Journal of Machine Learning Research 23 (2022) 1-47 Submitted 8/19; Revised 10/21; Published 3/22

Additive nonlinear quantile regression in ultra-high dimension

Ben Sherwood BEN.SHERWOOD@KU.EDU
School of Business
University of Kansas
Lawrence, KS 66045, USA

Adam Maidman ABMAIDMAN@GMAIL.COM
Microsoft
One Microsoft Way
Redmond, WA 98052, USA

Editor: Zaid Harchaoui

Abstract

We propose a method for simultaneous estimation and variable selection of an additive quantile regression model that can be used with high dimensional data. Quantile regression is an appealing method for analyzing high dimensional data because it can correctly model heteroscedastic relationships, is robust to outliers in the response, sparsity levels can change with quantiles, and it provides a thorough analysis of the conditional distribution of the response. An additive nonlinear model can capture more complex relationships, while avoiding the curse of dimensionality. The additive nonlinear model is fit using B-splines and a nonconvex group penalty is used for simultaneous estimation and variable selection. We derive the asymptotic properties of the estimator, including an oracle property, under general conditions that allow for the number of covariates, p_n, and the number of true covariates, q_n, to increase with the sample size, n. In addition, we propose a coordinate descent algorithm that reduces the computational cost compared to the linear programming approach typically used for solving quantile regression problems. The performance of the method is tested using Monte Carlo simulations, an analysis of fat content of meat conditional on a 100 channel spectrum of absorbances and predicting TRIM32 expression using gene expression data from the eyes of rats.

Keywords: Quantile Regression; Oracle Property; Nonparametric Regression; Splines; nonconvex penalty.

1. Introduction

We consider the sample $\{y_i, \mathbf{z}_i\}_{i=1}^n$ where $y_i \in \mathbb{R}$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip_n})^\top \in \mathbb{R}^{p_n}$. The τth conditional quantile, $\tau \in (0,1)$, of y given $\mathbf{z}$ is defined as $Q_{y|\mathbf{z}}(\tau) = \inf\{t : F(t|\mathbf{z}) \geq \tau\}$, where $F(\cdot|\mathbf{z})$ is the conditional distribution function of y given $\mathbf{z}$. There are p_n potential variables, but only $q_n(\tau)$ of these variables are needed to model the τth conditional quantile. Without loss of generality we assume the first $q_n(\tau)$ of these variables are active and the remaining $p_n - q_n(\tau)$ are inactive. The index n allows the set of active and inactive variables to increase with n, including the ultra-high dimensional case where p_n increase at an exponential rate

of n. For a given τ we consider the following sparse model for the conditional quantile

$$Q_{y|\mathbf{z}}(\tau) = g_0(\mathbf{z}, \tau) = \alpha_0(\tau) + \sum_{j=1}^{p_n} g_{0j}(z_j, \tau) = \alpha_0(\tau) + \sum_{j=1}^{q_n(\tau)} g_{0j}(z_j, \tau), \tag{1}$$

and for identifiability we assume $E[g_{0j}(z_j, \tau)] = 0$ for all $j \in \{1, \ldots, q_n(\tau)\}$. The model is a high-dimensional, sparse, nonparametric model that provides great flexibility. We assume an additive model to avoid the curse of dimensionality. The active variables, intercept, and additive functions are indexed by τ as the model allows for these values to change with τ. For simplicity of notation the τ symbol will be dropped throughout the remaining of the paper, but we emphasize here that the model allows for the nonlinear relationships and sparsity structure to change with τ. We propose using B-splines to model the nonlinear relationships and a group nonconvex penalty to correctly identify the covariates that have a relationship with the response at the given conditional quantile.

Tibshirani (1996) proposed the lasso penalty for simultaneous estimation and model selection, but strong conditions are required for model selection consistency (Zhao and Yu, 2006). Our focus is on model selection and our results will depend on using nonconvex penalty functions such as SCAD (Fan and Li, 2001) and MCP Zhang (2010) functions, which provide oracle estimators, a stronger result than model selection consistency, under milder conditions. When using splines multiple coefficients will be associated with a single covariate and thus we will use a group penalty, see Huang et al. (2012) for a review of group penalties in high-dimensional models. Previous works have proposed using splines with a group penalty for estimating an additive conditional mean function (Huang et al., 2010; Lin and Zhang, 2006; Meier et al., 2009; Xue, 2009). The work most similar to ours is Xue (2009) and Huang et al. (2010). Xue (2009) proposed using a group SCAD penalty and derived model consistency results for fixed q and p. Huang et al. (2010) proposed using a group adaptive lasso (Zou, 2006) and proved model selection consistency with fixed q, but allowing p to increase with n. Unlike these works, our focus is on estimating (1) instead of an additive conditional mean function.

Since Koenker and Bassett (1978) proposed linear quantile regression there have been many extensions, including work on nonlinear quantile regression. For a univariate covariate He and Shi (1994) demonstrated that using B-splines for nonlinear quantile regression has the same optimal rate of convergence as nonlinear mean regression (Stone, 1982). Motivated by the work of Stone (1985) (additive mean regression) and Stone (1986) (generalized additive models), De Gooijer and Zerom (2003) proposed a kernel based method for estimating an additive nonlinear conditional quantile model and demonstrated that for fixed p, additive quantile regression achieves the same rate of convergence found in He and Shi (1994) and thus theoretically alleviates the curse of dimensionality, although the proposed method requires bias correction for $p \geq 5$. Horowitz and Lee (2005) proposed a two-stage estimator for additive quantile regression that achieves the optimal rate of convergence and does not require a bias correction. Takeuchi et al. (2006) provided finite sample bounds for nonparametric quantile regression and discussed how to handle constraints such as monotonicity and non-crossing quantiles. Splines offer great flexibility in modeling conditional quantiles and have been proposed in a variety of conditional quantile models including, but not limited to, varying coefficients (Kim, 2007), growth curves (Wei et al., 2006) and

semiparametric models (He and Shi, 1996; He et al., 2002; Maidman and Wang, 2018; Wang et al., 2009b).

Quantile regression is a robust method which estimates a conditional quantile of interest. Our proposed method estimates conditional quantiles while allowing the sparsity structure to vary with τ, which has been stated as another example of the flexibility quantile regression provides for analyzing high-dimensional data (He et al., 2013; Wang et al., 2012). Previous work in penalized quantile regression includes using the lasso penalty (Belloni and Chernozhukov, 2011) and the nonconvex penalties MCP and SCAD (Wang et al., 2012) for estimating linear quantile regression with high-dimensional covariates. The high-dimensional linear quantile model has been relaxed to a partially linear model, where variable selection is only done on the high-dimensional linear terms (Sherwood, 2016; Sherwood and Wang, 2016). Other work proposed using splines with a group penalty for simultaneous estimation and variable selection of additive quantile regression. Kato (2012) proposed using a group lasso penalty, their work focused on convergence rates, while our work focuses on deriving an estimator with the oracle property which is asymptotically equivalent to the estimator that would be fit if we *a priori* knew the active covariates. Zhao and Lian (2016) considered the case where p is fixed and proposed using a nonconvex group penalty with the L_2 norm, while we allow p to increase with n and use the L_1 norm in our group penalty. Lin et al. (2013) proposed a smoothing spline ANOVA method that focuses on computational aspects and does not contain asymptotic results. Lv et al. (2018) considered estimation of (1) where the univariate functions reside in a reproducing kernel Hilbert space. Their work focused on estimation bounds, while our work focuses on deriving an oracle estimator, and they proposed a different penalty function than the one presented here.

Penalized quantile regression is not as commonly used as penalized least squares, but recent work has shown an interest in simultaneously estimating a conditional quantile and performing model selection. Essl et al. (2017) used penalized quantile regression to model extremes for the reserve capacity in the Australian electricity market, using time of day, year and week variables along with other forecast variables. Palma et al. (2020) modeled the age of a brain using MRI data for cognitively normal patients to better understand brain decay for cognitively impaired individuals. Quantile regression was used to model the .05, .5 and .95 quantiles, while the penalty was used to select the useful information from the MRI data. Motivated by the desire to identify counterfeit drugs, Ibrahim et al. (2020) used penalized quantile regression as a robust approach to model the amount of a certain chemical in a drug using spectroscopy data. Nonlinear or partially linear additive models with penalties have been used to simultaneously perform model selection and provide nonparametric estimates of conditional means. Examples include modeling stock returns given firm characteristics (Freyberger et al., 2020), predicting gene expression using DNA motifs (Lian et al., 2012), and constructing graphical models for frillice lettuce attributes and average environmental data during the cultivation period (Fujimoto et al., 2019). These are some examples of applications of penalized quantile regression and penalized additive models, but is by no means complete. In this paper we propose the penalized additive quantile model as a useful model for complex data. We demonstrate that this is a robust, theoretically sound model with few assumptions. To the best of our knowledge there are not many, or any, public applications of penalized additive nonlinear quantile models. However, given the flexibility of additive nonlinear quantile regression, we believe this can be a useful

tool for data analysis. To bridge the gap between theory and application, we discuss how to compute this model. In addition, our implementation is publicly available on CRAN (Sherwood and Maidman, 2020).

Theoretical challenges include dealing with a nonsmooth loss function, a nonconvex penalty function, approximation of nonlinear functions, and the number of covariates increasing with the sample size. Our asymptotic results allow for q_n to increase with n, which is challenging to deal with because both the number of predictors and basis functions increase to infinity with n. In addition, previous work on deriving oracle results for high dimensional quantile regression estimators have used the fact that a quantile regression objective function with a SCAD or MCP penalty can be written as a difference of convex functions (Sherwood, 2016; Sherwood and Wang, 2016; Wang et al., 2012). The theoretical results depend on demonstrating that asymptotically the oracle estimators satisfy properties about local minimizers of difference of convex functions provided by Tao and An (1997). Our proofs are more akin to the general approach taken by Fan and Lv (2011), where we only use some very general conditions about the penalty function. Results from Fan and Lv (2011) were for likelihood based methods and assumed that the objective function was differentiable. In their proofs they used a Taylor approximation of the penalized objective function, which is not possible for the non-differentiable quantile objective function. In this paper we show that the approach of Fan and Lv (2011) can be extended to quantile regression by replacing the Taylor approximation with Knight's identity, a common tool for theoretical results about quantile regression (Knight, 1998; Koenker, 2005). It is worth noting that previous work for adaptive lasso quantile regression, which has a convex objective function, used Knight's identity when establishing oracle properties (Wang et al., 2007; Zheng et al., 2015). We believe the approach provided here will be useful for future theoretical results because working with Knight's identity is easier than dealing with the subdifferential functions, which is required when using the properties of difference of convex functions. In addition, the results provided here are more general because they work with a large class of non-convex penalty functions and are not limited to the SCAD or MCP functions.

In addition to being theoretically challenging, high-dimensional quantile regression is a challenging computational problem. Koenker and Bassett (1978) showed that quantile regression can be solved by linear programming and many have found that minimizing penalized quantile regression objective functions can be framed as linear programming problems (Belloni and Chernozhukov, 2011; Sherwood and Wang, 2016; Wang et al., 2012; Wu and Liu, 2009). However, recent work has shown that in high dimensions alternative approaches can sacrifice little in terms of accuracy, while providing large computational gains. Peng and Wang (2015) proposed a coordinate descent algorithm for quantile regression with a non-convex penalty. Gu et al. (2018) proposed an alternating direction method of multiplier (ADMM) algorithm for quantile regression with lasso, adaptive lasso or a folded concave penalty. Yu et al. (2017) proposed an ADMM algorithm for nonconvex penalized quantile regression that can be computed in parallel. Yi and Huang (2017) proposed semismooth Newton coordinate descent algorithm for elastic-net penalized quantile regression that approximates the quantile loss function with a Huber loss function, creates a strong rule for discarding covariates, and uses a coordinate descent algorithm to update the remaining coefficients. None of the algorithms discussed use group penalties. We contribute to the

literature by proposing a coordinate descent algorithm for quantile regression with a nonconvex group penalty. Lv et al. (2018) also proposed a coordinate descent algorithm for penalized quantile regression. They approximated the quantile loss function with a smooth function and their penalty function is convex. In contrast, we do not approximate the quantile loss function and have a nonconvex penalty function.

The rest of the article is organized as follows. In Section 2 we discuss estimating the additive quantile regression model, when the active covariates are known *a priori*. We refer to this model as the oracle model and asymptotic properties of the oracle model are presented in Section 2. In Section 3 we present a group nonconvex penalty and present a theorem demonstrating that under reasonable conditions the group penalized method is asymptotically equivalent to the oracle model. In Section 4 we propose our new algorithm. In Section 5 we compare the proposed method using Monte Carlo simulations and in Section 6 we implement the proposed method to model fat content of meat using a 100 channel spectrum of absorbances and model TRIM32 expression using other gene expression data from the eyes of 120 twelve-week old male rats. We conclude with a summary in Section 7.

2. Oracle Model

To estimate (1) we propose first transforming the covariates using B-splines and then applying a group penalty method to simultaneously perform estimation and variable selection. In this section we assume that the active covariates are known *a priori* and thus are only estimating the model

$$Q_{y|\mathbf{z}}(\tau) = g_0(\mathbf{z}) = \alpha_0 + \sum_{j=1}^{q_n} g_{0j}(z_j). \tag{2}$$

The work in this section establishes convergence rates for the optimal local minimum of the penalized estimator. To estimate the nonlinear functions we use B-splines of order $m+1$ (degree m) with k_n quasi-uniform internal knots for (k_n+m+1) spline functions. Let $J_n = k_n + m$, for $j \in \{1,\ldots,p_n\}$ the jth covariate has J_n+1 corresponding functions of $[b_{j,0}(\cdot),\ldots,b_{j,J_n}(\cdot)]$ of order $m+1$ with k_n quasi-uniform internal knots on $[0,1]$ for a total of $2(m+1)+k_n$ knots, $(t_{j,-m},\ldots,t_{j,k_n+m+1})$. Define $h_j = \max_s |t_{j,s} - t_{j,s+1}|$, the largest distance between knots for the jth covariate, and $h = \max_j h_j$, the largest distance between knots for all covariates. A property of spline functions is that for any z_{ij} it follows that $\sum_{s=0}^{J_n} b_{j,s}(z_{ij}) = 1$ and to avoid collinearity we drop the first term when fitting the model. See Schumaker (1981) for more details about the construction of B-splines. The ith observation of the jth covariate will have a corresponding vector of $\boldsymbol{\pi}_j(z_{ij}) = [b_{j,1}(z_{ij}),...,b_{j,J_n}(z_{ij})]^\top \in \mathbb{R}^{J_n}$. Define

$$\boldsymbol{\Pi}_A(\mathbf{z}_i) = \left[1, \boldsymbol{\pi}_1(z_{i1})^\top, \ldots, \boldsymbol{\pi}_{q_n}(z_{iq_n})^\top\right]^\top \in \mathbb{R}^{J_n q_n + 1},$$

as the B-splines vector of active covariates and

$$\boldsymbol{\Pi}(\mathbf{z}_i) = \left[1, \boldsymbol{\pi}_1(z_{i1})^\top, \ldots, \boldsymbol{\pi}_{p_n}(z_{ip_n})^\top\right]^\top \in \mathbb{R}^{J_n p_n + 1},$$

as the B-splines vector for all covariates. B-splines can be used to approximate smooth functions and the following definitions help provide a formal definition of the class of functions for $g_0(\mathbf{z})$.

Definition 1 *Let $r \equiv m + v$, where m is a positive integer and $v \in (0, 1]$. Define $\mathcal{H}_r$ as the collection of functions $h(\cdot)$ on $[0, 1]$ whose mth derivative $h^{(m)}(\cdot)$ satisfies the Hölder condition of order v. That is, for any $h(\cdot) \in \mathcal{H}_r$, there exists some positive constant C such that*

$$\left|h^{(m)}(z') - h^{(m)}(z)\right| \leq C \left|z' - z\right|^v, \ \forall \ \ 0 \leq z', z \leq 1. \tag{3}$$

Definition 2 *Given $\mathbf{z} = (z_1, ..., z_{q_n})^\top$, the function $g(\mathbf{z})$ is said to belong to the class of functions $\mathcal{G}_r$ if it has the representation $g(\mathbf{z}) = \alpha + \sum_{j=1}^{q_n} g_j(z_j)$ where $\alpha \in \mathbb{R}$ and for all $j \in \{1, \ldots, q_n\}$, $g_j \in \mathcal{H}_r$, $E[g_j(\mathbf{z}_j)] = 0$ and $E[g_j(\mathbf{z}_j)^2] < M$, for some positive constant M.*

Definition 3 *Denote $\tilde{\mathcal{G}}_n$ as the space of additive functions spanned by $[\mathbf{\Pi}(\mathbf{z}_i)]_{i=1}^n$.*

B-splines can approximate any function $h(\cdot) \in \mathcal{H}_r$. That is, there exists $\boldsymbol{\gamma}_{0j} \in \mathbb{R}^{J_n}$ such that $\sup_{z \in [0,1]} \left|g_{0j}(z) - \boldsymbol{\pi}_j(z)^\top \boldsymbol{\gamma}_{0j}\right| = O\left(k_n^{-r}\right)$ (Schumaker, 1981). Thus, a function $g_0(\cdot) \in \mathcal{G}_r$ can be approximated by a function from $\tilde{\mathcal{G}}_n$, specifically there exists $\boldsymbol{\gamma}_{A0} = (\alpha_0, \boldsymbol{\gamma}_{01}^\top, \ldots, \boldsymbol{\gamma}_{0q_n}^\top)^\top \in \mathbb{R}^{q_n J_n + 1}$ such that

$$\sup_{\mathbf{z} \in [0,1]^{q_n}} \left|g_0(\mathbf{z}) - \mathbf{\Pi}_A(\mathbf{z})^\top \boldsymbol{\gamma}_{A0}\right| = O\left(q_n k_n^{-r}\right). \tag{4}$$

The quantile loss function is defined as $\rho_\tau(u) = u[\tau - I(u < 0)]$. The oracle estimator only relies on the active covariates and is defined as

$$\hat{\boldsymbol{\gamma}}_A = \underset{\boldsymbol{\gamma}_A \in \mathbb{R}^{q_n J_n + 1}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau[y_i - \mathbf{\Pi}_A(\mathbf{z}_i)^\top \boldsymbol{\gamma}_A]. \tag{5}$$

The estimator of $g_0(\mathbf{z}_i)$ is $\hat{g}(\mathbf{z}_i) = \mathbf{\Pi}_A(\mathbf{z}_i)^\top \hat{\boldsymbol{\gamma}}_A$. The proposed estimator is a robust estimator and will be robust to outliers in the response. Similar to a univariate estimate of a quantile, if values of $\{y_i\}_{i=1}^n$ are changed but the signs of residuals remain the same then the estimator $\hat{g}(\mathbf{z}_i)$ remains unchanged. See Theorem 2.4 and the discussion surrounding this theorem from Koenker (2005) for a detailed discussion of this property of quantile regression.

The following conditions were used to prove the rate of convergence of $n^{-1} \sum_{i=1}^n [\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)]^2$.

Condition 1 *(Conditions on the random error) The random error ϵ_i has the conditional distribution function $F_i(\cdot \mid \mathbf{z}_i)$, continuous conditional density function $f_i(\cdot \mid \mathbf{z}_i)$, and $f_i'(\cdot \mid \mathbf{z}_i)$ is the derivative of the conditional density function. The density functions are uniformly bounded away from 0 and infinity in a neighborhood of zero and there exists a positive constant c_f such that $|f_i'(\cdot \mid \mathbf{z}_i)| \leq c_f$ for all $i \in \{1, \ldots, n\}$.*

Condition 2 *(Conditions on the covariates) Let $z_{ij} \in [0, 1]$ for all $i \in \{1, \ldots, n\}$ and for all $j \in \{1, \ldots, p_n\}$. The joint density of the predictors is absolutely continuous and the density, $f_{\mathbf{z}}(\mathbf{z})$, is bounded away from zero and infinity by positive constants. In addition, define $f_{z_j}(z)$ as the density function for the jth covariate. There exist positive constants c_1 and c_2 such that $c_1 < f_{z_j}(z) < c_2$ for all $z \in [0, 1]$ and $j \in \{1, \ldots, p_n\}$.*

Condition 3 *(Conditions on the splines) There exists a positive constant c_3 such that*

$$\max_{j\in\{1,\ldots,p_n\}} \frac{\max_s(t_{j,s+1}-t_{j,s})}{\min_s(t_{j,s+1}-t_{j,s})} \leq c_3.$$

For the internal knots $k_n \approx (q_n n)^{1/(2r+1)}$, $h \approx (q_n n)^{-1/(2r+1)}$ and for all $j \in \{1,\ldots,p_n\}$ $\frac{h_j}{h} \approx 1$, where $a \approx b$ means both a and b have the same order.

Condition 4 *(Condition on the size of the model) For the active variables $q_n = o[\log(n)]$.*

Condition 5 *(Condition on the unknown functions) For $r = m + v > 3$ we assume $g_0(\cdot) \in \mathcal{G}_r$.*

Condition 1 has been used for asymptotic results of a fixed dimensional linear quantile regression model (Koenker, 2005) and is a weaker condition than the Gaussian or sub-Gaussian conditions that are common for penalized, high-dimensional models (Negahban et al., 2012). Condition 1 reflects the robust properties of quantile regression as it does not assume that any moments exist for the distribution of the errors, and thus the results will hold for heavy tailed distributions that have no moments such as the Cauchy distribution. Under Condition 2 there is no collinearity between the predictors. Stone (1985) introduced Condition 2 to provide a lower bound for the standard deviation of the additive function. It is also used to provide a lower bound for the minimum eigenvalue for the covariance matrix of the B-splines transformation of the active predictors (Chen et al., 2018b; Zhou et al., 1998). In addition, standard B-spline results depend on the covariates having a bounded support and without loss of generality, Condition 2 assumes the support to be the interval $[0, 1]$. The assumption that the density functions have a common lower and upper bound is frequent in work involving splines because the bounds allow for a direct application of Theorem 5.4.2 from Devore and Lorentz (2005). Condition 3 assumes that the distance between the internal knots are not drastically different, which holds in practice as long as the distributions of the covariates are not greatly skewed, and is a common assumption in work with splines (Huang, 1998a,b; Xue and Yang, 2006). In addition for fixed q_n, the rate for k_n is equivalent to the optimal rate found in Stone (1985). Condition 4 governs the rate at which q_n can increase with n. Though the rate is slow because k_n also needs to increase with n, most work in additive models assume q is fixed. The rate in Condition 4 is the same rate used by others that have considered an increasing number of true covariates when estimating additive models (Wang et al., 2014a). Condition 5 provides that only reasonably smooth functions can be estimated by the proposed method. The above conditions are used to prove the following theorem about the rate of convergence of $\hat{g}(\cdot)$. These conditions are sufficient for proving our results but are not necessarily the weakest conditions needed.

Define $\Pi_A = [\mathbf{\Pi}_A(\mathbf{z}_1), \ldots, \mathbf{\Pi}_A(\mathbf{z}_n)]^\top \in \mathbb{R}^{n\times q_n J_n+1}$. Note that our conditions lack an explicit assumption about bounds on the eigenvalues for the sample covariance matrix of the active predictors, $\frac{1}{n}\sum_{i=1}^n \mathbf{\Pi}_A(\mathbf{z}_i)\mathbf{\Pi}_A(\mathbf{z}_i)^\top$, which is common in work that derives an oracle property for high-dimensional data (Fan and Lv, 2011; Loh and Wainwright, 2015; Wang et al., 2012; Zheng et al., 2015). Using the properties of B-splines and Conditions 1 - 3, the following lemma provides these bounds and insight into why the rate of q_n provided in Condition 4 is so small.

Lemma 4 *Assume Conditions 1-3 hold. For* $\mathbf{a} \in \mathbb{R}^{q_n J_n + 1}$ *where* $||\mathbf{a}||_2 = 1$*, there exist positive constants* $b_1 > 0$*,* $B_1 > 0$ *and* $\delta \in (0,1)$ *with* $\delta_{q_n} = [(1-\delta)/2]^{q_n/2}$ *such that for sufficiently large* n *that* $b_1 \delta_{q_n}^2 k_n^{-1} \leq \mathbf{a}^\top \frac{1}{n} \mathbf{\Pi}_A^\top \mathbf{\Pi}_A \mathbf{a} \leq B_1 q_n$*.*

Proof of Lemma 4 is provided in the Appendix. Stone (1985) first introduced a lower bound that depended on a term similar to δ_{q_n} and is very common in the additive model literature. If q_n is fixed then the term δ_{q_n} is a constant and can be easily dealt within the asymptotic analysis. However, in the setting of q_n increasing with n, the term has to be dealt with more care. Specifically, the proof of the the next theorem depends on $n^{-1/2}(q_n k_n)^{1/2} \delta_{q_n}^{-1}$ converging to zero and thus the need for Condition 4.

Theorem 5 *If Conditions 1-5 hold, then*

$$n^{-1} \sum_{i=1}^{n} [\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)]^2 = O_P\left(q_n k_n / n + q_n^2 k_n^{-2r}\right).$$

Thus, under Condition 3 and for fixed q, the estimator $\hat{g}(\mathbf{z}_i)$ reaches the optimal rate of convergence found by Stone (1985) and extends to fixed dimensional quantile regression additive models (De Gooijer and Zerom, 2003; He and Shi, 1994; Horowitz and Lee, 2005). To the best of our knowledge, this is the first result that considers the rate of convergence for an additive quantile model with q_n increasing with n. Proof of Theorem 5 is provided in the Appendix.

3. Variable Selection

In the previous section we established that the oracle estimator is an optimal estimator but it requires *a priori* knowledge about the covariates that may not be known in practice. Define $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_{p_n}^\top)^\top \in \mathbb{R}^{J_n p_n + 1}$, where $\boldsymbol{\gamma}_j \in \mathbb{R}^{J_n}$ is the coefficient vector for the B-spline functions of the jth covariate. For a vector $\mathbf{a}$ we define $||\mathbf{a}||_q$ as the L_q norm of $\mathbf{a}$. To fit a sparse model that accounts for the groups of spline functions, we propose the following objective function

$$Q(\boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau[y_i - \mathbf{\Pi}(\mathbf{z}_i)^\top \boldsymbol{\gamma}] + \sum_{j=1}^{p_n} p_{\lambda,a}\left(||\boldsymbol{\gamma}_j||_1\right). \tag{6}$$

A group penalty is used to incorporate the group structure of the splines. Similar penalties have been used for mean additive models (Huang et al., 2010; Xue, 2009). Zhao and Lian (2016) consider a similar model for additive quantile regression, but use an L_2 norm inside the penalty function and their theoretical results assume a fixed q and p. The L_1 norm is used instead of the L_2 norm for computational convenience. The L_1 norm fits naturally with quantile regression and in the next section we discuss some of the computational conveniences it provides. Whether an L_1 or L_2 norm is used the oracle properties for group concave penalties are similar (Sherwood et al., 2020). Define the oracle estimator as $\hat{\boldsymbol{\gamma}} = \left[\hat{\boldsymbol{\gamma}}_A^\top, \mathbf{0}_{J_n(p_n - q_n)}^\top\right]^\top$, the estimator that only uses relevant groups. To derive an oracle property we use a general class of nonconvex functions for p_λ and will prove that

with probability going to one that $\hat{\boldsymbol{\gamma}}$ is a local minimum of $Q(\boldsymbol{\gamma})$. Two commonly used penalty functions are SCAD and MCP. For the SCAD penalty function

$$\begin{aligned} p_\lambda(x) &= \lambda|x|I(0 \le |x| < \lambda) + \frac{a\lambda|x| - \left(x^2 + \lambda^2\right)/2}{a-1} I(\lambda \le |x| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|x| > a\lambda), \text{ for some } a > 2, \end{aligned}$$

and for the MCP penalty function,

$$p_\lambda(x) = \lambda \left(|x| - \frac{x^2}{2a\lambda} \right) I(0 \le |x| < a\lambda) + \frac{a\lambda^2}{2} I\left(|x| \ge a\lambda\right), \text{ for some } a > 1.$$

For both penalty functions, the tuning parameter λ controls the complexity of the selected model and goes to zero as n increases to infinity. The other tuning parameter a controls how quickly $p'_{\lambda,a}(x)$ goes to zero, but is considered fixed in the asymptotics.

Our model selection consistency proofs use the following conditions about the penalty function and the strength of signal from a group of B-spline coefficients for an active covariate.

Condition 6 *The function $p_{\lambda,a}(x)$ is increasing, concave, and has a continuous positive derivative, $p'_{\lambda,a}(x)$ for $x \in [0,\infty)$. Also, $p'_{\lambda,a}(x)$ is increasing with respect to λ for $\lambda > 0$, $p'_{\lambda,a}(0+) = \lambda$ and $p'_{\lambda,a}(x) = 0$ for $|x| > a\lambda$.*

Condition 7 *(Condition on the signal) There exist positive constants c_4 and c_5 such that $4/(2r+1) < c_4 < (2r-1)/(2r+1)$ and $n^{(1-c_4)/2} \min_{j \in \{1,\ldots,q_n\}} ||\boldsymbol{\gamma}_{0j}||_1 \ge c_5$.*

Both the SCAD and MCP satisfy Condition 6, which is very similar to Condition 1 from Fan and Lv (2011). Condition 7 is a strength of signal condition that is very common in high-dimensional linear models, for instance see Kim et al. (2008), Kim et al. (2012) and Wang et al. (2012). The upper and lower bounds for c_4 are sensible by Condition 5. Again, these conditions are sufficient for proving our results, but are not necessarily the weakest conditions needed.

Theorem 6 *Assume Conditions 1 - 7 hold, $\lambda = o\left[n^{-(1-c_4)/2}\right]$, $\log(p_n) = o(n\lambda^2 k_n^{-1})$, $n^{-1/2}k_n^2 \log(n) = o(\lambda)$ and $n\lambda^2 \to \infty$. Let $\mathcal{M}_n(\lambda)$ be the set of local minima of the penalized objective function $Q(\boldsymbol{\gamma})$, defined in (6), for tuning parameter value λ then*

$$P\left[\hat{\boldsymbol{\gamma}} \in \mathcal{M}_n(\lambda)\right] \to 1.$$

The conditions on λ are satisfied for $\lambda = Cn^{-1/2+b}$, where $b \in \left(\frac{2}{2r+1}, \frac{c_4}{2}\right)$ and C is any positive constant, where $2/(2r+1) < c_4/2$ is guaranteed by Condition 7. The motivation for using concave penalties is that with probability approaching one, for the correct value of λ, the oracle estimator is a local minimum of the penalized objective function. Thus, the optimal value of λ needs to properly balance over-fitting and under-fitting the model. The upper bound depends on c_4, which depends on the function that provides the smallest signal. Smaller values of c_4 indicate a weaker minimum signal and thus smaller values

of λ are needed to avoid under-fitting. The lower bound depends on r and for smoother functions λ can be smaller. The intuition here is that for smoother functions it should be easier to separate the signal from the noise and thus smaller values of λ are needed. Finally, the oracle property holds for ultra-high dimensional predictors because the rates allow for $p_n = o\left\{\exp\left[n^{b-1/(2r+1)}\right]\right\}$.

Theorem 6 proves that with probability going to one the oracle estimator is a local minimizer of $Q(\boldsymbol{\gamma})$, but provides no guarantee that the oracle estimator is the global minimizer. Nor does it provide any guarantees about other potential local minimizers. The next theorem provides a bound on the l_2 norm of the difference between a sufficiently sparse local minimizer and the oracle estimator. However, an additional condition is used for that proof.

Condition 8 *If $\bar{\boldsymbol{\gamma}}$ is a local minimizer of $Q(\boldsymbol{\gamma})$, where $||\bar{\boldsymbol{\gamma}}||_0 = u_n$ then $\sum_{i=1}^{n} I[y_i = \mathbf{\Pi}(\mathbf{z}_i)^\top \bar{\boldsymbol{\gamma}}] = O(u_n)$ and there exists $\bar{\boldsymbol{\gamma}}_0$ such that $||\bar{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}_0||_2 = o_P(1)$.*

Condition 8 protects against pathological cases. It assumes that the local minimizer $\bar{\boldsymbol{\gamma}}$, converges in probability to some fixed value $\bar{\boldsymbol{\gamma}}_0$, but does not assume that $\bar{\boldsymbol{\gamma}}_0$ is equal to $\boldsymbol{\gamma}_0$, the coefficients that provide the best approximation to the unknown additive function. In addition, it provides a bound on the number of zero-valued residuals. Consider an unpenalized linear quantile regression estimator with p predictors. Of the n residuals corresponding to this estimator, with probability one there will be exactly $p+1$ zero-valued residuals if the errors have a density with respect to a Lebesgue measure. See section 2.2.2 of (Koenker, 2005) for a more detailed discussion of this topic. Using the same notation as Condition 8, let u_n be the number of nonzero coefficients for a weighted lasso estimator. Then the weighted lasso quantile regression model will have at most u_n zero-valued residuals with probability one. This is because the weighted lasso quantile regression problem has a linear programming formulation similar to the standard quantile regression problem. Therefore, we believe the assumption is reasonable for the SCAD and MCP for two reasons. First, it will hold for unpenalized quantile regression and the motivation for both the SCAD and MCP is to approximate an unpenalized estimator. Second, it holds for the weighted lasso estimator which is the approximation we use for (6), see (7) in Section 4.

Theorem 7 *Define $\bar{\boldsymbol{\gamma}}$ as a local minimizer of $Q(\boldsymbol{\gamma})$ and define $\mathcal{E} = \{j \in \{1, \ldots, p_n\} | ||\bar{\boldsymbol{\gamma}}_j||_\infty \neq 0 \text{ or } ||\hat{\boldsymbol{\gamma}}_j||_\infty \neq 0\}$ as the set of groups that have a non-zero entry in $\bar{\boldsymbol{\gamma}}$ or $\hat{\boldsymbol{\gamma}}$. Let $w_n = |\mathcal{E}| = o[\log(n)]$, assume Conditions 1-8 hold and that $\lambda = n^{-1/2+b}$, where $b \in \left(\frac{2}{2r+1}, \frac{c_4}{2}\right)$, where $\frac{2}{2r+1} < \frac{c_4}{2}$, then*

$$||\bar{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}||_2 = O_P\left[\log(n)\delta_{w_n}^{-2} k_n \left(\sqrt{\frac{w_n}{n}} + \lambda\sqrt{w_n k_n} + k_n w_n n^{-1}\sqrt{1+w_n}\right)\right].$$

Corollary 8 *Under the conditions of Theorem 7 with $\lambda = n^{-1/2+b}$ where $b \in \left(\frac{2}{2r+1}, \frac{r-1}{2r+1}\right)$ then*

$$||\bar{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}||_2 = o_P(1) \text{ and } ||\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_2 = o_P(1).$$

Corollary 8 provides that any sufficiently sparse local minimizer of $Q(\boldsymbol{\gamma})$ will be a consistent estimator.

4. Algorithm

The objective function $Q(\boldsymbol{\gamma})$ is non-convex and for high-dimensional data a grid search approach is not reasonable. Algorithms exist for finding estimators with good statistical properties for a wide class of nonconvex problems, but they assume the loss function is differentiable, which is not the case for quantile regression (Loh and Wainwright, 2015; Wang et al., 2014b). Zou and Li (2008) proposed a local linear approximation (LLA) that provides a convex approximation to a non-convex objective function. Let $\hat{\boldsymbol{\gamma}}_j^t$ represent the estimate of $\boldsymbol{\gamma}_{0j}$ at iteration t, with $\hat{\boldsymbol{\gamma}}^0 = 0$, then the LLA of $Q(\boldsymbol{\gamma})$ is

$$Q_t(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}] + \sum_{j=1}^{p_n} p'_\lambda\left(||\boldsymbol{\gamma}_j^{t-1}||_1\right)||\boldsymbol{\gamma}_j||_1. \tag{7}$$

For $\tau \in (0,1)$, $\rho_\tau(u) + \rho_\tau(-u) = |u|$ and thus the above problem can be restated as a weighted quantile regression problem with augmented data (Sherwood, 2016; Sherwood and Wang, 2016; Wang et al., 2012). In this approach the final estimates are derived once the estimates converge or a maximum number of iterations has been made. At each iteration minimizing (7) becomes a linear programming problem, however linear programming can be quite slow for high dimensional problems. If the traditional L_2 norm was used then solving (7) becomes a second-order cone programming problem, but these tend to be even slower than linear programming problems. In addition, using the L_1 norm allows us to build on existing computational approaches for penalized quantile regression. Peng and Wang (2015) proposed the quantile iterative coordinate descent (QICD) algorithm for solving (7) for the standard SCAD or MCP penalty, where $\boldsymbol{\gamma}_j$ is a scalar, which greatly reduces computational complexity without sacrificing estimation in accuracy. We propose an extension of the QICD algorithm for the group penalty setting, where $\boldsymbol{\gamma}_j$ is a vector.

The QICD algorithm is a two-step process that first majorizes the objective function and then uses a coordinate descent algorithm to solve each iteration of the majorization step. The coordinate descent algorithm is responsible for faster convergence. The key difference is our algorithm includes an L_1 grouping of coefficients and to minimize (6), we modify the QICD algorithm to allow for group penalties. Let $\gamma_{js}^{(k)}$ denote the value of γ_{js} after the kth iteration, $k = 1, 2, \ldots$ and $\boldsymbol{\gamma}_j^{(k)} = \left[\gamma_{j1}^{(k)}, \ldots, \gamma_{J_n}^{(k)}\right]^\top$ for $j \in \{1, \ldots, p_n\}$ and $s \in \{1, \ldots, J_n\}$. Furthermore, let $p'_\lambda(x+)$ be the limit of $p'_\lambda(y)$ as $y \to x$ from above. Then, in the kth iteration,

$$\begin{aligned}\phi_{\boldsymbol{\gamma}_j^{(k-1)}}\left(\boldsymbol{\gamma}_j\right) = {} & p'_\lambda\left(||\boldsymbol{\gamma}_j^{(k-1)}||_1+\right)\sum_{s=1}^{J_n}|\gamma_{js}| - p'_\lambda\left(||\boldsymbol{\gamma}_j^{(k-1)}||_1+\right)\sum_{s=1}^{J_n}|\gamma_{js}^{(k-1)}| \\ & + p_\lambda\left(||\boldsymbol{\gamma}_j^{(k-1)}||_1\right)\end{aligned} \tag{8}$$

majorizes the penalty function $p_\lambda\left(||\boldsymbol{\gamma}_j||_1\right)$ for $k \in \{1, 2, \ldots, K\}$, where K is a user defined value for the maximum number of iterations, and $j \in \{1, \ldots, p_n\}$. More specifically, $\phi_{\boldsymbol{\gamma}_j^{(k-1)}}\left(\boldsymbol{\gamma}_j\right) \geq p_\lambda\left(||\boldsymbol{\gamma}_j||_1\right)$ for all $\boldsymbol{\gamma}_j$ with equality when $\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_j^{(k-1)}$. Thus, the objective

function $Q(\boldsymbol{\gamma})$ defined in (6) is majorized by

$$Q_{\boldsymbol{\gamma}^{(k-1)}}(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}] + \sum_{j=1}^{p_n}\phi_{\boldsymbol{\gamma}_j^{(k-1)}}\left(\boldsymbol{\gamma}_j\right). \tag{9}$$

The majorizing function in (9) is similar to the majorizing function in Peng and Wang (2015). However, in our setting, coefficients for spline functions associated with the same covariate all have the same weight $p'_\lambda\left(||\boldsymbol{\gamma}_j^{(k-1)}||_1+\right)$.

For each $k = 1, 2, \ldots$, the update for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}^{(k)} = \arg\min_{\boldsymbol{\gamma}} Q_{\boldsymbol{\gamma}^{(k-1)}}(\boldsymbol{\gamma}). \tag{10}$$

This iteration scheme decreases the value of the objective function in (6) for each $\boldsymbol{\gamma}^{(k)}$. Additionally, the solution to the original nonconvex minimization problem can now be found by solving a sequence of convex minimization problems.

Coordinate descent can be used to solve the convex minimization problem in (10). In the following coordinate descent algorithm, each coefficient γ_{js} is updated one-at-a-time until convergence. For the dth iteration of the coordinate descent step and the kth iteration of the majorization step, let

$$\boldsymbol{\omega}_{js}^{(k)(d)} = \left(\gamma_{11}^{(k)(d+1)}, \ldots, \gamma_{js-1}^{(k)(d+1)}, \gamma_{js}^{(k)(d)}, \ldots, \gamma_{p_n J_n}^{(k)(d)}\right)^\top,$$

be the vector of coefficients that contains updates for the first $js-1$ coefficients, but not the remaining ones. We update each coefficient in the coordinate descent step as

$$\gamma_{js}^{(k)(d)} = \arg\min_{\gamma_{js}} Q_{\boldsymbol{\gamma}^{(k-1)}}\left(\boldsymbol{\omega}_{js}^{(k)(d)}\right). \tag{11}$$

We omit the complete derivation of the coordinate descent algorithm as it is very similar to Peng and Wang (2015). The algorithm converges when for some specified tolerance ϵ, $||\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}||_2 < \epsilon$ or k equals K, the maximum number of iterations allowed. In the following data analysis we used $K = 20$ and $\epsilon = .00001$.

It is important to have appropriate starting values for the algorithm to converge. We recommend using the estimates from lasso penalized quantile regression with the lasso penalty applied individually to each coefficient (i.e., ignoring the group penalty) as the starting values for $\gamma_{js}^{(0)}$. The algorithm is implemented in the R package **rqPen** (Sherwood and Maidman, 2020).

5. Simulations

We consider three different simulation settings. In the first setting the response is generated from an additive model where each function is nonlinear. The purpose of this setting is to demonstrate the effectiveness of the proposed method compared to other approaches for modeling nonlinear functions. This setting also includes comparisons of the QICD algorithm to a linear programming approach. In the second setting we use the proposed

approach where it might not be optimal. This setting includes a linear model, partially linear model and a non-additive model. In this setting we compare the proposed approach to linear models to test if the proposed approach is competitive with simpler methods. In the last setting, to verify results of Theorem 6, we test the model selection properties of the proposed approach for varying values of n, q_n, J_n and p_n.

In all settings the covariates are generated in two steps. For each observation a p-dimensional vector is generated by $\mathbf{x} \sim N(\mathbf{0}_p, \Sigma_p)$ with σ_{jk} being the entry for the jth row and kth column of Σ_p and $\sigma_{jk} = .5^{|j-k|}$. For a vector $\mathbf{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$ define $\Phi(\mathbf{a}) = \{\Phi(a_1), \ldots, \Phi(a_p)\}^\top \in \mathbb{R}^p$, where $\Phi(\cdot)$ is the normal CDF. Then the p-dimensional covariate vector is generated by $\mathbf{z} = \Phi(\mathbf{x})$. In the first two sections we consider a sample size of $n = 500$ and the number of covariates as $p = 100$, 300 or 600. More details about the third simulation will be provided later which includes different values of n, p_n, J_n and q_n.

For the first two settings, models are fit using 500 training samples. Then 1000 testing samples are generated from the same model. All models are fit using B-splines with the training and testing covariates transformed using cubic B-splines with $J_n = 3$. Let y_i^* and $\hat{y}_i^*(\tau)$ represent the observed value and the predicted τth quantile for the ith testing sample, where the prediction comes from a model that was fit only using the training data. A covariate is considered selected if any of its corresponding spline coefficients are non-zero. Models are compared using the following criteria.

1. Mean squared prediction error (MSPE), $\frac{1}{1000}\sum_{i=1}^{1000}[y_i^* - \hat{y}_i^*(\tau)]^2$.
2. Mean absolute prediction error (MAPE), $\frac{1}{1000}\sum_{i=1}^{1000}|y_i^* - \hat{y}_i^*(\tau)|$.
3. Mean check prediction error (MCPE), $\frac{1}{1000}\sum_{i=1}^{1000}\rho_\tau[y_i^* - \hat{y}_i^*(\tau)]$.
4. True positives (TP), the number of active covariates selected.
5. False positives (FP), the number of nonactive covariates selected.
6. Proportion smaller (PS), the proportion of testing responses smaller than their predicted value.

For consistent methods the value of PS should be close to τ. When modeling the median, MAPE and MCPE differ only by a multiple of 2. Thus, we only report MCPE in settings where we fit models for non-median quantiles. In the first two settings 100 replications are run for each simulation setting, while in the last setting 50 replications are run for each setting.

Setting I: Additive Model

In Setting I we consider the proposed quantile additive model where $p_{\lambda,a}(\cdot)$ is the SCAD penalty function. We implement both the coordinate descent (QA-SCAD CD) and linear programing (QA-SCAD LP) algorithms. We compare the method to the quantile additive model with the lasso penalty (QA-LASSO), QA-LASSO minimizes (6) with $p_{\lambda,a}(x) = \lambda|x|$. In addition, we consider the mean additive model with the group SCAD (MA-SCAD) and

group lasso (MA-LASSO) penalty. The mean regression methods use the same B-spline transformation and minimize

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \mathbf{\Pi}(\mathbf{z}_i)^\top \boldsymbol{\gamma}]^2 + \sum_{j=1}^{p} p_{\lambda,a}\left(||\boldsymbol{\gamma}_j||_2\right), \tag{12}$$

where for MA-LASSO $p_{\lambda,a}(x) = \lambda|x|$ and for MA-SCAD $p_{\lambda,a}(\cdot)$ is the SCAD penalty function. For both MA-SCAD and QA-SCAD we set $a = 3.7$. Let, $\tilde{\boldsymbol{\gamma}}_\lambda$ be the coefficient vector for a given value of λ and $\tilde{q}_\lambda$ be the number of nonzero coefficients. For the quantile regression methods λ is selected by minimizing,

$$\log\left\{\sum_{i=1}^{n}\rho_\tau[y_i - \mathbf{\Pi}(\mathbf{z}_i)^\top\tilde{\boldsymbol{\gamma}}_\lambda]\right\} + \tilde{q}_\lambda\frac{\log(n)}{2n}.$$

Let $\ell(\boldsymbol{\gamma})$ represent the Gaussian log-likelihood evaluated at $\boldsymbol{\gamma}$. For the mean regression methods λ is selected by minimizing

$$-2\ell(\tilde{\boldsymbol{\gamma}}_\lambda) + \tilde{q}_\lambda\log(n).$$

The quantile and mean regression models are fit using the R packages **rqPen** (Sherwood and Maidman, 2020) and **grpreg** (Breheny and Zeng, 2017), respectively. Theoretically, using BIC may not be optimal for high-dimensional variables. There exist challenges to demonstrating that BIC will select the true model when the number of predictors grows with the sample size that remain unsolved (Wang et al., 2009a; Lee et al., 2014). For additive quantile regression models, Lee et al. (2014) suggested replacing $\tilde{q}_\lambda\frac{\log(n)}{2n}$ with $C_n\tilde{q}_\lambda\frac{\log(n)}{2n}$, where $C_n \to \infty$, and provide theoretical justifications. The R package **rqPen** allows for implementing this high dimensional BIC. We used the standard BIC as our preliminary results found that approach to work better, but the approach of Lee et al. (2014) has been shown to be superior in other settings.

The response is generated under three different settings. For the first two settings we consider the model

$$y = -1 + 2z_1^3 + \sin(2\pi z_2) + 8(z_3 - .5)^2 + \epsilon, \tag{13}$$

with homoscedastic errors of $\epsilon \sim N(0,1)$ (Setting IA) or $\epsilon \sim T_3$ (Setting IB). In the third setting we consider the following heteroscedastic errors model

$$y = \sin(2\pi z_2) + 8(z_3 - .5)^2 + (.5 + z_1^3)\epsilon, \tag{14}$$

with $\epsilon \sim N(0,1)$ (Setting IC).

For Settings IA and IB the methods estimate the median, $\tau = .5$, while in Setting IC the methods estimate the .9 quantile, $\tau = .9$. The quantile regression methods can directly model the .9 quantile, but the mean regression methods do not directly provide non-median estimates. To estimate $\hat{y}_i^*(\tau)$ for the mean regression methods we propose a naive estimate of the conditional quantile based on estimation of the conditional quantile in the linear mean model when the error terms are normally distributed and p is small. Define $\mathbf{Y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ as the vector of observed responses. For an estimate $\hat{\boldsymbol{\gamma}}$,

define $\hat{\mathbf{y}}(\hat{\boldsymbol{\gamma}}) \in \mathbb{R}^n$ as the vector of fitted values, $q(\hat{\boldsymbol{\gamma}})$ as the number of nonzero entries in $\hat{\boldsymbol{\gamma}}$, $\boldsymbol{\Pi}(\mathbf{z}_i, \hat{\boldsymbol{\gamma}}) \in \mathbb{R}^{q(\hat{\boldsymbol{\gamma}})}$ as the vector of basis functions that have non-zero coefficients in $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}(\hat{\boldsymbol{\gamma}}) = \frac{1}{\sqrt{n-q(\hat{\boldsymbol{\gamma}})}} \left|\left|\mathbf{Y} - \hat{\mathbf{Y}}(\hat{\boldsymbol{\gamma}})\right|\right|_2$. Let $t^*_{d,\tau}$ be the τ quantile of a T-distribution with d degrees of freedom. For a vector of covariates $\mathbf{z}^*_i$ with estimate of the conditional mean, $\hat{y}^*_i$, the naive estimate of the τth conditional quantile is

$$\hat{y}^*_i(\tau) = \hat{y}^*_i + t^*_{n-q(\hat{\boldsymbol{\gamma}}),\tau}\hat{\sigma}(\hat{\boldsymbol{\gamma}})\sqrt{1 + \boldsymbol{\Pi}(\mathbf{z}^*_i, \hat{\boldsymbol{\gamma}})^\top \left[\sum_{i=1}^{n} \boldsymbol{\Pi}(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})\boldsymbol{\Pi}(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})^\top\right]^{-1} \boldsymbol{\Pi}(\mathbf{z}^*_i, \hat{\boldsymbol{\gamma}})}. \quad (15)$$

The above estimator is the standard prediction interval estimator of a conditional quantile using ordinary least squares. If the errors are i.i.d. and normally distributed then the estimator is the MLE for the conditional quantile. However, it is naive because it is a fixed dimensional solution to a high-dimensional problem. Even in the fixed dimension setting it will be an inconsistent estimate if the errors are not normally distributed or if there is nonconstant variance.

The means (and standard deviations) across the 100 simulations for the previously defined six summary statistics are provided in Tables 1-3 for Settings IA-IC. For $p = 100$ and $p = 300$ the tables include two versions of QA-SCAD. The linear programming (LP) approach to solving (7), using the Barrodale and Roberts (1974) algorithm for regression quantiles (Koenker and D'Orey, 1987, 1994), and the coordinate descent (CD) approach described in Section 4. The summary statistics between the two algorithms are almost identical in Settings IA and IB, except that the coordinate descent approach tends to select more false positives. In Setting IC the LP approach provides better results. Figure 1 compares the computational speed for the two algorithms, in the different settings when $p = 100$ or $p = 300$. For Settings IA and IB the QICD algorithm is noticeably faster for $p = 100$ or $p = 300$. For Setting IC, the QICD algorithm is slower at $p = 100$, but faster at $p = 300$. For $p = 600$ only the QICD algorithm was used, due to the excessive computational time of the linear programming approach. The rest of the simulations only consider the QICD algorithm because of the computational advantages over the linear programming approach.

Results in Table 1 show that the group SCAD approaches are fitting smaller models that all contain the true covariates and are doing a better job in terms of prediction. For the different values of p the MA-SCAD approach does the best in terms of prediction error, but this is not surprising as we expect a method using a squared error loss function to perform well when the errors are homoscedastic and normally distributed. In Setting IB, presented in Table 2, the QA-SCAD methods perform the best in terms of model selection and prediction accuracy. The linear mean methods are selecting more false positives because of the extra noise from the heavier tailed error distribution, T_3. For Setting IC, presented in Table 3, the largest difference can be seen in terms of TP. In this model z_1 is only an active variable for $\tau \neq .5$ and thus the mean regression methods do not consistently select this variable. Also, the PS results are slightly better for QA-SCAD approaches than the mean regression methods, which we expect because the mean regression methods for estimating the .9 quantile do not correctly account for the nonconstant variance. In all the results we see the SCAD methods picking smaller models than their LASSO counterparts, but still getting accurate results in terms of selecting the correct number of active covariates.

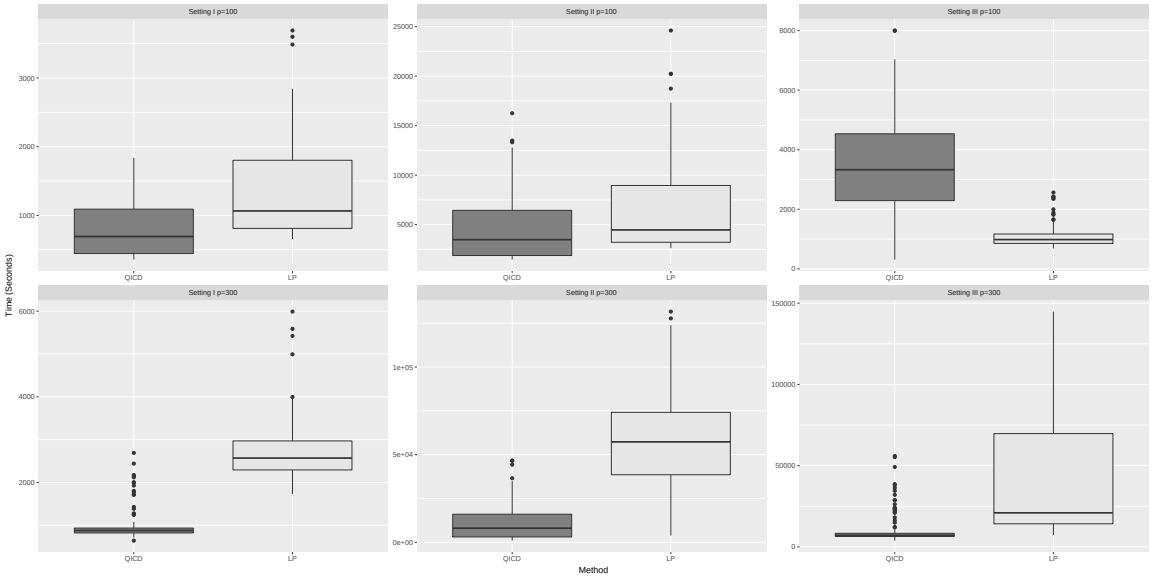

Figure 1: Computation Time comparison of coordinate descent method (QICD) and linear programming (LP) for 100 simulations with $p = 100$.

Method	p	MSPE	MAPE	TP	FP	PS
MA-LASSO	100	1.08 (0.06)	0.83 (0.02)	3 (0)	9.16 (2.4)	0.5 (0.02)
MA-SCAD	100	1.03 (0.05)	0.81 (0.02)	3 (0)	2.38 (2.48)	0.5 (0.02)
QA-LASSO	100	1.2 (0.09)	0.88 (0.03)	3 (0)	11.79 (5.45)	0.5 (0.03)
QA-SCAD CD	100	1.05 (0.06)	0.82 (0.02)	3 (0)	1.23 (0.69)	0.5 (0.03)
QA-SCAD LP	100	1.04 (0.06)	0.82 (0.02)	3 (0)	1.23 (0.78)	0.5 (0.03)
MA-LASSO	300	1.09 (0.06)	0.83 (0.02)	3 (0)	18.14 (2.67)	0.51 (0.02)
MA-SCAD	300	1.03 (0.05)	0.81 (0.02)	3 (0)	10.29 (3.15)	0.51 (0.02)
QA-LASSO	300	1.29 (0.1)	0.91 (0.04)	3 (0)	14.17 (6.91)	0.5 (0.03)
QA-SCAD CD	300	1.05 (0.05)	0.82 (0.02)	3 (0)	2.01 (1.55)	0.51 (0.03)
QA-SCAD LP	300	1.04 (0.05)	0.81 (0.02)	3 (0)	1.6 (1.34)	0.51 (0.03)
MA-LASSO	600	1.11 (0.06)	0.84 (0.02)	3 (0)	28.06 (4.23)	0.5 (0.02)
MA-SCAD	600	1.04 (0.05)	0.81 (0.02)	3 (0)	20 (4.89)	0.5 (0.02)
QA-LASSO	600	1.43 (0.13)	0.95 (0.04)	3 (0)	13.58 (8.26)	0.49 (0.03)
QA-SCAD CD	600	1.07 (0.05)	0.82 (0.02)	3 (0)	2.7 (1.9)	0.5 (0.02)

Table 1: Simulation results for homoscedastic N(0,1) errors (Setting IA)

Setting II: Non-optimal Models

The previous section demonstrated the computational advantages of the CD algorithm so for this setting we only consider the coordinate descent implementation of the proposed approach (QA-SCAD CD). In this section the response is generated from a model where

Method	p	MSPE	MAPE	TP	FP	PS
MA-LASSO	100	3.19 (2.2)	1.18 (0.06)	2.99 (0.1)	7.66 (2.12)	0.51 (0.03)
MA-SCAD	100	3.1 (2.19)	1.15 (0.06)	2.99 (0.1)	6.76 (2.92)	0.5 (0.03)
QA-LASSO	100	3.28 (2.2)	1.21 (0.06)	3 (0)	8.48 (4.47)	0.51 (0.03)
QA-SCAD CD	100	3.03 (2.19)	1.12 (0.06)	3 (0)	1.59 (1.22)	0.5 (0.03)
QA-SCAD LP	100	3.02 (2.19)	1.12 (0.06)	3 (0)	1.25 (0.77)	0.51 (0.03)
MA-LASSO	300	3.05 (0.64)	1.18 (0.06)	3 (0)	15.74 (2.8)	0.5 (0.03)
MA-SCAD	300	2.97 (0.65)	1.15 (0.06)	3 (0)	15.16 (4.09)	0.5 (0.03)
QA-LASSO	300	3.31 (0.65)	1.26 (0.07)	2.96 (0.24)	7.24 (5.32)	0.51 (0.03)
QA-SCAD CD	300	2.89 (0.62)	1.13 (0.05)	3 (0)	2.52 (2.28)	0.5 (0.03)
QA-SCAD LP	300	2.88 (0.63)	1.12 (0.06)	3 (0)	2.04 (1.77)	0.5 (0.03)
MA-LASSO	600	3.16 (0.96)	1.19 (0.07)	2.98 (0.14)	25.96 (4.69)	0.5 (0.03)
MA-SCAD	600	4.99 (5.42)	1.45 (0.73)	2.98 (0.14)	40.37 (35.56)	0.5 (0.03)
QA-LASSO	600	3.52 (0.95)	1.3 (0.08)	2.82 (0.5)	5.67 (5.68)	0.5 (0.03)
QA-SCAD CD	600	2.99 (0.95)	1.14 (0.07)	3 (0)	3.32 (3)	0.5 (0.03)

Table 2: Simulation results for homoscedastic T_3 errors (Setting IB)

Method	p	MSPE	MAPE	MCPE	TP	FP	PS
MA-LASSO	100	1.88 (0.14)	1.17 (0.05)	0.15 (0.01)	2.22 (0.42)	6.86 (2.19)	0.92 (0.01)
MA-SCAD	100	1.78 (0.13)	1.14 (0.05)	0.15 (0.01)	2.04 (0.2)	2.08 (2.53)	0.92 (0.01)
QA-LASSO	100	1.85 (0.36)	1.09 (0.12)	0.16 (0.01)	2.97 (0.17)	28.88 (16.28)	0.86 (0.03)
QA-SCAD CD	100	1.89 (0.35)	1.1 (0.11)	0.15 (0.01)	2.92 (0.27)	2.9 (2.78)	0.89 (0.02)
QA-SCAD LP	100	1.76 (0.21)	1.05 (0.06)	0.14 (0.01)	3 (0)	2.32 (2.19)	0.89 (0.02)
MA-LASSO	300	1.97 (0.17)	1.21 (0.06)	0.15 (0.01)	2.17 (0.38)	14.92 (2.98)	0.93 (0.01)
MA-SCAD	300	1.85 (0.15)	1.17 (0.06)	0.15 (0.01)	2.1 (0.3)	9.93 (3.78)	0.93 (0.01)
QA-LASSO	300	1.9 (0.25)	1.11 (0.09)	0.16 (0.01)	3 (0)	45.61 (12.11)	0.86 (0.02)
QA-SCAD CD	300	2.05 (0.49)	1.15 (0.14)	0.15 (0.02)	2.83 (0.4)	4.13 (4.39)	0.89 (0.02)
QA-SCAD LP	300	1.75 (0.19)	1.04 (0.06)	0.14 (0.01)	3 (0)	6.13 (5.65)	0.88 (0.02)
MA-LASSO	600	2.06 (0.18)	1.24 (0.06)	0.15 (0.01)	2.12 (0.33)	25.23 (4.49)	0.93 (0.01)
MA-SCAD	600	1.92 (0.17)	1.2 (0.06)	0.15 (0.01)	2.08 (0.27)	19.94 (4.55)	0.93 (0.01)
QA-LASSO	600	2 (0.3)	1.14 (0.1)	0.18 (0.01)	3 (0)	54.23 (16.41)	0.85 (0.02)
QA-SCAD CD	600	2.06 (0.52)	1.16 (0.16)	0.16 (0.02)	2.74 (0.46)	5.87 (9.1)	0.88 (0.02)

Table 3: Simulation results for heteroscedastic errors (Setting IC)

the proposed approach is not optimal, either because it is too complex or not complex enough. The settings are

Setting IIA (linear model) $y = 1 + z_1 - z_2 + 3z_3 + \epsilon$;

Setting IIB (partially linear model) $y = 1 + z_1 + 2\sin(4\pi z_2) + 2(z_3 - .5)^3 + \epsilon$;

Setting IIC (nonadditive model) $y = -1 + 2z_1^3[\sin(2\pi z_2) + 1] + 8(z_3 - .5)^2 + \epsilon$.

In each setting $\epsilon \sim T_3$. The QA-SCAD CD model is compared to simpler linear models. We consider linear mean and quantile regression with the SCAD (ML-SCAD, QL-SCAD)

and lasso (ML-LASSO, QL-LASSO) penalty. For the linear models the objective functions are

$$\frac{1}{n}\sum_{i=1}^{n} m_\tau(y_i - \mathbf{z}_i^\top \boldsymbol{\beta}) + \sum_{j=1}^{p} p_{\lambda,a}\left(|\beta_j|\right). \tag{16}$$

Where $m_\tau(x) = x^2$ for the mean models and $m_\tau(x) = \rho_\tau(x)$ for the quantile regression models and $p_{\lambda,a} = \lambda|x|$ for lasso and for SCAD $p_{\lambda,a}(\cdot)$ is the SCAD penalty function. For these simulations we consider the case of $\tau = .5$. For all models λ is selected using BIC similar to what was described in the previous section and for the SCAD penalties a is fixed to 3.7. For Setting IIB J_n was set to 5, while in all other settings we fixed $J_n = 3$. The ML-LASSO and ML-SCAD models were fit using **glmnet** (Friedman et al., 2008) and **ncvreg** (Breheny and Huang, 2011), respectively. All the quantile regression models were fit using **rqPen** (Sherwood and Maidman, 2020).

Results for the simulations are reported in Tables 4-6. The QA-SCAD approach is competitive for both the linear and partially linear setting. In setting IIB it does best at selecting the true variables when p is large, $p \in \{300, 600\}$. For Setting IIC it dominates with respect to all metrics, except PS, which all do well at, and FP, where QL-LASSO does the best. This demonstrates that the proposed additive model has benefits compared to simpler models even when the true model is not additive.

Method	p	MSPE	MAPE	TP	FP	PS
ML-LASSO	100	3.09(2.21)	1.14(0.06)	1.73(0.62)	1.84(2.42)	0.5(0.03)
ML-SCAD	100	3.06(2.21)	1.13(0.05)	1.66(0.83)	2.25(3.85)	0.5(0.03)
QA-SCAD CD	100	3.07(2.21)	1.14(0.06)	2.21(0.41)	1.06(0.34)	0.51(0.03)
QL-LASSO	100	3.08(2.21)	1.14(0.05)	1.71(0.61)	1.88(1.7)	0.5(0.03)
QL-SCAD	100	3.05(2.2)	1.13(0.05)	2.42(0.81)	2.97(1.76)	0.51(0.03)
ML-LASSO	300	2.95(0.63)	1.14(0.05)	1.51(0.52)	2.36(3.61)	0.5(0.03)
ML-SCAD	300	2.91(0.64)	1.13(0.06)	1.39(0.58)	1.87(3.89)	0.5(0.03)
QA-SCAD CD	300	2.91(0.63)	1.13(0.05)	2.3(0.46)	1.3(0.96)	0.5(0.03)
QL-LASSO	300	2.93(0.63)	1.14(0.05)	1.7(0.48)	1.98(1.28)	0.5(0.03)
QL-SCAD	300	2.9(0.63)	1.13(0.05)	1.9(0.72)	2.05(1.31)	0.5(0.03)
ML-LASSO	600	3.05(0.81)	1.16(0.05)	1.5(0.51)	1.84(2.85)	0.5(0.03)
ML-SCAD	600	3.01(0.82)	1.14(0.05)	1.16(0.37)	0.94(1.73)	0.5(0.04)
QA-SCAD CD	600	3(0.84)	1.14(0.05)	2.26(0.44)	1.12(0.48)	0.5(0.03)
QL-LASSO	600	3.02(0.82)	1.15(0.05)	1.7(0.51)	2.7(1.81)	0.5(0.03)
QL-SCAD	600	2.99(0.82)	1.14(0.05)	1.22(0.46)	1.04(0.2)	0.5(0.03)

Table 4: Simulation results for Setting IIA (linear model).

Setting III: Model Selection Performance

To validate the results of Theorem 6 we consider the model selection performance of QA-SCAD CD for q_n, J_n, and p_n increasing with n. The responses in this section are generated from the model

$$\begin{aligned} y \;=\; & -1 + 2z_1^3 + \sin(2\pi z_2) I(q_n > 1) + 8(z_3 - .5)^2 I(q_n > 2) + 2z_{11}^3 I(q_n > 3) \\ & + \sin(2\pi z_{12}) I(q_n > 4) + 8(z_{13} - .5)^2 I(q_n = 6) + \epsilon, \end{aligned}$$

Method	p	MSPE	MAPE	TP	FP	PS
ML-LASSO	100	3.09(2.21)	1.14(0.06)	1.73(0.62)	1.84(2.42)	0.5(0.03)
ML-SCAD	100	3.06(2.21)	1.13(0.05)	1.66(0.83)	2.25(3.85)	0.5(0.03)
QA-SCAD CD	100	3.07(2.21)	1.14(0.06)	2.21(0.41)	1.06(0.34)	0.51(0.03)
QL-LASSO	100	3.08(2.21)	1.14(0.05)	1.71(0.61)	1.88(1.7)	0.5(0.03)
QL-SCAD	100	3.05(2.2)	1.13(0.05)	2.42(0.81)	2.97(1.76)	0.51(0.03)
ML-LASSO	300	2.95(0.63)	1.14(0.05)	1.51(0.52)	2.36(3.61)	0.5(0.03)
ML-SCAD	300	2.91(0.64)	1.13(0.06)	1.39(0.58)	1.87(3.89)	0.5(0.03)
QA-SCAD CD	300	2.91(0.63)	1.13(0.05)	2.3(0.46)	1.3(0.96)	0.5(0.03)
QL-LASSO	300	2.93(0.63)	1.14(0.05)	1.7(0.48)	1.98(1.28)	0.5(0.03)
QL-SCAD	300	2.9(0.63)	1.13(0.05)	1.9(0.72)	2.05(1.31)	0.5(0.03)
ML-LASSO	600	3.05(0.81)	1.16(0.05)	1.5(0.51)	1.84(2.85)	0.5(0.03)
ML-SCAD	600	3.01(0.82)	1.14(0.05)	1.16(0.37)	0.94(1.73)	0.5(0.04)
QA-SCAD CD	600	3(0.84)	1.14(0.05)	2.26(0.44)	1.12(0.48)	0.5(0.03)
QL-LASSO	600	3.02(0.82)	1.15(0.05)	1.7(0.51)	2.7(1.81)	0.5(0.03)
QL-SCAD	600	2.99(0.82)	1.14(0.05)	1.22(0.46)	1.04(0.2)	0.5(0.03)

Table 5: Simulation results for Setting IIB (partially linear model).

Method	p	MSPE	MAPE	TP	FP	PS
ML-LASSO	100	3.61(2.2)	1.3(0.05)	1.12(0.57)	1.67(2.58)	0.52(0.03)
ML-SCAD	100	3.6(2.19)	1.3(0.05)	1.26(0.65)	3.02(4.3)	0.52(0.03)
QA-SCAD CD	100	3.21(2.2)	1.18(0.06)	2.95(0.22)	1.51(1.11)	0.5(0.03)
QL-LASSO	100	3.63(2.21)	1.31(0.06)	1.86(0.49)	0.49(1)	0.51(0.03)
QL-SCAD	100	3.62(2.19)	1.31(0.06)	2.51(0.52)	2.01(1.85)	0.51(0.03)
ML-LASSO	300	3.48(0.62)	1.3(0.05)	0.79(0.43)	1.56(2.76)	0.52(0.03)
ML-SCAD	300	3.47(0.62)	1.3(0.05)	0.85(0.54)	2.18(3.84)	0.52(0.03)
QA-SCAD CD	300	3.07(0.63)	1.18(0.06)	2.88(0.33)	1.84(1.72)	0.5(0.03)
QL-LASSO	300	3.48(0.62)	1.31(0.05)	1.67(0.47)	0.5(0.93)	0.51(0.03)
QL-SCAD	300	3.54(0.63)	1.32(0.06)	2.38(0.58)	4.38(3.41)	0.5(0.03)
ML-LASSO	600	3.55(0.93)	1.31(0.06)	0.86(0.45)	2.22(3.98)	0.51(0.03)
ML-SCAD	600	3.55(0.93)	1.31(0.06)	0.85(0.48)	2.36(3.95)	0.51(0.03)
QA-SCAD CD	600	3.19(0.93)	1.2(0.07)	2.82(0.41)	2.73(3.14)	0.5(0.03)
QL-LASSO	600	3.56(0.92)	1.31(0.06)	1.59(0.49)	0.41(1.2)	0.5(0.03)
QL-SCAD	600	3.53(0.93)	1.3(0.06)	2.01(0.44)	1.93(2.84)	0.5(0.03)

Table 6: Simulation results for Setting IIC (nonadditive model).

where $\epsilon \sim N(0,1)$. We fit the QA-SCAD model in 3 different settings where J_n, q_n, or p_n vary. In all the settings we fit models with sample size of 100, 300, 600 and 1000. In Setting IIIA we fit the model with $J_n \in \{3,4,5\}$. In setting IIIB models are fit with $q_n \in \{1,2,3,4,5,6\}$. Finally, in Setting IIIC models are fit with $p_n \in \{100, 300, 500, 1000, 2000\}$. When they are not varying we fix $J_n = 3$, $q_n = 3$ and $p_n = 300$. For instance, in Setting IIIA we fix $q_n = 3$ and $p_n = 300$ and consider performance of QA-SCAD CD with different values of J_n and n. The purpose of these simulations is to corroborate the model selection

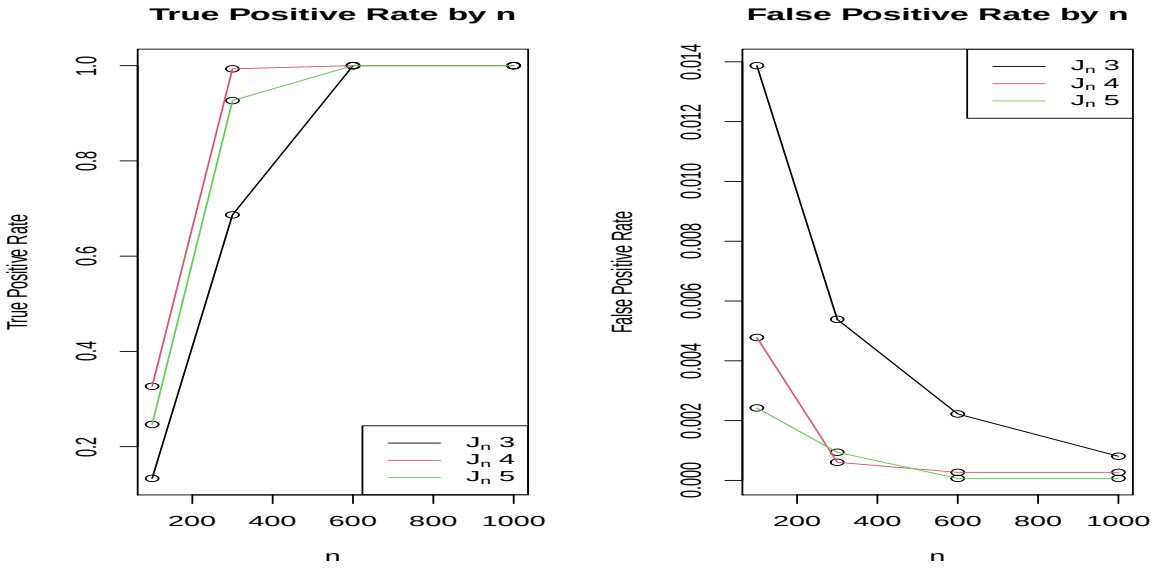

Figure 2: True positive and false positive rates by n with varying J_n from Setting IIIA

properties presented in Theorem 6 and thus we only consider the QA-SCAD CD model. In addition, performance is evaluated with respect to the TP and FP rates, number of true or false covariates selected divided by the number of true or false potential covariates, to account for the fact that q_n and p_n can vary. Again to duplicate the settings of Theorem 6 we fix $\lambda = n^{-1/10}/4$ which for the given additive functions satisfies the potential valid choices of λ outlined after Theorem 6. Figures 2-5 present how the average true positive and false positive rates vary with n in the different settings across the 50 replications. Figure 2 presents how the false and true positives vary with n and J_n. The relationship between J_n and correct model selection is not straightforward. From a theoretical perspective larger values of J_n will provide a better approximation of the functions and thus we could reason that for larger value of J_n the true positive rate should go up and false positive rate should go down. However, from a practical perspective larger values of J_n cause the size of the grouped coefficients to increase where the intuition is this could lead to an increase in the false positive rate. Figure 2 reflects some of this uncertainty, but also shows that as n increases, no matter the value of J_n, the active covariates tend to be selected and the inactive covariates tend to be dropped. Figure 3 demonstrates that the larger the value of q_n the harder it is to select the correct variables. We also see that for large values of n that, no matter the value of q_n, with high probability the active covariates are selected and the inactive covariates are discarded. Figures 4 and 5 present the false and true positives, respectively, as functions of n and $n/\log(p_n)$. The $n/\log(p_n)$ is used to verify that p_n can grow exponentially with n. Both figures demonstrate that settings with large values of p_n are doing worse for smaller n, but for the larger values of n the true covariates tend to be selected and the noise covariates are removed from the models.

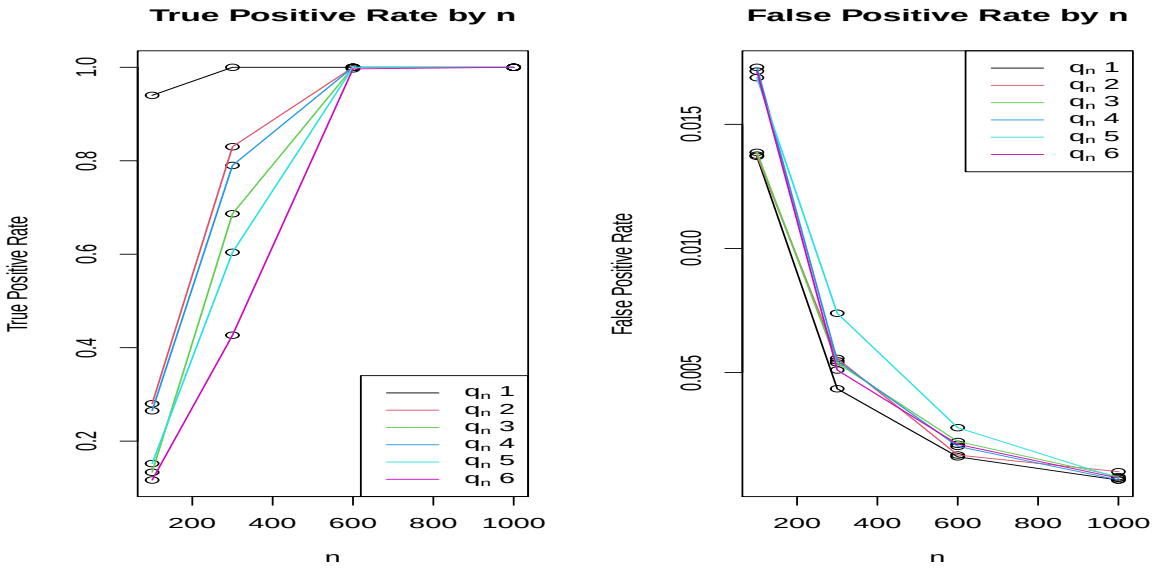

Figure 3: True positive and false positive rates by n with varying q_n from Setting IIIB

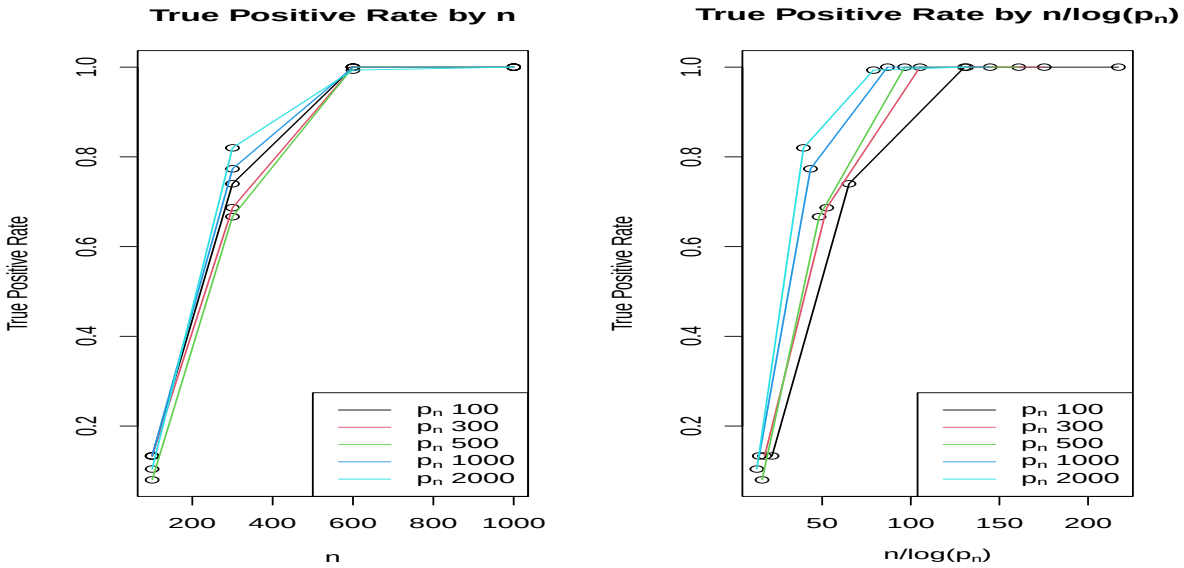

Figure 4: True positive rates by n and $n/log(p_n)$ for varying p_n from Setting IIIC

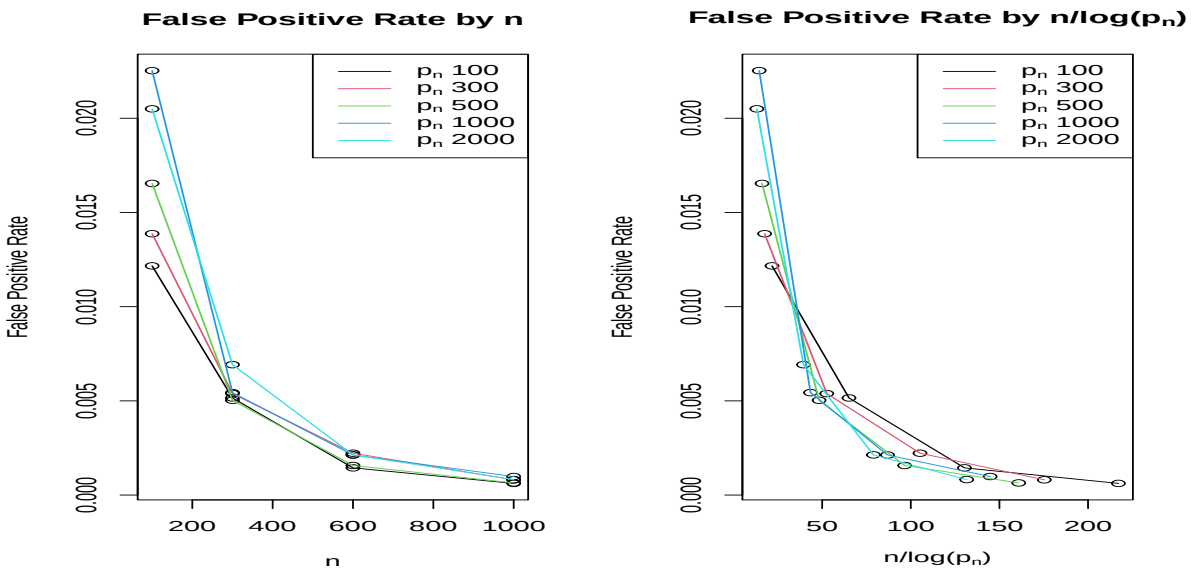

Figure 5: False positive rates by n and $n/log(p_n)$ for varying p_n from Setting IIIC

6. Data Analysis

6.1 Fat content of ground pork

Borggaard and Thodberg (1992) measured the fat content and 100 channel spectrum of absorbances from 240 ground pork meat samples. Our analysis is limited to the 215 samples available from the R package **faraway** (Faraway, 2016) and we removed 5 observations with outlying values leaving us with 210 samples. We analyze this data using the QA-SCAD model and consider both the CD and LP algorithms. We compare these approaches to three other models: (1) MA-SCAD discussed in simulation Setting I; (2) ML-SCAD discussed in simulation Setting II; and (3) QL-SCAD discussed in simulation Setting II. For the linear models no B-spline transformation is done and each coefficient is penalized individually using the SCAD penalty. For the nonlinear additive models a B-spline transformation is used and group penalties are used for coefficients. The channel spectrum is scaled and centered to have a mean of zero and a standard deviation of one. The fat content data is first log transformed and then scaled and centered to have a mean of zero and a standard deviation of one.

To compare the methods we randomly sample 200 of the 210 samples as training data and the other 10 samples are used as testing data. The channel spectrum data is highly correlated. Following an approach similar to the one outlined in Meier et al. (2009), we transform the predictors by using the first 30 principal components. The principal components are centered and scaled to have mean zero and a standard deviation of one. The five models are fit using the 30 principal components as covariates to model log of fat content. For the nonlinear models the principal components are transformed using cubic B-splines with $J_n = 3$.

Fitting a quantile regression model requires a choice of τ. Choosing $\tau = .5$ provides a robust estimate of central tendencies. Two values of τ can be used to create a prediction interval that does not require a parametric assumption about the errors. For instance, models with $\tau = .1$ and $\tau = .9$ can be used to create an 80% prediction interval. If the whole conditional distribution is of interest then a wide range values of τ could be used. To estimate the whole conditional distribution, and test the proposed method for several values of τ, models are fit for values of $\tau \in \mathcal{T} \equiv \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$. The tuning parameter λ is selected using BIC, as outlined in the previous section, and we set $a = 3.7$. The mean models estimate the conditional quantiles using the naive procedure outlined in the previous section. The testing covariates are transformed using the rotation defined by the first 30 principal components from the testing data. The 30 covariates are then centered and scaled by the sample mean and standard deviation from the testing data. For the nonlinear models the testing covariates are transformed by the B-spline functions used on the training data. Using this transformed data, predictions from the six methods are made for the log fat content. This process is repeated 100 times.

Let y_{ij} represent the scaled and centered log fat content of the ith sample from the jth testing data set and $\hat{y}_{ij}^{\tau}$ represent its corresponding estimate for the τth quantile. Let $I(a \leq b)$ take a value of one if $a \leq b$ and zero otherwise. The models are compared using

1. MSPE, $\frac{1}{10}\sum_{i=1}^{10}(y_{ij} - \hat{y}_{ij}^{.5})^2$.
2. MAPE, $\frac{1}{10}\sum_{i=1}^{10}|y_{ij} - \hat{y}_{ij}^{.5}|$.
3. MCPE, $\frac{1}{10}\sum_{\tau\in\mathcal{T}}\sum_{i=1}^{10}\rho_{\tau}(y_{ij} - \hat{y}_{ij}^{\tau})$.
4. Quantile Bias (QB), $\sum_{\tau\in\mathcal{T}}\left|\frac{1}{1000}\sum_{j=1}^{100}\sum_{i=1}^{10}I(y_{ij} \leq \hat{y}_{ij}^{\tau}) - \tau\right|$.

Methods that correctly model the τth quantile will have

$$\frac{1}{1000}\sum_{j=1}^{100}\sum_{i=1}^{10}I(y_{ij} \leq \hat{y}_{ij}^{\tau}) \approx \tau.$$

Thus, QB is providing a summary of how accurate the conditional quantile estimates are across all partitions and all values of τ. The statistic QB and means (and standard deviations) of the other three statistics are reported in Table 7.

In this data set we are comparing the linear and nonlinear approaches to see if there is justification for fitting the more complex nonlinear models. In addition, quantile and mean models are compared to see if the quantile models are providing a better description of the conditional distribution. One of the two nonlinear quantile algorithms has the best average results for MAPE, MCPE and QB. For MSPE the linear quantile approach does the best and the linear mean approach also does better than the nonlinear quantile approach. The superiority of the nonlinear quantile approach in terms of MCPE and QB suggests that the more complex nonlinear quantile models are providing useful predictions for non-central tendencies. Performance of the CD and LP algorithms is similar.

Method	MSPE	MAPE	MCPE	QB
ML-SCAD	0.66(0.41)	0.59(0.19)	0.24(0.07)	0.37
QL-SCAD	0.65(0.41)	0.58(0.19)	0.24(0.07)	0.32
MA-SCAD	0.96(1.56)	0.66(0.25)	0.28(0.1)	0.41
QA-SCAD CD	0.7(0.68)	0.58(0.22)	0.24(0.07)	0.21
QA-SCAD LP	0.71(0.61)	0.57(0.22)	0.23(0.08)	0.22

Table 7: Means (and standard deviations) of statistics from the Monte Carlo randomization results for the ground pork data.

6.2 Modeling TRIM32 expression levels

The previous example provides some evidence that nonlinear quantile regression can provide a less biased estimate of a conditional quantile, but does not demonstrate a dramatic difference between linear and nonlinear quantile regression in terms of prediction accuracy. This section presents an example where the additive quantile regression model outperforms the linear quantile regression model in terms of prediction accuracy. Huang et al. (2010) presented an analysis of modeling high-dimensional genomics data, from Scheetz et al. (2006), where a nonlinear additive mean model improved upon the prediction performance of a linear mean model. We consider the same data set for modeling the conditional median. Scheetz et al. (2006) used 31,042 different probe sets to analyze RNA from the eyes of 120 twelve-week old male rats. Similar to Huang et al. (2010) we model the expression of gene TRIM32, because Chiang et al. (2006) identified TRIM32 as a Bardet-Biedl syndrome gene and one symptom of Bardet-Biedl syndrome is retinal degeneration. Scheetz et al. (2006) note that many of the probes were not expressed in the eye. Thus, following Huang et al. (2010) we limit our analysis to the 500 genes that have the highest absolute Pearson's correlation with the TRIM32 expression.

To demonstrate that this is a setting where the nonlinear quantile model improves prediction accuracy over the linear counterpart we consider the QL-SCAD and QA-SCAD CD approaches using Monte Carlo randomization. All variables are log transformed and the predictors are further transformed to have a minimum value of zero and maximum value of one. First the data is randomly partitioned into a training set of 100 observations and a testing set of 20 observations. We fit the models using the 100 training observations and make prediction of TRIM32 expression on the remaining 20 testing observations. For the nonlinear model we set $J_n = 4$. This process was repeated 100 times and the MAPE recorded at each iteration. Figure 6 presents the MAPE of the two methods, demonstrating that the nonlinear model tends to be more accurate. In addition, in 69 of the 100 iterations the nonlinear model had a lower MAPE than the linear model.

7. Conclusions

We proposed an additive nonlinear model to provide a flexible model. However, it is possible that too complex a model will be fit. For instance, if some of the true functions are linear than the model being fit will be more complex than necessary. To balance model

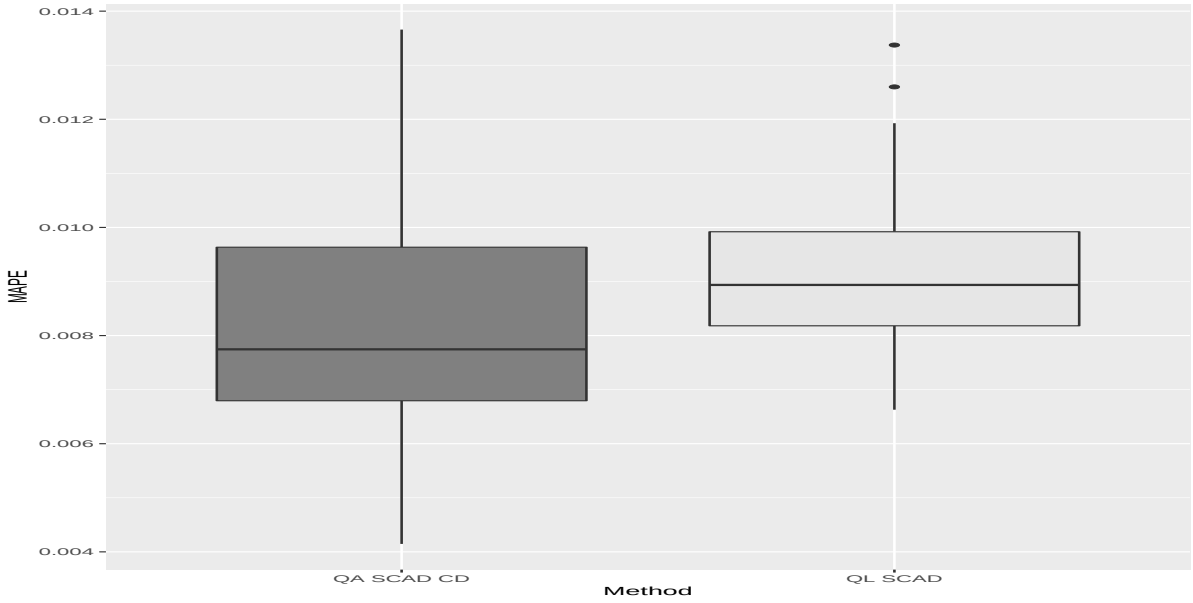

Figure 6: Monte Carlo randomization MAPE of TRIM32 for the genomics data. The nonlinear model tends to perform better and in 69 of the 100 Monte Carlo iterations had the lower MAPE.

complexity and ease of interpretation Lou et al. (2016) proposed a penalized approach for mean regression that does both variable selection and automatic assignment of a covariate to a linear or nonlinear term. However, even this approach has some rigidity as, similar to our work, it requires preset definitions of the basis splines including the number and placement of knots. Desire for flexibility has resulted in methods which use adaptive knots (Petersen et al., 2016; Sadhanala and Tibshirani, 2019) and adaptive knot assignment and classification of predictors as linear or nonlinear (Petersen and Witten, 2019). However, all the cited work has focused on mean regression. Developing adaptive methods for quantile regression would be a useful contribution to this line of research.

Acknowledgments

The authors thank the anonymous reviewers for their time and effort, including finding errors in the proofs of previous versions of this work.

8. Appendix

Throughout the proofs C is used to represent a generic positive constant that can change in value from line to line. We start by presenting some useful equalities that are used

throughout the proofs. For $u \neq 0$, Knight (1998) introduced the equality of

$$|u-v| - |u| = -u[I(u>0) - I(u<0)] + 2\int_0^v [I(u \leq s) - I(u \leq 0)]ds. \tag{17}$$

Define $\psi_\tau(u) = \tau - I(u<0)$. As $\rho_\tau(u) = 1/2[|u| + (2\tau-1)u]$, Koenker (2005) generalized (17) and for $u \neq 0$ presented Knight's identity as

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v I(u \leq s) - I(u \leq 0)ds. \tag{18}$$

Knight's identity also provides that

$$\rho_\tau(u-a-v) - \rho_\tau(u-v) = \int_{-v}^{-(a+v)} \psi_\tau(u+s)ds = \int_v^{a+v} [I(u<s) - \tau]ds, \tag{19}$$

which is an intuitive result as $\psi_\tau(u)$ is the derivative of $\rho_\tau(u)$ where it is defined.

The following definitions are used throughout the proof

$$\begin{aligned}
u_{ni} &= \mathbf{\Pi}_A(\mathbf{z}_i)^\top \boldsymbol{\gamma}_{A0} - g_0(\mathbf{z}_i), \\
D_n &= \text{diag}[f_1(0 \mid \mathbf{z}_1), \ldots, f_n(0 \mid \mathbf{z}_n)] \in \mathbb{R}^{n \times n}, \\
W_D^2 &= \mathbf{\Pi}_A^\top D_n \mathbf{\Pi}_A \in \mathbb{R}^{q_n J_n + 1 \times q_n J_n + 1}, \\
\boldsymbol{\theta} &= W_D(\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}), \\
\tilde{\mathbf{W}}(\mathbf{z}_i) &= W_D^{-1}\mathbf{\Pi}_A(\mathbf{z}_i), \\
Q_i(\boldsymbol{\theta}) &= \rho_\tau[\epsilon_i - \tilde{\mathbf{W}}(\mathbf{z}_i)^\top \boldsymbol{\theta} - u_{ni}], \\
E_z(x) &= E\left(x \mid z\right), \\
D_i(\boldsymbol{\theta}) &= Q_i(\boldsymbol{\theta}) - Q_i(0) - E_{\mathbf{z}_i}\left[Q_i(\boldsymbol{\theta}) - Q_i(0)\right] + \tilde{\mathbf{W}}(\mathbf{z}_i)^\top \boldsymbol{\theta}\psi_\tau(\epsilon_i), \\
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta} \in \mathbb{R}^{q_n J_n + 1}}{\text{argmin}} \sum_{i=1}^n \rho_\tau[\epsilon_i - \tilde{\mathbf{W}}(\mathbf{z}_i)^\top \boldsymbol{\theta} - u_{ni}].
\end{aligned}$$

Notice that $\hat{\boldsymbol{\theta}} = W_D(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0})$ and $\tilde{\mathbf{W}}(\mathbf{z}_i)^\top \hat{\boldsymbol{\theta}} = \mathbf{\Pi}_A(\mathbf{z}_i)^\top \hat{\boldsymbol{\gamma}}_A$ for all $i \in \{1, \ldots, n\}$. Define $d_{j,s} = (t_{j,s+1} - t_{j,s-m})/(m+1)$. The following Lemma is restating Corollary 1 from de Boor (1976) and provided here for convenience.

Lemma 9 *(Corollary 1 from de Boor (1976).) For $1 \leq p < \infty$ and for all $j \in \{1, \ldots, q_n\}$ and $s \in \{0, \ldots, J_n\}$*

$$\frac{(m+1)^{1/p}}{m+1} \leq \left[\int_0^1 d_{j,s}^{-1} \left|b_{j,s}(z)\right|^p dx\right]^{1/p} \leq 1.$$

8.1 Proof of Lemma 4

Proof The majority of the proof for the lower bound follows work done in proof of Lemma 1 from Chen et al. (2018a). The major difference is the constant term, corresponding to the intercept, is accounted for in these results while the results from Chen et al. (2018a) ignore the intercept because it can be removed in mean regression by centering the predictors and

response to be mean zero. However, for quantile regression an intercept is still required after such transformations. Define $\Pi_j \in \mathbb{R}^{n\times J_n}$, as the matrix of splines associated with the jth predictor, such that $\Pi_A = [\mathbf{1}_n, \Pi_1, \ldots, \Pi_{q_n}]$ and $\mathbf{a} = (a_0, \mathbf{a}_1^\top, \ldots, \mathbf{a}_{q_n}^\top)^\top$, where $\mathbf{a}_j \in \mathbb{R}^{J_n}$ for all $j \in \{1, \ldots, q_n\}$. Notice that $\mathbf{a}^\top \Pi_A^\top \Pi_A \mathbf{a} = ||a_0\mathbf{1}_n + \Pi_1\mathbf{a}_1 + \ldots + \Pi_{q_n}\mathbf{a}_{q_n}||_2^2$. By Lemmas S.5 from Chen et al. (2018b) and that $||\Pi_j \mathbf{a}_j||_2 \geq 0$ it follows that

$$||a_0\mathbf{1}_n + \Pi_1\mathbf{a}_1 + \ldots + \Pi_{q_n}\mathbf{a}_{q_n}||_2^2 \geq \left(\frac{1-\delta}{2}\right)^{q_n}\left(na_0^2 + \sum_{j=1}^{q_n}||\Pi_j\mathbf{a}_j||_2^2\right).$$

From Lemma 6.2 of Zhou et al. (1998), for any $j \in \{1, \ldots, q_n\}$ there exists a positive constant C such that $\lambda_{\min}(\Pi_j^\top \Pi_j) \geq CJ_n^{-1}n$. Therefore,

$$\begin{aligned}\left(\frac{1-\delta}{2}\right)^{q_n}\left(na_0^2 + \sum_{j=1}^{q_n}||\Pi_j\mathbf{a}_j||_2^2\right) &\geq C\left(\frac{1-\delta}{2}\right)^{q_n}\left(na_0^2 + CJ_n^{-1}n\sum_{j=1}^{q_n}\mathbf{a}_j^\top\mathbf{a}_j\right)\\ &\geq C\left(\frac{1-\delta}{2}\right)^{q_n}J_n^{-1}n.\end{aligned}$$

It immediately follows that there exists a positive constant b_1 such that $b_1\delta_{q_n}^2 k_n^{-1} \leq \mathbf{a}^\top \frac{1}{n}\mathbf{\Pi}_A^\top \mathbf{\Pi}_A \mathbf{a}$.

For the upper bound, using the Cauchy-Schwarz inequality and that $\sum_{s=0}^{J_n} b_{j,s}(z_{ij}) = 1$ and that $b_{j,s}(z) \geq 0$ for all $z \in [0,1]$, $j \in \{1, \ldots, p_n\}$ and $s \in \{0, \ldots, J_n\}$,

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbf{a}^\top\Pi_A(\mathbf{z}_i)]^2 \leq ||\mathbf{a}||_2^2\frac{1}{n}\sum_{i=1}^{n}\left[1 + \sum_{j=1}^{q_n}\sum_{s=1}^{J_n}b_{j,s}(z_{ij})^2\right] \leq q_n + 1.$$

Set $B_1 = 2$ and the result immediately follows. ■

The following lemma provides some bounds on the vector and matrices of B-splines.

Lemma 10 *Under Conditions 1-4 and for sufficiently large n the following properties hold.*

(1) For $\mathbf{a} \in \mathbb{R}^{q_n J_n + 1}$ where $||\mathbf{a}||_2 = 1$, there exist positive constants b_2 and B_2 such that for sufficiently large n that $b_2\delta_{q_n}^2 k_n^{-1} \leq \mathbf{a}^\top \frac{1}{n}W_D^2\mathbf{a} \leq B_2 q_n$.

(2) There exists a constant b_3 such that $\max_i ||\tilde{\mathbf{W}}(\mathbf{z}_i)||_2 \leq b_3\delta_{q_n}^{-1}\sqrt{\frac{k_n q_n}{n}}$.

(3) There exist constants $m_1 < M_1$ such that for all $j \in \{1, \ldots, q_n\}$ and $s \in \{0, \ldots, J_n\}$

$$m_1 k_n^{-1} \leq \int_0^1 b_{j,s}^2(z)dz \leq M_1 k_n^{-1}.$$

(4) For all $j \in \{1, \ldots, q_n\}$ and all $s \in \{0, \ldots, J_n\}$ there exist positive constants $m_2 < M_2$ such that

$$m_2 k_n^{-1} \leq E[b_{j,s}(z_{ij})^2] \leq M_2 k_n^{-1}.$$

Proof

(1) Follows from Condition 1, providing uniform upper and lower bounds for $f_i(0)$ for all $i \in \{1, \ldots, n\}$, and Lemma 4.

(2) By Lemma 10 (1), it follows that

$$||\tilde{\mathbf{W}}(\mathbf{z}_i)||_2^2 \leq b_2 \delta_{q_n}^{-2} k_n n^{-1} ||\mathbf{\Pi}_A(\mathbf{z}_i)||_2^2 = b_2 \delta_{q_n}^{-2} k_n n^{-1} \left[1 + \sum_{j=1}^{q_n} \sum_{s=1}^{J_n} b_{j,s}(z_{ij})^2 \right] \leq C \delta_{q_n}^{-2} k_n n^{-1} q_n.$$

(3) Using Lemma 9 with $p = 2$, squaring all terms and moving $d_{j,s}$ to the upper and lower bounds it follow that

$$(m+1)^{-1} d_{j,s} \leq \int_0^1 |b_{j,s}(z)|^2 \, dx \leq d_{j,s}. \tag{20}$$

By Condition 3 and the definition of h there exist positive constants $c^* < C^*$ such that for all $j \in \{1, \ldots, q_n\}$ and $s \in \{0, \ldots, J_n\}$ that

$$c^* k_n^{-1} \leq d_{j,s} \leq C^* k_n^{-1}. \tag{21}$$

Proof is complete by combining equations (20) and (21).

(4) Using c_1 and c_2 from Condition 2 it follows that for all $j \in \{1, \ldots, q_n\}$ and $s \in \{0, \ldots, J_n\}$

$$c_1 \int_0^1 b_{j,s}^2(z) dz \leq \int_0^1 b_{j,s}^2(z) f_{z_j}(z) dz \leq c_2 \int_0^1 b_{j,s}^2(z) dz. \tag{22}$$

Proof is complete by combining (22) with Lemma 10 (3).

■

Lemma 11 *Under Condition 4, for any positive constants a and b,* $\delta_{q_n}^{-a} = o\left(n^b\right)$.

Proof Condition 4 provides that $q_n = o[\log(n)]$. Therefore,

$$\frac{\delta_{q_n}^{-a}}{n^b} = \exp\left(\log(n) \left\{ a q_n / [2 \log(n)] \log[2/(1-\delta)] - b \right\}\right) = o(1).$$

■

The following lemma is central to our proof of Theorem 5.

Lemma 12 *For some positive constant L*

$$\sup_{||\boldsymbol{\theta}||_2 \leq L, \boldsymbol{\theta} \in \mathbb{R}^{q_n J_n + 1}} (q_n k_n)^{-1} \left| \sum_{i=1}^n D_i(\sqrt{q_n k_n} \boldsymbol{\theta}) \right| = o_P(1).$$

Proof Define $\tilde{\mathbf{W}}_n = \max_i ||\tilde{\mathbf{W}}(\mathbf{z}_i)||_2$. Let F_{n1} denote the event $\tilde{\mathbf{W}}_n < C_1 n^{-1/2}(q_n k_n)^{1/2}\delta_{q_n}^{-1}$ and F_{n2} denote the event $\max_i |u_{ni}| < C_2 q_n k_n^{-r}$. Combining Lemma 10 and (4) it follows that there exist positive constants C_1 and C_2 such that $P(F_{n1}, F_{n2}) \to 1$. Thus, to prove Lemma 12 it is sufficient to show that $\forall\, \epsilon > 0$

$$P\left[\sup_{||\boldsymbol{\theta}||_2 \le 1,\, \boldsymbol{\theta}\in\mathbb{R}^{q_n J_n+1}} (q_n k_n)^{-1}\left|\sum_{i=1}^n D_i(L\sqrt{q_n k_n}\boldsymbol{\theta})\right| > \epsilon,\, F_{n1}\cap F_{n2}\right] \to 0.$$

Define $\Theta \equiv \left\{\boldsymbol{\theta} \mid ||\boldsymbol{\theta}||_2 \le 1,\, \boldsymbol{\theta}\in\mathbb{R}^{q_n J_n+1}\right\}$. We can partition Θ as a union of disjoint regions $\Theta_1, ..., \Theta_{M_n}$, such that the diameter of each region does not exceed $m_0 = \frac{\epsilon\delta_{q_n}}{4C_1 L\sqrt{n}}$. Then, following the proof of Lemma 3.2 in He and Shi (1994), $M_n \le \left(2\sqrt{q_n J_n + 1}/m_0 + 1\right)^{q_n J_n+1}$. Let $\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{M_n}^*$ be arbitrary points in $\Theta_1, ..., \Theta_{M_n}$. Then

$$P\left[\sup_{||\boldsymbol{\theta}||_2\le 1}(q_n k_n)^{-1}\left|\sum_{i=1}^n D_i(L\sqrt{q_n k_n}\boldsymbol{\theta})\right| > \epsilon, F_{n1}\cap F_{n2}\right] \le \sum_{k=1}^{M_n} P\Bigg\{(q_n k_n)^{-1}\left|\sum_{i=1}^n D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right|$$
$$+\sup_{\boldsymbol{\theta}\in\Theta_k}(q_n k_n)^{-1}\left|\sum_{i=1}^n\left[D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}) - D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right]\right| > \epsilon, F_{n1}\cap F_{n2}\Bigg\}.$$

We will next show that

$$\sup_{\boldsymbol{\theta}\in\Theta_k}(q_n k_n)^{-1}\left|\sum_{i=1}^n\left[D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}) - D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right]\right| I\left(F_{n1}\cap F_{n2}\right) \le \epsilon/2.$$

From definition of $D_i(\boldsymbol{\theta})$ and $Q_i(\boldsymbol{\theta})$ and that $\rho_\tau(u) = \frac{1}{2}|u| + \left(\tau - \frac{1}{2}\right)u$ for fixed $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$

$$D_i(\sqrt{q_n k_n}\boldsymbol{\theta}) - D_i(\sqrt{q_n k_n}\boldsymbol{\theta}^*) = \frac{1}{2}\left[\left|\epsilon_i - \sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta} - u_{ni}\right| - \left|\epsilon_i - \sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}^* - u_{ni}\right|\right]$$
$$-\quad \frac{1}{2}E_{\mathbf{z}_i}\left[\left|\epsilon_i - \sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta} - u_{ni}\right| - \left|\epsilon_i - \sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}^* - u_{ni}\right|\right] + \sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top[\boldsymbol{\theta} - \boldsymbol{\theta}^*]\psi_\tau(\epsilon_i).$$

Then using the above equality, the triangle inequality and the definition of m_0

$$\sup_{\boldsymbol{\theta}\in\Theta_k}(q_n k_n)^{-1}\left|\sum_{i=1}^n\left[D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}) - D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right]\right| I\left(F_{n1}\cap F_{n2}\right)$$
$$\le\quad 2nLm_0(q_n k_n)^{-1/2}\tilde{\mathbf{W}}_n I\left(F_{n1}\cap F_{n2}\right) \le 2\sqrt{n}Lm_0 C_1\delta_{q_n}^{-1} = \epsilon/2.$$

The proof is complete if it can be shown that

$$\sum_{k=1}^{M_n} P\left(\left|\sum_{i=1}^n D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right| > q_n k_n\epsilon/2, F_{n1}\cap F_{n2}\right) \to 0. \tag{23}$$

We will use Bernstein's inequality to prove the above result. First we need upper bounds for the maximum and the second moment for the left side of the above inequality. Note that for any $\boldsymbol{\theta}$

$$D_i(\boldsymbol{\theta}) = \frac{1}{2}\left[\left|\epsilon_i - \tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] - \frac{1}{2}E_{\mathbf{z}_i}\left[\left|\epsilon_i - \tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] + \tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\psi_\tau(\epsilon_i).$$

Then using the triangle inequality and the above equality we have,

$$\max_i \left|D_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*)\right| I\left(F_{n1} \cap F_{n2}\right) \leq 2L\sqrt{q_n k_n}\tilde{\mathbf{W}}_n I\left(F_{n1} \cap F_{n2}\right) \leq 2LC_1\delta_{q_n}^{-1} q_n k_n n^{-1/2}.$$

Define $V_i(\boldsymbol{\theta}) = Q_i(\boldsymbol{\theta}) - Q_i(\mathbf{0}) + \tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\psi_\tau(\epsilon_i)$. Notice that $D_i(\boldsymbol{\theta}) = V_i(\boldsymbol{\theta}) - E_{\mathbf{z}}[V_i(\boldsymbol{\theta})]$, and that $\sum_{i=1}^n \mathrm{Var}\left[D_i(\boldsymbol{\theta})I\left(F_{n1} \cap F_{n2}\right) \mid \mathbf{z}_i\right] \leq \sum_{i=1}^n E\left[V_i^2(\boldsymbol{\theta})I\left(F_{n1} \cap F_{n2}\right) \mid \mathbf{z}_i\right]$. Using Knight's identity

$$\begin{aligned} V_i(L\sqrt{q_n k_n}\boldsymbol{\theta}_k^*) &= L\sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*\left[I(\epsilon_i - u_{ni} < 0) - I(\epsilon_i < 0)\right] \\ &+ \int_0^{L\sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*} \left[I(\epsilon_i - u_{ni} < s) - I(\epsilon_i - u_{ni} < 0)\right] ds \equiv V_{i1} + V_{i2}. \end{aligned}$$

We have

$$\begin{aligned} &\sum_{i=1}^n E_{\mathbf{z}_i}\left[V_{i1}^2 I\left(F_{n1} \cap F_{n2}\right)\right] \leq Cq_n k_n \sum_{i=1}^n E_{\mathbf{z}_i}\left[\tilde{\mathbf{W}}_n^2 I\left(0 < |\epsilon_i| < |u_{ni}|\right) I\left(F_{n1} \cap F_{n2}\right)\right] \\ \leq\ & C\delta_{q_n}^{-2}(q_n k_n)^2 n^{-1} \sum_{i=1}^n \int_{-|u_{ni}|}^{|u_{ni}|} f_i(s \mid \mathbf{z}_i) ds I\left(F_{n1} \cap F_{n2}\right) \leq C\delta_{q_n}^{-2} q_n^3 k_n^2 k_n^{-r}, \end{aligned}$$

where the last inequality uses Condition 1. Noting that V_{i2} is always non-negative and that there exists a positive constant C such that $\max_i \left|\sqrt{q_n k_n}L\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*\right| I\left(F_{n1} \cap F_{n2}\right) \leq C\delta_{q_n}^{-1} q_n k_n n^{-1/2}$, we have

$$\begin{aligned} &\sum_{i=1}^n E_{\mathbf{z}_i}\left[V_{i2}^2 I\left(F_{n1} \cap F_{n2}\right)\right] \leq \max_i \left|\sqrt{q_n k_n}L\left[\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*\right]\right| \\ &\times \sum_{i=1}^n E\left\{\int_0^{\sqrt{q_n k_n}L\left[\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*\right]} \left[I(\epsilon_i - u_{ni} < s) - I(\epsilon_i - u_{ni} < 0)\right] ds \middle| \mathbf{z}_i\right\} I\left(F_{n1} \cap F_{n2}\right) \\ \leq\ & C\delta_{q_n}^{-1} q_n k_n n^{-1/2} \sum_{i=1}^n \int_0^{\sqrt{q_n k_n}L\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*} \left[F_i(s + u_{ni} \mid \mathbf{z}_i) - F_i(u_{ni} \mid \mathbf{z}_i)\right] I\left(F_{n1} \cap F_{n2}\right) ds \\ \leq\ & C\delta_{q_n}^{-1} q_n k_n n^{-1/2} \sum_{i=1}^n \int_0^{\sqrt{q_n k_n}L\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}_k^*} \left[f_i(0 \mid \mathbf{z}_i) + o(1)\right]\left[s + O(s^2)\right] ds \\ \leq\ & C\delta_{q_n}^{-1}(q_n k_n)^2 n^{-1/2}\boldsymbol{\theta}_k^{*\top}\left[\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)\tilde{\mathbf{W}}(\mathbf{z}_i)\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\right]\boldsymbol{\theta}_k^*[1 + o(1)] \\ =\ & C\delta_{q_n}^{-1}(q_n k_n)^2 n^{-1/2}||\boldsymbol{\theta}_k^*||_2^2[1 + o(1)] \leq C\delta_{q_n}^{-1}(q_n k_n)^2 n^{-1/2}[1 + o(1)]. \end{aligned}$$

Where the last equality follows because $\sum_{i=1}^{n} f_i(0 \mid \mathbf{z}_i)\tilde{\mathbf{W}}(\mathbf{z}_i)\tilde{\mathbf{W}}(\mathbf{z}_i)^\top = W_D^{-1}W_D^2W_D^{-1} = I$. Therefore, for sufficiently large n,

$$\sum_{i=1}^{n} \text{Var}\left[D_i(\boldsymbol{\theta})I\left(F_{n1} \cap F_{n2}\right)\right] \quad \leq \quad C\delta_{q_n}^{-1}(q_nk_n)^2\left(\delta_{q_n}^{-1}q_nk_n^{-r} + n^{-1/2}\right).$$

By Bernstein's inequality, for all n sufficiently large,

$$\begin{aligned}
&\sum_{k=1}^{M_n} P\left[\left|\sum_{i=1}^{n} D_i\left(\boldsymbol{\theta}_k^*, L\sqrt{q_nk_n/n}\right)\right| > q_nk_n\epsilon/2, F_{n1} \cap F_{n2}\middle|\mathbf{z}_i\right] \\
\leq \quad & 2\sum_{k=1}^{M_n} \exp\left\{\frac{-(q_nk_n)^2\epsilon^2/4}{C\delta_{q_n}^{-1}(q_nk_n)^2\left[\delta_{q_n}^{-1}q_nk_n^{-r} + n^{-1/2}\right] + C\epsilon\delta_{q_n}^{-1}q_nk_nn^{-1/2}}\right\} \\
\leq \quad & 2M_n\text{exp}\left(\frac{-\epsilon^2}{C\delta_{q_n}^{-2}q_nk_n^{-r}}\right) \leq C\left[C\sqrt{nq_nJ_n}(\epsilon\delta_{q_n})^{-1} + 1\right]^{q_nJ_n+1}\exp\left(\frac{-\epsilon^2}{C\delta_{q_n}^{-2}q_nk_n^{-r}}\right) \\
\leq \quad & C\text{exp}\left\{Cq_nk_n\log\left[C\sqrt{nq_nk_n}(\epsilon\delta_{q_n})^{-1}\right] - C\epsilon^2\delta_{q_n}^2q_n^{-1}k_n^r\right\} \\
\leq \quad & C\text{exp}\left[Cq_nk_n\log(n) - C\epsilon^2\delta_{q_n}^2q_n^{-1}k_n^r\right].
\end{aligned}$$

By taking the expected value of the initial conditional probability and the final upper bound it follows that

$$\sum_{k=1}^{M_n} P\left(\left|\sum_{i=1}^{n} D_i(\boldsymbol{\theta}_k^*, L\sqrt{q_nk_n/n})\right| > q_nk_n\epsilon/2, F_{n1} \cap F_{n2}\right) \leq C\text{exp}\left[Cq_nk_n\log(n) - C\epsilon^2\delta_{q_n}^2q_n^{-1}k_n^r\right].$$

Where the upper bound goes to zero because by Conditions 3-4 and Lemma 11 it follows that $\delta_{q_n}^2q_n^{-1}k_n^r \to \infty$ and $\frac{q_n^2\delta_{q_n}^{-2}k_n\log(n)}{k_n^r} \to 0$. ■

8.2 Proof of Theorem 5

Proof We will first prove that for all $\eta > 0$, there exists an $L > 0$ such that

$$P\left\{\inf_{\substack{\boldsymbol{\theta}\in\mathbb{R}^{q_nJ_n+1} \\ \|\boldsymbol{\theta}\|_2=L}} \frac{1}{q_nk_n}\sum_{i=1}^{n}\left[Q_i(\sqrt{q_nk_n}\boldsymbol{\theta}) - Q_i(0)\right] > 0\right\} \geq 1 - \eta. \tag{24}$$

Define

$$\begin{aligned}
G_{n1}(\boldsymbol{\theta}) &= (q_nk_n)^{-1}\sum_{i=1}^{n} D_i\left(\sqrt{q_nk_n}\boldsymbol{\theta}\right), \\
G_{n2}(\boldsymbol{\theta}) &= (q_nk_n)^{-1}\sum_{i=1}^{n} E_{\mathbf{z}_i}\left[Q_i\left(\sqrt{q_nk_n}\boldsymbol{\theta}\right) - Q_i(0)\right], \\
G_{n3}(\boldsymbol{\theta}) &= -(q_nk_n)^{-1/2}\sum_{i=1}^{n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\psi_\tau(\epsilon_i),
\end{aligned}$$

and note that $(q_n k_n)^{-1}\sum_{i=1}^n \left[Q_i(\sqrt{q_n k_n}\boldsymbol{\theta}) - Q_i(0)\right] = \sum_{k=1}^3 G_{nk}(\boldsymbol{\theta})$. From Lemma 12 we have that $\sup_{||\boldsymbol{\theta}||_2 \leq L} |G_{n1}| = o_P(1)$. For G_{n3}, first notice that $E(G_{n3}) = 0$. From Condition 1 there exists a positive constant c^* such that $\min_i f_i(0 \mid \mathbf{z}_i) \geq c^*$ and thus

$$\begin{aligned} E[G_{n3}^2] &\leq C(q_n k_n)^{-1}\sum_{i=1}^n E\left\{\frac{f_i(0 \mid \mathbf{z}_i)}{c^*}\left[\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\right]^2\right\} \\ &\leq C(q_n k_n)^{-1}\boldsymbol{\theta}^\top E\left[W_D^{-1}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)\boldsymbol{\Pi}_A(\mathbf{z}_i)\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top W_D^{-1}\right]\boldsymbol{\theta} = C(q_n k_n)^{-1}||\boldsymbol{\theta}||_2^2. \end{aligned}$$

Therefore, $\sup_{||\boldsymbol{\theta}||_2 \leq L} G_{n3}(\boldsymbol{\theta}) = O_P\left[(q_n k_n)^{-1/2}\,||\boldsymbol{\theta}||_2\right]$. We will complete the proof by proving that $\inf_{||\boldsymbol{\theta}||_2 \leq L} G_{n2}(\boldsymbol{\theta})$ has a positive asymptotic lower bound that does not converge to zero. Applying (19)

$$\begin{aligned} G_{n2}(\boldsymbol{\theta}) &= (q_n k_n)^{-1}\sum_{i=1}^n E_{\mathbf{z}_i}\left\{\int_{u_{ni}}^{\sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}+u_{ni}} [I(\epsilon_i \leq s) - I(\epsilon_i \leq 0)]ds\right\} \\ &= (q_n k_n)^{-1}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)\frac{1}{2}\left\{q_n k_n\left[\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\right]^2 + u_{ni}\sqrt{q_n k_n}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\right\}[1+o(1)] \\ &= C||\boldsymbol{\theta}||_2^2[1+o(1)] + (q_n k_n)^{-1/2}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)u_{ni}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}[1+o(1)]. \end{aligned}$$

Define $\mathbf{u}_n = (u_{n1}, \ldots, u_{nn})^\top \in \mathbb{R}^n$, by Condition 3 it follows that

$$\begin{aligned} &\sup_{||\boldsymbol{\theta}||_2 \leq L}\left|(q_n k_n)^{-1/2}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)u_{ni}\tilde{\mathbf{W}}(\mathbf{z}_i)^\top\boldsymbol{\theta}\right| \\ \leq &\sup_{||\boldsymbol{\theta}||_2 \leq L}(q_n k_n)^{-1/2}||\mathbf{u}_n^\top D_n^{1/2}||_2||D_n^{1/2}\Pi_A W_D^{-1}||_2||\boldsymbol{\theta}||_2 = O_P(||\boldsymbol{\theta}||_2). \end{aligned}$$

Proof of (24) is completed by noticing that for sufficiently large L, $\inf_{||\boldsymbol{\theta}||_2 \leq L} G_{n2}(\boldsymbol{\theta})$ has a dominating, positive lower bound of $||\boldsymbol{\theta}||_2^2$. By the corollary to Theorem 25 in Eggleston (1958) (p.47) and the convexity of $Q_i(\cdot)$, (24) implies $||\hat{\boldsymbol{\theta}}||_2 = O_P(\sqrt{q_n k_n})$. From the definition of $\hat{\boldsymbol{\theta}}$, it follows that $||W_D(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0})||_2 = O_P\left(\sqrt{q_n k_n}\right)$. Condition 5 and (4) guarantee that $u_{ni} = O(q_n k_n^{-r})$ and therefore

$$\begin{aligned} n^{-1}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)\left[\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)\right]^2 &= n^{-1}\sum_{i=1}^n f_i(0 \mid \mathbf{z}_i)\left[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0}) - u_{ni}\right]^2 \\ &\leq 2n^{-1}\left(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0}\right)^\top W_D^2\left(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0}\right) + O_P\left(q_n^2 k_n^{-2r}\right) \\ &= O_P\left(n^{-1}q_n k_n + q_n^2 k_n^{-2r}\right). \end{aligned}$$

By Condition 1, which provides a constant uniform lower bound for $f_i(0)$ for all $i \in \{1, \ldots, n\}$, $n^{-1}\sum_{i=1}^n \left[\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)\right]^2 = O_P\left(n^{-1}q_n k_n + q_n^2 k_n^{-2r}\right)$. ■

The following lemmas are used to prove Theorem 6

Lemma 13 *If Conditions 1-4 hold, then*

$$||\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{A0}||_2 = O_P\left(k_n \delta_{q_n}^{-1}\sqrt{\frac{q_n}{n}}\right).$$

Proof The proof of Theorem 5 shows $||W_D(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{A0})||_2 = O_P(\sqrt{q_n k_n})$. While from Lemma 10 it follows that $||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{A0}||_2 \leq b_2^{-1/2}\sqrt{\frac{k_n}{n}}\delta_{q_n}^{-1}||W_D(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{A0})||_2$. ■

Lemma 14 *If the Conditions of Theorem 6 hold then*

$$P\left(\max_{q_n+1\leq j\leq p_n}\frac{1}{n}\left|\left|\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}\right|\right|_1 > \lambda/4\right) \to 0.$$

Proof Recall that $\boldsymbol{\pi}_j(z_{ij}) = [b_{j,1}(z_{ij}), ..., b_{j,J_n}(z_{ij})]^\top$. Note, $\max_{j,s,i}|b_{j,s}(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}| \leq 1$ and $E\left[b_{j,s}^2(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}^2\right] \leq Ck_n^{-1}$, see Theorem 10 (4) for the latter. For sufficiently large n, using Bernstein's inequality,

$$\begin{aligned} P\left(\left|\sum_{i=1}^{n} b_{j,s}(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}\right| > nk_n^{-1}\lambda/4\right) &\leq 2\exp\left(-\frac{\lambda^2 n^2 k_n^{-2}/32}{Cnk_n^{-1} + \lambda nk_n^{-1}/12}\right) \\ &\leq 2\exp\left(-C\lambda^2 nk_n^{-1}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} & P\left(\max_{q_n+1\leq j\leq p_n}\frac{1}{n}\left|\left|\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}\right|\right|_1 > \lambda/4\right) \\ \leq\ & P\left(Ck_n\max_{q_n+1\leq j\leq p_n}\max_{1\leq s\leq J_n} n^{-1}\left|\sum_{i=1}^{n} b_{j,s}(z_{ij})\left\{I\left[Y_i - g_0(\mathbf{z}_i)\leq 0\right] - \tau\right\}\right| > \lambda/4\right) \\ \leq\ & Cp_n k_n \exp(-Cnk_n^{-1}\lambda^2) = C\exp(\log p_n + \log k_n - Cnk_n^{-1}\lambda^2) \to 0. \end{aligned}$$

Where the limit holds using the rates of p_n and λ provided in Theorem 6.

■

Lemma 15 *Assume the Conditions of Theorem 6 hold*

$$\begin{aligned} & P\Bigg(\max_{q_n+1\leq j\leq p_n}\sup_{||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}||_2 \leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}}\Bigg|\Bigg|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\Big\{I[Y_i - \boldsymbol{\Pi}(\mathbf{z}_i)_A^\top\boldsymbol{\gamma}_A \leq 0] \\ & -I[Y_i - g_0(\mathbf{z}_i)\leq 0] - P[Y_i - \boldsymbol{\Pi}(\mathbf{z}_i)_A^\top\boldsymbol{\gamma}_A \leq 0 \mid \mathbf{z}_i] + P[Y_i - g_0(\mathbf{z}_i)\leq 0 \mid \mathbf{z}_i]\Big\}\Bigg|\Bigg|_1 > \lambda/8\Bigg) \to 0. \end{aligned}$$

Proof Extending results from Welsh (1989) and Wang et al. (2012), we consider the set $\mathcal{Z} \equiv \left\{\boldsymbol{\gamma}_A : ||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}||_2 \leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}\right\}$. The set $\mathcal{Z}$ can be covered by balls with radii $Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}$ and cardinality $N \equiv |\mathcal{Z}| \leq \tilde{C}n^{4k_n^2\delta_{q_n}^{-2}q_n}$, for some positive constant $\tilde{C}$. Denote the N balls by $\boldsymbol{\gamma}_A(\mathbf{u}_1), ..., \boldsymbol{\gamma}_A(\mathbf{u}_N)$, where the ball $\boldsymbol{\gamma}_A(\mathbf{u}_l)$ is centered at $\mathbf{u}_l$ for $l \in \{1, \ldots, N\}$. Let $\epsilon_i(\boldsymbol{\gamma}_A) = Y_i - \boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\boldsymbol{\gamma}_A$, $\epsilon_i = Y_i - g_0(\mathbf{z}_i)$ and $m_i(a,b) = I(a \leq 0) - I(b \leq 0)$. Then

$$
\begin{aligned}
& P\left[\sup_{||\boldsymbol{\gamma}_A-\boldsymbol{\gamma}_{A0}||_2\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}}\left|\left|\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij})\Big(m_i[\epsilon_i(\boldsymbol{\gamma}_A),\epsilon_i] - E\{m_i[\epsilon_i(\boldsymbol{\gamma}_A),\epsilon_i] \mid \mathbf{z}_i\}\Big)\right|\right|_1 > n\lambda/8\right] \\
\leq & \sum_{l=1}^N P\left[\left|\left|\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij})\Big(m_i[\epsilon_i(\mathbf{u}_l),\epsilon_i] - E\{m_i[\epsilon_i(\mathbf{u}_l),\epsilon_i] \mid \mathbf{z}_i\}\Big)\right|\right|_1 > n\lambda/16\right] \\
+ & \sum_{l=1}^N P\Bigg[\sup_{||\tilde{\boldsymbol{\gamma}}_A-\mathbf{u}_l||_2\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}}\Bigg|\Bigg|\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij})\Big\{m_i\left[\epsilon_i(\tilde{\boldsymbol{\gamma}}_A),\epsilon_i(\mathbf{u}_l)\right] \\
& -E\left\{m_i\left[\epsilon_i(\tilde{\boldsymbol{\gamma}}_A),\epsilon_i(\mathbf{u}_l)\right] \mid \mathbf{z}_i\right\}\Big\}\Bigg|\Bigg|_1 > n\lambda/16\Bigg] \\
\equiv & I_{nj1} + I_{nj2}.
\end{aligned}
$$

Notice that

$$
I_{nj1} \leq \sum_{l=1}^N\sum_{s=1}^{J_n} P\left[\left|\sum_{i=1}^n b_{j,s}(z_{ij})\left(m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] - E\left\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] \mid \mathbf{z}_i\right\}\right)\right| > \frac{n\lambda}{16J_n}\right].
$$

To evaluate I_{nj1}, define $\nu_{ijls} = b_{j,s}(z_{ij})\left(m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] - E\left\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] \mid \mathbf{z}_i\right\}\right)$, which are bounded, independent mean-zero random variables. Note that

$$
\mathrm{Var}\left(\nu_{ijls} \mid \mathbf{z}_i\right) = b_{j,s}(z_{ij})^2\left(E\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right]^2 \mid \mathbf{z}_i\} - E\left\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] \mid \mathbf{z}_i\right\}^2\right).
$$

Then using Condition 1

$$
\begin{aligned}
& E\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right]^2 \mid \mathbf{z}_i\} - E\{m_i\left[\epsilon_i(\mathbf{u}_l),\epsilon_i\right] \mid \mathbf{z}_i\}^2 \\
= & F_i(0 \mid \mathbf{z}_i)[1 - F_i(0 \mid \mathbf{z}_i)] + 2F_i(0 \mid \mathbf{z}_i)F_i[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni} \mid \mathbf{z}_i] \\
& +F_i[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni} \mid \mathbf{z}_i]\left\{1 - F_i[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni} \mid \mathbf{z}_i]\right\} \\
& -2F_i\left\{\min[0, \boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni}] \mid \mathbf{z}_i\right\} \leq C\left|\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni}\right|.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality and Lemma 10 it follows that

$$
\begin{aligned}
\sum_{i=1}^n \mathrm{Var}(\nu_{ijl} \mid \mathbf{z}_i) &\leq Cn\left\{n^{-1}\sum_i\left[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni}\right]^2\right\}^{1/2} \leq C\left(n\sqrt{||\mathbf{u}_l - \boldsymbol{\gamma}_{A0}||_2^2 q_n} + nq_nk_n^{-r}\right) \\
&\leq C\left(k_n\delta_{q_n}^{-1}q_nn^{1/2} + nq_nk_n^{-r}\right).
\end{aligned}
$$

Recall that ν_{ijl} is bounded, then applying Bernstein's inequality and using the assumed rates of λ, k_n and q_n it follows that

$$P\left[\left|\sum_{i=1}^{n}\nu_{ijl}\right| > n\lambda/(16J_n)\middle|\mathbf{z}_i\right] \leq \exp\left[-C\frac{n^2\lambda^2k_n^{-2}}{\left(k_n\delta_{q_n}^{-1}q_nn^{1/2}+nq_nk_n^{-r}\right)+n\lambda k_n^{-1}}\right] \leq \exp\left(-Cn\lambda k_n^{-1}\right).$$

The term $n\lambda k_n^{-1}$ dominates the denominator in the first inequality by combining Conditions 3 and 4, Lemma 11 and the assumption that $n^{-1/2}k_n^2\log(n) = o(\lambda)$. Note, the upper bound does not depend on $\mathbf{z}_i$ and taking expectations on both sides we get

$$P\left[\left|\sum_{i=1}^{n}\nu_{ijl}\right| > n\lambda/(2k_n)\right] \leq \exp\left(-Cn\lambda k_n^{-1}\right).$$

Therefore,

$$\begin{aligned} I_{nj1} &\leq CNk_n\exp\left(-Cn\lambda k_n^{-1}\right) = Cn^{4k_n^2\delta_{q_n}^{-2}q_n+1/(2r+1)}q_n^{1/(2r+1)}\exp\left(-Cn\lambda k_n^{-1}\right) \\ &\leq C\exp\left\{\left[4k_n^2\delta_{q_n}^{-2}q_n+1/(2r+1)\right]\log(n)+1/(2r+1)\log(q_n)-Cn\lambda k_n^{-1}\right\}. \end{aligned}$$

To evaluate I_{nj2}, note that $I[\epsilon_i(\tilde{\boldsymbol{\gamma}}_A)\leq 0] = I\left[\epsilon_i(\mathbf{u}_l)\leq \boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\tilde{\boldsymbol{\gamma}}_A-\mathbf{u}_l)\right]$. First, we will derive an upper bound for the sum in the probability statement. Since $I(x\leq a)$ is an increasing function of a, we have

$$\begin{aligned} &\sup_{||\tilde{\boldsymbol{\gamma}}_A-\mathbf{u}_l||_2\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}}\left|\left|\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\Big\{m_i\left[\epsilon_i(\tilde{\boldsymbol{\gamma}}_A),\epsilon_i(\mathbf{u}_l)\right]-E\{m_i\left[\epsilon_i(\tilde{\boldsymbol{\gamma}}_A),\epsilon_i(\mathbf{u}_l)\right]\mid\mathbf{z}_i\}\Big\}\right|\right|_1 \\ \leq\ & \sum_{i=1}^{n}||\boldsymbol{\pi}_j(z_{ij})||_1\left\{I\left[\epsilon_i\left(\mathbf{u}_l\right)\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2\right]-I\left[\epsilon_i\left(\mathbf{u}_l\right)\leq 0\right]\right. \\ &\left.-P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2\middle|\mathbf{z}_i\right]+P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq 0|\mathbf{z}_i\right]\right\} \\ &+\sum_{i=1}^{n}||\boldsymbol{\pi}_j(z_{ij})||_1\left\{P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2k_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}\middle|\mathbf{z}_i\right]\right. \\ &\left.-P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq -Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2\middle|\mathbf{z}_i\right]\right\}. \end{aligned}$$

First, the second sum will be examined. Using Condition 1, Taylor series expansion and that the elements of $\boldsymbol{\pi}_j(z_{ij})$ are bounded,

$$\begin{aligned} &\sum_{i=1}^{n}||\boldsymbol{\pi}_j(z_{ij})||_1\left\{P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2\middle|\mathbf{z}_i\right]\right. \\ &\left.-P\left[\epsilon_i\left(\mathbf{u}_l\right)\leq -Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2}||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2\middle|\mathbf{z}_i\right]\right\} \\ \leq\ & C\sum_{i=1}^{n}||\boldsymbol{\pi}_j(z_{ij})||_1\,||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2k_n\delta_{q_n}^{-1}q_n^{1/2}n^{-5/2} \leq Ck_n^{5/2}\delta_{q_n}^{-1}q_nn^{-3/2} = o(n\lambda). \end{aligned}$$

Define

$$\begin{aligned}\alpha_{ijl} &= ||\boldsymbol{\pi}_j(z_{ij})||_1 \Bigg\{ I\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2}\right] - I\left[\epsilon_i(\mathbf{u}_l) \le 0\right] \\ &- P\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2} \mid \mathbf{z}_i\right] + P\left[\epsilon_i(\mathbf{u}_l) \le 0 \mid \mathbf{z}_i\right] \Bigg\}.\end{aligned}$$

Then for n sufficiently large, $I_{nj2} \le \sum_{l=1}^{N} P\left(\sum_{i=1}^{n} \alpha_{ijl} \ge \frac{n\lambda}{32}\right)$ and again Bernstein's inequality will be used to provide an upper bound for this probability. To evaluate α_{ijl}, define

$$\begin{aligned}\omega_{ijls} &= |b_{j,s}(z_{ij})| \Bigg\{ I\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2}\right] - I\left[\epsilon_i(\mathbf{u}_l) \le 0\right] \\ &-P\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2} \mid \mathbf{z}_i\right] + P\left[\epsilon_i(\mathbf{u}_l) \le 0 \mid \mathbf{z}_i\right] \Bigg\},\end{aligned}$$

which are bounded, independent mean-zero random variables. Using that the elements of $||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2$ are bounded for all i, it follows that

$$\begin{aligned}\mathrm{Var}(\omega_{ijkl} \mid \mathbf{z}_i) &\le C\max_i \left| I\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2}\right] - I\left[\epsilon_i(\mathbf{u}_l) \le 0\right]\right| \\ &\quad \times E\left\{ \left| I\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2}\right] - I\left[\epsilon_i(\mathbf{u}_l) \le 0\right]\right| \Big| \mathbf{z}_i \right\} \\ &\le E\left\{ I\left[\epsilon_i(\mathbf{u}_l) \le C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2}\right] - I\left[\epsilon_i(\mathbf{u}_l) \le 0\right] \Big| \mathbf{z}_i \right\} \\ &= CF_i\left[\mathbf{W}(\mathbf{z}_i)_A^{\top}(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni} + C||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2} \Big| \mathbf{z}_i\right] \\ &\quad -CF_i\left[\mathbf{W}(\mathbf{z}_i)_A^{\top}(\mathbf{u}_l - \boldsymbol{\gamma}_{A0}) + u_{ni} \Big| \mathbf{z}_i\right] \\ &\le C\max_i ||\boldsymbol{\Pi}_A(\mathbf{z}_i)||_2 k_n n^{-5/2} \le C\sqrt{q_n} k_n^{3/2} n^{-5/2}.\end{aligned}$$

Notice,

$$\sum_{l=1}^{N} P\left(\sum_{i=1}^{n} \alpha_{ijl} \ge \frac{n\lambda}{32}\right) \le \sum_{l=1}^{N}\sum_{s=1}^{J_n} P\left(\left|\sum_{i=1}^{n} \omega_{ijkl}\right| \ge \frac{n\lambda}{32J_n}\right).$$

Applying Bernstein's inequality, for some positive constants C_1, C_2 and C_3,

$$\begin{aligned}&\sum_{l=1}^{N}\sum_{s=1}^{J_n} P\left(\left|\sum_{i=1}^{n} \omega_{ijkl}\right| \ge \frac{n\lambda}{32J_n}\right) \le CNk_n \exp\left(-\frac{C_1 n^2\lambda^2 k_n^{-2}}{C_2\sqrt{q_n}k_n^{3/2}n^{-3/2} + C_3\lambda n k_n^{-1}}\right) \\ \le\ & C\exp\left\{\left[4k_n^2\delta_{q_n}^{-2}q_n + 1/(2r+1)\right]\log(n) + 1/(2r+1)q_n - Cn\lambda k_n^{-1}\right\}.\end{aligned}$$

Note that $n\lambda k_n^{-1}$ dominates $\sqrt{q_n}k_n^{3/2}n^{-3/2}$ because

$$\frac{\sqrt{q_n}k_n^{3/2}n^{-3/2}}{n\lambda k_n^{-1}} = \frac{\sqrt{q_n}k_n^{5/2}}{n^2\lambda} = \frac{k_n}{n}\frac{\sqrt{q_n}k_n^2}{n\lambda}.$$

Where both fractions are $o(1)$ by assumed rates. Note that under Condition 7, $\lambda = o[n^{-(1-c_4)/2}]$ implies $\lambda = o(k_n^{-1})$. To complete the proof, notice there exist positive constants C_1, C_2, C_3 and C_4 such that for all n sufficiently large, the probability of interest in the lemma is bounded by

$$\begin{aligned}
&\sum_{j=q_n+1}^{p_n} (I_{nj1} + I_{nj2}) \\
\leq\ & C_1 \exp\left\{\log(p_n) + C_2\left[k_n^2 \delta_{q_n}^{-2} q_n + (2r+1)^{-1}\right]\log(n) + C_3/(2r+1)\log(q_n) - C_4 n\lambda k_n^{-1}\right\}.
\end{aligned}$$

This upper bound converges to zero under the assumptions of this lemma and using Lemma 11. ■

Lemma 16 *Assume the conditions of Theorem 6 hold, then*

$$P\left(\max_{q_n+1\leq j\leq p_n} \sup_{||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}||_2 \leq C k_n \delta_{q_n}^{-1} q_n^{1/2} n^{-1/2}} \left|\left| n^{-1}\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij}) \Big\{ P\left[Y_i - \boldsymbol{\Pi}_A(\mathbf{z}_i)^\top \boldsymbol{\gamma}_A \leq 0 \mid \mathbf{z}_i\right] \right.\right.\right.$$
$$\left.\left.\left. - P[Y_i - g_0(\mathbf{z}_i) \leq 0 \mid \mathbf{z}_i] \Big\} \right|\right|_1 > \lambda/8 \right) \to 0.$$

Proof Define $v_n = k_n \delta_{q_n}^{-1} q_n^{1/2} n^{-1/2}$, using the Cauchy-Schwarz inequality and Lemma 10

$$\begin{aligned}
&\max_{q_n+1\leq j\leq p_n} \sup_{||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}||_2 \leq C v_n} \left|\left| n^{-1}\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij}) \left\{ P\left[Y_i - \boldsymbol{\Pi}_A(\mathbf{z}_i)^\top \boldsymbol{\gamma}_A \leq 0 \mid \mathbf{z}_i\right] - P\left[Y_i - g_0(\mathbf{z}_i) \leq 0 \mid \mathbf{z}_i\right] \right\} \right|\right|_1 \\
\leq\ & \max_{q_n+1\leq j\leq p_n} \sup_{||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}|| \leq C v_n} \sum_{s=1}^{J_n} \sqrt{\frac{1}{n}\sum_{i=1}^n b_{j,s}^2(z_{ij})} \sqrt{\frac{1}{n}\sum_{i=1}^n \left\{F_i\left[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}) - u_{ni} \mid \mathbf{z}_i\right] - F_i(0 \mid \mathbf{z}_i)\right\}^2} \\
\leq\ & \max_{q_n+1\leq j\leq p_n} \sup_{||\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0}|| \leq C v_n} C\sum_{s=1}^{J_n} \sqrt{E\left[b_{j,s}^2(z_{ij})\right] + O_P(n^{-1/2})} \sqrt{\frac{2}{n}\sum_{i=1}^n \left[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top(\boldsymbol{\gamma}_A - \boldsymbol{\gamma}_{A0})\right]^2 + u_{ni}^2} \\
\leq\ & C\sum_{s=1}^{J_n} \sqrt{M_4 k_n^{-1} + O_P(n^{-1/2})} \sqrt{C(k_n^2 \delta_{q_n}^{-2} q_n^2 n^{-1} + q_n^2 k_n^{-2r})} \leq C\left(k_n^{3/2}\delta_{q_n}^{-1} q_n n^{-1/2} + q_n k_n^{1/2-r}\right)[1 + o_P(1)].
\end{aligned}$$

From the conditions on λ, k_n and q_n it can be derived that $k_n^{3/2}\delta_{q_n}^{-1} q_n n^{-1/2} + q_n k_n^{1/2-r} = o(\lambda)$, thus completing the proof. ■

The following lemma is an extension of Lemma 2.2 and 2.3 from Wang et al. (2012) which considered the linear model.

Lemma 17 *Assume the Conditions of Theorem 6 hold. Then for the oracle estimator, $\hat{\boldsymbol{\gamma}}$,with probability approaching one*

$$\min_{j\in\{1,\ldots,q_n\}} \left|\left|\hat{\boldsymbol{\gamma}}_j\right|\right|_1 \geq (a+1/2)\lambda, \tag{25}$$

$$\min_{j\in\{q_n+1,\ldots,p_n\}} \left|\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\psi_\tau[y_i-\boldsymbol{\Pi}(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}]\right|\right|_1 \leq \lambda/2. \tag{26}$$

Proof Proof of (25): Note that

$$\min_{j\in\{1,\ldots,q_n\}}||\hat{\boldsymbol{\gamma}}_j||_1 \geq \min_{j\in\{1,\ldots,q_n\}}||\boldsymbol{\gamma}_{0j}||_1 - ||\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_{A0}||_1. \tag{27}$$

By Lemmas 11 and 13 and Conditions 3, 4 and 7, $||\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_{A0}||_1 \leq \sqrt{J_nq_n+1}||\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_{A0}||_2 = O_P\left(k_n^{3/2}q_nn^{-1/2}\delta_{q_n}^{-1}\right) = o_P\left[n^{-(1-c_4)/2}\right]$. Condition 7 guarantees that there exists a positive constant c_5 such that $\min_{j\in\{1,\ldots,q_n\}}||\boldsymbol{\gamma}_{0j}||_1 \geq c_5n^{-(1-c_4)/2}$. Finally, $\lambda = o\left[n^{-(1-c_4)/2}\right]$ and therefore $P\left[\min_{j\in\{1,\ldots,q_n\}}||\hat{\boldsymbol{\gamma}}_j||_1 > (a+1/2)\lambda\right] \to 1$.

Proof of (26): Define $s_j(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\psi_\tau[y_i-\boldsymbol{\Pi}(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}]$ and $\mathcal{D} = \{i : Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A = 0\}$. For $j \in \{q_n+1,...,p_n\}$,

$$s_j(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I\left[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A \leq 0\right]-\tau\right\} - \frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{\pi}_j(z_{ij})[a_i^*+(1-\tau)],$$

where $a_i^* \in [\tau-1,\tau]$ with $i \in \mathcal{D}$ such that $s_j(\hat{\boldsymbol{\gamma}}) = \mathbf{0}_{J_n}$ for $j \in \{1,\ldots,q_n\}$ and

$$\frac{1}{n}\sum_{i=1}^{n}\left\{I\left[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A \leq 0\right]-\tau\right\} - \frac{1}{n}\sum_{i\in\mathcal{D}}[a_i^*+(1-\tau)] = 0.$$

From Section 2.2 of Koenker (2005) it follows that with probability one $|\mathcal{D}| \leq q_nJ_n+1$. Then by Conditions 2-4 and the assumptions about the rate of λ it follows that

$$\max_{q_n+1\leq j\leq p_n}\left|\left|n^{-1}\sum_{i\in\mathcal{D}}\boldsymbol{\pi}_j(z_{ij})[a_i^*+(1-\tau)]\right|\right|_1 = O_P\left(q_nk_nn^{-1}\right) = o_P(\lambda).$$

Thus, it is sufficient to show that

$$P\left(\max_{q_n+1\leq j\leq p_n}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A \leq 0]-\tau\right\}\right|\right|_1 > \lambda/2\right) \to 0.$$

Using Lemma 14 for the second inequality it follows that,

$$
\begin{aligned}
& P\left(\max_{q_n+1\le j\le p_n}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A\le 0]-\tau\right\}\right|\right|_1>\lambda/2\right)\\
\le\ & P\Big(\max_{q_n+1\le j\le p_n}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\hat{\boldsymbol{\gamma}}_A\le 0]-I[Y_i-g_0(\mathbf{z}_i)\le 0]\right\}\right|\right|_1>\lambda/4\Big)\\
& +P\left(\max_{q_n+1\le j\le p_n}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\left\{I[Y_i-g_0(\mathbf{z}_i)\le 0]-\tau\right\}\right|\right|_1>\lambda/4\right)\\
\le\ & P\left(\max_{q_n+1\le j\le p_n}\sup_{||\boldsymbol{\gamma}_A-\boldsymbol{\gamma}_{A0}||_2\le Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\Big\{I[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\boldsymbol{\gamma}_A\le 0]\right.\right.\\
& \left.\left.-I[Y_i-g_0(\mathbf{z}_i)\le 0]\Big\}\right|\right|_1>\lambda/4\right)+o_P(1)\\
\le\ & P\left(\max_{q_n+1\le j\le p_n}\sup_{||\boldsymbol{\gamma}_A-\boldsymbol{\gamma}_{A0}||_2\le Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\Big\{I\left[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\boldsymbol{\gamma}_A\le 0\right]\right.\right.\\
& \left.\left.-I[Y_i-g_0(\mathbf{z}_i)\le 0]-P[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\boldsymbol{\gamma}_A\le 0\mid\mathbf{z}_i]+P[Y_i-g_0(\mathbf{z}_i)\le 0\mid\mathbf{z}_i]\Big\}\right|\right|_1>\lambda/8\right)\\
& +P\left(\max_{q_n+1\le j\le p_n}\sup_{||\boldsymbol{\gamma}_A-\boldsymbol{\gamma}_{A0}||_2\le Ck_n\delta_{q_n}^{-1}q_n^{1/2}n^{-1/2}}\left|\left|n^{-1}\sum_{i=1}^{n}\boldsymbol{\pi}_j(z_{ij})\Big\{P\left[Y_i-\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top\boldsymbol{\gamma}_A\le 0\mid\mathbf{z}_i\right]\right.\right.\right.\\
& \left.\left.\left.-P[Y_i-g_0(\mathbf{z}_i)\le 0\mid\mathbf{z}_i]\Big\}\right|\right|_1>\lambda/8\right)+o_P(1).
\end{aligned}
$$

The two probability statements go to zero by Lemmas 15 and 16. This completes the proof. ■

8.3 Proof of Theorem 6

Proof Define the neighborhood $\mathcal{X}_\phi=\left\{\boldsymbol{\gamma}\in\mathbb{R}^{J_np_n+1}\middle|||\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}||_1<\phi<\lambda/2\right\}$. In this proof we show that for sufficiently large n there exists a ϕ such that $Q(\hat{\boldsymbol{\gamma}})\le Q(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma}\in\mathcal{X}_\phi$. Define

$$\mathcal{W}=\left\{\boldsymbol{\gamma}=(\gamma_0,\ldots,\gamma_{J_np_n})^\top\in\mathbb{R}^{J_np_n+1}\middle|\gamma_j=0\text{ for }j\in\{J_nq_n+1,\ldots,J_np_n\}\right\}$$

and $\mathcal{F}_\phi=\mathcal{W}\cap\mathcal{X}_\phi$. For any $\boldsymbol{\gamma}\in\mathcal{X}_\phi$ and for any $j\in\{1,\ldots,q_n\}$ it follows from Lemma 17 and the definition of $\mathcal{X}_\phi$ that with probability approaching one

$$||\boldsymbol{\gamma}_j||_1\ge||\hat{\boldsymbol{\gamma}}_j||_1-||\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j||_1\ge(a+1/2)\lambda-\lambda/2=a\lambda.$$

By Condition 6 and Lemma 17 it follows, with probability approaching one, that for any $\boldsymbol{\gamma}\in\mathcal{F}_\phi$ that $p_{\lambda,a}(||\hat{\boldsymbol{\gamma}}_j||_1)=p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1)$ for all $j\in\{1,\ldots,p_n\}$. By definition of the or-

acle estimator and $\mathcal{F}_\phi$ it follows that for any $\boldsymbol{\gamma} \in \mathcal{F}_\phi$ that $\frac{1}{n}\sum_{i=1}^n \rho_\tau\left[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \hat{\boldsymbol{\gamma}}\right] \leq \frac{1}{n}\sum_{i=1}^n \rho_\tau\left[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \boldsymbol{\gamma}\right]$. Therefore, for any $\boldsymbol{\gamma} \in \mathcal{F}_\phi$ it holds that $Q(\hat{\boldsymbol{\gamma}}) \leq Q(\boldsymbol{\gamma})$.

For any vector $\boldsymbol{\gamma} \in \mathcal{X}_\phi$ let $\tilde{\boldsymbol{\gamma}}$ represent the projection of $\boldsymbol{\gamma}$ into $\mathcal{F}_\phi$. For sufficiently large n, and thus sufficiently small λ, $Q(\hat{\boldsymbol{\gamma}}) \leq Q(\tilde{\boldsymbol{\gamma}})$ and thus the proof will be complete if it can be shown that $Q(\tilde{\boldsymbol{\gamma}}) \leq Q(\boldsymbol{\gamma})$. Let $\boldsymbol{\gamma}_A$ represent the first $q_n J_n + 1$ entries of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_N$ the remaining $J_n(p_n - q_n)$ entries such that $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_A^\top, \boldsymbol{\gamma}_N^\top\right)^\top$ and $\tilde{\boldsymbol{\gamma}} = \left[\boldsymbol{\gamma}_A^\top, \mathbf{0}_{J_n(p_n-q_n)}^\top\right]^\top$. Similarly define $\boldsymbol{\Pi}_N(\mathbf{z}_i)$ such that $\boldsymbol{\Pi}(\mathbf{z}_i) = \left[\boldsymbol{\Pi}_A(\mathbf{z}_i)^\top, \boldsymbol{\Pi}_N(\mathbf{z}_i)^\top\right]^\top$. By Knight's identity

$$\begin{aligned}
Q(\boldsymbol{\gamma}) - Q(\tilde{\boldsymbol{\gamma}}) &= \frac{1}{n}\sum_{i=1}^n \rho_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \boldsymbol{\gamma}] - \rho_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}] + \sum_{j=q_n+1}^{p_n} \left[p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1) - p_{\lambda,a}(0)\right] \\
&= -\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}(\mathbf{z}_i)^\top (\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}] \\
&\quad + \frac{1}{n}\sum_{i=1}^n \int_0^{\boldsymbol{\Pi}(\mathbf{z}_i)^\top(\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}})} I[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}} \leq s] - I[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}} \leq 0]\, ds \\
&\quad + \sum_{j=q_n+1}^{p_n} \left[p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1) - p_{\lambda,a}(0)\right].
\end{aligned}$$

As $\sum_{i=1}^n \int_0^{\boldsymbol{\Pi}(\mathbf{z}_i)^\top(\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}})} I[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}} \leq s] - I[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}} \leq 0] ds$ is non-negative for all i, it will be sufficient to show that

$$\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}(\mathbf{z}_i)^\top (\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}]\right| \leq \sum_{j=q_n+1}^{p_n} \left[p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1) - p_{\lambda,a}(0)\right].$$

Notice,

$$\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}(\mathbf{z}_i)^\top (\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}]\right| \leq \sum_{j=q_n+1}^{p_n} ||\boldsymbol{\gamma}_j||_1 \left|\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}]\right|\right|_1.$$

By the mean value theorem, for some $c_j^* \in (0, ||\boldsymbol{\gamma}_j||_1)$

$$p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1) - p_{\lambda,a}(0) = p'_\lambda(c_j^*)||\boldsymbol{\gamma}_j||_1.$$

By Lemma 17 and Condition 6 there exists a sufficiently small ϕ such that for all $j \in \{q_n + 1, \ldots, p_n\}$

$$\left|\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\pi}_j(z_{ij})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}]\right|\right|_1 \leq p'_{\lambda,a}(\phi).$$

Note that $c_j^* < \phi$, for all j, and therefore by the assumption that $p_{\lambda,a}(\cdot)$ is concave in $[0, \infty)$, from Condition 6, it follows that $p'_{\lambda,a}(\phi) \leq p'_{\lambda,a}(c_j^*)$ for all $j \in \{q_n + 1, \ldots, p_n\}$. Therefore,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}(\mathbf{z}_i)^\top (\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})\psi_\tau[y_i - \boldsymbol{\Pi}(\mathbf{z}_i)^\top \tilde{\boldsymbol{\gamma}}]\right| \leq \sum_{j=q_n+1}^{p_n} ||\boldsymbol{\gamma}_j|| p'_{\lambda,a}(\phi) \leq \sum_{j=q_n+1}^{p_n} p'_\lambda(c_j^*)||\boldsymbol{\gamma}_j||_1 \\
= &\sum_{j=q_n+1}^{p_n} \left[p_{\lambda,a}(||\boldsymbol{\gamma}_j||_1) - p_{\lambda,a}(0)\right].
\end{aligned}$$

■

8.4 Proof of Theorem 7

Proof Define the vector functions of

$$\begin{aligned}
\mathbf{Q}'(\boldsymbol{\gamma},\mathbf{a},\mathbf{v}) &= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)\{\tau - I[y_i \leq \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}]\} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)(1-\tau+a_i)I[y_i = \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}] + \mathbf{v},\\
\mathbf{r}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)\{I[y_i \leq \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}] - \tau\} - E\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)\{I[y_i \leq \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}] - \tau\}\right),\\
\tilde{\pi}(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\mathbf{a}_1,\mathbf{a}_2) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)(1-\tau+a_{2i})I[y_i = \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}_2] - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}(\mathbf{z}_i)(1-\tau+a_{1i})I[y_i = \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma}_1],
\end{aligned}$$

and $\tilde{\mathbf{r}}(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2) = \mathbf{r}(\boldsymbol{\gamma}_1) - \mathbf{r}(\boldsymbol{\gamma}_2)$. For $\mathbf{a}\in\mathbb{R}^n$ define the subgradient of $||\mathbf{a}||_1$ as

$$\partial||\mathbf{a}||_1 = \{\mathbf{b}\in\mathbb{R}^n | b_k = \text{sgn}(a_k) \text{ for } a_k \neq 0 \text{ and } b_k \in [-1,1] \text{ otherwise}\},$$

and define the sets

$$\begin{aligned}
\mathcal{V}(\boldsymbol{\gamma}) &= \{\mathbf{b} = (0,\mathbf{b}_1^\top,\ldots,\mathbf{b}_{p_n}^\top)^\top \in \mathbb{R}^{p_nJ_n+1} | \mathbf{b}_j = p'_\lambda(||\boldsymbol{\gamma}_j||)\mathbf{c}_j, \text{ where } \mathbf{c}_j \in \partial||\boldsymbol{\gamma}_j||_1, \text{ for all } j\in\{1,\ldots,p_n\}\},\\
\mathcal{A}(\boldsymbol{\gamma}) &= \{\mathbf{b} = (b_1,\ldots,b_n)^\top \in \mathbb{R}^n | b_i = 0 \text{ if } y_i \neq \boldsymbol{\Pi}(\mathbf{z}_i)^\top\boldsymbol{\gamma} \text{ and } b_i \in [-1,1] \text{ otherwise}\}.
\end{aligned}$$

By first order conditions if $\boldsymbol{\gamma}$ is a local minimizer of $Q(\boldsymbol{\gamma})$ then there exists $\mathbf{v}\in\mathcal{V}(\boldsymbol{\gamma})$ and $\mathbf{a}\in\mathcal{A}(\boldsymbol{\gamma})$ such that $Q'(\boldsymbol{\gamma},\mathbf{a},\mathbf{v}) = \mathbf{0}_{p_nJ_n+1}$. Thus, there exists $\bar{\mathbf{v}}\in\mathcal{V}(\bar{\boldsymbol{\gamma}})$ and $\bar{\mathbf{a}}\in\mathcal{A}(\bar{\boldsymbol{\gamma}})$ such that $Q'(\bar{\boldsymbol{\gamma}},\bar{\mathbf{a}},\bar{\mathbf{v}}) = \mathbf{0}_{p_nJ_n+1}$. Similarly, with probability approaching one, by Theorem 6, there exists $\hat{\mathbf{v}}\in\mathcal{V}(\hat{\boldsymbol{\gamma}})$ and $\hat{\mathbf{a}}\in\mathcal{A}(\hat{\boldsymbol{\gamma}})$ such that $Q'(\hat{\boldsymbol{\gamma}},\hat{\mathbf{a}},\hat{\mathbf{v}}) = \mathbf{0}_{p_nJ_n+1}$. By Condition 6 and that the first derivatives of differentiable concave functions are decreasing $|p'_\lambda(||\boldsymbol{\gamma}_j||)| \leq \lambda$ for all $j\in\{1,\ldots,p_n\}$ and thus $||\mathbf{v}||_\infty \leq \lambda$ for all $\mathbf{v}\in\mathcal{V}(\boldsymbol{\gamma})$ and any $\boldsymbol{\gamma}\in\mathbb{R}^{p_nJ_n+1}$.

For any vector $\mathbf{a}\in\mathbb{R}^{p_nJ_n+1}$ define $\mathbf{a}_\mathcal{E}\in\mathbb{R}^{w_nJ_n+1}$ as the sub-vector from the element of $\mathcal{E}$ similar to how we have defined $\mathbf{a}_A\in\mathbb{R}^{q_nJ_n+1}$. For some $\tilde{m}_i$ between $u_{ni}+\boldsymbol{\Pi}(\mathbf{z}_i)_\mathcal{E}^\top(\bar{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)_\mathcal{E}$ and $u_{ni}+\boldsymbol{\Pi}(\mathbf{z}_i)_\mathcal{E}^\top(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)_\mathcal{E}$ and using Conditions 1 and 8 there exists a positive constant C such that with probability approaching one

$$\begin{aligned}
0 &= \left|\left[Q'_\mathcal{E}(\bar{\boldsymbol{\gamma}},\bar{\mathbf{a}},\bar{\mathbf{v}}) - Q'_\mathcal{E}(\hat{\boldsymbol{\gamma}},\hat{\mathbf{a}},\hat{\mathbf{v}})\right]^\top \frac{(\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E})}{||\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E}||_2}\right|\\
&= \left|\frac{\left\{(\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E})^\top E\left[\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{m}_i)\boldsymbol{\Pi}_\mathcal{E}(\mathbf{z}_i)\boldsymbol{\Pi}_\mathcal{E}(\mathbf{z}_i)^\top\right] + \tilde{\mathbf{r}}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}}) + \bar{\mathbf{v}} - \hat{\mathbf{v}} - \tilde{\boldsymbol{\pi}}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}},\bar{\mathbf{a}},\hat{\mathbf{a}})\right\}_\mathcal{E}^\top(\bar{\boldsymbol{\gamma}}-\hat{\boldsymbol{\gamma}})_\mathcal{E}}{||\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E}||_2}\right|\\
&\geq C\delta_{w_n}^2 k_n^{-1}||\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E}||_2 - ||\tilde{\mathbf{r}}_\mathcal{E}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}})||_2 - 2\lambda\sqrt{w_nJ_n} - ||\tilde{\boldsymbol{\pi}}_\mathcal{E}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}},\bar{\mathbf{a}},\hat{\mathbf{a}})||_2.
\end{aligned}$$

Note for any $\boldsymbol{\gamma}$, $||\mathbf{r}_\mathcal{E}(\boldsymbol{\gamma})||_2 = O_P\left[\sqrt{\frac{w_n}{n}}\right]$ by Lemma 10 (4) and thus $||\tilde{\mathbf{r}}_\mathcal{E}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}})||_2 = O_P\left[\sqrt{\frac{w_n}{n}}\right]$. By Condition 8 $||\tilde{\boldsymbol{\pi}}_\mathcal{E}(\bar{\boldsymbol{\gamma}},\hat{\boldsymbol{\gamma}},\bar{\mathbf{a}},\hat{\mathbf{a}})||_2 = O_P(k_nw_nn^{-1}\sqrt{1+w_n})$. If with probability approaching one $||\bar{\boldsymbol{\gamma}}_\mathcal{E}-\hat{\boldsymbol{\gamma}}_\mathcal{E}||_2$ has a lower bound of order

$$\log(n)\delta_{w_n}^{-2}k_n\left(\sqrt{\frac{w_n}{n}} + \lambda\sqrt{w_nk_n} + k_nw_nn^{-1}\sqrt{1+w_n}\right),$$

then with probability approaching one $\left|[Q'_{\mathcal{E}}(\bar{\gamma}, \bar{\mathbf{a}}, \bar{\mathbf{v}}) - Q'_{\mathcal{E}}(\hat{\gamma}, \hat{\mathbf{a}}, \hat{\mathbf{v}})]^\top \frac{(\bar{\gamma}_{\mathcal{E}} - \hat{\gamma}_{\mathcal{E}})}{||\bar{\gamma}_{\mathcal{E}} - \hat{\gamma}_{\mathcal{E}}||_2}\right|$ has a positive lower bound, which is a contradiction. Therefore,

$$||\bar{\gamma}_{\mathcal{E}} - \hat{\gamma}_{\mathcal{E}}||_2 = O_P\left[\log(n)\delta_{w_n}^{-2} k_n \left(\sqrt{\frac{w_n}{n}} + \lambda\sqrt{w_n k_n} + k_n w_n n^{-1}\sqrt{1+w_n}\right)\right].$$

■

References

I. Barrodale and F.D.K. Roberts. Solution of an overdetermined system of equations in the ℓ_1 norm. *Communications of the ACM*, 17:319–320, 1974.

Alexandre Belloni and Victor Chernozhukov. L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

Claus Borggaard and Hans Henrik Thodberg. Optimal minimal neural interpretation of spectra. *Analytical Chemistry*, 64(5):545–551, 1992.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.

Patrick Breheny and Yaohui Zeng. grpreg: Regularization paths for regression models with grouped covariates 3.1-2, 2017. URL `https://cran.r-project.org/web/packages/grpreg/index.html`.

Zhao Chen, Jianqing Fan, and Runze Li. Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113(521):315–327, 2018a.

Zhao Chen, Jianqing Fan, and Runze Li. Supplemental material of error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113(521):315–327, 2018b.

Annie P. Chiang, John S. Beck, Hsan-Jan Yen, Marwan K. Tayeh, Todd E. Scheetz, Ruth E. Swiderski, Darryl Y. Nishimura, Terry A. Braun, Kwang-Youn A. Kim, Jian Huang, Khalil Elbedour, Rivka Carmi, Diane C. Slusarski, Thomas L. Casavant, Edwin M. Stone, and Val C. Sheffield. Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292, 2006.

Carl de Boor. Splines as linear combinations of b-splines. In *Approximation Theory II*, pages 1–47. Academic Press (New York), 1976.

Jan G. De Gooijer and Dawit Zerom. On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, 98:135–146, 2003.

Ronald A. Devore and George G. Lorentz. *Constructive Approximation*. Cambridge University Press, 2005.

Harold Gordon Eggleston. *Convexity, Cambridge Tracts in Mathematics and Mathematical Physics, No.47*. Cambridge University Press, 1958.

A. Essl, A. Ortner, R. Haas, and P. Hettegger. Machine learning analysis for a flexibility energy approach towards renewable energy integration with dynamic forecasting of electricity balancing power. In *2017 14th International Conference on the European Energy Market (EEM)*, pages 1–6, 2017.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Julian Faraway. faraway: Functions and datasets for books by julian faraway 1.0.7, 2016. URL `https://cran.r-project.org/web/packages/faraway/index.html`.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 04 2020. ISSN 0893-9454.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularized paths for genearlized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2008.

Y. Fujimoto, S. Murakami, N. Kaneko, H. Fuchikami, T. Hattori, and Y. Hayashi. Machine learning approach for graphical model-based analysis of energy-aware growth control in plant factories. *IEEE Access*, 7:32183–32196, 2019.

Yuwen Gu, Jun Fan, Lingchen Kong, Shiqian Ma, and Hui Zou. Admm for high-dimensional sparse penalized quantile regression. *Technometrics*, 60(3):319–331, 2018.

Xuming He and Peide Shi. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3:299–308, 1994.

Xuming He and Peide Shi. Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58(2):162–181, 1996.

Xuming He, Zhong-Yi Zhu, and Wing-Kam Fung. Estimation in semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590, 2002.

Xuming He, Lan Wang, and Hyokyoung Grace Hong. Quantile-adaptive model-free nonlinear feature screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41 (1):342–369, 2013.

Joel L. Horowitz and Sokbae Lee. Nonparametric estimation of additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249, 2005.

Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 2010.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499, 2012.

Jianhua Z. Huang. Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics*, 26(1):242–272, 1998a.

Jianhua Z. Huang. Functional anova models for generalized regression. *Journal of Multivariate Analysis*, 67:49–71, 1998b.

Ahmed M. Ibrahim, Hassan A.M. Hendawy, Wafaa S. Hassan, Abdalla Shalaby, and Manal S. ElMasry. Determination of terazosin in the presence of prazosin: Different state-of-the-art machine learning algorithms with uv spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 236:1386–1425, 2020.

Kengo Kato. Group lasso for high dimensional sparse quantile regression models. https://arxiv.org/pdf/1103.1458, March 2012.

Mi-Ok Kim. Quantile regression with varying coefficients. *The Annals of Statistics*, 35(1): 92–108, 2007.

Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057, 2012.

Keith Knight. Limiting distributions for l_1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770, 1998.

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

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Roger Koenker and Vasco D'Orey. A remark on algorithm as 229: Computing dual regression quantiles and regression rank score. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(2):410–414, 1994.

Roger W. Koenker and Vasco D'Orey. Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):383–393, 1987.

Eun Ryung Lee, Hohsuk Noh, and Byeong U. Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

Heng Lian, Xin Chen, and Jian-Yi Yang. Identification of partially linear structure in additive models with an application to gene expression prediction from sequences. *Biometrics*, 68(2):437–445, 2012.

Chen-Yen Lin, Howard Bondell, Hao Helen Zhang, and Hui Zou. Variable selection for nonparametric quantile regression via smoothing spline analysis of variance. *Stat*, 2: 255–268, 2013.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140, 2016.

Shaogao Lv, Huazhen Lin, Heng Lian, and Jian Huang. Oracle inequalities for sparse additive quantile regression in reproducing kernel hilbert space. *Ann. Statist.*, 46(2): 781–813, 04 2018.

Adam Maidman and Lan Wang. New semiparametric method for predicting high-cost patients. *Biometrics*, 74(3):1104–1111, 2018.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis fo m-estimators with decomposable regualrizers. *Statistical Science*, 27(4):538–557, 2012.

Marco Palma, Shahin Tavakoli, Julia Brettschneider, and Thomas E. Nichols. Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. *NeuroImage*, 219:1–14, 2020.

Bo Peng and Lan Wang. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24(3):676–694, 2015.

Ashley Petersen and Daniela Witten. Data-adaptive additive modeling. *Statistics in Medicine*, 38(4):583–600, 2019.

Ashley Petersen, Daniela Witten, and Noah Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025, 2016.

Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models with trend filtering. *Ann. Statist.*, 47(6):3032–3068, 12 2019.

Todd E. Scheetz, Kwang-Youn A. Kim, Ruth E. Swiderski, Alisdair R. Philp, Terry A. Braun, Kevin L. Knudtson, Anne M. Dorrance, Gerald F. DiBona, Jian Huang, Thomas L. Casavant, Val C. Sheffield, and Edwin M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

Larry L. Schumaker. *Spline Functions: Basic Theory*. Wiley, New York, 1981.

Ben Sherwood. Variable selection for additive partial linear quantile regression with missing covariates. *Journal of Multivariate Analysis*, 152:206–223, 2016.

Ben Sherwood and Adam Maidman. rqpen: Penalized quantile regression 2.2.2, 2020. URL `https://cran.r-project.org/package=rqPen`.

Ben Sherwood and Lan Wang. Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317, 2016.

Ben Sherwood, Aaron J. Molstad, and Sumanta Singha. Asymptotic properties of concave l1-norm group penalties. *Statistics and Probability Letters*, 157:108631, 2020.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

Charles J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986.

Ichiro Takeuchi, Quoc V. Le, Tim Sears, and Alexander J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3), 2007.

Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009a.

Huixia Judy Wang, Zhongyi Zhu, and Jianhui Zhou. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B):3841–3866, 2009b.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression of analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Li Wang, Lan Xue, Annie Qu, and Hua Liang. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, 42(2):592–624, 2014a.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6): 2164–2201, 2014b.

Ying Wei, Anneli Pere, Roger Koenker, and Xuming He. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25:1369–1382, 2006.

A.H. Welsh. On m-processes and m-estimation. *The Annals of Statistics*, 17(1):337–361, 1989.

Yicaho Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19 (2):801–817, 2009.

Lan Xue. Consistent variable selection in additive models. *Statistica Sinica*, 19:1281–1296, 2009.

Lan Xue and Lijian Yang. Additive coefficient modeling via polynomial spline. *Statistica Sinica*, 16(4):1423–1446, 2006.

Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

Liqun Yu, Nan Lin, and Lan Wang. A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 26(4):935–939, 2017.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Kaifeng Zhao and Heng Lian. Variable selection in additive quantile regression using non-concave penalty. *Statistics*, 50(6):1276–1289, 2016.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Qi Zheng, Limin Peng, and Xuming He. Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics*, 43(5):2225–2258, 2015.

S. Zhou, X. Shen, and D.A. Wolfe. Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782, 1998.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.