*Submitted to the Annals of Statistics*
arXiv: arXiv:0000.0000

# ORACLE INEQUALITIES FOR SPARSE ADDITIVE QUANTILE REGRESSION IN REPRODUCING KERNEL HILBERT SPACE*

By Shaogao Lv†, Huazhen Lin†, Heng Lian‡ and Jian Huang§

*Southwestern University of Finance and Economics, Chengdu, China†*
*University of New South Wales, Sydney, Australia 2052‡*
*University of Iowa, Iowa City, Iowa, USA§*

This paper considers the estimation of the sparse additive quantile regression (SAQR) in high-dimensional settings. Given the nonsmooth nature of the quantile loss function and the nonparametric complexities of the component function estimation, it is challenging to analyze the theoretical properties of ultrahigh-dimensional SAQR. We propose a regularized learning approach with a two-fold Lasso-type regularization in a reproducing kernel Hilbert space (RKHS) for SAQR. We establish nonasymptotic oracle inequalities for the excess risk of the proposed estimator without any coherent conditions. If additional assumptions including an extension of the restricted eigenvalue condition are satisfied, the proposed method enjoys sharp oracle rates without the light tail requirement. In particular, the proposed estimator achieves the minimax lower bounds established for sparse additive mean regression. As a by-product, we also establish the concentration inequality for estimating the population mean when the general Lipschitz loss is involved. The practical effectiveness of the new method is demonstrated by competitive numerical results.

**1. Introduction.** High dimensional sparse models arise in situations where many predictors are available and the regression function is well-approximated by a few relevant, yet unknown covariates. Sparsity assumption leads to more interpretable and stable models, less computational cost, and allows for model identifiability even when the number of covariates is much bigger than the sample size. Over the last decade, high dimensional sparse models have been extensively studied in conditional mean regression settings, which lead to several important regularized least squares approaches (see, for example, [47, 35, 38, 39] and the references therein).

*We sincerely thank Kenji Fukumizu and Taiji Suzuki for their helpful personal communications. The research of Lv and Lin is partially supported by the National Natural Science Foundation of China (Grant No. 11571282 and 11528102), the Ministry of Education of China (KLAS-130026507) and the Fundamental Research Funds for the Central Universities (Grants No. JBK120509, 14TD0046). Corresponding to: linhz@swufe.edu.cn.

*Keywords and phrases:* quantile regression, additive models, sparsity, regularization methods, reproducing kernel Hilbert space

However, high-dimensional data often display heterogeneity in practice, due to either heteroscedastic variance or other forms of non-location-scale covariate effects. This type of heterogeneity usually conceals a piece of important information that tends to be ignored by the mean of the conditional distribution. In addition, the active sets of covariates may be different for different quantile points of the conditional distribution. These can not be addressed by the mean regression models, but can be handled well by the quantile regression. The quantile regression method, first proposed by Koenker and Bassett [26], has been widely used in various disciplines, including finance, economics, medicine, and biology. We refer to [25] for a comprehensive introduction, and to [17] for a general overview of many interesting recent developments. In contrast to the conditional mean regression models, the quantile regression is able to capture heterogeneity [10] and different active sets of covariates for different quantile points, possesses certain robustness properties to outliers [25], has comparable computational efficiency [27], and requires relatively weak assumptions on the noise terms.

Several authors have investigated the problem of variable selection in linear quantile regression models in recent years [53, 6, 51]. [51] proposed a penalized quantile approach with a non-convex SCAD penalty [16] and established the corresponding model selection property under a ultra high-dimensional setting. Their method is based on a non-convex programme, but they did not give oracle rates of the estimators. Under some mild conditions, [53] established the oracle properties of the SCAD and adaptive-Lasso penalized quantile regressions. [6] studied the variable selection, estimation and prediction properties of the Lasso-type methods. All these works assume a linear or other parametric forms for the regression model.

In practice, there is often little prior information indicating that the effects of the predictors take a linear form or belong to any other finite-dimensional parametric family. Substantial improvements are possible by allowing a data-analytic transform of the predictors [20], which leads to quantile additive regression model. Nonetheless, the literature on sparse additive quantile regression (SAQR) is limited ([31, 23, 34]). Using an argument similar to that in [53], [31] studied several statistical properties of semiparametric quantile regression in the finite dimensional setting. In high dimensions, [23] proposed the group Lasso method by finite splines approximation to estimate additive components, and derived $\ell_2$-estimation errors of the estimator under quite different settings from what are given in this article. By contrast, our proposed approach belongs to an infinite dimensional Lasso-type scheme and enables us to estimate each additive component directly within RKHSs. Thus, no approximation error of the proposed estimator is present. Besides,

our method is quite flexible, since the use of RKHS includes linear functions, polynomial functions and Sobolev/Besov spaces as special cases. Moreover, another advantage of the RKHS is that only few tuning parameters are chosen, and in general it suffices to specify the kernel function. By contrast, the commonly-used spline methods in nonparametric estimation require to specify the number of basis functions and the sequence of knots. Also, the excess risk of the estimator has not been studied in [23], which is significantly different from estimation error. Analogous to our current paper, reference [34] proposed a group-type Lasso scheme for the SAQR within the RKHSs. However, they only established a low rate of convergence in terms of the excess risk of the estimator.

Despite the aforementioned developments, a rigorous analysis of penalization-based methods for SAQR in high-dimensional settings is still lacking. In particular, it is unclear whether the rates derived in each of the above papers are sharp under certain conditions, nor whether the rates are near minimax-optimal in some sense. Also, it is interesting to explore different conditions under which different-order oracle rates are established, so as to indicate the applicable scope of these $\ell_1$-type penalized methods. In practice, these $\ell_1$-type penalized methods often work better in terms of prediction performance than sparsity recovery. It is also interesting to justify this observation from theoretical aspects. Investigating such theoretical problems mentioned above is of significance, in view of the recent advances in understanding the performance of regularization methods in the mean regression and classification models [33, 38, 35].

In this paper we consider a regularized approach in SAQR with two Lasso-type regularization terms under the framework of RKHS. The proposed approach has two features. First, the proposed quantile approach with two Lasso-type regularization terms is defined directly on the infinite dimensional RKHSs and avoids approximation error. Second, as mentioned above, compared with the traditional spline-based methods or other dictionary-based tools, the RKHS has fewer tuning parameters to be determined and are quite flexible, including linear functions, various spline spaces and Sobelev/Besov class as special cases, see [37] for some discussions on RKHS and spline tools.

In theoretic development, however, it is challenging to simultaneously deal with the nonsmoothnees of the quantile loss function, the nonparametric nature of the component functions and the high-dimensionality of the predictor vector in a SAQR model. In particular, unlike any dictionary-based tool, the RHKS does not have an explicit form and is generally an infinite dimensional Hilbert space, which makes most of traditional analysis techniques not

applicable. We thus need to develop certain empirical process techniques to establish the oracle inequalities. We also note that [38, 29] developed some techniques to derive oracle inequalities or minimax rates based on the RHKS for sparse additive mean regression model. However, their techniques can not be used directly in our case since the quantile loss functions are nonsmooth and non-strongly-convex. To deal with the non-smoothness problem in the quantile loss, we apply the subgradient tools to guarantee that the error bound induced by the loss function is dominated by that of the penalty terms. In this case, a weighted empirical process associated with general random variables should be employed to deal with sequences of nonsymmetric random variables, involved in the theoretical analysis for SAQR. Note also that, our analysis does not require a global boundedness condition on the candidate space, which is required in [29]. Besides, our theoretical results allow for mutual dependence among covariates, by contrast, [38] assumes that the covariates are drawn from a product probability measure, which is equivalent to assuming that the covariates are mutually independent. This is an unrealistic assumption in the context of regression analysis, particularly in high-dimensional settings.

Our results are obtained under two different settings, so that we shed new light on high dimensional oracle properties for SAQR models. We first establish oracle inequalities of the excess risk under some quite mild conditions. In particular, we may dispense with the usual incoherence conditions required by $\ell_1$-based methods. If some additional suitable assumptions including an extension of the restricted eigenvalue condition are satisfied, the proposed method enjoys sharp oracle rates with proper choices of the regularization parameters. Furthermore, oracle rates for the excess risk and estimation error of the proposed estimator are shown to be adaptive to the sparsity of the model. These results show that, up to a logarithmic factor of the ambient dimension, the upper bound for the estimation error in Theorem 2 coincides with the minimax lower bound obtained recently in [38] for the least square regression with Gaussian noise. By contrast, our results allow for possibility of heavy-tailed noise term.

Even though our analysis is conducted only for SAQR, our proof techniques can be extended to general regularized approaches with Lipschitz loss functions, for example, the hinge loss for support vector machines studied recently in [55]. In particular, we establish the concentration inequality for estimating the population mean when the general Lipschitz loss is involved. In addition, our proofs also reveal some new features for the classical fixed-dimensional settings. We shall show that, the oracle rate (see Corollary 1 below) coincides with a rate of convergence for the fixed dimensional case

established in [19]. However, tackling proofs for our main results in Hilbert spaces are more challenging and are based on the structures of sparsity and complexity and advanced empirical processes, which is of a significant difference between analyzing high dimensional problems and analyzing fixed dimensional problems.

The paper is organized as follows. In Section 2 we introduce some basic notation and model assumptions. Section 3 includes the description of our method and statements of main theoretical results. The upper bounds on the convergence rate are provided under weak and strong conditions respectively, as well as model selection property is given. In Section 4, simulation results are implemented for numerically exploring the robustness and efficiency of the estimators. A real example analysis is displayed in Section 5. Proofs and some useful lemmas are located in Section 6.

**2. Model and Assumptions.** Let $(x_i, y_i), i = 1, ..., n$, be random vectors that are independent and identically distributed as $(x, y)$, where $y$ is a response variable and $x = (x_1, \ldots, x_d)'$ is a $d$-dimensional covariate vector. Suppose that $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is endowed with an underlying joint distribution $\mathbb{P}$. For theoretical analysis, we assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. The SAQR has the following form

$$y_i = \mu_\tau + \sum_{j \in S_\tau} f^*_{\tau,j}(x_{ij}) + \epsilon_{\tau,i}, \quad i = 1, ..., n, \tag{2.1}$$

where $\mu_\tau$ is an intercept, and $f^*_{\tau,j}$ is an unknown univariate component function in some specified RKHS. $S_\tau \subseteq \{1, 2, \ldots, d\}$ is some unknown subset with cardinality $|S_\tau| = s_\tau\, (s_\tau \ll d)$, and $\epsilon_{\tau,i}$ is the random error with $P(\epsilon_{\tau,i} \leq 0\,|x_i) = \tau$. Note that the distribution of $\epsilon_{\tau,i}$ is not specified and heterogeneous errors are not excluded since $\epsilon_{\tau,i}$ can depend on the predictors. In addition, the active set $S_\tau$ may vary with different quantile points $\tau$. This allows for different sparsity patterns for different quantile $\tau$. For simplicity, we will omit $\tau$ in the expressions wherever there is no confusion from the context. For identifiability, we assume $\mathbb{E}[f^*_j(x_j)] = 0$ for $j \in S$.

Given a compact subset $\mathrm{X} \subset \mathbb{R}$, let $K : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ be a bounded, symmetric, and positive semi-definite function. The RKHS denoted by $\mathcal{H}_K$ associated with the kernel $K$ is the completion of the linear span of functions $\{K_x := K(x, \cdot), x \in \mathrm{X}\}$ with the inner product given by $\langle K_x, K_y \rangle_K = K(x, y)$. For more detailed discussions on RKHS, we refer the reader to two standard references [1, 41]. The so-called reproducing property of $\mathcal{H}_K$ is critical in both theoretical analysis and computation, which states that $f(x) = \langle f, K(x, ) \rangle_K,\ x \in \mathrm{X}, \forall\, f \in \mathcal{H}_K$. This property implies that $\|f\|_\infty \leq$

$\kappa\|f\|_K$ for all $f \in \mathcal{H}_K$, where $\kappa = \sup_{x\in \mathrm{X}} \sqrt{K(x,x)}$. For notational simplicity, we assume that $\kappa = 1$ in the rest of the paper. For each $j = 1, \ldots, d$, the marginal distribution of $x_j$ is denoted by $\mathbb{Q}_j$ . Let $\mathcal{H}_j \subset L^2(\mathbb{Q}_j)$ be a RKHS of univariate functions on $\mathcal{X}_j$ with kernel $K_j$. We assume that $\mathbb{E}[f_j(x)] = \int_{\mathcal{X}_j} f_j(x) d\mathbb{Q}_j(x) = 0$, for any $f_j \in \mathcal{H}_j$. These centering constraints is to facilitate the identifying restrictions $\mathbb{E}[f_j^*(x_j)] = 0$, $j \in S$. In the model (2.1), we also assume that $|\mu_\tau| \leq 1$ and $\|f_j^*\|_{K_j} \leq 1$ for the ease of notation.

For any fixed $j = 1, \ldots, d$, denote by $\mathbb{B}_{\mathcal{H}_j}(1)$ the unit ball of $\mathcal{H}_j$, the hypothesis space we consider in this paper is defined by

$$\mathcal{F} := \Big\{ f = \sum_{j=1}^{d} f_j : f_j \in \mathbb{B}_{\mathcal{H}_j}(1), j = 1, ..., d \Big\},$$

which corresponds to the class of functions $f : \mathcal{X} \to \mathbb{R}$ that decompose as sums of univariate functions on each coordinates. Note that $\mathcal{F}$ is a subset of a RKHS [1], with the norm $\|f\|_{\mathcal{F}}^2 = \inf \Big\{ \sum_{j=1}^d \|f_j\|_{K_j}^2, \text{ with every form } f = \sum_{j=1}^d f_j \Big\}$, where $\|\cdot\|_{\mathcal{H}_j}$ denotes the norm on the univariate Hilbert space $\mathcal{H}_j$. Let $\mathbb{Q}$ be the marginal distribution on $\mathcal{X}$ induced by the joint distribution $\mathbb{P}$. We write the usual $L^q(\mathbb{Q})$ norm ($1 \leq q \leq 2$), $\|f\|_{L^q(\mathbb{Q})}^q = \int_{\mathcal{X}} |f|^q d\mathbb{Q}$. Analogously, we define the empirical $L^2(\mathbb{Q}_n)$-norm associated with the sample $\{x_i\}_{i=1}^n$, $\|f\|_{L^2(\mathbb{Q}_n)}^2 = \frac{1}{n}\sum_{i=1}^n f^2(x_i)$. For simplicity, we frequently use the notation $\|f\|_q = \|f\|_{L^q(\mathbb{Q})}$ and $\|f\|_n = \|f\|_{L^2(\mathbb{Q}_n)}$. For a univariate function $f_j \in \mathcal{H}_j$, we also use the shorthand $\|f_j\|_q = \|f_j\|_{L^q(\mathbb{Q}_j)}$ and $\|f_j\|_n = \|f_j\|_{L^2(\mathbb{Q}_{n,j})}$.

**3. Two-Fold Penalization and Main Results.** For ease of presentation, we state our results in the special case where the univariate Hilbert space $\mathcal{H}_j$, $j = 1, \ldots, d$ are all identical, still denoted by $\mathcal{H}_K$. For any function of the form $f = \sum_{j=1}^d f_j$, the $(L^2(\mathbb{Q}_n), 1)$ and $(\mathcal{H}_K, 1)$-norms are given by $\|f\|_{n,1} = \sum_{j=1}^d \|f_j\|_n$ and $\|f\|_{K,1} = \sum_{j=1}^d \|f_j\|_K$, respectively. We define the empirical and population risk respectively associated with the quantile loss as $\mathcal{E}_n(\mu, f) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\left(y_i - \mu - f(x_i)\right)$ and $\mathcal{E}(\mu, f) = \mathbb{E}[\rho_\tau\left(Y - \mu - f(X)\right)]$. The cost functional with a sparsity penalty $\|f\|_{n,1}$ and a smoothness penalty $\|f\|_{K,1}$ is

$$\mathcal{L}(\mu, f) = \mathcal{E}_n(\mu, f) + \lambda_1\|f\|_{n,1} + \lambda_2\|f\|_{K,1},$$

where $\rho_\tau(t) = t(\tau - I_{\{t\le 0\}})$ is the quantile loss function( also call the pinball loss). We consider the minimization problem

$$(\hat{\mu}, \hat{f}) = \arg \min_{\mu\in[-1,1], f} \mathcal{L}(\mu, f), \tag{3.1}$$

$$\text{subject to} \quad f = \sum_{j=1}^{d} f_j \text{ and } \|f_j\|_K \le 1 \text{ for } j = 1, \dots, d.$$

Here $(\lambda_1, \lambda_2)$ is a pair of positive regularization parameters, whose choice in our theory is to be specified as follows: $\lambda_2 = \lambda_1^2 = \zeta\gamma_n^2$, where $\gamma_n$ is given in (3.5) and $\zeta$ is given in Theorem 1 below. From the choices of $(\lambda_1, \lambda_2)$, there exists a close relationship between them in theory and $\lambda_1$ is larger than $\lambda_2$, since $\gamma_n \to 0$ as $n$ increases. The reasoning behind this is that it is more important to control sparsity than control smoothness in high dimensions. In practice, cross validation for choosing the regularization parameter seems to be the most reasonable way to go.

Though the formulation (3.1) is defined directly on infinite-dimensional Hilbert spaces, it can be shown that the solution to (3.1) is finite-dimensional by the reproducing property of RKHS, and has the form $\hat{f}(z_1, ..., z_d) = \sum_{i=1}^{n}\sum_{j=1}^{d} \hat{\alpha}_{ij}K(z_j, x_{ij})$. Let $\hat{\alpha}_j = (\hat{\alpha}_{1j}, ..., \hat{\alpha}_{nj})(j = 1, ..., d)$ and the empirical matrix $\mathbb{K}^j \in \mathbb{R}^{n\times n}$, with entries $\mathbb{K}^j_{i,\ell} = K(x_{ij}, x_{\ell j})$. The optimal coefficients $(\hat{\mu}, \hat{\alpha}_1, ..., \hat{\alpha}_d)$ are solution to the convex optimization problem

$$\min_{\mu\in[-1,1], \alpha_j\in\mathbb{R}^n, \alpha_j^T\mathbb{K}^j\alpha_j\le 1} \Big\{ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\Big(y_i - \mu - \Big[\sum_{j=1}^{d}(\mathbb{K}^j\alpha_j)\Big]_i\Big) + \tag{3.2}$$

$$\frac{\lambda_1}{\sqrt{n}}\sum_{j=1}^{d}\sqrt{\alpha_j^T(\mathbb{K}^j)^2\alpha_j} + \lambda_2\sum_{j=1}^{d}\sqrt{\alpha_j^T\mathbb{K}^j\alpha_j}\Big\}.$$

This problem can be transformed easily into a second-order cone program (SOCP) [23] and the existence of an optimal solution is guaranteed, however, the SOCP is computationally intensive for large-scale problems. To address this issue, we propose a majorization-minimization algorithm in Section 4, which will be repeatedly used to practically solve the optimization problem (3.2) in high dimensions.

*Remark 1.* As mentioned above, the computational tasks in RKHS's can be reduced to finite-dimensional optimization problems whose solutions are spanned naturally by the kernelized functions $\{K_{x_i}\}_{i=1}^n$, rather than some basis functions defined by specified knots.

*Remark 2.* In the literature, there are various combinations of sparsity and smoothness penalties with the quadratic loss for the mean regression

models. For example, the square norm $\sum_{j=1}^{d}\|f_j\|_K^2$ in [30] is only used to control functional smoothness and can not capture sparsity structure. [32, 28] combine the least squares loss and the penalty $\sum_{j=1}^{d}\|f_j\|_K$. [35] used the penalty $\sum_{j=1}^{d}\sqrt{\lambda_1\|f_j\|_n^2+\lambda_2\|f_j\|_K^2}$ for the least squares approach. The penalty we consider here belongs to two-fold Group Lasso scheme [54] and was also used in [29, 38] for mean regression problems. Compared with the other types of penalties, the two-fold penalization can lead to sharp rates of convergence. Moreover, the growth rates of the penalty parameters induced by this scheme are adaptive, without involving any prior information on the active set $S$.

*Remark 3.* Computationally, the nonsmooth quantile loss and either of two penalties can be approximated well by a series of quadratic functions, then the approximated $\mathcal{L}(\mu, f)$ can be solved by many methods, such as the proximal methods and block-coordinate descent algorithms [5, 48]. For more details, we refer the reader to a survey on sparsity-induced algorithms [2].

In this paper, we focus on the theoretical properties of the estimator, including the upper bounds on the excess risk and estimation error of $\hat{f}$, that is, $\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*)$ and $\sum_{j=1}^{d}\|\hat{f}_j - f_j^*\|_2$. These bounds are obtained based on the empirical process theory [4, 3]. We first introduce the following notations and assumptions.

A key property of RKHS is the *spectral theorem*, which says that $K$ associated to $\mathcal{H}_K$ admits the following eigenvalue decomposition [1]:

$$K(x, x') = \sum_{\ell \geq 1} b_\ell \psi_\ell(x)\psi_\ell(x'), \quad x,\, x' \in \mathrm{X}, \tag{3.3}$$

where $b_1 \geq b_2 \geq \ldots \geq 0$ are its eigenvalues and $\{\psi_\ell : \ell \geq 1\}$ are the corresponding eigenfunctions, consisting of an orthogonal basis in $L^2(\nu(X))$, where $\nu$ is some underlying measure on $X$. In this paper $\nu$ refers in particular to each $\mathbb{Q}_j$', $j = 1, \ldots, d$. Since the complexity of RKHS is determined by the rate of decay of eigenvalues $b_\ell$'s [42], we introduce the following spectral assumption for our analysis.

ASSUMPTION 1. **(Spectral Assumption)** *There exists some* $0 < \alpha < 1$ *and a universal constant* $C_0 > 0$*, such that*

$$b_\ell \leq C_0 \ell^{-1/\alpha}, \quad \forall\, \ell \geq 1. \tag{3.4}$$

Throughout this paper, $C$ with various subscripts denote universal constants independent of $n$, $d$ or $s$, and may be different from line to line. Note that, $\alpha < 1$ is a quite weak condition, due to the relation $\sum_{\ell=1} b_\ell =$

$\mathbb{E}[K(x,x)] \leq 1$. For example, if $\mathbb{Q}_j$ is the Lebesgue measure on $[0,1]$, it is known that $b_\ell \asymp \ell^{-2h}$ for the Sobolev class $\mathcal{H}_K = \mathcal{W}_2^h$ $(h > \frac{1}{2})$. Also, *spectral assumption* has a close quantitative relationship with entropy number of the RKHS [42].

We next impose the following technical assumption concerning the sup-norm of members in the RKHS.

ASSUMPTION 2. **(Sup-norm Assumption)**. *For some $0 < \alpha < 1$ given in Assumption 1, there exists some universal constant $C_1 > 0$, such that*

$$\|f\|_\infty \leq C_1 \|f\|_2^{1-\alpha} \|f\|_K^\alpha, \quad \forall\, f \in \mathcal{H}_K.$$

In general, this assumption is slightly stronger than Spectral Assumption, but when the eigenfunctions $\psi_\ell$'s defined as above are bounded uniformly, Sup-norm Assumption is equivalent to the spectral decay, as stated in Assumption 1. See the related details in [42], as well as Assumption 4 in [44].

3.1. *A Slow Oracle Inequality.* We first state an oracle inequality for the SAQR with a slow rate of convergence. This inequality is obtained under a weak assumption that only concerns the complexity of $\mathcal{H}_K$, thereby giving us an understanding of the statistical performance of the proposed estimator without any "correlatedness" conditions on covariates.

To this end, we introduce the notation of Rademacher complexity as another measurement of functional complexity. Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of variables from X distributed as some underlying measure, we define the Rademacher complexity in RKHS by $R_n(f) := \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)$, $f \in \mathcal{H}_K$, where $\{\sigma_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables that take the values $\{\pm 1\}$ with probability $1/2$.

For any given $A \geq 2$ and an arbitrary $\tilde{d} \geq d$ such that $\log \tilde{d} \geq 2 \log\log n$, we define the critical quantity $\gamma_n$ as

(3.5)
$$\gamma_n = \gamma_n(K) := \inf\Big\{\gamma \geq \sqrt{A \log \tilde{d}/n},\, \mathbb{E}\big[\sup_{\substack{\|f\|_K=1 \\ \|f\|_2 \leq t}} \big|R_n(f)\big|\big] \leq \gamma t + \gamma^2,\ \forall\, t \in (0,1]\Big\}.$$

The quantity $\gamma_n$, as the critical univariate rate, plays a crucial role in the error bounds on the excess risk in various empirical risk minimization problems in nonparametric context [4, 3]. In order to establish the relationship between $\alpha$ in Spectral Assumption and $\gamma_n$, we need the following conclusion. In the case of a univariate Hilbert space $\mathcal{H}_K$ with eigenvalues $\{b_\ell\}_{\ell=1}^\infty$, [36] has proved that for all $t^2 \geq 1/n$, we have $\mathbb{E}R_n(\{\|g\|_K = 1,\, \|g\|_2 \leq$

$t\}) \asymp \frac{1}{\sqrt{n}}\big[\sum_{\ell=1}^{\infty}\min\{t^2, b_\ell\}\big]^{1/2}$, which follows from Spectral Assumption (Assumption 1) that

$$\gamma_n \asymp \max\Big(\sqrt{\frac{A\log\tilde{d}}{n}}, \big(\frac{1}{n}\big)^{\frac{1}{2(1+\alpha)}}\Big). \tag{3.6}$$

Under very weak conditions, we are now in a position to provide a slow rate of the excess risk $\mathcal{E}(\hat{\mu},\hat{f}) - \mathcal{E}(\mu_\tau, f^*)$ that characterizes the prediction accuracy of the proposed estimator $(\hat{\mu},\hat{f})$.

THEOREM 1. *Let $(\hat{\mu},\hat{f})$ be the minimizer of the convex program (3.1) with regularization parameters $\lambda_1 = \sqrt{\zeta}\gamma_n$ and $\lambda_2 = \zeta\gamma_n^2$. Suppose that Assumption 1 and Assumption 2 hold, for any $\tilde{d} \geq d$ such that $\log d \leq \sqrt{n}$ and $\log\tilde{d} \geq 2\log\log n$, then for some constant $A \geq 2$, such that with probability at least $1 - 2\tilde{d}^{-A}$,*

$$\mathcal{E}(\hat{\mu},\hat{f}) - \mathcal{E}(\mu_\tau, f^*) \leq 9s\sqrt{\zeta}\gamma_n + 14\sqrt{\frac{3A\log\tilde{d}}{n}},$$

*where $t_0 = 2\log(2\sqrt{3}/\log 2) + A\log\tilde{d} + 2\log\tilde{d}$, $\eta(t) = \max(1, \sqrt{t}, t/\sqrt{n})$ and $\zeta = \max\{2C_*\eta(t_0)c, 4\}$.*

We provide the proof of Theorem 1 in the supplement material due to space limitation. This result in Theorem 1 means that the excess risk $\mathcal{E}(\hat{\mu},\hat{f}) - \mathcal{E}(\mu_\tau, f^*)$ is essentially controlled by $\gamma_n$, up to the logarithm factor $\log\tilde{d}$. To the best of our knowledge, this non-asymptotic prediction error bound is a new result for SAQR under such weak conditions. The condition is weaker than those in the existing literature [6, 50, 23]. In particular, the existing oracle inequalities depend heavily on some coherent conditions among the predictors such as "irrepresentable condition" [56], the "restricted isometry constants" condition [11] and "restricted eigenvalue" condition [7]. Furthermore, due to allowing $\sqrt{n} \geq \log d \geq 2\log\log n$ by taking $d = \tilde{d}$, the proposed estimator can handle a nonpolynomially growing dimension of predictors as high as $d = o(e^{\sqrt{n}})$. Meanwhile, the dimension of the true sparse model can be $s = o(n^{1/4})$ in the worst case ( $\alpha \to 1$) and $s = o(n^{1/2})$ in the ideal case ( $\alpha \to 0$). Finally, the commonly used sub-gaussian assumption for the noise term is not needed here. Thus heavy-tailed distributions are allowed in Theorem 1.

The results in this paper only requires that $\tilde{d} \to \infty$ as $n \to \infty$, and there is no similar constraint on the number of ambient dimension $d$. In other words, our results cover the fixed-dimensional case, which has been studied

extensively, see [19, 31, 14] and among others. Interestingly, the result of the excess risk via the proposed method shows that the Lasso-type method is more favorable in term of prediction performance than sparsity recovery, since the latter requires strong incoherent conditions to guarantee sparsity recovery consistency, see the related details in [56].

We notice that, for mean regression model with Gaussian noise term, [40] obtained the oracle rate in $\|\cdot\|_2^2$ with the order $(\log \tilde{d}/n)^{1/2}$ without any "correlatedness "assumption on covariates. Since there is no approximation error in our setting and $\alpha \to 0$ for the linear case, the oracle rate derived in Theorem 1 is the same as that in Theorem 4.1 in [40]. However, there is no light-tail assumption on the noise error for obtaining our result in Theorem 1. Under similar setting to that adopted in this paper, [34] derived the same rate of the excess risk as ours in Theorem 1, nevertheless, the result in [34] rules out the classical fixed dimensional case.

3.2. *A Fast Oracle Inequality.* Before providing sharp convergence rates in terms of the excess risk and estimation error, we show that the proposed estimator $(\hat{\mu}, \hat{f})$ belongs to a small restricted subset of $\mathcal{F}_{S,\mu}$, defined by

$$\mathcal{F}_{S,\mu} := \Big\{(\mu, f),\ (\mu, f) \in [-1,1] \times \mathcal{F} \text{ such that } \sum_{j=1}^{d} \|f_j - f_j^*\|_K \le 4 \sum_{j \in S} \|f_j - f_j^*\|_K$$
$$\text{or } \sum_{j=1}^{d} \|f_j - f_j^*\|_n \le 4 \sum_{j \in S} \|f_j - f_j^*\|_n + C_3|\mu - \mu_\tau|\Big\}. \tag{3.7}$$

This is a subset of $[-1,1] \times \mathcal{F}$ whose elements satisfy the inequality in (3.7). It will be shown in Lemma 2 below that, under mild conditions and appropriate constraints for regularization parameters, the estimator $(\hat{\mu}, \hat{f})$ belongs to $\mathcal{F}_{S,\mu}$ with high probability. Thus, it suffices to conduct our refined analysis over the restricted subset $\mathcal{F}_{S,\mu}$.

Since the proposed method (3.1) has the form of nonsmooth loss plus $L_1$ penalization, our proofs have some essential differences from those with a quadratic loss. Particularly, to establish strong oracle rates, we need the following self-calibration inequality, which shows that convergence in a weak form implies strong convergence in norms under certain conditions.

ASSUMPTION 3. **(Self-Calibration Assumption)** *There exist universal constants $c_1 > 0$ and $q \in [1,2]$, such that $\mathcal{E}(\mu, f) - \mathcal{E}(\mu_\tau, f^*) \ge c_1 \|\mu + f - (\mu_\tau + f^*)\|_q^2$ for all $(\mu, f) \in \mathcal{F}_{S,\mu}$.*

Assumption 3, as an identifiability condition, characterizes the concentration of the conditional distribution of $y$ given $x$ near $\mu_\tau + f^*(x)$. To satisfy

Assumption 3 for $q \in [1,2)$, a sufficient condition concerning the underlying distribution is provided by Theorem 2.7 in [43]. To further clarify this sufficient condition, we state a simple version of Theorem 2.7 in [43] in the supplementary material. In the specific case of $q = 2$, Assumption 3 under finite dimensional setting has been verified under the so-called RNI condition imposed in [6, 23]. Considering that the RNI condition can not be verified in the general cases under the infinite dimensional setting, we introduce a relaxed RNI condition under which Assumption 3 still holds in our additive RKHSs. Moreover, our relaxed RNI condition can be verified when each kernel is appropriately smooth and Assumption 4 below holds. See the related discussions in the last part of our supplementary file.

Let $\beta_q(S)$ be defined as follows

$$\beta_q(S) := \sup\Big\{\beta : 0 \le \beta \le \frac{\|(\mu+f)-(\mu_\tau+f^*)\|_q}{\sum_{j\in S\cup\{0\}}\|f_j - f_j^*\|_2},\ (\mu,f)\in\mathcal{F}_{S,\mu}\Big\},$$

where we denote $f_0 := \mu$ and $f_0^* := \mu_\tau$ for unified representation of notation.

ASSUMPTION 4. **(Incoherence Assumption)** *$\beta_q(S)$ is strictly bounded from zero below, that is, $0 < \beta_q(S)$.*

Assumption 4, as a coherence condition in infinite spaces, is analogous to Condition D.4 of [6] and Condition C5 of [23], respectively, for the quantile linear regression models, as well as Assumption C in [50]. In fact, it is frequently used in the theory of sparse recovery with $q = 2$ as a standard coherent condition, including the so called "restricted isometry constants" [11] and "restricted eigenvalue" [7], as well as in [44, 29] for learning kernels in an infinite dimensional framework. In the supplementary material, we give a lower bound of a simple version of $\beta_q(S)$, which consists of two interpretable geometric quantities. Our result can be viewed as a generalization of Proposition 1 in [29] with $q = 2$.

*Remark 4.* By Assumptions 3-4, we can check that $\sqrt{\mathcal{E}(\mu,f)-\mathcal{E}(\mu_\tau,f^*)} \ge c_1\beta_q(S)\sum_{j\in S\cup\{0\}}\|f_j - f_j^*\|_2$ for all $(\mu,f)\in\mathcal{F}_{S,\mu}$, which is sufficient for most of our main results except Theorem 3. Here the separate introduction of Assumptions 3-4 is to make the exposition more transparent and interpretable, and tie them to those existing familiar conditions in the literature [43, 44, 29].

*Remark 5.* If we consider all candidate estimators within bounded domain, that is, $\|\mu+f-(\mu_\tau+f^*)\|_\infty$ is upper bounded by some constant $C_b$, already used in [29, 46] in high dimensions, then we have $\|\mu+f-(\mu_\tau+f^*)\|_q^2 \le C_b^{2(q-1)/q}\|\mu+f-(\mu_\tau+f^*)\|_1^{2/q}$ for any $q\in[1,2]$. On the other hand, this boundedness conclusion also yields that $\|\mu+f-(\mu_\tau+f^*)\|_2^2 \le C_b^{2-q}\|\mu+f-$

$(\mu_\tau + f^*)\|_q^q$, which means that, for Assumption 4, it also suffices to consider the well-understood case $q = 2$.

The following result states sharp upper bounds on the excess risk and estimation error of the estimator (3.1), based on $n$ i.i.d. samples $(x_i, y_i)_{i=1}^n$ from the observation model (2.1).

THEOREM 2. *Let $(\hat{\mu}, \hat{f})$ be the minimizer of the convex program (3.1) with regularization parameters $\lambda_1 = \sqrt{\zeta}\gamma_n$ and $\lambda_2 = \zeta\gamma_n^2$. Suppose that Assumptions 1-4 hold. If $\log d \le \sqrt{n}$, then for some constant $A \ge 2$ with probability at least $1 - 3\tilde{d}^{-A}$, we have*

$$\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*) \le \Big(16c\sqrt{\zeta} + \frac{64c^2\zeta}{c_1\beta_q^2(S)}\Big)s\gamma_n^2 + 156c_1^{-1}\frac{t_0}{n}, \quad \textit{and}$$

$$\sum_{j=1}^d \|\hat{f}_j - f_j^*\|_2 \le 32c^2\Big(1 + \frac{4c\sqrt{\zeta}}{c_1\beta_q^2(S)}\Big)s\gamma_n + 28cc_1^{-1/2}\big(3\sqrt{2} + 14\sqrt{3}c_1^{-1/2}\big)s\sqrt{\frac{t_0}{n}},$$

*where $\zeta$, $t_0$ and $\tilde{d}$ are defined as those in Theorem 1, $c_1$ was given in Assumption 3 and constant $c$ is given Lemma 4 below.*

We provide the proof of Theorem 2 in Section 6. Up to the logarithmic factor $\log\tilde{d}$, Theorem 2 shows that the convergence rate of the excess risk is of the order $\max\big(\frac{sA\log d}{n}, s\big(\frac{1}{n}\big)^{1/(1+\alpha)}\big)$ by taking $\tilde{d} = d$, hence allows for non-polynomial dimension growing as $n$ increases. This rate also reveals the degree of the influence of the functional complexity, the effective dimension $s$ and the number of dimensions $d$ on the prediction accuracy, to be exponential, linear and logarithmic, respectively. The functional complexity hence has the strongest influence, whereas the number of dimension $d$ has the weakest influence on the prediction accuracy. Similar conclusions still hold true for the estimation error $\sum_{j=1}^d \|\hat{f}_j - f_j^*\|_2$.

When $q = 2$ in Assumption 3 is satisfied, the rate of the excess risk in Theorem 2 implies the same rate of the estimation error, which is also the lower bound established for the additive mean regression in [38], up to some logarithmic factors. Without requirement of light-heavy tail, we achieve the same oracle rates for estimating additive components as that for the additive mean regression with Gaussian noise. Using finite splines approximation to additive components, a group Lasso approach for estimating each additive components of the SAQR is proposed in [23]. In the case of Sobolev class and $q = 2$, the oracle rates of the estimation error derived in [23] is the same as ours in Theorem 2, up to the logarithmic factor. However, their settings are quite different from our study, for example, they imposed a stronger

condition than Assumption 3 with $q = 2$; see their lemma A.2. In the case of $q = 2$, the excess risk of the estimator is similar to those based on quadratic losses. In contrast, our Assumption 3 allows for the possibility of $q < 2$, under which the estimation error derived in Theorem 2 is still optimal as long as the sparsity parameter $s$ is fixed. This result indicates that, the quadratic property of the loss function is not essential to the optimal rates. In addition the method in [23] is a standard group Lasso scheme, while our proposed method is a two-fold Lasso scheme within infinite-dimensional Hilbert spaces.

For fixed $d$ and $s$, the rate of $\sum_{j=1}^{d} \|\hat{f}_j - f_j^*\|_2$ simplifies to the order $(1/n)^{1/(1+\alpha)}$. This is the optimal rate in the classical literature of statistical learning for finitely many predictors (see Theorem 3 of [44]). Under the framework of RKHS, reference [14] gave the same learning rates of regularized kernel based methods for additive but fixed dimensional models. Nonetheless, their rates require a variance-expectation bound, different from our sparse $\ell_1$-regularized method.

Theorem 2 is a weak form of the oracle inequality. To derive explicit rates with the $\|\cdot\|_2$-norm for any $1 \le q \le 2$, we need further analysis which gives the following conclusions. Denote $\widehat{S} := \{j : \|\hat{f}_j\|_2 \neq 0\}$, following our oracle rate, we can state the model selection properties of (3.1).

THEOREM 3. *Suppose all the conditions in Theorem 2 hold. Then with the same probability as stated in Theorem 2, we have*

$$\Big(\sum_{j\in S} \|\hat{f}_j - f_j^*\|_2\Big)^2 \le r_0\Big(\frac{A(s+3)\log \tilde{d}}{n} + s\big(\frac{1}{n}\big)^{\frac{1}{1+\alpha}}\Big),$$

*where* $r_0 := \frac{1}{c_1\beta_q(S)} \max\Big(16c\sqrt{\zeta} + \frac{64c^2\zeta}{c_1\beta_q^2(S)}, 156c_1^{-1}\Big)$. *In addition, if the value of the minimal true function is separated well from zero, to be precise,* $\min_{j\in S} \|f_j^*\|_2 > \sqrt{r_0}\sqrt{\frac{A(s+3)\log \tilde{d}}{n} + s\big(\frac{1}{n}\big)^{\frac{1}{1+\alpha}}}$, *we have* $\widehat{S} \supseteq S$ *with the same probability as stated in Theorem 2.*

The proof of Theorem 3 is provided in Section 6 as well. These results parallel those of [6] for linear sparse quantile models. Although we can also obtain a similar conclusion for model selection in Theorem 2, the rate in Theorem 3 relative to the index $S$ is sharper than that in Theorem 2 since $s$ is replaced by $\sqrt{s}$. We notice that the same rate was derived in the linear model [6], however, we do not require any additional conditions on the conditional density and the covariates. The second result of Theorem 3 implies that with high probability, the estimator selects a sup-set of the active functions,

provided that the active functions have enough signal. On the other hand, Theorems 2 and 3 show that the oracle rates are closely related to the structure and complexity of the additive components. In order to correctly identify the active set $S$, much stronger signal is needed when the additive component is more complex.

At the end of this section, we consider the optimal rate within the minimax framework given in [38]. For the additive mean regression with Gaussian noise, [38] has shown that, with a probability at least $1/2$, the lower bounds of $\|\hat{f}_n - f^*\|_2^2$ is of the order $\left\{\frac{s\log(d/s)}{n} + s(\frac{1}{n})^{1/(1+\alpha)}\right\}$ for any measurable estimator $\hat{f}_n$. Surprisingly, up to the sparse parameter $s$, our convergence rates of Theorem 2 with $q = 2$ can attain this lower bound under weaker assumptions, particularly, we do not require the Gaussian tails for the noise term. An interesting and fundamental question is the lower bound of the SAQR in terms of the excess risk and estimation error, which will be our future work.

We finally present a corollary for the Sobolev kernel classes. When the design distribution is uniform on $[-1, 1]$, it is known from [1] that the spectral decay of eigenvalues in the Sobolev space $W^h$ with smoothness $h > 1/2$ has the following asymptotic behavior $b_\ell \asymp \ell^{-2h}$, which implies $\gamma_n \asymp n^{-\frac{h}{2h+1}}$. Then the following corollary can be derived immediately from Theorem 2.

COROLLARY 1. *Suppose all the conditions in Theorem 2 hold. Consider an univariate Sobolev class $W^h[-1, 1]$ with some smoothness $h > 1/2$. For some constant $A \geq 2$, such that the kernel estimator defined in (3.1) with $\lambda_1 = \sqrt{\zeta}\gamma_n$ and $\lambda_2 = \zeta\gamma_n^2$ satisfies*

$$\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*) \leq \widetilde{C} s \max\left(\frac{A\log\tilde{d}}{n}, n^{-\frac{2h}{2h+1}}\right), \quad \textit{and}$$

$$\sum_{j=1}^{d} \|\hat{f}_j - f_j^*\|_2 \leq \widetilde{C} s \max\left(\sqrt{\frac{A\log\tilde{d}}{n}}, n^{-\frac{h}{2h+1}}\right),$$

*with probability at least $1 - 3\tilde{d}^{-A}$, where $\widetilde{C}$ is some universal constant.*

Note that up to constant factors, the achievable rate from Corollary 1 is the same as that in Theorem 3.3 in [19] for the fix dimensional case.

## 4. Numerical Results.

4.1. *Implementation details.* We use the Majorization-Minimization (M-M) algorithm to solve (3.2), which is a general technique for solving complicated optimization problems (see [22] for a nice review). Here we briefly describe the method. More details on MM algorithms can be found in [21].

First, the loss function $\rho_\tau(u)$ is approximated by its perturbation for some small $\epsilon > 0$, $\rho_\tau^\epsilon(u) := \rho_\tau(u) - \frac{\epsilon}{2}\ln(\epsilon + |u|)$. The function $\tilde{\rho}(u|u^k) = \frac{1}{4}\left[\frac{u^2}{\epsilon+|u^k|} + (4\tau-2)u + c\right]$ can be shown to majorize $\rho_\tau^\epsilon(u)$ at $u^k$ (which simply means that $\tilde{\rho}(u|u^k) \geq \rho_\tau^\epsilon(u)$ for all $u$ and $\tilde{\rho}(u^k|u^k) = \rho_\tau^\epsilon(u^k)$) for an appropriately chosen constant $c$. Without the penalty, at iteration $k+1$, the MM algorithm works by minimizing the majorizer $\frac{1}{n}\sum_i \tilde{\rho}_\tau(u_i|u_i^k)$ with respect to $\alpha_j, j = 1, \ldots, d$, where $u_i = y_i - [\sum_{j=1}^d (\mathbb{K}^j\alpha_j)]_i$, and $u_i^k = y_i - [\sum_{j=1}^d (\mathbb{K}^j\alpha_j^k)]_i$ is the residual at iteration $k$. The minimizer is the new estimate $\alpha_j^{k+1}, j = 1, \ldots, d$.

With the two penalties, the implementation is only slightly more complicated. For example, given the current estimate $\alpha_j^k$, the first penalty can be approximated as $\sqrt{\alpha_j^T(\mathbb{K}^j)^2\alpha_j} \approx \sqrt{\alpha_j^{kT}(\mathbb{K}^j)^2\alpha_j^k} + \frac{1}{2\sqrt{\alpha_j^T(\mathbb{K}^j)^2\alpha_j}}(\alpha_j^T(\mathbb{K}^j)^2\alpha_j - \alpha_j^{kT}(\mathbb{K}^j)^2\alpha_j^k)$. Note that this is similar to the majorizer for the loss function when $\tau = 0.5$, and is just the same as local quadratic approximation advocated in [16].

After these approximations, the loss function in each iteration becomes a quadratic function and the minimization problem can be solved in closed form. However, with the number of parameters $nd$, directly solving the minimization problem is time-consuming. Thus we adopt the group coordinate descent approach that updates $\alpha_j$ for each $j$ in turn with others fixed. Group coordinate descent has been used previously in [52, 9] without quadratic approximation. However, since there are two penalty terms here, it seems necessary to approximate the penalty functions first. In our implementation, we set $\epsilon = 10^{-3}$. Besides, if $\|\alpha_j\|$ falls below a small number ($10^{-4}$ in our implementation), we take it to be zero.

For the kernel, we use the kernel for the Sobelev space of order 3 (the space containing functions whose 3rd derivative is in $L_2$) given by $K(s,t) = 1 + st + 0.25(st)^2 + 0.25(\min(s,t)^3\max(s,t)^2/3 - \min(s,t)^4\max(s,t)/6 + \min(s,t)^5/30)$. We use five-fold cross-validation to select $\lambda_n$ and we set $\rho_n = \lambda_n^2$ as suggested by our theory.

4.2. *Simulations.* We conducted Monte Carlo studies for the following i.i.d. and heteroscedastic error model

$$y_i = 0.5 + \sum_{j=1}^{p} g_j(x_{ij}) + 0.5(1 + g(x_{i3}))e_i \tag{4.1}$$

with $g_1(x) = -6(x(1-x) - 1/6)$, $g_2(x) = \sin(2\pi x)/(2 - \sin(2\pi x)) - 0.1547$ and $g(x) = 6((x(1-x) - 1/6)$. Thus in our generating model, the number of nonzero nonparametric components is $s = 2$ at $\tau = 0.5$, while $s = 3$ for other values of $\tau$. More specifically, the conditional quantile function is $0.5 + 0.5F_e^{-1}(\tau) + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3})$ with $f_1 = g_1, f_2 = g_2, f_3 = 0.5F_e^{-1}(\tau)g$, where $F_e(.)$ denotes the distribution function of the mean zero error $e_i$. Several simulation scenarios are considered. For sample size, we set $n = 100$ or 200, for $\tau$ we consider $\tau = 0.5$ and 0.75, for $e_i$ we consider a standard normal distribution and a student-t distribution with degrees of freedom 3. To generate the covariates, we first let $x_{ij}$ be marginally standard normal with correlations given by $Cov(x_{ij_1}, x_{ij_2}) = 0.3^{|j_1 - j_2|}$, and then apply the cumulative distribution function of the standard normal distribution to transform $x_{ij}$ to be marginally uniform on $[0, 1]$. For all scenarios, 200 datasets are generated.

We generate data from (4.1) and fit our RKHS model to the data, as well as the group lasso estimator of [23] and a linear quantile model with lasso penalty for comparison, at both $\tau = 0.5$ and $\tau = 0.75$. We also fit a RKHS mean regression model [38] to the same data. For the spline estimator of [23], we use cubic splines with the number of internal knots together with the tuning parameter for the penalty selected by five-fold cross-validation. When performing cross-validation, the prediction error is defined based on the loss function used in the corresponding estimator. More specifically, for mean regression, the prediction error is $\sum_i (\hat{y}_i - y_i)^2$ while for quantile regression the prediction error is $\sum_i \rho_\tau(\hat{y}_i - y_i)$ where $\hat{y}_i$ is the fitted response value and $y_i$ is the true response value in the hold-out data.

First we compare the different estimators at $\tau = 0.5$, as well as the RKHS mean regression estimator. For functions $f_j, j = 1, 2, 3$ (note $f_3$ is a zero function at $\tau = 0.5$), we compute the mean squared error: $mse_j = \int (\hat{f}_j(x) - f_j(x))^2 dx$, with integral approximated by the Riemannian sum over a grid of 500 points, and the total mean squared error $mse = \sum_{j=1}^{p} mse_j$. Furthermore, we define the following two prediction errors: $pred_{LS} = \sum_{i=1}^{500} (\hat{y}_i - y_i)^2/500$ and $pred_{LAD} = \sum_{i=1}^{500} |\hat{y}_i - y_i|/500$, calculated on independently generated 500 observations. We also present the estimated number of nonzero components (#NZ) and the estimated number of nonzero components that are truly nonzero (#NZC). The results are summarized in Tables

1 and 2, for the two error distributions respectively. We see that for normal error, least squares regression is similar or better than RKHS median estimator. While for student's t error, median regression is better probably due to its robustness. The error decreases when sample size is increased from 100 to 200. The variable selection results also look reasonable in all additive models. The nonlinear models performs better than linear quantile model as expected, and RKHS estimator is often, though not always, better than the spline estimator. Linear models often miss some important variables as expected.

TABLE 1

*Simulation results for $\tau = 0.5$ with normal error using RKHS additive quantile regression (RKHS), spline additive quantile regression (SPLINE), least squares additive regression (LSQ), and linear quantile regression (LIN). The numbers in the brackets are the standard errors computed on the same 200 datasets.*

| $(n, p)$ | method | $mse_1$ | $mse_2$ | $mse_3$ | $mse$ | $pred_{LS}$ | $pred_{LAD}$ | #NZ | #NZC |
|---|---|---|---|---|---|---|---|---|---|
| $(100, 100)$ | RKHS | 0.0145 | 0.0172 | 0.0068 | 0.0861 | 0.1497 | 0.1509 | 3.67 | 2 |
| | | (0.0084) | (0.0085) | (0.0048) | (0.0208) | (0.0252) | (0.0120) | (1.08) | (0) |
| | SPLINE | 0.0170 | 0.0214 | 0.0053 | 0.0962 | 0.1517 | 0.1534 | 3.54 | 2 |
| | | (0.0186) | (0.0099) | (0.0068) | (0.0310) | (0.0370) | (0.0186) | (1.43) | (0) |
| | LSQ | 0.0182 | 0.0201 | 0.0064 | 0.0927 | 0.1476 | 0.1519 | 3.33 | 2 |
| | | (0.0131) | (0.0082) | (0.0088) | (0.0282) | (0.0270) | (0.0142) | (1.26) | (0) |
| | LIN | 0.1889 | 0.1057 | 0.0061 | 0.3172 | 0.4033 | 0.2571 | 1.79 | 1.15 |
| | | (0.0073) | (0.0150) | (0.0118) | (0.0351) | (0.0397) | (0.0128) | (1.07) | (0.30) |
| $(100, 200)$ | RKHS | 0.0351 | 0.0460 | 0.0065 | 0.1712 | 0.2322 | 0.1893 | 3.96 | 2 |
| | | (0.0250) | (0.0337) | (0.0054) | (0.0548) | (0.0528) | (0.0207) | (1.84) | (0) |
| | SPLINE | 0.0474 | 0.0509 | 0.0065 | 0.2002 | 0.2735 | 0.1901 | 6.93 | 2 |
| | | (0.0177) | (0.0247) | (0.0065) | (0.0873) | (0.0954) | (0.0363) | (2.66) | (0) |
| | LSQ | 0.0327 | 0.0480 | 0.0037 | 0.1579 | 0.2149 | 0.1838 | 3.80 | 2 |
| | | (0.0212) | (0.0346) | (0.0025) | (0.0452) | (0.0450) | (0.0205) | (1.54) | (0) |
| | LIN | 0.1973 | 0.1320 | 0.0010 | 0.3790 | 0.4501 | 0.2700 | 2.33 | 0.92 |
| | | (0.0345) | (0.0462) | (0.0032) | (0.0996) | (0.1044) | (0.0276) | (2.94) | (0.64) |
| $(200, 100)$ | RKHS | 0.0150 | 0.0124 | 0.0017 | 0.0314 | 0.1072 | 0.1281 | 2.27 | 2 |
| | | (0.0103) | (0.0081) | (0.0015) | (0.0160) | (0.0192) | (0.0121) | (0.55) | (0) |
| | SPLINE | 0.0155 | 0.0133 | 0.0026 | 0.0393 | 0.1096 | 0.1298 | 2.25 | 2 |
| | | (0.0036) | (0.0039) | (0.0018) | (0.0133) | (0.0172) | (0.0107) | (0.44) | (0) |
| | LSQ | 0.0158 | 0.0135 | 0.0019 | 0.0380 | 0.1057 | 0.1272 | 2.18 | 2 |
| | | (0.0098) | (0.0073) | (0.0017) | (0.0135) | (0.0177) | (0.0111) | (0.36) | (0) |
| | LIN | 0.1872 | 0.1149 | 0.0006 | 0.3160 | 0.4034 | 0.2569 | 1.52 | 0.92 |
| | | (0.0027) | (0.0290) | (0.0013) | (0.0311) | (0.0351) | (0.0109) | (0.94) | (0.39) |
| $(200, 200)$ | RKHS | 0.0188 | 0.0154 | 0.0025 | 0.0874 | 0.1502 | 0.1523 | 2.61 | 2 |
| | | (0.0046) | (0.0046) | (0.0021) | (0.0206) | (0.0262) | (0.0134) | (0.68) | (0) |
| | SPLINE | 0.0193 | 0.0171 | 0.0019 | 0.0880 | 0.1537 | 0.1597 | 2.53 | 2 |
| | | (0.0125) | (0.0082) | (0.0020) | (0.0243) | (0.0325) | (0.0175) | (0.68) | (0) |
| | LSQ | 0.0193 | 0.0177 | 0.0017 | 0.0854 | 0.1496 | 0.1476 | 2.46 | 2 |
| | | (0.0124) | (0.0073) | (0.0023) | (0.0217) | (0.0284) | (0.0157) | (0.59) | (0) |
| | LIN | 0.1899 | 0.1052 | 0.0007 | 0.3063 | 0.3878 | 0.2518 | 1.20 | 1.13 |
| | | (0.0125) | (0.0191) | (0.0021) | (0.0217) | (0.0317) | (0.0093) | (0.61) | (0.32) |

Next we compare different quantile estimators at $\tau = 0.75$, with results reported in Tables 3 and 4. We present the $mse_j, j = 1, 2, 3$ and total mse

$mse = \sum_{j=1}^{p} mse_j$, the prediction error based on quantile loss defined by $pred_{QL} = \sum_{i=1}^{500} \rho_{0.75}(\hat{y}_i - y_i)/500$, as well as the variable selection results (#NZ and #NZC). As expected, the performance is better with normal noise than with t noise, and the errors are smaller when $n = 200$. Again, the RKHS estimator is generally better than the spline estimator, and the linear estimator is the worst. Finally, we note that least squares estimator never picks up $x_{i3}$ in our simulations, which is a significant variable at $\tau = 0.75$. The additive quantile methods can identify all three significant variables most of the time as seen from the tables.

TABLE 2

*Simulation results for $\tau = 0.5$ with $t(3)$ error using RKHS additive quantile regression (RKHS), spline additive quantile regression (SPLINE), least squares additive regression (LSQ), and linear quantile regression (LIN). The numbers in the brackets are the standard errors computed on the same 200 datasets.*

| $(n, p)$ | method | $mse_1$ | $mse_2$ | $mse_3$ | $mse$ | $pred_{LS}$ | $pred_{LAD}$ | #NZ | #NZC |
|---|---|---|---|---|---|---|---|---|---|
| $(100, 100)$ | RKHS | 0.0180 | 0.0232 | 0.0066 | 0.0982 | 0.2864 | 0.1930 | 3.82 | 2 |
| | | (0.01059) | (0.0129) | (0.0079) | (0.0347) | (0.0792) | (0.0264) | (1.32) | (0) |
| | SPLINE | 0.0204 | 0.0233 | 0.0073 | 0.1002 | 0.3051 | 0.2087 | 4.77 | 2 |
| | | (0.0204) | (0.0375) | (0.0085) | (0.0579) | (0.0651) | (0.0229) | (1.65) | (0) |
| | LSQ | 0.0285 | 0.0584 | 0.0071 | 0.1444 | 0.3631 | 0.2587 | 3.98 | 2 |
| | | (0.0198) | (0.0374) | (0.0088) | (0.0442) | (0.0546) | (0.0183) | (1.57) | (0) |
| | LIN | 0.1870 | 0.1266 | 0.0040 | 0.3445 | 0.5325 | 0.2856 | 1.88 | 0.85 |
| | | (0.0022) | (0.0314) | (0.0076) | (0.0449) | (0.0765) | (0.0183) | (1.42) | (0.36) |
| $(100, 200)$ | RKHS | 0.0579 | 0.0625 | 0.0083 | 0.2638 | 0.4304 | 0.2489 | 7.14 | 1.94 |
| | | (0.0510) | (0.0452) | (0.0062) | (0.085) | (0.1130) | (0.0317) | (3.16) | (0.30) |
| | SPLINE | 0.0578 | 0.0631 | 0.0083 | 0.2933 | 0.4723 | 0.2633 | 7.99 | 1.85 |
| | | (0.0431) | (0.0474) | (0.0108) | (0.1093) | (0.1131) | (0.0343) | (4.01) | (0.61) |
| | LSQ | 0.0726 | 0.0817 | 0.0045 | 0.3602 | 0.4986 | 0.2796 | 7.52 | 1.75 |
| | | (0.0626) | (0.0451) | (0.0065) | (0.0750) | (0.1162) | (0.0292) | (3.59) | (0.55) |
| | LIN | 0.1864 | 0.1313 | 0.0003 | 0.3423 | 0.5292 | 0.2853 | 1.75 | 0.83 |
| | | (0.0003) | (0.0434) | (0.0013) | (0.0483) | (0.0683) | (0.0166) | (1.71) | (0.41) |
| $(200, 100)$ | RKHS | 0.0156 | 0.0155 | 0.0027 | 0.0464 | 0.2089 | 0.1618 | 2.35 | 2 |
| | | (0.0040) | (0.0049) | (0.0024) | (0.0158) | (0.0349) | (0.0114) | (0.65) | (0) |
| | SPLINE | 0.0149 | 0.0150 | 0.0041 | 0.0616 | 0.2365 | 0.1756 | 3.27 | 2 |
| | | (0.0157) | (0.0115) | (0.0044) | (0.0391) | (0.0393) | (0.0173) | (1.25) | (0) |
| | LSQ | 0.0284 | 0.0333 | 0.0034 | 0.1015 | 0.2701 | 0.2019 | 2.94 | 2 |
| | | (0.0188) | (0.0301) | (0.0042) | (0.0454) | (0.0504) | (0.0219) | (0.94) | (0) |
| | LIN | 0.1884 | 0.1105 | 0.0004 | 0.3153 | 0.4900 | 0.2737 | 1.41 | 1.13 |
| | | (0.0043) | (0.0208) | (0.0009) | (0.0305) | (0.0511) | (0.0125) | (0.94) | (0.32) |
| $(200, 200)$ | RKHS | 0.0254 | 0.0275 | 0.0035 | 0.1197 | 0.2918 | 0.1947 | 3.01 | 1.93 |
| | | (0.0385) | (0.0426) | (0.0034) | (0.0695) | (0.0838) | (0.0305) | (1.35) | (0.44) |
| | SPLINE | 0.0257 | 0.0289 | 0.0039 | 0.1201 | 0.3341 | 0.2116 | 5.13 | 1.93 |
| | | (0.0396) | (0.0349) | (0.0035) | (0.0809) | (0.1054) | (0.0356) | (3.11) | (0.22) |
| | LSQ | 0.0295 | 0.0482 | 0.0035 | 0.1619 | 0.3433 | 0.2476 | 4.20 | 2 |
| | | (0.0417) | (0.0369) | (0.0047) | (0.0735) | (0.0890) | (0.0315) | (1.73) | (0) |
| | LIN | 0.1865 | 0.1083 | 0.0003 | 0.3034 | 0.4836 | 0.2703 | 1.23 | 1 |
| | | (0.0007) | (0.0152) | (0.0012) | (0.0147) | (0.0533) | (0.0111) | (0.44) | (0) |

**5. Breast Cancer Data.** We use the breast cancer data from The Cancer Genome Atlas project [45] to illustrate the proposed method. We focus on

the gene expression data obtained using Agilent mRNA expression microarrays. In this dataset, expression measurements of 17814 genes, including BRCA1, from 536 patients are available at http://cancergenome.nih.gov/. BRCA1 is the first gene identified that increases the risk of early onset breast cancer. Because BRCA1 is likely to interact with many other genes, including tumor suppressors and regulators of the cell division cycle, it is of interest to find genes with expression levels related to that of BRCA1. These genes may be functionally related to BRCA1 and are useful candidates for further studies.

TABLE 3
*Simulation results for $\tau = 0.75$ with normal error. The numbers in the brackets are the standard errors computed on the same 200 datasets.*

| $(n, p)$ | method | $mse_1$ | $mse_2$ | $mse_3$ | $mse$ | $pred_{QL}$ | #NZ | #NZC |
|---|---|---|---|---|---|---|---|---|
| $(100, 100)$ | RKHS | 0.0178 | 0.0233 | 0.0831 | 0.1692 | 0.1374 | 4.57 | 3 |
| | | (0.0073) | (0.0076) | (0.0226) | (0.0316) | (0.0217) | (1.27) | (0) |
| | SPLINE | 0.0179 | 0.0284 | 0.0929 | 0.1802 | 0.1343 | 4.09 | 3 |
| | | (0.0189) | (0.0417) | (0.0247) | (0.0528) | (0.0203) | (1.19) | (0) |
| | LIN | 0.1807 | 0.1210 | 0.0904 | 0.4135 | 0.2132 | 1.64 | 0.92 |
| | | (0.0024) | (0.0404) | (0.0124) | (0.0449) | (0.0136) | (1.27) | (0.44) |
| $(100, 200)$ | RKHS | 0.0390 | 0.0620 | 0.0932 | 0.2632 | 0.1965 | 4.82 | 3 |
| | | (0.0248) | (0.0431) | (0.0223) | (0.0522) | (0.0192) | (1.60) | (0) |
| | SPLINE | 0.0462 | 0.0640 | 0.0823 | 0.3175 | 0.1973 | 5.25 | 2.94 |
| | | (0.0378) | (0.0366) | (0.0308) | (0.1022) | (0.0444) | (3.10) | (0.30) |
| | LIN | 0.1885 | 0.1305 | 0.0861 | 0.4319 | 0.2146 | 1.63 | 0.93 |
| | | (0.0067) | (0.0346) | (0.0019) | (0.0387) | (0.0117) | (1.30) | (0.44) |
| $(200, 100)$ | RKHS | 0.0159 | 0.0125 | 0.0739 | 0.1161 | 0.1096 | 3.49 | 3 |
| | | (0.0045) | (0.0037) | (0.0184) | (0.0243) | (0.0168) | (0.68) | (0) |
| | SPLINE | 0.0177 | 0.0139 | 0.0896 | 0.1346 | 0.1140 | 3.26 | 3 |
| | | (0.0096) | (0.0055) | (0.0124) | (0.0189) | (0.0115) | (0.52) | (0) |
| | LIN | 0.1876 | 0.1236 | 0.0855 | 0.4069 | 0.2117 | 1.16 | 0.80 |
| | | (0.0030) | (0.0286) | (0.0004) | (0.0278) | (0.0119) | (0.98) | (0.52) |
| $(200, 200)$ | RKHS | 0.0226 | 0.0166 | 0.0830 | 0.1640 | 0.1353 | 3.94 | 3 |
| | | (0.0042) | (0.0059) | (0.0135) | (0.0310) | (0.0156) | (1.05) | (0) |
| | SPLINE | 0.0223 | 0.0187 | 0.0872 | 0.1664 | 0.1374 | 3.28 | 3 |
| | | (0.0123) | (0.0103) | (0.0147) | (0.0290) | (0.0139) | (0.44) | (0) |
| | LIN | 0.1886 | 0.1048 | 0.0855 | 0.3966 | 0.2105 | 1.64 | 1.13 |
| | | (0.0064) | (0.0152) | (0.0006) | (0.0287) | (0.0087) | (0.98) | (0.30) |

We only include genes with sufficient expression levels and variations across the subjects in the analysis. So we first do an initial screen according to the following requirements: the coefficient of variation is greater than 1; and, the standard deviation is greater than 0.6. Finally, we do a marginal screening using nonparametric regression spline estimation as proposed in [18] to select the top 200 genes and feed these into our proposed model.

We use the same Sobolev kernel of order 3 as used in our simulations,

and consider three quantile levels $\tau = 0.25, 0.5, 0.75$, and report the selected genes in Table 5. Linear lasso methods are also applied to the same 200 genes. It is seen that additive model and linear model selected entirely differently set of genes. These identified genes may be worth further investigation in genomic studies. Different genes are identified for different quantile levels suggesting heterogeneity of gene effects at different quantile levels.

TABLE 4

*Simulation results for $\tau = 0.75$ with $t(3)$ error. The numbers in the brackets are the standard errors computed on the same 200 datasets.*

| $(n, p)$ | method | $mse_1$ | $mse_2$ | $mse_3$ | $mse$ | $pred_{QL}$ | #NZ | #NZC |
|---|---|---|---|---|---|---|---|---|
| $(100, 100)$ | RKHS | 0.0373 | 0.0384 | 0.1013 | 0.2300 | 0.1752 | 5.16 | 2.91 |
| | | (0.0419) | (0.0472) | (0.0156) | (0.0784) | (0.0266) | (1.63) | (0.44) |
| | SPLINE | 0.0366 | 0.0465 | 0.1042 | 0.2413 | 0.1783 | 4.78 | 2.95 |
| | | (0.0391) | (0.0329) | (0.0238) | (0.0706) | (0.0230) | (1.74) | (0.22) |
| | LIN | 0.190 | 0.1398 | 0.1133 | 0.4759 | 0.2418 | 1.92 | 0.76 |
| | | (0.0064) | (0.0450) | (0.0135) | (0.0409) | (0.0137) | (1.66) | (0.71) |
| $(100, 200)$ | RKHS | 0.0781 | 0.0792 | 0.1096 | 0.3654 | 0.2167 | 6.41 | 2.73 |
| | | (0.0680) | (0.0490) | (0.0113) | (0.0803) | (0.0299) | (4.31) | (0.57) |
| | SPLINE | 0.0828 | 0.0842 | 0.1021 | 0.3805 | 0.2231 | 6.44 | 2.45 |
| | | (0.0642) | (0.0731) | (0.0188) | (0.1302) | (0.0386) | (5.12) | (0.88) |
| | LIN | 0.1878 | 0.1228 | 0.1135 | 0.4422 | 0.2723 | 1.37 | 0.73 |
| | | (0.0030) | (0.0374) | (0.0170) | (0.0400) | (0.0115) | (0.87) | (0.44) |
| $(200, 100)$ | RKHS | 0.0190 | 0.0232 | 0.0879 | 0.1439 | 0.1391 | 4.18 | 3 |
| | | (0.0072) | (0.0092) | (0.0286) | (0.0369) | (0.0121) | (1.34) | (0) |
| | SPLINE | 0.0228 | 0.0242 | 0.1052 | 0.1897 | 0.1451 | 4.04 | 3 |
| | | (0.0145) | (0.0141) | (0.0295) | (0.0454) | (0.0143) | (1.12) | (0) |
| | LIN | 0.1879 | 0.1084 | 0.1127 | 0.4325 | 0.2283 | 1.76 | 1.07 |
| | | (0.0031) | (0.0206) | (0.0064) | (0.0364) | (0.0111) | (1.20) | (0.32) |
| $(200, 200)$ | RKHS | 0.0310 | 0.0300 | 0.0951 | 0.2147 | 0.1688 | 4.32 | 2.90 |
| | | (0.0378) | (0.0427) | (0.0232) | (0.0760) | (0.0253) | (1.78) | (0.44) |
| | SPLINE | 0.0378 | 0.0448 | 0.1062 | 0.2401 | 0.1692 | 4.48 | 3 |
| | | (0.0403) | (0.0379) | (0.0189) | (0.0642) | (0.0249) | (1.23) | (0) |
| | LIN | 0.1881 | 0.1113 | 0.1105 | 0.4339 | 0.2301 | 1.87 | 1.03 |
| | | (0.0035) | (0.0205) | (0.0056) | (0.0239) | (0.0156) | (0.87) | (0.32) |

The two additive methods (RKHS and SPLINE) selected similar sets of genes. The numbers in the bracket after the gene names indicate the ranking of the selected genes in the marginal screening stage, which suggests that the genes selected by the additive model are usually top ranking genes in the screening stage, which is as expected since the screening stage used a nonparametric model and thus linear models may fail to select these genes. The list of selected genes clearly indicate the existence of heterogeneity in the dataset, with only one gene, C17orf53, selected in all three quantile levels. After the genes are selected, we also re-fit a RKHS additive regression model using a ridge smoothing penalty with the results shown in Figures 1-3.

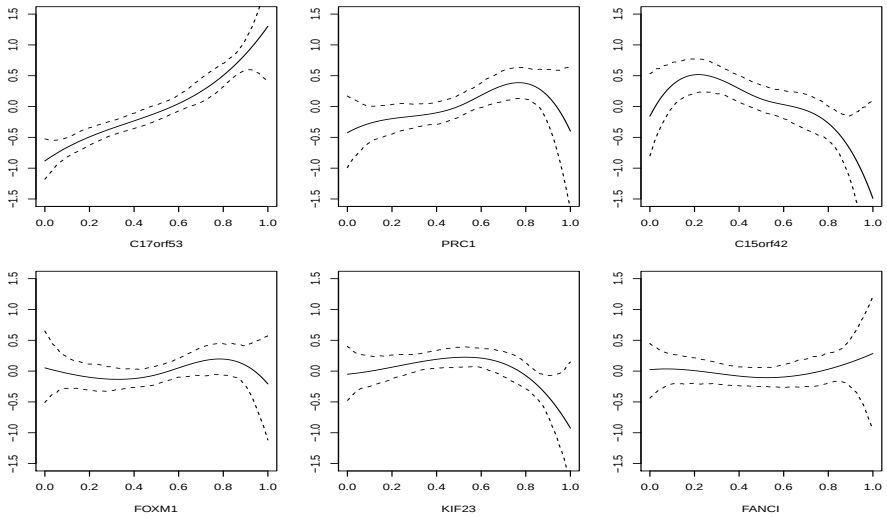


FIG 1. *Estimated curves for the selected genes when* $\tau = 0.25$*, with* 95% *pointwise confidence interval.*

The 95% pointwise confidence interval is obtained by calculating the 0.025 and 0.975 quantiles from estimates based on 500 bootstrap samples. The fitted curves suggest the existence of nonlinear effect of some of the genes selected. We note that such intervals are certainly exploratory in nature, without theoretical guarantees. Inferences for penalized estimators based on bootstrap is a challenging problem (see some recent works such as [12, 13] for parametric models) and we are not aware of any well-developed statistical approach for high-dimensional penalized semiparametric models.

Finally, we compare the cross-validation errors for different methods by randomly partitioning the data into training data and test data, with 400 observations used in training. Average cross-validation error based on 100 random partitions are reported in Table 6. Here the cross-validation error is the average quantile loss between the fitted response and the true response $\sum_i \rho_\tau(\hat{y}_i - y_i)/n$, over hold-out data. For $\tau = 0.5$, we also computed the least squares loss $\sum_i (\hat{y}_i - y_i)^2/n$. We see again that the performance of the two quantile additive models are similar, both better than the linear quantile models. The median additive models is only slightly better than least squares regression. However, the least squares regression can only target the mean response, instead of the tails of the distribution.

**6. Main Proofs.** The main idea of the proof is to first define an event with high probability and then analyze the behavior of the regularized es-

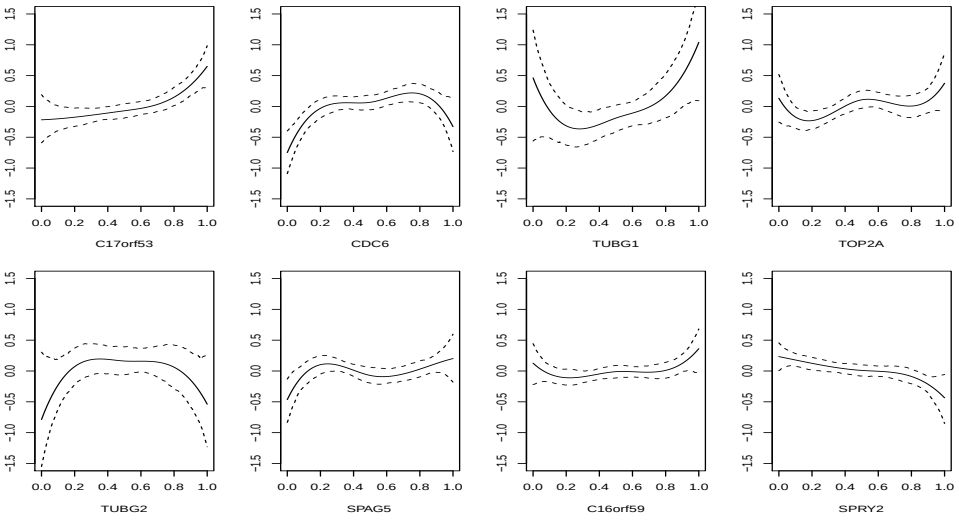


FIG 2. *Estimated curves for the selected genes when* $\tau = 0.5$.

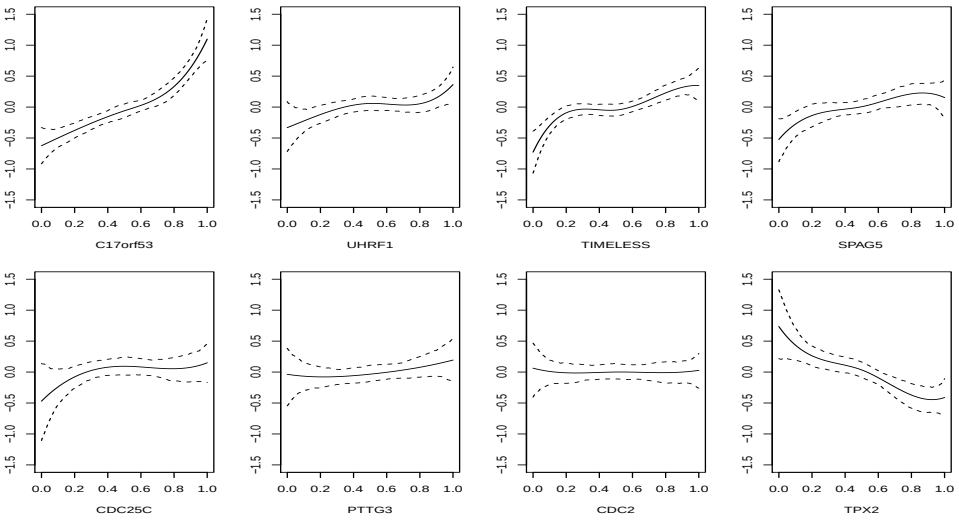


FIG 3. *Estimated curves for the selected genes when* $\tau = 0.75$.

TABLE 5
*Selected genes for the breast cancer data.*

| Method | Genes | | | |
|---|---|---|---|---|
| RKHS ($\tau = 0.25$) | C17orf53 (1) | PRC1 (2) | C15orf42 (4) | FOXM1 (18) |
| | KIF23 (24) | FANCI (25) | | |
| RKHS ($\tau = 0.5$) | C17orf53 (1) | CDC6 (2) | TUBG1 (3) | TOP2A(4) |
| | TUBG2 (6) | SPAG5 (9) | C16orf59 (39) | SPRY2 (170) |
| RKHS ($\tau = 0.75$) | C17orf53 (3) | UHRF1 (5) | TIMELESS (6) | SPAG5 (7) |
| | CDC25C(8) | PTTG3 (15) | CDC2 (16) | TPX2 (24) |
| SPLINE ($\tau = 0.25$) | C17orf53 (1) | PRC1 (2) | C15orf42 (4) | FOXM1 (18) |
| | TOP2A (20) | KIF23 (24) | FANCI (25) | |
| SPLINE ($\tau = 0.5$) | C17orf53 (1) | CDC6 (2) | TUBG1 (3) | TUBG2 (6) |
| | SPAG5 (9) | CENPM (18) | C16orf59 (39) | |
| SPLINE ($\tau = 0.75$) | CDC6 (1) | C17orf53 (3) | UHRF1 (5) | TIMELESS (6) |
| | CDC25C(8) | PTTG3 (15) | TPX2 (24) | RNASEH2A (32) |
| | KIF20A (33) | | | |
| LSQ | C17orf53 (1) | CDC6 (2) | TUBG1 (3) | TOP2A(4) |
| | DTL(8) | SPAG5 (9) | UHRF1 (14) | C16orf59 (39) |
| | TPX2 (148) | KPNB1 (194) | | |
| Linear ($\tau = 0.25$) | BLM (10) | RDM1 (14) | CCDC56(28) | DTL(32) |
| | MAST4 (44) | CDC14B (46) | XRCC2 (66) | CENPE (92) |
| | DDX39(107) | KIF2C (109) | PPAP2B (141) | FAM54A (153) |
| | SIGIRR(154) | MCM6 (165) | RACGAP1(167) | MSH6 (171) |
| | HCN3 (179) | | | |
| Linear ($\tau = 0.5$) | RDM1 (7) | CENPE(26) | NEIL3 (43) | FANCA (47) |
| | CENPQ (48) | FEN1 (54) | C15orf42 (64) | CDCA5 (65) |
| | MCM7(83) | KIF11 (99) | CCDC56 (118) | MKI67 (155) |
| | SPRY2 (170) | | | |
| Linear ($\tau = 0.75$) | CENPE (10) | DTL (11) | FAM54A (100) | ANLN (130) |
| | RBL1 (141) | CENPA (146) | MAD2L1 (172) | CCDC56 (197) |

The number in the bracket after the gene names indicate the ranking of the selected genes in the marginal screening stage.

timator $\hat{f}$ conditional on that event. The first part involves concentration inequality. In view of our two $\ell_1$-type penalties, we will introduce a weighted empirical process for asymmetric random sequences, since asymmetric noises should be considered in the quantile regression models.

For any $t > 0$, define the function $\eta(t) := \max(1, \sqrt{t}, t/\sqrt{n})$. For any given $\lambda > 0$, let $\xi_n := \xi_n(\lambda) = \max\big(\lambda^{-\alpha/2}/\sqrt{n}, \lambda^{-1/2}/n^{1/(1+\alpha)}, \sqrt{\log d/n}\big)$. Let $\varepsilon_i$ be zero-mean i.i.d. random variables with $|\varepsilon_i| \le L$, and we introduce the event $\Theta(t) = \Big\{\big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\big| \le C_\alpha \eta(t)\xi_n\big(\|f\|_2 + \lambda^{1/2}\|f\|_K\big), \forall\, f \in \mathcal{H}_K\Big\}$, where $C_\alpha$ is a constant depending on $\alpha$ and $L$.

Now we describe a key inequality on weighted empirical processes associated with bounded non-symmetric random variables, which can be found in Theorem 10 of the supplementary file in [44].

TABLE 6
*Cross-validation error for different methods. For $\tau = 0.25$ and 0.75, the error is the average quantile loss $\sum_i \rho_\tau(\hat{y}_i - y_i)/n$ over hold-out data. For $\tau = 0.5$, the first number in each cell is the average quantile loss while the second number is the average mean squared error $\sum_i (\hat{y}_i - y_i)^2/n$.*

| | RKHS | SPLINE | LSQ | LIN |
|---|---|---|---|---|
| $\tau = 0.25$ | 0.443 | 0.475 | NA | 0.826 |
| $\tau = 0.5$ | 0.406/0.249 | 0.400/0.256 | 0.426/0.274 | 0.734/0.651 |
| $\tau = 0.75$ | 0.338 | 0.349 | NA | 0.740 |

LEMMA 1. *Let $\varepsilon_i$ be zero-mean i.i.d. random variables with $|\varepsilon_i| \le L$. Under the Spectral Assumption and Sup-norm Assumption, when $\frac{\log d}{\sqrt{n}} \le 1$, we have for all $\lambda > 0$ and all $t \ge 1$*

$$\mathbb{P}\big(\Theta(t)\big) \ge 1 - \exp(-t).$$

Let $\delta = \mu - \mu_\tau$, $\Delta = f - f^*$, and $\Delta_S = \sum_{j \in S}(f_j - f_j^*)$. The next lemma shows that the quantities $\|\widehat{\Delta}\|_{n,1}$ and $\|\widehat{\Delta}\|_{K,1}$ can be controlled by the corresponding norms as applied to the function $\widehat{\Delta}_S$. This allows us to exploit the sparsity assumption.

LEMMA 2. *Under the Spectral Assumption and Sup-norm Assumption, with the choices $\lambda_1 \ge 2C_\alpha c\eta(t)\xi_n$ and $\lambda_2 \ge 2C_\alpha \eta(t)\xi_n(c\gamma_n + \lambda^{1/2})$, then the following bounds holds with probability at least $1 - 2\exp(-t) - \tilde{d}^{-A}$*

$$\lambda_1\|\widehat{\Delta}\|_{n,1} + \lambda_2\|\widehat{\Delta}\|_{K,1} \le 4\lambda_1\|\widehat{\Delta}_S\|_{n,1} + 4\lambda_2\|\widehat{\Delta}_S\|_{K,1} + 2\sqrt{\frac{2t}{n}}\hat{\delta}. \tag{6.1}$$

PROOF. See the supplement material. □

Remark that, since we choose $(\lambda, t)$ as $n^{-\frac{1}{1+\alpha}}$ and $t = t_0$ respectively in this paper, it is easy to check that the constraint of $(\lambda_1, \lambda_2)$ in Lemma 2 does not contradict their choices in Theorem 1-3. In addition, it is easy to check that, the result in Lemma 2 implies that $\hat{f}$ lies in the subset $\mathcal{F}_{\mu,S}$ defined in (3.7) with high probability. In fact, it is trivial if $\|\widehat{\Delta}\|_{K,1} \le 4\|\widehat{\Delta}_S\|_{K,1}$. Otherwise, we have $\lambda_1\|\widehat{\Delta}\|_{n,1} \le 4\lambda_1\|\widehat{\Delta}_S\|_{n,1} + 2\sqrt{\frac{2t}{n}}\hat{\delta}$, which implies $\|\widehat{\Delta}\|_{n,1} \le 4\|\widehat{\Delta}_S\|_{n,1} + C_3\hat{\delta}$ due to the relation $\lambda_1 \ge \sqrt{\frac{2t_0}{n}}$.

6.1. *Uniform Concentration Inequality for the Quantile Loss.* We will use several basic facts of the empirical processes theory for our analysis, including symmetrization inequalities and particularly concentration inequality for empirical process. In this paper, we employ the following Talagrand's

concentration inequality [8, 4]. The result shows that the supremum of any empirical process is concentrated near its mean. The amount of concentration depends only on the maximal sup norm and the maximal variance.

LEMMA 3. *(Concentration Theorem [Bousquet (2002)]) Let $Z_1, ..., Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\Gamma$ be a class of real-valued functions on $\mathcal{Z}$, satisfying for some positive constants $\eta_n$ and $\tau_n$,*

$$\|\gamma\| \le \eta_n \quad \text{and} \ \frac{1}{n}\sum_{i=1}^{n} var(\gamma(Z_i)) \le \tau_n^2, \quad \forall\, \gamma \in \Gamma.$$

*Define* $\boldsymbol{Z} := \sup_{\gamma\in\Gamma} \left|\frac{1}{n}\sum_{i=1}^{n}(\gamma(Z_i) - \mathbb{E}\gamma(Z_i))\right|$. *Then for* $t > 0$

$$\mathbb{P}\Big(\boldsymbol{Z} \ge \mathbb{E}(\boldsymbol{Z}) + t\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}(\boldsymbol{Z}))} + \frac{2\eta_n t^2}{3}\Big) \le \exp(-nt^2).$$

Furthermore, for the case of a single RKHS $\mathcal{H}_K$, we need the relationship between the empirical and $\|\cdot\|_2$ norms for function in $\mathcal{H}_K$. The following conclusion is derived immediately from Theorem 4 of [29] with a slight change.

LEMMA 4. *Suppose that $A \ge 1$ and any given $\tilde{d} \ge d$ with $\log\tilde{d} \ge 2\log\log n$. Then there exists a universal constant $c > 0$ such that with probability at least $1-\tilde{d}^{-A}$, for all $f \in \mathcal{H}_K$, $\|f\|_2 \le c(\|f\|_n + \gamma_n\|f\|_K)$ and $\|f\|_n \le c(\|f\|_2 + \gamma_n\|f\|_K)$.*

Recall that Theorem 4 of [29] is stated as the special case $\tilde{d} = d$, which excludes the fixed dimensional case. Fortunately, we check the proof of Theorem 4 there and conclude that Lemma 4 still holds if one replaces $d$ by an arbitrary $\tilde{d} \ge d$ such that $\log\tilde{d} \ge 2\log\log n$. Indeed, this observation was also pointed out in Theorem 3 of [29].

For any given $\Delta_\mu$, $\Delta_-$ and $\Delta_+ > 0$, we define

$$\mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+) := \{(\mu, f) : |\mu - \mu_\tau| \le \Delta_\mu,\ \gamma_n\|f - f^*\|_{2,1} \le \Delta_-, \gamma_n^2\|f - f^*\|_{K,1} \le \Delta_+\}.$$

Based on this notation, we prove a refined uniform convergence rate. More interestingly, the following conclusion still holds for general Lipschitz-type loss, including the hinge loss for SVM. We believe that it will play an important role in studying sparse additive SVM in the high-dimensional setting.

PROPOSITION 1. *Let $\mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+)$ be as defined as above. Suppose that Spectral Assumption and Sup-norm Assumption hold for each univariate*

$\mathcal{H}_K$. *For any given* $A \geq 2$, *with probability at least* $1 - \tilde{d}^{-A}$, *the following bound holds uniformly on* $\Delta_- \leq e^{\tilde{d}}$ *and* $\Delta_+ \leq e^{\tilde{d}}$,

$$[\mathcal{E}(\mu, f) - \mathcal{E}(\mu_\tau, f^*)] - [\mathcal{E}_n(\mu, f) - \mathcal{E}_n(\mu_\tau, f^*)]$$
$$\leq C_* \eta(t_0)(\Delta_- + \Delta_+) + \exp(-\tilde{d}) + 7\Delta_\mu \sqrt{\frac{t_0}{n}}, \quad \forall\, (\mu, f) \in \mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+),$$

*where* $t_0 = 2\log(2\sqrt{3}/\log 2) + A\log\tilde{d} + 2\log\tilde{d}$ *and* $\lambda = n^{-\frac{1}{1+\alpha}}$.

Proof. To apply Lemma 3, denote $\Gamma = \{\gamma(z),\, \gamma(z) = [\rho_\tau(y - \mu - f(x)) - \rho_\tau(y - \mu_\tau - f^*(x))],\ (\mu, f) \in \mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+)\}$. We can write $[\mathcal{E}(\mu, f) - \mathcal{E}(\mu_\tau, f^*)] - [\mathcal{E}_n(\mu, f) - \mathcal{E}_n(\mu_\tau, f^*)] = \mathbb{E}[\gamma(z)] - \frac{1}{n}\sum_{i=1}^n \gamma(z_i), \gamma \in \Gamma$. Then, by Bousquet concentration inequality, with probability at least $1 - \exp(-t)$

$$\boldsymbol{Z} \leq \mathbb{E}(\boldsymbol{Z}) + \sqrt{\frac{2t(\tau_n^2 + 2\eta_n \mathbb{E}\boldsymbol{Z})}{n}} + \frac{2\eta_n t}{3n}. \tag{6.2}$$

The sub-additivity of $\sqrt{\cdot}$ implies that

$$\sqrt{\frac{2t(\tau_n^2 + 2\eta_n \mathbb{E}\boldsymbol{Z})}{n}} \leq \sqrt{\frac{2t}{n}\tau_n^2} + 2\sqrt{\frac{\eta_n}{n}\mathbb{E}(\boldsymbol{Z})} \leq \sqrt{\frac{2t}{n}\tau_n^2} + \mathbb{E}\boldsymbol{Z} + \frac{\eta_n}{n},$$

where we used the basic inequality $\sqrt{xy} \leq (x + y)/2$ for any $x, y \geq 0$. Meanwhile, the contraction property of $\rho_\tau$ implies $\mathbb{E}(\gamma(Z))^2 \leq 2\|f - f^*\|_2^2 + 2|\mu - \mu_\tau|^2$ for any $f \in \mathcal{F}(\Delta_-, \Delta_+)$ and $\mu \in \mathbb{R}$, that is, $\tau_n^2 \leq 2\sup_{f \in \mathcal{F}(\Delta_-, \Delta_+)} \|f - f^*\|_2^2 + 2\sup_{\mu \in [-1,1]} |\mu - \mu_\tau|^2$. Plugging the above quantities into (6.2) yields $\boldsymbol{Z} \leq 2\mathbb{E}(\boldsymbol{Z}) + 2\sqrt{\frac{t}{n}}\big(\sup_{f \in \mathcal{F}} \|f - f^*\|_2 + \sup_{\mu \in [-1,1]} |\mu - \mu_\tau|\big) + \frac{(1+t)\eta_n}{n}$. Thus, by the contraction property of $\rho_\tau$ and noting $\kappa = 1$, we get $\|\gamma\|_\infty \leq \|f - f^*\|_\infty + |\mu - \mu_\tau| \leq \|f - f^*\|_{K,1} + |\mu - \mu_\tau| \leq \frac{\Delta_+}{\gamma_n^2} + \Delta_\mu,\ \forall\, (\mu, f) \in \mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+)$, which means $\eta_n = \Delta_+/\gamma_n^2 + \Delta_\mu$. In addition, for any $f \in \mathcal{F}(\Delta_\mu, \Delta_-, \Delta_+)$, $\|f - f^*\|_2 \leq \sum_{j=1}^d \|f_j - f_j^*\|_2 \leq \frac{\Delta_-}{\gamma_n}$. In summary, with probability at least $1 - \exp(-t)$ we have

$$\boldsymbol{Z} \leq 2\mathbb{E}(\boldsymbol{Z}) + 2\big(\frac{\Delta_-}{\gamma_n} + \Delta_\mu\big)\sqrt{\frac{t}{n}} + \big(\frac{\Delta_+}{\gamma_n^2} + \Delta_\mu\big)\frac{(1+t)}{n}. \tag{6.3}$$

To bound $\mathbb{E}(\boldsymbol{Z})$, we use a symmetrization technique [49]. Then we have $\mathbb{E}(\boldsymbol{Z}) \leq 2\mathbb{E}[R_n(\Gamma)] \leq 2\mathbb{E}[R_n(\mathcal{F} - f^*)] + 2\mathbb{E}[R_n(\mu - \mu_\tau)] \leq 2\mathbb{E}[R_n(\mathcal{F} - f^*)] + 2\Delta_\mu/\sqrt{n}$, where the second inequality follows from the contraction property of Rademacher process. Applying Talagrand's concentration inequality in Lemma 3 for $R_n(\mathcal{F} - f^*)$, we get that,

$$\mathbb{E}[R_n(\mathcal{F} - f^*)] \leq 2\Big(R_n(\mathcal{F} - f^*) + \Big(\frac{\Delta_-}{\gamma_n}\Big)\sqrt{\frac{2t}{n}} + \Big(\frac{\Delta_+}{\gamma_n^2}\Big)\frac{(1+t)}{n}\Big),$$

with probability at least $1-\exp(-t)$. By Lemma 1, on the event $\Theta(t)$ we have

$$\big|R_n(f)\big| \le C_\alpha \eta(t)\xi_n\big(\|f\|_2 + \lambda^{1/2}\|f\|_K\big), \quad \forall\, f \in \mathcal{H}_K,\ \forall\, \lambda > 0.$$

Hence, with probability at least $1-3\exp(-t)$, we have

$$\begin{aligned}
\boldsymbol{Z} &\le 8R_n(\mathcal{F}-f^*) + \frac{18\Delta_-}{\gamma_n}\sqrt{\frac{t}{n}} + \frac{9\Delta_+}{\gamma_n^2}\frac{(1+t)}{n} + 7\Delta_\mu\sqrt{\frac{t}{n}} \\
&\le 8\sum_{j=1}^{d} R_n(\mathcal{H}_K - f_j^*) + \frac{18\Delta_-}{\gamma_n}\sqrt{\frac{t}{n}} + \frac{9\Delta_+}{\gamma_n^2}\frac{(1+t)}{n} + 7\Delta_\mu\sqrt{\frac{t}{n}} \\
&\le 8C_\alpha\eta(t)\xi_n \sup_{f\in\mathcal{F}}\Big\{\|f-f^*\|_{2,1} + \lambda^{1/2}\|f-f^*\|_{K,1}\Big\} + \frac{18\Delta_-}{\gamma_n}\sqrt{\frac{t}{n}} \\
&\quad + \frac{9\Delta_+}{\gamma_n^2}\frac{(1+t)}{n} + 7\Delta_\mu\sqrt{\frac{t}{n}},
\end{aligned}$$

where the second inequality follows from the subadditivity of Rademacher process. Then, with the choices of $\lambda = n^{-\frac{1}{1+\alpha}}$, we can check that $\xi_n \le c\gamma_n$ and $\xi_n\lambda^{1/2} \le c\gamma_n^2$. Hence we conclude that

$$\boldsymbol{Z} \le 8C_\alpha c\eta(t)(\Delta_- + \Delta_+) + \frac{18\Delta_-}{\gamma_n}\sqrt{\frac{t}{n}} + \frac{9\Delta_+}{\gamma_n^2}\frac{(1+t)}{n} + 7\Delta_\mu\sqrt{\frac{t}{n}}, \tag{6.4}$$

which holds on some event $\Theta(\Delta_\mu, \Delta_-, \Delta_+)$, where $\mathbb{P}(\Theta(\Delta_\mu, \Delta_-, \Delta_+)) \ge 1-3e^{-t}$. By the definition of $\gamma_n$, $\gamma_n \ge \sqrt{A\log\tilde{d}/n}$. So, the inequality in (6.4) can be further bounded by

$$\boldsymbol{Z} \le 8C_\alpha c\eta(t)(\Delta_- + \Delta_+) + 18\Delta_-\sqrt{\frac{t}{A\log\tilde{d}}} + 18\Delta_+\frac{t}{A\log\tilde{d}} + 7\Delta_\mu\sqrt{\frac{t}{n}}.$$

Under the choice of $t = 2\log(2\sqrt{3}/\log 2) + A\log\tilde{d} + 2\log\tilde{d}$, we will obtain a bound that uniformly over

$$e^{-\tilde{d}} \le \Delta_- \le e^{\tilde{d}} \ \text{ and } e^{-\tilde{d}} \le \Delta_+ \le e^{\tilde{d}}. \tag{6.5}$$

For this purpose, we consider $(2\tilde{d}+1)^2$-different discrete pairs $\Delta_-^j = \Delta_+^j := 2^{-j}$, $\quad j = -\tilde{d}, ..., \tilde{d}$. Then, on the event $\bigcap_{k,j}\Theta(\Delta_\mu, \Delta_-^j, \Delta_+^k)$ with at most $(2/\log 2)^2\tilde{d}^2$ intersection terms, we have that $\boldsymbol{Z} \le C_*\eta(t)\big(\Delta_-^j + \Delta_+^k\big) +$

$7\Delta_\mu\sqrt{\frac{t}{n}}$ for all $j,k$, since $A\geq 2$. Moreover,

$$\begin{aligned}\mathbb{P}\big(\bigcap_{k,j}\Theta(\Delta_-^j,\Delta_+^k,t)\big) &\geq 1-3(2/\log 2)^2\tilde{d}^2\exp\big(-2\log(2\sqrt{3}/\log 2)\\ &- A\log\tilde{d}-2\log\tilde{d}\big)\geq 1-\tilde{d}^{-A},\end{aligned}$$

which tends to 1 as $\tilde{d}$ increases. Besides, using monotonicity of the functions $\Delta_-$, $\Delta_+$ involved in the inequalities, the result can be extended to the whole range of $\Delta_-$, $\Delta_+$ satisfying (6.5).

If $\Delta_-\leq e^{-\tilde{d}}$ or, $\Delta_+\leq e^{-\tilde{d}}$, it is trivial to derive the desired result with the same probability. This completes the proof of Proposition 1. □

The proof of Proposition 1 is inspired partially by Lemma 9 in [29] for the quadratic loss. Remark that, global boundedness condition is required for deriving Lemma 9 in [29]. However, this condition leads to sub-optimal convergence rates, which has been discussed in [38]. In the case of Lipschitz-type loss, we show that global boundedness condition is not necessary and this may lead to better rates.

6.2. *Proof of Sharp Oracle Inequalities..* The main procedure of the proofs for Theorem 2 is as follows. We first fully exploit sparsity pattern using the additivity of the Lasso penalties. Then, we make use of the concentration inequality given in Proposition 1, whose upper bound is controlled by the regularization terms. Furthermore, by Assumptions 3-4, the upper bound on $\sum_{j\in S}\|\hat{f}_j-f_j^*\|_n$ can be absorbed by the root of the excess risk. Thus, a tight oracle rates on the excess risk and estimation error are obtained easily.

**Proof of Theorem 2.** By the definition of $(\hat{\mu},\hat{f})$, it follows that $\mathcal{E}_n(\hat{\mu},\hat{f})+\lambda_1\|\hat{f}\|_{n,1}+\lambda_2\|\hat{f}\|_{K,1}\leq\mathcal{E}_n(\mu_\tau,f^*)+\lambda_1\|f^*\|_{n,1}+\lambda_2\|f^*\|_{K,1}$, or equivalently,

$$\begin{aligned}&\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)+\lambda_1\|\hat{f}\|_{n,1}+\lambda_2\|\hat{f}\|_{K,1}\leq\\ &[\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)]-[\mathcal{E}_n(\hat{\mu},\hat{f})-\mathcal{E}_n(\mu_\tau,f^*)]+\lambda_1\|f^*\|_{n,1}+\lambda_2\|f^*\|_{K,1}.\end{aligned}\tag{6.6}$$

Similar to the proof of Theorem 1, provided that $(\lambda_1,\lambda_2)$ is chosen with properly large constant $\zeta$ as that in Theorem 1, with probability at least $1-2\tilde{d}^{-A}$, we have

$$\begin{aligned}&\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)+\frac{c^{-1}}{2}\lambda_1\|\hat{f}-f^*\|_{2,1}+\frac{\lambda_2}{4}\|\hat{f}-f^*\|_{K,1}\\ &\leq 2\lambda_1\sum_{j\in S}\|\hat{f}_j-f_j^*\|_n+2\lambda_2\sum_{j\in S}\|\hat{f}_j-f_j^*\|_K+7|\hat{\mu}-\mu_\tau|\sqrt{\frac{t_0}{n}}+e^{-\tilde{d}}.\end{aligned}\tag{6.7}$$

We first consider the case when

$$2\lambda_1 \sum_{j\in S} \|\hat{f}_j - f_j^*\|_n \geq 2\lambda_2 \sum_{j\in S} \|\hat{f}_j - f_j^*\|_K + 7|\hat{\mu} - \mu_\tau| \sqrt{\frac{t_0}{n}} + e^{-\tilde{d}}.$$

In this case, (6.7) implies that

$$\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*) + \frac{c^{-1}\lambda_1}{2}\|\hat{f} - f^*\|_{2,1} + \frac{\lambda_2}{4}\|\hat{f} - f^*\|_{K,1} \leq 4\lambda_1 \sum_{j\in S} \|\hat{f}_j - f_j^*\|_n, \tag{6.8}$$

Note that by Lemma 4 and Assumptions 3-4, with probability at least $1-\tilde{d}^{-A}$ there holds

$$\begin{aligned}
\sum_{j\in S} \|\hat{f}_j - f_j^*\|_n &\leq c\sum_{j\in S} \|\hat{f}_j - f_j^*\|_2 + 2cs\gamma_n \leq c \sum_{j\in S\cup\{0\}} \|\hat{f}_j - f_j^*\|_2 + 2cs\gamma_n \\
&\leq \frac{c}{\beta_q(S)} \|\hat{\mu} + \hat{f} - (\mu_\tau + f^*)\|_q + 2cs\gamma_n \\
&\leq \frac{c}{\sqrt{c_1}\beta_q(S)} \sqrt{\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*)} + 2cs\gamma_n.
\end{aligned}$$

The first inequality follows from Lemma 4, the third inequality follows from Assumption 4, and the last one is derived from Assumption 3. Then we obtain from (6.8) that

$$\begin{aligned}
&\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*) + \frac{c^{-1}\lambda_1}{2}\|\hat{f} - f^*\|_{2,1} + \frac{\lambda_2}{4}\|\hat{f} - f^*\|_{K,1} \\
&\leq \frac{4c\lambda_1}{\sqrt{c_1}\beta_q(S)} \sqrt{\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)} + 8cs\lambda_1\gamma_n.
\end{aligned} \tag{6.9}$$

By a simple calculation, with probability at least $1 - 3\tilde{d}^{-A}$ there holds

$$\begin{aligned}
\mathcal{E}(\hat{\mu}, \hat{f}) - \mathcal{E}(\mu_\tau, f^*) &\leq \max\left\{16cs\lambda_1\gamma_n, \tfrac{64c^2}{c_1\beta_q^2(S)}\lambda_1^2\right\} \\
&\leq \left(16c\sqrt{\zeta} + \tfrac{64c^2\zeta}{c_1\beta_q^2(S)}\right) s\gamma_n^2,
\end{aligned} \tag{6.10}$$

and based on the above conclusion, this furthermore follows from (6.9)

$$\|\hat{f} - f^*\|_{2,1} \leq 32c^2\left(1 + \frac{4c\sqrt{\zeta}}{c_1\beta_q^2(S)}\right) s\gamma_n. \tag{6.11}$$

It remains to consider the other case when

$$2\lambda_1 \sum_{j\in S} \|\hat{f}_j - f_j^*\|_n < 2\lambda_2 \sum_{j\in S} \|\hat{f}_j - f_j^*\|_K + 7|\hat{\mu} - \mu_\tau| \sqrt{\frac{t_0}{n}} + e^{-\tilde{d}}.$$

In this case, it follows easily from (6.7) that

$$\begin{aligned}&\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)+\frac{c^{-1}\lambda_1}{2}\|\hat{f}-f^*\|_{2,1}+\frac{\lambda_2}{4}\|\hat{f}-f^*\|_{K,1}\\&\le 4\lambda_2\sum_{j\in S}\|\hat{f}_j-f_j^*\|_K+14|\hat{\mu}-\mu_\tau|\sqrt{\frac{t_0}{n}}+2e^{-\tilde{d}}.\end{aligned}$$

Note that $\|\hat{f}_j\|_K\le 1$ and $\|f_j^*\|_K\le 1$, we have $\sum_{j\in S}\|\hat{f}_j-f_j^*\|_K\le 2s$. Then it follows that

$$\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)+\frac{c^{-1}\lambda_1}{2}\|\hat{f}-f^*\|_{2,1}+\frac{\lambda_2}{4}\|\hat{f}-f^*\|_{K,1}\le 8s\lambda_2+14|\hat{\mu}-\mu_\tau|\sqrt{\frac{t_0}{n}}+2e^{-\tilde{d}}.$$

As verified in the proof of Theorem 1, $e^{-\tilde{d}}\le n^{-2}\le\gamma_n^2$, which implies that

$$\begin{aligned}&\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)+\tfrac{c^{-1}\lambda_1}{2}\|\hat{f}-f^*\|_{2,1}+\tfrac{\lambda_2}{4}\|\hat{f}-f^*\|_{K,1}\\&\qquad\le 9\zeta s\gamma_n^2+14|\hat{\mu}-\mu_\tau|\sqrt{\tfrac{t_0}{n}},\end{aligned}\tag{6.12}$$

with probability at least $1-3\tilde{d}^{-A}$. Since $\mathbb{E}[\hat{f}]=\mathbb{E}[f^*]=0$ by assumption, a simple calculation yields that, for $\forall\, q\ge 1$,

$$|\hat{\mu}-\mu_\tau|=|\mathbb{E}[\hat{\mu}+\hat{f}]-\mathbb{E}[\mu_\tau+f^*]|\le\|\hat{\mu}+\hat{f}-(\mu_\tau+f^*)\|_q\le c_1^{-1/2}\sqrt{\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)},\tag{6.13}$$

where the second inequality follows from the Cauchy–Schwartz inequality and the third one follows from Assumption 3. Therefore, this together with (6.12) and (6.13) implies that

$$\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)\le 18\zeta s\gamma_n^2+156c_1^{-1}\frac{t_0}{n},\tag{6.14}$$

and furthermore by (6.13), we have $|\hat{\mu}-\mu_\tau|\le c_1^{-1/2}\Big(3\sqrt{2\zeta s}\gamma_n+14c_1^{-1/2}\sqrt{\frac{t_0}{n}}\Big)$. Similarly, we also get

$$\|\hat{f}-f^*\|_{2,1}\le 28cc_1^{-1/2}\big(3\sqrt{2}s+14\sqrt{3}c_1^{-1/2}\big)\sqrt{\frac{t_0}{n}}+18cs\gamma_n.\tag{6.15}$$

Finally, combining (6.10), (6.11) with (6.14), (6.15), our desired results follow immediately. □

Theorem 3 is easily derived from the combination of Theorem 2 and Assumptions 3 and 4, whose proof is stated as follows.

**Proof of Theorem 3.** By Assumptions 3 and 4, we have

$$\Big(\sum_{j\in S}\|\hat{f}_j-f_j^*\|_2\Big)^2 \le \frac{1}{\beta_q(S)}\|\hat{\mu}+\hat{f}-(\mu_\tau+f^*)\|_q^2 \le \frac{1}{c_1\beta_q(S)}\big[\mathcal{E}(\hat{\mu},\hat{f})-\mathcal{E}(\mu_\tau,f^*)\big].$$

This together with the result of Theorem 2 yields the first conclusion. Furthermore, let $\bar{r}_0 = \sqrt{r_0}\sqrt{\frac{A(s+3)\log\tilde{d}}{n}+s\big(\frac{1}{n}\big)^{\frac{1}{1+\alpha}}}$, for any given $j\in S$, the first result we just derived implies that

$$\|f_j^*\|_2 - \|\hat{f}_j\|_2 \le \|\hat{f}_j - f_j^*\|_2 \le \bar{r}_0.$$

Hence if $\min_{j\in S}\|f_j^*\|_2 > \bar{r}^0$, we have $\|\hat{f}_j\|_2 \ge \|f_j^*\|_2 - \bar{r}_0 > 0$. That is, $\widehat{S}\supseteq S$ with the same probability as that in Theorem 2. □

**7. Supplementary Material.** To highlight the nature and usefulness of Assumptions 3-4, we state some simple sufficient conditions to verify them respectively in the supplementary material. Besides, due to space limitation, we also give the proofs of Theorem 1 and Lemma 2 in the supplementary material.

## REFERENCES


[1] N. Aronszajn. (1950). Theory of reproducing kernels. *Tran. Am. Math. Soc.* **68**, 337–404.

[2] F. Bach, R. Jenatton, J. Mairal and G. Obozinski. (2011). "Convex optimization with sparsity-inducing norms", Chapter 1 of *Optimization for Machine Learning*, NIPS, Springer.

[3] P. Bartlett, O. Bousquet and S. Mendelson. (2005). Local Rademacher complexities. *Ann. Statist.* **33**, 1497–1537.

[4] P. Bartlett and S. Mendelson. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482.

[5] A. Beck and M. Teboulle. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* **2**, 183–202.

[6] A. Belloni and V. Cheronzhukov. (2011). $l^1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39**, 82–130.

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

[8] O. Bousquet. (2002). A Bennet concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris **334**, 495–550.

[9] P. Breheny and J. Huang. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, *Statist. Comput.* **25**, 173-187.

[10] M. Buchinsky. (1994). Changes in the U.S. wage structure 1963-1987: Application of Quantile regression. *Econometrica* **62**, 405–458.

[11] E. J. Candes and T. Tao. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *Ann. Statist.* **35**, 2313–2351.

[12] A. Chatterjeea and S. N. Lahiri. (2011), Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106**, 608–625.

[13] A. Chatterjeea and S. N. Lahiri. (2013), Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41**,1232–1259.

[14] A. Christmann and D. X. Zhou. (2016). Learning rates for the risk of kernel based quantile regression estimators in additive models. *Anal. Appl.* **14**, 449–477.

[15] F. Cucker and S. Smale. (2001). On the mathematical foundations of learning. *Bull. Amer. Soc.* **39**, 1–49.

[16] J. Fan and R. Li. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

[17] X. He (2009). Modeling and inference by Quantile regression, Technical Report, University of Illinois at Urbana-Champaign, Dept. of Statistics.

[18] X. He, L. Wang and H. G. Hong. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.*, **41**, 342–369.

[19] J. L. Horowitz and S. Lee. (2005). Nonparametric estimation of an additive quantile regression model. *J. Amer. Statist. Assoc.* **37**, 1238–1249.

[20] J. Huang, Horowitz, J. L. and Wei F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

[21] D. R. Hunter and K. Lange. (2000). Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.*, **9**, 60-77.

[22] D. R. Hunter and K. Lange. (2004). A tutorial on MM algorithms, *Amer. Statist.* **58**, 30-37.

[23] K. Kato. (2016). Group Lasso for high dimensional sparse quantile regression models. ArXiv:1103.1458.

[24] M. O. Kim. (2007). Quantile regression with varying coefficients. *Ann. Statist.* **35**, 92–108.

[25] R. Koenker. (2005). Quantile Regression, *New York: Cambridge University Press.*

[26] R. Koenker and G. Basset. (1978). Regression quantiles, *Econometrica* **46**, 33–50.
[27] R. Koenker, W. Roger and V. D'Orey. (1987). Algorithm AS 229: Computing regression quantiles *J. Roy. Statist. Soc.. Ser. C* **36**, 383–384.
[28] V. Koltchinskii and M. Yuan. (2008). Sparse recovery in large ensembles of kerenl machines. In proceedings of COLT.
[29] V. Koltchinskii and M. Yuan. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38**, 3660–3695.
[30] Y. J. Li, Y. F. Liu, and J. Zhu. (2007). Quantile regression in reproducing kernel Hilbert spaces. *J. Amer. Statist. Assoc.* **102**, 255–268.
[31] H. Lian. (2012). Semiparametric estimation of additive quantile regression models by two-fold penalty. *J. Bus. Econ. Statist.* **30**, 337–350.
[32] Y. Lin and H. H. Zhang. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–2297.
[33] J. Lv and Y. Fan. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.
[34] S. G. Lv, X. He and J.H. Wang. (2016). A unified penalized method for sparse additive quantile models: An RKHS approach. it Ann. Int. Statist. Math. . Accepted.
[35] L. Meier, S. van de Geer and P. Bühlmann. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779–3821.
[36] S. Mendelson. (2002). Geometric parameters of kernel machines. *In proceedings of COLT*, 29–43.
[37] N. D. Pearce and M. P. Wand (2006). Penalized splines and reproducing kernel methods. *Amer. Statist.* **60**, 233–240.
[38] G. Raskutti, M. J. Wainwright and B. Yu. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13**, 389–427.
[39] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. (2009). SpAM: Sparse additive models. *J. Roy. Statist. Soc., Ser. B* **71**, 1009–1030.
[40] P. Rigollet and A. Tsybakov. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–771.
[41] B. Scholköpf and A. Smola. (2002). Learning with Kernels: Support Vector Machine, Regularization, Optimization, and Beyond. *MIT Press*, Cambridge.
[42] I. Steinwart and A. Christmann. (2008). Support Vector Machines. Springer, New York.
[43] I. Steinwart and A. Christmann. (2011). Estimating conditional quantiles with the help of pinball loss. *Bernoulli* **17**, 211–225.

[44] T. J. Suzuki and M. Sugiyama. (2013). Fast learning rate of multiple kernel learning: trade off between sparsity and smoothness. *Ann. Statist.* **3**, 1381–1405.
[45] The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.
[46] B. Tarigan and S. V. D. Geer. (2006). Classifiers of support vector machine type with $\ell_1$ complexity regularization. *Bernoulli*, **12**, 1045–1076.
[47] R. Tibshirani. (1996). Regression selection and shrinkage via the lasso, *J. R. Statist. Soc. Ser. B*, **58**, 267–288.
[48] P. Tseng and S. Yun. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program., Ser. B*, **117**, 387–423.
[49] S. van de Geer. (2002). Empirical Processes in M-estimation. *Cambridge University Press*, Cambridge.
[50] S. van de Geer. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36**, 614–645.
[51] L. Wang, Y. C. Wu and R. Z. Li. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107**, 214–222.
[52] F. Wei, J. Huang and H. Z. Li. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21**, 1515-1540.
[53] Y. C. Wu and Y. F. Liu. (2009). Variable selection in quantile regression. *Statist. Sinica.* **19**, 801–817.
[54] M. Yuan and Y. Lin. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc., Ser. B* **68**, 49–67.
[55] X. Zhang, Y. Wu, L. Wang and R. Li. (2016). Variable selection for support vector machines in moderately high dimensions. *J. Roy. Stat. Soc., Ser. B* **78**, 53–76.
[56] P. Zhao and B. Yu. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Shaogao LV and Huazhen Lin
Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China
E-mail: lvsg716@swufe.edu.cn
E-mail: linhz@swufe.edu.cn

Heng Lian
School of Mathematics and Statistics
University of New South Wales, Sydney, Australia 2052
E-mail: heng.lian@unsw.edu.au

Jian Huang
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, USA
E-mail: jian-huang@uiowa.edu