseteq \left\{v \in \tilde{S} : \Delta^2(v^I - \theta^{*,I}) \le c_1 t^2\right\}. \tag{25}$$

Lemma 9. *Under Assumptions 1-5, we have that*

$$\mathbb{E}\left[\sup_{v\in\tilde{S}:\Delta^2(v^I-\theta^{*,I})\le t^2} \sum_{i=1}^{n} \xi_i(v_i - \theta_i^*)\right] \le \text{poly}(\log n)\mathbb{E}\left[\sup_{v\in\tilde{S}:D^2(v^I-\theta^{*,I})\le t^2} \sum_{j=1}^{N^d} \tilde{\xi}_j(v_j^I - \theta_j^{*,I})\right],$$

where $\tilde{\xi} \in \mathbb{R}^{N^d}$ is a 1-subgaussian vector whose coordinates are independent.

Proof. We notice that

$$\begin{aligned}\xi^\top(v-\theta^*) &= \xi^\top(v-v_I) + \xi^\top(v_I - \theta_I^*) + \xi^\top(\theta_I^* - \theta^*)\\ &\le 2\left(\|\nabla_G v\|_1 + \|\nabla_G \theta^*\|_1\right) + \xi^\top(v_I - \theta_I^*).\end{aligned}$$

Moreover,

$$\xi^\top(v_I - \theta_I^*) = \sum_{j=1}^{N^d}\sum_{l\in I_j} \xi_l(v_{i_j} - \theta_{i_j}^*) = \left\{\max_{u\in I}|C(u)|\right\}^{1/2} \tilde{\xi}^\top(v^I - \theta^{*,I})$$

where

$$\tilde{\xi}_j = \left\{\max_{u\in I}|C(u)|\right\}^{-1/2} \sum_{l\in I_j}\xi_l.$$

Clearly, the $\tilde{\xi}_1, ..., \tilde{\xi}_{N^d}$ are independent and also 1-subgaussian as the original Rademacher random variables $\xi_1, ..., \xi_n$. Then, following from (22), we have

$$\xi^\top(v-\theta^*) \le \text{poly}(\log n)\tilde{\xi}^\top(v^I - \theta^{*,I}).$$

Hence, the desired inequality holds. □

Lemma 10. *Let $\delta \in \mathbb{R}^{N^d}$ with $\Delta^2(\delta) \le t^2$. Then*

$$\|\Pi\delta\|_\infty \le \frac{t^2}{N^d} + \frac{t}{\sqrt{N^d}}.$$

Proof. Notice that $\text{Ker}(D) = \text{span}(1_{N^d})$, then $\Pi\delta = \delta^\top v \cdot v$, where $v = \frac{1}{\sqrt{N^d}}1_{N^d}$. Hence

$$\|\Pi\delta\|_\infty = \|\delta^\top v \cdot v\|_\infty \le |\delta^\top v| \cdot \|v\|_\infty = \frac{|\delta^\top v|}{\sqrt{N^d}}. \tag{26}$$

Now,

$$\begin{aligned}|\delta^\top v| &\le \sum_{i=1}^{N^d} |\delta_i||v_i|\\ &= \sum_{i=1}^{N^d} |\delta_i||v_i|1_{\{|\delta_i|>L\}} + \sum_{i=1}^{N^d} |\delta_i||v_i|1_{\{|\delta_i|\le L\}}\\ &\le \|v\|_\infty \sum_{i=1}^{N^d} |\delta_i|1_{\{|\delta_i|>L\}} + \|v\|\left(\sum_{i=1}^{N^d}\delta_i^2 1_{\{|\delta_i|\le L\}}\right)^{1/2}\\ &\le \frac{t^2}{\sqrt{N^d}} + t,\end{aligned} \tag{27}$$

where the first inequality follows from the triangle inequality, the second from Hölder's and Cauchy Schwarz inequalities. The claim follows combining (26) with (27). □

Lemma 11. *Under Assumptions 1-5, for $N \asymp n^{1/d}$, we have that*

$$SGW\left(\left\{\delta : \delta \in \tilde{S}, \Delta^2(\delta) \leq c_1 t^2\right\}\right) \leq C_d (\log n)^{1/2} \left(\frac{c_1 t^2}{\sqrt{n}} + \sqrt{c_1} t\right) + V n^{1-1/d} \text{poly}(\log n).$$

Proof. Recall that the projection on $\text{Ker}(D)^{\perp}$, $D^{\dagger} D = I - \Pi$, which yields

$$\tilde{\xi}^{\top} \delta = \tilde{\xi}^{\top} \Pi \delta + \tilde{\xi}^{\top} D^{\dagger} D \delta.$$

Then we have

$$\begin{aligned} \mathbb{E}\left[\sup_{\delta \in \tilde{S}: \Delta^2(\delta) \leq t^2} \tilde{\xi}^{\top} \delta\right] &\leq \mathbb{E}\left[\sup_{\delta \in \tilde{S}: \Delta^2(\delta) \leq t^2} \tilde{\xi}^{\top} \Pi \delta\right] + \mathbb{E}\left[\sup_{\delta \in \tilde{S}: \Delta^2(\delta) \leq t^2} \tilde{\xi}^{\top} D^{\dagger} D \delta\right] \\ &=: T_1 + T_2 \end{aligned} \tag{28}$$

We first bound T_1. Notice that Π is idempotent, i.e., $\Pi^2 = \Pi$, thus

$$\begin{aligned} \tilde{\xi}^{\top} \Pi \delta &= \tilde{\xi}^{\top} \Pi \Pi \delta \\ &\leq \|\tilde{\xi}^{\top} \Pi\|_1 \|\Pi \delta\|_{\infty} \\ &= \|\tilde{\xi}^{\top} v \cdot v\|_1 \|\Pi \delta\|_{\infty} \\ &\leq N^d \|\tilde{\xi}\|_{\infty} \|v\|_{\infty}^2 \|\Pi \delta\|_{\infty} \\ &= \|\tilde{\xi}\|_{\infty} \|\Pi \delta\|_{\infty} \end{aligned} \tag{29}$$

where the first two inequalities follow from Hölder's inequality. Then, from Lemma 10 and Gaussian maximal inequality (see Chapter 1 in Rigollet and Hütter (2015)), we get

$$T_1 \leq \mathbb{E}\left[\|\tilde{\xi}\|_{\infty}\right] \left(\frac{c_1 t^2}{\sqrt{N^d}} + \sqrt{c_1} t\right) \leq C_d (\log n)^{1/2} \left(\frac{c_1 t^2}{\sqrt{n}} + \sqrt{c_1} t\right). \tag{30}$$

for some positive constant C_d.
Next, we bound T_2. Notice that

$$\begin{aligned} T_2 &\leq \mathbb{E}\left[\sup_{\delta \in \tilde{S}: \Delta^2(\delta) \leq t^2} \|\tilde{\xi}^{\top} D^{\dagger}\|_{\infty} \|D\delta\|_1\right] \\ &\leq V n^{1-1/d} \text{poly}(\log n) \mathbb{E}\left[\|\tilde{\xi}^{\top} D^{\dagger}\|_{\infty}\right], \end{aligned} \tag{31}$$

thus we only need to bound $\mathbb{E}\left[\|\tilde{\xi}^{\top} D^{\dagger}\|_{\infty}\right]$. From Section 3 in Hütter and Rigollet (2016), we write $D^{\dagger} = [s_1, ..., s_m]$, and $\max_{j=1,...,m} \|s_j\|_2$ is bounded above by bounded by $(\log n)^{1/2}$ for $d = 2$ or by a constant for $d > 2$. Then,

$$\begin{aligned} \mathbb{E}\left[\|\tilde{\xi}^{\top} D^{\dagger}\|_{\infty}\right] &= \mathbb{E}\left[\max_{j=1,...,m} |s_j^{\top} \tilde{\xi}|\right] \\ &= \mathbb{E}\left[\max_{j=1,...,m} |\tilde{s}_j^{\top} \tilde{\xi}|\right] \max_{j=1,...,m} \|s_j\|_2, \end{aligned} \tag{32}$$

where

$$\tilde{s}_j = \frac{s_j}{\max_{j=1,...,m} \|s_j\|_2}.$$

Moreover, $\tilde{s}_j^{\top} \tilde{\xi}$ is also sub-Gaussian but with parameter at most 1. Combining (31) with (32), we obtain that

$$T_2 \leq C_d V n^{1-1/d} \text{poly}(\log n),$$

for some positive constant C_d. The conclusion follows then. □

Theorem 12. *Suppose that*

$$2SGW\left(\{\delta \in \tilde{S} : \Delta^2(\delta^I) \leq c_1 \eta^2\}\right) \leq \kappa(\eta),$$

for a function $\kappa : \mathbb{R} \to \mathbb{R}$. Then for all $\eta > 0$ we have that

$$\mathbb{P}\left(\Delta^2(\hat{\delta}) > \eta^2\right) \leq \frac{\text{poly}(\log n)\kappa(\eta)}{c_\tau \eta^2},$$

where c_τ is the constant from Lemma 4. Furthermore, if $\{r_n\}$ is a sequence such that

$$\lim_{t\to\infty} \sup_n \frac{\text{poly}(\log n)\kappa(t r_n n^{1/2})}{t^2 r_n^2 n} \to 0,$$

then

$$\frac{1}{n}\Delta^2(\hat{\theta} - \theta^*) = O_{\mathbb{P}}(r_n^2).$$

Proof. Let $\hat{\delta} = \hat{\theta} - \theta^*$ and suppose that

$$\frac{1}{n}\delta^2 > \frac{\eta^2}{n}. \tag{33}$$

Next, let $q^2 = \Delta^2(\hat{\delta})$. Then define $g : [0,1] \to \mathbb{R}$ as $g(t) = \Delta^2(t\hat{\delta})$. Clearly, g is a continuous function with $g(0) = 0$ and $g(1) = q^2$. Therefore, there exists $t_{\hat{\delta}}$ such that $g(t_{\hat{\delta}}) = \eta^2$. Hence, letting $\tilde{\delta} = t_{\hat{\delta}}$ we observe that by the basic inequality $\hat{M}(\theta^* + \tilde{\delta}) \leq 0, \delta \in S$ by convexity of S, and $\Delta^2(\tilde{\delta}) = \eta^2$ by construction. This implies, along with Lemma 4, that

$$\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v) \geq M(\theta^* + \tilde{\delta}) - \hat{M}(\theta^* + \tilde{\delta}) \geq M(\theta^* + \tilde{\delta}) \geq c_\tau \eta^2.$$

Therefore, combing the results of Lemma 6-9, we have

$$\begin{aligned}
\mathbb{P}\left\{\Delta^2(\hat{\delta}) > \eta^2\right\} &\leq \mathbb{P}\left\{\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v) \geq c_\tau \eta^2\right\} \\
&\leq \frac{1}{c_\tau \eta^2}\mathbb{E}\left\{\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v)\right\} \\
&\leq \frac{1}{c_\tau \eta^2}\mathbb{E}\left\{\sup_{v \in \tilde{S} : \Delta^2(v-\theta^*) \leq c_1\eta^2} M(v) - \hat{M}(v)\right\} \\
&\leq \frac{2\text{poly}(\log n)}{c_\tau \eta^2} SGW\left(\{\delta \in \tilde{S} : \Delta^2(\delta^I) \leq c_1 \eta^2\}\right) \\
&\leq \frac{\text{poly}(\log n)\kappa(\eta)}{c_\tau \eta^2},
\end{aligned}$$

where the second inequality follows from Markov's inequality. This completes the proof. □

E.4 Proof of Theorem 1

Proof. The claim follows immediately from Lemma 11 and Theorem 12 by setting

$$r_n \asymp n^{-1/2d}\text{poly}(\log n).$$

□

F Theorem 2

F.1 Auxiliary lemmas for Proof of Theorem 2

Throughout we assume that Assumptions 1-5 hold, and all notations follow the same as in the proof of Theorem 1.

Lemma 13. *Let $\epsilon \in (0,1)$, then there exists a choice*

$$\lambda = \begin{cases} \Theta\{\log n\} & \text{for } d = 2, \\ \Theta\left\{(\log n)^{1/2}\right\} & \text{for } d > 2, \end{cases}$$

such that for a constant $C_0 > 0$, we have that, with probability at least $1 - \epsilon/4$,

$$\kappa(\hat{\theta} - \theta^*) \in \mathcal{A},$$

with

$$\mathcal{A} := \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{c_1 \Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \sqrt{c_1} \Delta(\tilde{\theta} - \theta^*) \right] \right) \right\}$$

for all $\kappa \in [0, 1]$, where

$$R_1 = C_d \log \left\{ \frac{n}{\epsilon} \right\}^{1/2},$$

$$R_2 = \begin{cases} C_d (\log n)^{1/2} \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d = 2, \\ C_d \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d > 2, \end{cases}$$

and C_0, C_d are positive constants.

Proof. Pick $\kappa \in [0, 1]$ fixed, and let $\tilde{\delta} = \kappa(\hat{\theta} - \theta^*)$. Then by the optimality of $\hat{\theta}$ and the convexity of (8), we have that

$$\sum_{i=1}^{n} \rho_\tau(y_i - \tilde{\theta}_i) + \lambda \|\nabla_G \tilde{\theta}\|_1 \leq \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i^*) + \lambda \|\nabla_G \theta^*\|_1,$$

where $\tilde{\theta} = \theta^* + \tilde{\delta}$. Then as in the proof of Lemma 3 from Belloni and Chernozhukov (2011),

$$0 \leq \lambda \left[\|\nabla_G \theta^*\|_1 - \|\nabla_G \tilde{\theta}\|_1 \right] + (\tilde{\theta} - \theta^*)^\top a^*, \tag{34}$$

where $a_i^* = \tau - 1\{y_i \leq \theta_i^*\}$ for $i = 1, ..., n$.

Next, we bound the second term of the right hand of (34). From Lemma 9, we know it is sufficient to bound

$$\tilde{a}^\top (\tilde{\theta}^I - \theta^{*,I}),$$

where

$$\tilde{a}_j = \left\{ \max_{u \in I} |C(u)| \right\}^{-1/2} \sum_{l \in I_j} a_l^*.$$

Now,

$$\begin{aligned} \tilde{a}^\top (\tilde{\theta}^I - \theta^{*,I}) &= \tilde{a}^\top \Pi (\tilde{\theta}^I - \theta^{*,I}) + \tilde{a}^\top D^\dagger D (\tilde{\theta}^I - \theta^{*,I}) \\ &=: A_1 + A_2 \end{aligned} \tag{35}$$

From Lemmas 8, 10 and (29), we obtain

$$\begin{aligned} A_1 &\leq \|\tilde{a}\|_\infty \left(\frac{\Delta^2(\tilde{\theta}^I - \theta^{*,I})}{n^{1/2}} + \Delta(\tilde{\theta}^I - \theta^{*,I}) \right) \\ &\leq \|\tilde{a}\|_\infty \left(\frac{c_1 \Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \sqrt{c_1} \Delta(\tilde{\theta} - \theta^*) \right) \end{aligned} \tag{36}$$

To bound A_2, we use the result of Lemma 8 to obtain

$$\begin{aligned} A_2 &\leq \|(D^\dagger)^\top \tilde{a}\|_\infty \left[\|D\tilde{\theta}^I\|_1 + \|D\theta^{*,I}\|_1 \right] \\ &\leq \|(D^\dagger)^\top \tilde{a}\|_\infty \left[\|\nabla_G \tilde{\theta}\|_1 + \|\nabla_G \theta^*\|_1 \right] \end{aligned} \tag{37}$$

Since a^* is Bernoulli with parameter τ and is thus sub-Gaussian with parameter $\frac{1}{4}$, $\tilde{a}$ is also sub-Gaussian. As in the proof of Theorem 2 from Hütter and Rigollet (2016), we have that the following two inequalities hold simultaneously on an event of probability at least $1 - 2\epsilon$,

$$\|\tilde{a}\|_\infty \leq R_1 := C_d \log \left\{ \frac{n}{\epsilon} \right\}^{1/2}, \quad \|(D^\dagger)^\top \tilde{a}\|_\infty \leq R_2 := \begin{cases} C_d (\log n)^{1/2} \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d = 2, \\ C_d \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d > 2, \end{cases}$$

for some constant c_k.

Then, with probability at least $1-2\epsilon$, By choosing $\lambda = 2R_2$, we obtain

$$\kappa(\hat{\theta}-\theta^*) \in \mathcal{A} := \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2}\left[\frac{c_1 \Delta^2(\tilde{\theta}-\theta^*)}{n^{1/2}} + \sqrt{c_1}\Delta(\tilde{\theta}-\theta^*) \right] \right) \right\},$$

for some positive constant C_0. □

F.2 Proof of Theorem 2

Proof. Let $\epsilon \in (0,1)$. By Lemma 13 we can suppose that the following event

$$\Omega = \left\{ \kappa(\hat{\theta}-\theta^*) \in \mathcal{A},\ \forall \kappa \in [0,1] \right\} \tag{38}$$

happen with probability at least $1-\epsilon/2$ with $\mathcal{A}$ as in Lemma 13. Then,

$$\mathbb{P}\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \leq \mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] + \frac{\epsilon}{2}.$$

Now suppose that the event

$$\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega$$

holds. As in the proof of Theorem 12, there exists $\tilde{\delta} = t_{\hat{\delta}}\hat{\delta}$ with $t_{\hat{\delta}} \in [0,1]$ such that $\tilde{\delta} \in \mathcal{A}$, $\Delta^2(\tilde{\delta}) = \eta^2$. Hence, by the basic inequality,

$$\hat{M}(\theta^*+\tilde{\delta}) + \lambda\left[\|\nabla_G(\theta^*+\tilde{\delta})\|_1 - \|\nabla_G\theta^*\|_1 \right] \leq 0.$$

Then,

$$\begin{aligned}\sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \left[M(\theta^*+\delta) - \hat{M}(\theta^*+\delta) + \lambda\left\{ \|\nabla_G\theta^*\|_1 - \|\nabla_G(\theta^*+\tilde{\delta})\|_1 \right\} \right] &\geq M(\theta^*+\tilde{\delta}) \\ &\geq c_\tau \eta^2,\end{aligned}$$

where the second inequality follows from Lemma 4. Therefore,

$$\begin{aligned}\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] &\leq \mathbb{P}\Bigg(\Bigg\{ \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \Big[M(\theta^*+\delta) - \hat{M}(\theta^*+\delta) \\ &\qquad + \lambda\Big\{ \|\nabla_G\theta^*\|_1 - \|\nabla_G(\theta^*+\tilde{\delta})\|_1 \Big\} \Big] \geq c_\tau\eta^2 \Bigg\} \cap \Omega \Bigg) \\ &\leq \frac{1}{c_\tau\eta^2}\mathbb{E}\Bigg(\mathbf{1}_\Omega \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \Big[M(\theta^*+\delta) - \hat{M}(\theta^*+\delta) \\ &\qquad + \lambda\Big\{ \|\nabla_G\theta^*\|_1 - \|\nabla_G(\theta^*+\tilde{\delta})\|_1 \Big\} \Big] \Bigg) \\ &\leq \frac{1}{c_\tau\eta^2}\mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \Big[M(\theta^*+\delta) - \hat{M}(\theta^*+\delta) \Big] \right) \\ &\quad + \frac{\lambda}{c_\tau\eta^2}\mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \Big[\|\nabla_G\theta^*\|_1 - \|\nabla_G(\theta^*+\tilde{\delta})\|_1 \Big] \right) \\ &\leq \frac{1}{c_\tau\eta^2}\mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \Big[M(\theta^*+\delta) - \hat{M}(\theta^*+\delta) \Big] \right) \\ &\quad + \frac{\lambda}{c_\tau\eta^2}\mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta\in\mathcal{A},\Delta^2(\delta)\leq\eta^2} \|\nabla_G\delta\|_1 \right),\end{aligned} \tag{39}$$

where the second inequality follows from Markov's inequality, and the last from the triangle inequality.

Next, define

$$\mathcal{H}(\eta) = \{\delta \in \mathcal{A} : \Delta(\delta) \leq \eta\}.$$

Hence, if $\delta \in \mathcal{H}(\eta)$ and Ω holds, then

$$\|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{c_1 \Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \sqrt{c_1} \Delta(\tilde{\theta} - \theta^*) \right] \right) \tag{40}$$

where the inequality follow from the definition of $\mathcal{H}(\eta)$ and Lemma 13.

We now define

$$\mathcal{L}(\eta) = \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left\{ \|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{c_1 \Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \sqrt{c_1} \Delta(\tilde{\theta} - \theta^*) \right] \right\}, \Delta(\delta) \leq \eta \right\},$$

$$\tilde{\mathcal{L}}(\eta) = \left\{ \delta : \|D\delta^I\|_1 \leq C_0 \left\{ \|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{c_1 \Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \sqrt{c_1} \Delta(\tilde{\theta} - \theta^*) \right] \right\}, \Delta(\delta^I) \leq \sqrt{c_1}\eta \right\}.$$

Then

$$\begin{aligned}
\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] &\leq \frac{1}{c_\tau \eta^2} \mathbb{E} \left(\sup_{\delta \in \mathcal{L}(\eta)} \left[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \right] \right) + \frac{\lambda}{c_\tau \eta^2} \sup_{\delta \in \mathcal{L}(\eta)} \|\nabla_G \delta\|_1, \\
&\leq \frac{1}{c_\tau \eta^2} \mathbb{E} \left(\sup_{\delta \in \tilde{\mathcal{L}}(\eta)} \left[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \right] \right) + \frac{\lambda}{c_\tau \eta^2} \sup_{\delta \in \mathcal{L}(\eta)} \|\nabla_G \delta\|_1 \\
&\leq \frac{2\text{poly}(\log n)}{c_\tau \eta^2} \mathbb{E} \left(\sup_{\delta \in \tilde{\mathcal{L}}(\eta)} \sum_{i=1}^{N^d} \tilde{\xi}_i \delta_i^I \right) + \frac{\lambda}{c_\tau \eta^2} \sup_{\delta \in \mathcal{L}(\eta)} \|\nabla_G \delta\|_1.
\end{aligned} \tag{41}$$

By Lemma 11, we have

$$\begin{aligned}
\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] \leq \frac{2\text{poly}(\log n)}{c_\tau \eta^2} \Bigg\{ & C_d \left(\frac{c_1 \eta^2}{n^{1/2}} + \sqrt{c_1}\eta \right) (\log n)^{1/2} \\
& + C_d C_0 \left[V n^{1-1/d} \text{poly}(\log n) + \frac{R_1}{R_2} \left(\frac{c_1 \eta^2}{n^{1/2}} + \sqrt{c_1}\eta \right) \right] \\
& + \frac{\lambda}{c_\tau \eta^2} C_0 \left[V n^{1-1/d} \text{poly}(\log n) + \frac{R_1}{R_2} \left(\frac{c_1 \eta^2}{n^{1/2}} + \sqrt{c_1}\eta \right) \right] \Bigg\}.
\end{aligned}$$

Hence given our choice of λ, by choosing

$$\eta = \ c_\gamma n^{\frac{1}{2}(1-1/d)} \text{poly}(\log n)$$

for some $c_\gamma > 1$, we conclude that

$$\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] \leq \epsilon,$$

provided that c_γ is large enough. □

References

Morteza Alamgir, Gábor Lugosi, and Ulrike von Luxburg. "Density-preserving quantization with application to graph downsampling". In: *Proceedings of the 27th Conference on Learning Theory*. Vol. 35. 2014, pp. 543–559.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. "Estimation with norm regularization". In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 1556–1564.

Alexandre Belloni and Victor Chernozhukov. "l_1-penalized quantile regression in high-dimensional sparse models". In: *The Annals of Statistics* 39.1 (2011), pp. 82–130.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. "Conditional quantile processes based on series or many regressors". In: *Journal of Econometrics* 213.1 (2019), pp. 4–29.

Jon Louis Bentley. "Multidimensional divide-and-conquer". In: *Communications of the ACM* 23.4 (1980), pp. 214–229.

Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends in Machine Learning* 3.1 (2011), pp. 1–122.

Halley L. Brantley, Joseph Guinness, and Eric C. Chi. "Baseline drift estimation for air quality data using quantile trend filtering". In: *The Annals of Applied Statistics* 14.2 (2020), pp. 585–604.

Zongwu Cai and Xiaoping Xu. "Nonparametric quantile estimations for dynamic smooth coefficient models". In: *Journal of the American Statistical Association* 103.484 (2008), pp. 1595–1608.

Antonin Chambolle and Jérome Darbo. "On total variation minimization and surface evolution using parametric maximum flows". In: *International Journal of Computer Vision* 84.3 (2009), pp. 288–307.

Jie Chen, Haw-ren Fang, and Yousef Saad. "Fast approximate kNN graph construction for high dimensional data via recursive Lanczos bisection". In: *Journal of Machine Learning Research* 10 (2009), pp. 1989–2012.

Dennis D. Cox. "Asymptotics for M-type smoothing splines". In: *The Annals of Statistics* 11.2 (1983), pp. 530–551.

Randall L. Eubank. *Spline smoothing and nonparametric regression*. Vol. 90. M. Dekker New York, 1988.

Jianqing Fan, Yingying Fan, and Emre Barut. "Adaptive robust variable selection". In: *The Annals of Statistics* 42.1 (2014), pp. 324–351.

Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Xuming He, Pin Ng, and Stephen Portnoy. "Bivariate quantile smoothing splines". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 60.3 (1998), pp. 537–550.

Xuming He and Peide Shi. "Convergence rate of b-spline estimators of nonparametric conditional quantile functions". In: *Journal of Nonparametric Statistics* 3.3–4 (1994), pp. 299–308.

Dorit S. Hochbaum and Cheng Lu. "A faster algorithm for solving a generalization of isotonic median regression and a class of fused lasso problems". In: *SIAM Journal on Optimization* 27.4 (2017), pp. 2563–2596.

David R. Hunter and Kenneth Lange. "Quantile regression via an MM algorithm". In: *Journal of Computational and Graphical Statistics* 9.1 (2000), pp. 60–77.

Jan-Christian Hütter and Philippe Rigollet. "Optimal rates for total variation denoising". In: *Proceedings of the 29th Annual Conference on Learning Theory*. Vol. 49. 2016, pp. 1115–1146.

Kengo Kato. *Group lasso for high dimensional sparse quantile regression models*. 2011. arXiv: 1103.1458.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. "l_1 trend filtering". In: *SIAM Review* 51.2 (2009), pp. 339–360.

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

Roger Koenker and Gilbert Bassett Jr. "Regression quantiles". In: *Econometrica* 46.1 (1978), pp. 33–50.

Roger Koenker, Pin Ng, and Stephen Portnoy. "Quantile smoothing splines". In: *Biometrika* 81.4 (1994), pp. 673–680.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Youjuan Li and Ji Zhu. "l_1-norm quantile regression". In: *Journal of Computational and Graphical Statistics* 17.1 (2008), pp. 163–185.

Enno Mammen and Sara van de Geer. "Locally adaptive regression splines". In: *The Annals of Statistics* 25.1 (1997), pp. 387–413.

Nicolai Meinshausen. "Quantile random forests". In: *Journal of Machine Learning Research* 7 (2006), pp. 983–999.

Kelley Pace and Ronald Barry. "Sparse spatial autoregressions". In: *Statistics & Probability Letters* 33.3 (1997), pp. 291–297.

Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. *Risk bounds for quantile trend filtering*. 2020. arXiv: 2007.07472.

Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M. Witten. "Adaptive nonparametric regression with the K-nearest neighbour fused lasso". In: *Biometrika* 107.2 (2020), pp. 293–310.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. *Quantile regression with ReLU networks: estimators and minimax rates*. 2020. arXiv: 2010.08236.

Ashley Petersen, Noah Simon, and Daniela Witten. "Convex regression with interpretable sharp partitions". In: *Journal of Machine Learning Research* 17.94 (2016), pp. 1–31.
Matthew Pietrosanu et al. "Advanced algorithms for penalized quantile and composite quantile regression". In: *Computational Statistics* (2020), pp. 1–14.
Philippe Rigollet and Jan-Christian Hütter. *High dimensional statistics*. Lecture notes for course 18S997. 2015.
Leonid I. Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: Nonlinear Phenomena* 60.1–4 (1992), pp. 259–268.
Gideon Schwarz. "Estimating the dimension of a model". In: *The Annals of Statistics* 6.2 (1978), pp. 461–464.
Vladimir Spokoiny, Weining Wang, and Wolfgang Karl Härdle. "Local quantile regression". In: *Journal of Statistical Planning and Inference* 143.7 (2013), pp. 1109–1129.
Michel Talagrand. *The Generic Chaining*. Springer, 2005.
Wesley Tansey, Jesse Thomason, and James Scott. "Maximum-variance total variation denoising for interpretable spatial smoothing". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018.
Robert Tibshirani and Michael Saunders. "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 67.1 (2005), pp. 91–108.
Ryan J. Tibshirani. "Adaptive piecewise polynomial estimation via trend filtering". In: *The Annals of Statistics* 42.1 (2014), pp. 285–323.
Ryan J. Tibshirani and Jonathan Taylor. "Degrees of freedom in lasso problems". In: *The Annals of Statistics* 40.2 (2012), pp. 1198–1232.
Florencio I. Utreras. "On computing robust splines and applications". In: *SIAM Journal on Scientific and Statistical Computing* 2.2 (1981), pp. 153–163.
Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. "Hitting and commute times in large random neighborhood graphs". In: *Journal of Machine Learning Research* 15.52 (2014), pp. 1751–1798.
Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2018.
Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J. Tibshirani. "Trend filtering on graphs". In: *Journal of Machine Learning Research* 17.105 (2016), pp. 1–41.
Rebecca Willett, Robert Nowak, and Rui M Castro. "Faster rates in regression via active learning". In: *Advances in Neural Information Processing Systems 18*. 2006, pp. 179–186.
Tong Tong Wu and Kenneth Lange. "Coordinate descent algorithms for lasso penalized regression". In: *The Annals of Applied Statistics* 2.1 (2008), pp. 224–244.
Keming Yu and M. C. Jones. "Local linear quantile regression". In: *Journal of the American Statistical Association* 93.441 (1998), pp. 228–237.
Keming Yu and Rana A. Moyeed. "Bayesian quantile regression". In: *Statistics & Probability Letters* 54.4 (2001), pp. 437–447.
William P Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*. Vol. 120. Springer Science & Business Media, 2012.