Ann Inst Stat Math (2017) 69:897–923
DOI 10.1007/s10463-016-0566-9

A unified penalized method for sparse additive quantile models: an RKHS approach

Shaogao Lv[1] · **Xin He**[2] · **Junhui Wang**[2]

Received: 11 June 2015 / Revised: 19 February 2016 / Published online: 6 June 2016
© The Institute of Statistical Mathematics, Tokyo 2016

Abstract This paper focuses on the high-dimensional additive quantile model, allowing for both dimension and sparsity to increase with sample size. We propose a new sparsity-smoothness penalty over a reproducing kernel Hilbert space (RKHS), which includes linear function and spline-based nonlinear function as special cases. The combination of sparsity and smoothness is crucial for the asymptotic theory as well as the computational efficiency. Oracle inequalities on excess risk of the proposed method are established under weaker conditions than most existing results. Furthermore, we develop a majorize-minimization forward splitting iterative algorithm (MMFIA) for efficient computation and investigate its numerical convergence properties. Numerical experiments are conducted on the simulated and real data examples, which support the effectiveness of the proposed method.

Keywords Additive models · Large p small n · Oracle inequality · Quantile regression · Reproducing kernel Hilbert space · Variable selection

✉ Junhui Wang
j.h.wang@cityu.edu.hk

Shaogao Lv
lvsg716@swufe.edu.cn

Xin He
xinhe6-c@my.cityu.edu.hk

[1] Center of Statistics, Southwestern University of Finance and Economics, 55 Guanghuacun St., Chengdu, Sichuan 610072, China

[2] Department of Mathematics, City University of Hong Kong, 83 Tat Chee Ave., Kowloon Tong, Hong Kong 999077, China

1 Introduction

We consider the problem of analyzing ultra-high-dimensional data, allowing dimension p to grow at an exponential order of sample size n; that is, $\log p = \mathcal{O}(n^{\nu})$ with $0 < \nu < 1$. In recent years, much effort has been devoted to tackle this challenging problem, motivated by modern applications in genomics, bioinformatics, chemometrics, among others. Considering that high-dimensional data often display heterogeneity, outliers and sparsity, we advocate a penalized quantile regression model as an alternative to the widely used penalized mean regression formulation. Besides, in view of the complex relationship between covariates and response, this paper considers a high-dimensional additive quantile regression model. More precisely, given the available training sample $\{(x_i, y_i)\}_{i=1}^n$, the high-dimensional additive quantile regression model is formulated as

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

where τ is a given quantile, S_τ is the active subset of $\{1, 2, \ldots, p\}$ that may change with τ, $\epsilon_{\tau,i}$ is the random error satisfying $\mathbb{P}(\epsilon_{\tau,i} \le 0 | x_i) = \tau$, and $f^*_{\tau,j} : \mathbb{R} \to \mathbb{R}$ is a smooth univariate function. Furthermore, $\mathbb{E}[f^*_{\tau,j}(x)] = 0$ is imposed for $j = 1, \ldots, p$ to circumvent the identifiability issue. Note that no distributional assumption for $\epsilon_{\tau,i}$ is needed, and it may depend on the covariates to account for the heterogeneous errors and heavy-tailed distributions.

Since its introduction by Koenker and Basset (1978), quantile regression (QR) has attracted great attention due to its interpretability and robustness. Compared to the mean regression formulation, by tuning different quantiles, QR provides a complete picture of the conditional distribution of the response given the covariates. This advantage enables us to better understand the intrinsic relationship between the covariates and the response. Recently, many researchers have demonstrated that high-dimensional data often display heterogeneity due to either heteroscedastic variance or covariate variety. Besides, it is usually difficult to check error distribution with high-dimensional data, and thus the validity of least square regression formulation can be problematic. These observations partially motivate researchers to study the sparse additive QR, as well as its interpretability and flexibility. Analogous to high-dimensional sparse mean regression models, estimation and variable selection properties of high-dimensional sparse QR models are intensively studied (Belloni and Chernozhukov 2011; Kato 2016; Lian 2012; Li and Zhu 2008; Van der Geer 2008; Wang et al. 2012; He et al. 2013). Among them, Belloni and Chernozhukov (2011) proposed a lasso-type penalized method for the linear quantile model, and established some nice statistical properties, including the oracle inequalities. Wang et al. (2012) proposed a penalized QR based on the SCAD penalty (Fan and Li 2001) in a high-dimensional setting, which established variable selection consistency for linear models. Kato (2016) presented a group-lasso penalized method for the linear QR model, leading to a second-order cone programming (SOCP) problem.

Note that the aforementioned work is all geared for linear or parametric QR models. In many applications, however, little prior justification can be made for the parametric

forms. To allow more flexible modeling while still avoiding the "curse of dimensionality", the additive model has become a natural and popular choice (Hastie and Tibshirani 1990). For example, Koenker (2011) proposed a twofold penalized method with a total variation roughness penalty, which leads to a sparse linear programming problem and is solved by the interior point method. Lian (2012) proposed a model selection and semi-parametric method with two SCAD-type penalties in the fixed p setting. Van der Geer (2008) investigated the non-asymptotic oracle inequities of the adaptive Lasso estimators for general Lipschitz loss functions, which includes the quantile loss as a special case. Kato (2016) also proved a non-asymptotic oracle inequality with the group-Lasso penalty for the sparse additive QR. However, it is noticed that the existing work often lead to computationally demanding numerical algorithms, including the interior point method (Koenker 2011), the local quadratic approximation (Lian 2012), or the SOCP (Kato 2016). Moreover, Koenker (2011) and Lian (2012) only considered fixed p settings, and Van der Geer (2008) did not address computational challenges and variable selection of the corresponding penalized approaches.

In this paper, we propose a new penalized QR approach with a combined smoothness-sparsity penalty under the RKHS (Wahba 1999) framework. The proposed method attains nice theoretical properties and allows for flexible modeling due to the properties of RKHS. Particularly, the oracle inequality is established under much weaker conditions than the existing results in the literature (Belloni and Chernozhukov 2011; Kato 2016; Van der Geer 2000; Wang et al. 2012), which often require restrictive conditions, such as the extended restricted eigenvalue assumption (Ravikumar et al. 2009). In terms of computation, we developed an efficient algorithm combining the majorize minimization (MM; Lian 2012) algorithm and the proximal gradient method. On one hand, the penalization of our method enables us to reduce computational cost significantly by reformulating into a group-Lasso type of formulation. On the other hand, to tackle the computational challenges, we use a smooth quadratic function to majorize the non-smooth quantile loss function. Then, we minimize the majorized loss function with a group-Lasso penalty, which then is to be optimized by the proposed MMFIA algorithm. The MMFIA algorithm enjoys fast computation and nice convergence properties, compared against a number of existing optimization algorithms developed for penalized QR models (Li and Zhu 2008; Pearce and Wand 2006). To our knowledge, our study is new under the high-dimensional QR models, taking both computational efficiency and theoretical properties simultaneously into account.

It is also worth pointing out that the proposed method is developed for a general RKHS, which contains most of the existing methods as special cases. For example, if the linear kernel is used, the proposed method reduces down to a linear QR method; if part of additive components are in a specified nonlinear kernel space, the proposed method becomes the partially linear model; the classical spline-based approaches are also included in our method, as penalized spline functions can be embedded in an RKHS (Pearce and Wand 2006). Moreover, compared with the finite-dimensional functional spaces, the RKHS is an infinite-dimensional space and gains much more flexibility and avoids the choice of basis functions as in many classical nonparametric models.

The rest of the article is organized as follows. Section 2 introduces some basic notations and the regularized QR model with the smoothness-sparsity penalty. Sec-

tion 3 presents an efficient numerical optimization for the proposed model, which is achieved by combing the MM algorithm and the proximal gradient method. In Sect. 4, oracle inequalities for the proposed method are established under weak conditions. Numerical experiments on simulated and real examples are conducted in Sect. 5 to examine the effectiveness of the proposed method. All technical proofs are relegated to the Appendix.

2 Proposed approach

2.1 Preambles

Given a compact subset of $\mathcal{X} \subset \mathbb{R}$, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded, symmetric, and positive semi-definite kernel function. The RKHS associated with the kernel function K, denoted as $\mathcal{H}_K$, is the completion of the linear span of functions $K_x := K(x, \cdot)$; $x \in \mathcal{X}$ with the inner product given by $< K_x, K_y >_{\mathcal{H}_K} = K(x, y)$. It satisfies the reproducing property:

$$f(x) =< K_x, f >_{\mathcal{H}_K}, \quad \text{for any } f \in \mathcal{H}_K.$$

This property implies that $\|f\|_\infty \le \kappa \|f\|_{\mathcal{H}_K}$ with $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$. For notational simplicity, we assume that $\kappa = 1$ in the sequel.

Assume that the training sample is drawn from $\mathcal{X}^p \subset [0, 1]^p$, endowed with an underlying probability measure $\mathbb{Q}$. Let $\mathcal{H}_j$; $j = 1, 2, \ldots, p$ denote an RKHS of univariate functions on the domain $\mathcal{X}$. As the intercept effect is accounted by μ_τ in (1), we assume that

$$\mathbb{E}[f_j(x)] = \int_{\mathcal{X}} f_j(x) d\mathbb{Q}(x) = 0, \quad \text{for any } f_j \in \mathcal{H}_j;\ j = 1, 2, \ldots, p,$$

to avoid the identifiability issue.

Furthermore, we define a composite RKHS by

$$\mathcal{F} := \left\{ f = \sum_{j=1}^{p} f_j : \ f_j \in \mathcal{H}_j \right\},$$

where the norm is defined as $\|f\|_{\mathcal{F}}^2 = \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}^2$, and the associated kernel is $K^c(x, u) = \sum_{j=1}^p K_j(x_j, u_j)$. Denote by $L^2(\mathbb{Q})$ the usual square integral norm on the space $\mathcal{F}$. In addition, we consider the empirical norm $L^2(\mathbb{Q}_n)$ by the sample $\{x_i\}_{i=1}^n$, defined as $\|f\|_{L^2(\mathbb{Q}_n)}^2 = \frac{1}{n} \sum_{i=1}^n f^2(x_i)$. For short hand, we frequently use $\|f\|_2 = \|f\|_{L^2(\mathbb{Q})}$ and $\|f\|_n = \|f\|_{L^2(\mathbb{Q}_n)}$ for a p-variate function $f \in \mathcal{F}$. For a univariate function $f_j \in \mathcal{H}_j$, we also use $\|f_j\|_2 = \|f_j\|_{L^2(\mathbb{Q}_j)}$ and $\|f_j\|_n = \|f_j\|_{L^2(\mathbb{Q}_{n,j})}$, where $\mathbb{Q}_j$ is the marginal distribution of the jth covariate induced by $\mathbb{Q}$ and $\mathbb{Q}_{n,j}$ is its empirical counterpart by the sample $\{x_{ij}\}_{i=1}^n$ as well.

2.2 Regularized QR with smoothness-sparsity penalty

For illustration, we assume all $\mathcal{H}_j$ values are identical, denoted by $\mathcal{H}$, and the proposed method can be extended to allow different $\mathcal{H}_j$ values. Indeed, if we consider different kernels, such that some kernels are linear and the others are nonlinear kernels, the assumed model in (1) reduces to the standard partially linear model.

Define the empirical risk concerning the quantile loss as

$$\mathcal{E}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - f(x_i)\right),$$

where $\rho_\tau(u) = u(\tau - I_{u\le 0})$ is the quantile loss function, also called the asymmetric absolute deviation function (Koenker and Basset 1978). Based on the identification condition, the intercept term μ_τ in (1) can be estimated separately; $\hat{\mu}_\tau$ is the τth quantile of the responses $\{y_i\}_{i=1}^n$. Then, we propose the following minimization problem with respect to the additive components:

$$\min_{f=\sum_{j=1}^p f_j,\, f_j\in\mathcal{H}} \left\{\mathcal{E}_n(\hat{\mu}+f) + \lambda_n I_n(f)\right\}, \tag{2}$$

where the regularization term is defined as

$$I_n(f) = \sum_{j=1}^{p} \sqrt{\|f_j\|_n^2 + \rho_n \|f_j\|_{\mathcal{H}}^2}.$$

Clearly, $I_n(f)$ consists of two penalties, where $\|f_j\|_n$ is used to control the sparsity and $\|f_j\|_{\mathcal{H}}$ is a smoothness penalty. The regularization parameters (λ_n, ρ_n) will be specified, respectively, in our theory and in practice.

Note that the penalty $I_n(f)$ was proposed originally by Meier et al. (2009) in the context of mean regression, which has achieved superior performance both theoretically and computationally. The main idea behind this combined penalty is that $\|f_j\|_n$, as an approximation of $\|f_j\|_2$, is able to measure nonparametric function, whereas $\|\cdot\|_{\mathcal{H}}$ is added to the penalty term to control the functional complexity. Meanwhile, considering the importance of decomposable property to the generated sparsity and fast convergence rates (Negahban et al. 2012), $I_n(f)$ is ultimately formulated as a mixed norm to enjoy the decomposable property.

In the literature, a variety of combinations of sparsity and smoothness penalties have been studied based on sparse additive models. In the context of additive mean regression, Ravikumar et al. (2009) proposed a regularized method with only the sparsity regularization, while putting the smoothness term as a constraint. Although the method in Ravikumar et al. (2009) can be solved by a backfitting procedure, it does not establish any theoretical guarantees, and may encounter algorithmic and computational instability. Koltchinskii and Yuan (2008) developed a regularized learning algorithm for selecting significant kernels under the multi-kernel setting. Their method is based on a single penalty term $\sum_{j=1}^p \|f_j\|_{\mathcal{H}}$. However, if $\mathcal{H}$ is a RKHS, this method leads

to a SOCP problem and can be computationally challenging. We also notice that there exist two popular combinations of the sparsity and smoothness penalties: (i) $I(f) = \sum_{j=1}^{p}(\lambda_1\|f_j\|_n + \lambda_2\|f_j\|_{\mathcal{H}}^2)$ in Rosasco (2013) and (ii) $I(f) = \sum_{j=1}^{p}(\lambda_1\|f_j\|_n + \lambda_2\|f_j\|_{\mathcal{H}})$ in Raskutti et al. (2012). While the first combined penalty leads to a group lasso formulation, it appears to be lack of theoretical justification. The second term has been proved to enjoy some theoretical properties (Lv et al. 2016; Raskutti et al. 2012); however, it still requires SOCP. To enjoy both theoretical properties and computational efficiency, we propose to equip the additive QR model with the new combination of the sparsity and smoothness penalty $I_n(f)$ in (2).

3 Computing algorithm

This section presents an efficient computing algorithm for solving (2). First, applying the representer theorem of the RKHS, (2) reduces to a finite-dimensional minimization problem.

Specifically, the representer theorem assures that the solution to (2) has the form

$$\hat{f}(z_1, \ldots, z_p) = \sum_{j=1}^{p}\sum_{i=1}^{n} \hat{\alpha}_{ij} K(z_j, x_{ij}) = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^T \Theta(z),$$

where $\hat{\alpha}_j = (\hat{\alpha}_{1j}, \ldots, \hat{\alpha}_{nj})$, $j = 1, \ldots, p$. The empirical np vector-valued function is defined as

$$\Theta(z) = (K(x_{11}, z_1), \ldots, K(x_{n1}, z_1), \ldots, K(x_{1p}, z_p), \ldots, K(x_{np}, z_p)), \quad z = (z_1, \ldots, z_p).$$

For every $j \in \{1, \ldots, p\}$, denote $\mathbb{K}^j \in \mathbb{R}^{n\times n}$, with entries $\mathbb{K}^j_{i,\ell} = K(x_{ij}, x_{\ell j})$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ be an np column vector. The optimal coefficients $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$ are any minimizer to the following convex optimization:

$$\arg\min_{\alpha_j \in \mathbb{R}^n} \{\Omega(\boldsymbol{\alpha})\}, \quad \text{where } \Omega(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{\alpha}^T \Theta(x_i)\right) + \lambda_n \sum_{j=1}^{p} \sqrt{\alpha_j^T M_j \alpha_j}, \tag{3}$$

and $M_j := \frac{(\mathbb{K}^j)^2}{n} + \rho_n \mathbb{K}^j$ for each j and some specified parameter ρ_n.

Traditionally, by introducing two slack variables, the quantile function ρ_τ can be linearized. Thus, the optimization problem of (2) is reduced to an SOCP accordingly, as suggested in Kato (2016). However, SOCP is known to be computationally expensive, and is not scalable for large problems.

Alternatively, this section develops a direct optimization scheme for solving (2), which consists of two parts. First, we propose to use MM algorithm to majorize the nonsmooth quantile loss function by a smooth one. Second, we will employ the proximal method to solve the resultant optimization.

In essence, the MM algorithm replaces a difficult optimization problem by a sequence of easier optimization problems. In our case, a quadratic function is used to majorize $\rho_\tau(u)$ for any u. We first note that ρ_τ can be approximated by its perturbed version with some small $\epsilon > 0$:

$$\rho_\tau^\epsilon(u) := \rho_\tau(u) - \frac{\epsilon}{2}\ln(\epsilon + |u|).$$

This leads to a intermediate convex optimization:

$$\min_{\alpha_j \in \mathbb{R}^n} \{\Omega_\epsilon(\boldsymbol{\alpha})\} \quad \text{where } \Omega_\epsilon(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^n \rho_\tau^\epsilon\Big(y_i - \boldsymbol{\alpha}^T\Theta(x_i)\Big) + \lambda_n \sum_{j=1}^p \sqrt{\alpha_j^T M_j \alpha_j}. \tag{4}$$

Then, at the kth step, ρ_τ^ϵ can be majorized by the quadratic function at u^k:

$$\tilde{\rho}_\tau(u|u^k) = \frac{1}{4}\left[\frac{u^2}{\epsilon + |u^k|} + (4\tau - 2)u + c\right].$$

The majorization holds true, as it can showed by direct calculation that

$$\tilde{\rho}_\tau\big(u|u^k\big) \geq \rho_\tau^\epsilon(u) \text{ for all } u, \text{ and } \tilde{\rho}_\tau\big(u^t|u^k\big) = \rho_\tau^\epsilon\big(u^k\big) \tag{5}$$

for an appropriately chosen constant c.

Next, we solve the resultant optimization problem with ρ_τ replaced by the majorized function $\tilde{\rho}_\tau$ to update $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_\epsilon^{k+1} = \arg\min_{\alpha_j \in \mathbb{R}^n} \left\{\Omega_\epsilon(\boldsymbol{\alpha}|\,\boldsymbol{\alpha}_\epsilon^k)\right\}, \tag{6}$$

where $\Omega_\epsilon(\boldsymbol{\alpha}|\,\boldsymbol{\alpha}_\epsilon^k) := \frac{1}{n}\sum_{i=1}^n \tilde{\rho}_\tau\Big(y_i - \boldsymbol{\alpha}^T\Theta(x_i)\Big|y_i - (\boldsymbol{\alpha}_\epsilon^k)^T\Theta(x_i)\Big) + \lambda_n \sum_{j=1}^p \sqrt{\alpha_j^T M_j \alpha_j}$. To solve the sub-optimization problem in (6), the proximal method is employed. The proximal method has been proved to be effective in solving sparse optimization problems (Beck and Teboulle 2009; Combettes and Wajs 2005) because of its fast convergence rates and the ability to deal with nonsmooth convex problems.

The general proximal methods can be described as follows. The proximal operator $\text{Prox}_{\lambda I_n} : \mathcal{G} \to \mathcal{G}$ given by Moreau (1962) is defined as the unique solution of any given function:

$$\mathrm{Prox}_{\lambda I_n}(f) = \arg\min_{g\in\mathcal{G}} \left\{ \frac{1}{2}\|f-g\|_{\mathcal{G}}^2 + \lambda I_n(g) \right\}. \tag{7}$$

At any given kth iteration, denote the functional $F^k(f) = \frac{1}{n}\sum_{i=1}^n \tilde{\rho}_\tau\Big(y_i - f(x_i)\big|\, y_i - f^k(x_i)\Big)$ with any $f \in \mathcal{G}$. The minimization of (3) can be done iteratively using the forward-backward splitting algorithm. Let $\tilde{f}^0 = f^0 = f^1 \in \mathcal{G}$ is an arbitrary initialization, $c_{1,t}$ and $c_{2,t}$ are suitable chosen positive sequences, when $t \geq 2$

$$\begin{aligned} f^t &= \mathrm{Prox}_{\frac{\lambda_n}{L} I_n}\big(\tilde{f}^t - \frac{1}{2L}\nabla F^k(\tilde{f}^t)\big), \\ \tilde{f}^t &= c_{1,t} f^{t-1} + c_{2,t} f^{t-2}, \end{aligned} \tag{8}$$

where L is a parameter which should essentially be an upper bound on the Lipschitz constant of $\nabla F^k/2$ and is typically set with a line search. An alternative choice of $c_{1,t}$ and $c_{2,t}$ leads to an accelerated version of the algorithm (8), sometimes called FISTA (fast iterative shrinkage thresholding algorithm Beck and Teboulle 2009; Tseng 2010), which is obtained by setting $s_0 = 1$:

$$s_t = \frac{1}{2}\left(1 + \sqrt{1+4s_{t-1}^2}\right), c_{1,t} = 1 + \frac{s_{t-1}-1}{s_t}, \quad \text{and} \quad c_{2,t} = \frac{1-s_{t-1}}{s_t}. \tag{9}$$

Using the above sequences, it is proved that the objective values generated by such a procedure have convergence of order $\mathcal{O}(1/t^2)$ in Beck and Teboulle (2009).

Computing the proximal operator efficiently and exactly is crucial to enjoying the fast convergence rates of proximal methods. We, therefore, discuss the properties of this operator and on its computation for our sparsity penalty. Since I_n is one-homogeneous, namely, $I_n(\theta f) = \theta I_n(f)$, for $\theta > 0$. The Moreau identity (Combettes and Wajs 2005) gives a useful equivalent relationship between the proximal operator, and the projection operator, that is,

$$\mathrm{Prox}_{\frac{\lambda_n}{L} I_n} = I - \pi_{\frac{\lambda_n}{L}\mathcal{C}_n}, \tag{10}$$

where $\mathcal{C}_n = (\partial I_n(0))$ is the subdifferential of I_n at the origin, and $\pi_{\frac{\lambda_n}{L}\mathcal{C}_n} : \mathcal{G} \to \mathcal{G}$ is the projection on $\frac{\lambda_n}{L}\mathcal{C}_n$, which is well defined, since $\mathcal{C}_n$ is a closed subset of $\mathcal{G}$. Hence, the central point of the proximal method is computing the projection operator $\pi_{\lambda\mathcal{C}_n}$ for any given parameter λ.

In our setting, $\mathcal{G} = \mathbb{R}^{np}$ with the usual Euclidean inner product. For notational simplicity, we use $\boldsymbol{\alpha}^k$ instead of $\boldsymbol{\alpha}_\epsilon^k$ in the following. At the kth iteration, denote the functional $F^k(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n \tilde{\rho}_\tau\Big(y_i - \boldsymbol{\alpha}^T\Theta(x_i)\big|\, y_i - (\boldsymbol{\alpha}^k)^T\Theta(x_i)\Big)$ with any $\boldsymbol{\alpha} \in \mathbb{R}^{np}$. For each $j \in \{1,\ldots,p\}$, we define a map $J_j : \mathbb{R}^{np} \to \mathbb{R}^n$ as $J_j(\boldsymbol{\alpha}) = \alpha_j$, and J is the identity map from $\mathbb{R}^{np}$ to $\mathbb{R}^{np}$. Note that we endow $\mathbb{R}^n$ with the weight inner product, that is, $\langle u, v\rangle_{w_j} = u^T M_j v$ for any $u,\ v \in \mathbb{R}^n$ and $j \in \{1,\ldots,p\}$. Thus, the penalty term of (3) can be rewritten as

$$I_n(\boldsymbol{\alpha}) = \sum_{j=1}^{p} \left\| J_j(\boldsymbol{\alpha}) \right\|_{w_j}.$$

Denote the adjoint of J by J^*, and it is easy to verify that $J^*(\boldsymbol{\alpha}) = \sum_{j=1}^{p} J_j^*(\alpha_j) = (M_1\alpha_1, \ldots, M_p\alpha_p)$, where we used the fact $J_j^*(\alpha_j) = (0, \ldots, M_j\alpha_j, \ldots, 0)$.

With these preparations, following the conclusion of Proposition 2 of Mosci et al. (2010), we have

$$\mathcal{C}_n = \left\{ J^*\mathbf{v} : \mathbf{v} \in \mathbb{R}^{np},\ \|v_j\|_{w_j} \le 1,\ \forall\, j \right\}.$$

Moreover, the projection of an element $\mathbf{u} = (u_1, \ldots, u_p) \in \mathbb{R}^{np}$ on the set $\lambda\mathcal{C}_n$ is given by $\lambda J^*\bar{\mathbf{v}}$, with

$$\bar{v}_j = \arg \min_{\|v_j\|_{w_j} \le 1} \|\lambda M_j v_j - u_j\|_{\mathbb{R}^n}^2,$$

which yields that

$$\bar{v}_j = \min\left\{1, \frac{\|(M_j^{-1})u_j\|_{w_j}}{\lambda}\right\} \frac{(M_j^{-1})u_j}{\|(M_j^{-1})u_j\|_{w_j}}.$$

Therefore, the nonlinear operation $(I - \pi_{\lambda\mathcal{C}_n})(\mathbf{u})$ acts on each block as

$$\begin{aligned}[(I - \pi_{\lambda\mathcal{C}_n})(\mathbf{u})]_j &= u_j - \min\left\{\lambda, \|(M_j^{-1})u_j\|_{w_j}\right\} \frac{u_j}{\|(M_j^{-1})u_j\|_{w_j}} \\ &= \left(\sqrt{u_j^T(M_j^{-1})u_j} - \lambda\right)_+ \frac{u_j}{\sqrt{u_j^T(M_j^{-1})u_j}}, \end{aligned} \tag{11}$$

since that $\|(M_j^{-1})u_j\|_{w_j} = \sqrt{u_j^T(M_j^{-1})u_j}$. This leads to the classical soft-thresholding operators introduced Donoho and Johnstone (1995).

Thus, the remaining work we need to do is computing the gradient function $\nabla F^k(\boldsymbol{\alpha})/2$ and the parameter L involved in (8). By a direct calculation, $\nabla F^k(\boldsymbol{\alpha})$ has an explicit form as

$$\nabla(F^k)(\boldsymbol{\alpha}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{\alpha}^T\Theta(x_i)}{2(\epsilon + \left|y_i - (\boldsymbol{\alpha}^k)^T\Theta(x_i)\right|)} + \frac{2\tau - 1}{2}\right)\Theta(x_i). \tag{12}$$

For the step size L_k at iteration k, it is not easily computable in our situation, since it depends on the maximum eigenvalue of $\nabla^2(F^k)$, which is an $np \times np$ matrix. Alternatively, we use a line search called backtracking to return an feasible constant for L_k.

As a consequence, the overall procedure can be stated as follows:

MMFIA Algorithm:
given: parameters λ_n, ρ_n, ϵ, c, and quantile point $\tau > 0$.
initialize: $\boldsymbol{\alpha}^1 = 0$, $s_1 = 1$, $k = t = 1$,
for $k \geq 1$
repeat
given: $\boldsymbol{\alpha}_k^0 = \boldsymbol{\alpha}_k^1 = 0$, $L_k > 0$
for $t \geq 2$ **repeat**

$$
\begin{aligned}
s_t &= \frac{1}{2}\left(1 + \sqrt{1 + 4s_{t-1}^2}\right) \\
\tilde{\boldsymbol{\alpha}}_k^t &= \left(1 + \frac{s_{t-1} - 1}{s_t}\right)\boldsymbol{\alpha}_k^{t-1} + \frac{1 - s_{t-1}}{s_t}\boldsymbol{\alpha}_k^{t-2} \\
\boldsymbol{\alpha}_k^t &= \left(I - \pi_{\frac{\lambda_n}{L_k}\mathcal{C}_n}\right)\left(\tilde{\boldsymbol{\alpha}}_k^t - \frac{1}{2L_k}\nabla F^k(\tilde{\boldsymbol{\alpha}}_k^t)\right) \\
t &\leftarrow t + 1
\end{aligned}
$$

until $\boldsymbol{\alpha}_k^t$ converges to some $\boldsymbol{\alpha}^*$
$\boldsymbol{\alpha}^{k+1} \leftarrow \boldsymbol{\alpha}^*$, then $k \leftarrow k + 1$
until $\boldsymbol{\alpha}^{k+1}$ converges to $\hat{\boldsymbol{\alpha}}_\epsilon$.
then return $(\hat{\boldsymbol{\alpha}}_\epsilon)$

Proposition 1 *(Convergence property of MMFIA) Consider the objective function (6), where the data (y, X) lie in a compact set and no two columns of X are identical. For a given (λ_n, ρ_n) and $\epsilon > 0$, we have a descent property with respect to Ω_ϵ, that is*

$$\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^{k+1}) \leq \Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^k), \quad k \geq 1.$$

If additionally, the kernel matrix $\widetilde{\Theta} = (\Theta(x_1), \ldots, \Theta(x_n))$ arranged by row is non-singular, the algorithm (6) converges to the unique minimizer of $\Omega_\epsilon(\boldsymbol{\alpha})$.

Proof By the descent property of (5), it follows that

$$\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^{k+1}) \leq \Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^{k+1} | \boldsymbol{\alpha}_\epsilon^k) \leq \Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^k | \boldsymbol{\alpha}_\epsilon^k) = \Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^k),$$

where the second inequality follows from the definition of $\boldsymbol{\alpha}_\epsilon^{k+1}$,and the last equality follows from (5) as well. Thus, we complete the first part of Proposition 1.

On the other hand, we define the iteration map $\Psi : \boldsymbol{\alpha}_\epsilon^k \to \boldsymbol{\alpha}_\epsilon^{k+1}$. Due to the strict convexity of $\Omega_{\epsilon,k}(\boldsymbol{\alpha})$, Ψ is a point-point map. Moreover, since $\Omega_\epsilon(\boldsymbol{\alpha} | \boldsymbol{\alpha}^k)$ is continuous in $\boldsymbol{\alpha}^k$, Ψ is continuous as well. Note that the derivatives

$$\frac{\partial}{\partial u}\rho_\tau^\epsilon(u) = \begin{cases} \tau - \frac{\epsilon}{2(\epsilon+u)}, & u \geq 0; \\ \tau - 1 + \frac{\epsilon}{2(\epsilon-u)}, & u < 0. \end{cases}$$

This furthermore implies that

$$\frac{\partial^2}{\partial \boldsymbol{\alpha}^2}\left\{\sum_{i=1}^{n}\rho_\tau^\epsilon\left(y_i-\boldsymbol{\alpha}^T\Theta(x_i)\right)\right\}=\sum_{i=1}^{n}\frac{\epsilon}{2(\epsilon+\left|y_i-\boldsymbol{\alpha}^T\Theta(x_i)\right|)^2}\Theta(x_i)^T\Theta(x_i)$$
$$=\frac{\epsilon}{2}\widetilde{\Theta}^T W_\epsilon(\boldsymbol{\alpha})\widetilde{\Theta},$$

where the matrix $W_\epsilon(\boldsymbol{\alpha})=\operatorname{diag}\left((\epsilon+\left|y_i-\boldsymbol{\alpha}^T\Theta(x_i)\right|)^{-2}\right)_{i=1}^n$. This shows that $\Omega_\epsilon(\boldsymbol{\alpha})$ is strictly convex when the kernel matrix $\widetilde{\Theta}$ is of full rank, and $\Omega_\epsilon(\boldsymbol{\alpha})$ has the unique minimizer denoted by $\hat{\boldsymbol{\alpha}}_\epsilon$. Given any convergent subsequence $\boldsymbol{\alpha}_\epsilon^{k_n}$ of $\boldsymbol{\alpha}_\epsilon^k$ with limit $\boldsymbol{\alpha}_\epsilon^*$. Rewriting the first derived result of this proof, we have

$$\Omega_\epsilon\left(\Psi\left(\boldsymbol{\alpha}_\epsilon^{k_n}\right)\right)\le\Omega_\epsilon\left(\boldsymbol{\alpha}_\epsilon^{k_n}\right).$$

Let $n\to\infty$, and based on the continuity of Ψ and $\Omega_\epsilon(\boldsymbol{\alpha})$, we claim that

$$\Omega_\epsilon\left(\Psi\left(\boldsymbol{\alpha}_\epsilon^*\right)\right)=\Omega_\epsilon\left(\boldsymbol{\alpha}_\epsilon^*\right). \tag{13}$$

Otherwise, there holds $\Omega_\epsilon(\Psi(\boldsymbol{\alpha}_\epsilon^*))<\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*)$, which implies that $d\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*)\neq 0$. In addition, it is also verified that $d\Omega_\epsilon(\boldsymbol{\alpha}|\,\boldsymbol{\alpha})=d\Omega_\epsilon(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha}\in\mathbb{R}^{np}$, and then $d\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*|\boldsymbol{\alpha}_\epsilon^*)\neq 0$. It is impossible from the definition of $\boldsymbol{\alpha}_\epsilon^*$ as a limiting point of $\boldsymbol{\alpha}_\epsilon^{k_n}$. As a similar argument as above, we see that

$$\Omega_\epsilon(\Psi(\boldsymbol{\alpha}_\epsilon^*)|\boldsymbol{\alpha}_\epsilon^*)\le\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*|\boldsymbol{\alpha}_\epsilon^*)=\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*)=\Omega_\epsilon(\Psi(\boldsymbol{\alpha}_\epsilon^*))\le\Omega_\epsilon(\Psi(\boldsymbol{\alpha}_\epsilon^*)|\boldsymbol{\alpha}_\epsilon^*),$$

that is

$$\Omega_\epsilon(\Psi(\boldsymbol{\alpha}_\epsilon^*)|\boldsymbol{\alpha}_\epsilon^*)=\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*|\boldsymbol{\alpha}_\epsilon^*).$$

By the definition of Ψ and the strict convexity of $\Omega_\epsilon(\boldsymbol{\alpha}|\boldsymbol{\alpha}_\epsilon^*)$, we conclude that $\Psi(\boldsymbol{\alpha}_\epsilon^*)=\boldsymbol{\alpha}_\epsilon^*$. Based on the equality $d\Omega_\epsilon(\boldsymbol{\alpha}|\,\boldsymbol{\alpha})=d\Omega_\epsilon(\boldsymbol{\alpha})$ again, we have $d\Omega_\epsilon(\boldsymbol{\alpha}_\epsilon^*)=0$. That is, $\boldsymbol{\alpha}_\epsilon^*=\hat{\boldsymbol{\alpha}}_\epsilon$ follows from the strict convexity of $\Omega_\epsilon(\boldsymbol{\alpha})$, which is proved as above. □

Proposition 2 (Convergence property of the algorithm (4)) *If $\hat{\boldsymbol{\alpha}}_\epsilon$ minimizes $\Omega_\epsilon(\boldsymbol{\alpha})$, then any limit point of $\{\hat{\boldsymbol{\alpha}}_\epsilon\}$ minimizes $\Omega(\boldsymbol{\alpha})$ as ϵ tends to zero. If $\Omega(\boldsymbol{\alpha})$ has a unique minimizer $\hat{\boldsymbol{\alpha}}$, then* $\lim_{\epsilon\to 0}\hat{\boldsymbol{\alpha}}_\epsilon=\hat{\boldsymbol{\alpha}}$.

The proof of Proposition 2 is omitted here, since it can be done following that of Proposition 4 of Hunter and Lange (2000).

4 Statistical theory

This section states our main results that provide upper bounds on the estimation error and the excess risk achieved by the estimator (2). Our main theorems are based on an appropriate adaptation to advanced empirical process theory under the non-parametric

settings. In contract to the parametric settings or spline-based approaches, our analysis is established from a more abstract form, since the RKHS is infinite dimensional. This requires us to make use of various concentration theorems, such as results on the covering number and the Rademacher complexity of kernel classes. In addition, another challenge technically comes from the quantile loss, which is non-smooth and non-quadratic. Most of the existing analysis tools concerning the quadratic loss are invalid for the quantile models under the RKHS framework.

For the theoretical analysis, we introduce some basis notations and assumptions as follows. First, the empirical covering number is needed to describe the functional complexity.

Definition 1 (*Empirical covering number with $L^2(\mathbb{Q}_n)$-norm*) Denote $\mathcal{G}$ a function space endowed with the empirical norm $\|\cdot\|_n$. For every $\epsilon > 0$, let $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_n)$ be the smallest integer N, such that there exists $\{g_j\}_{j=1}^N$ with

$$\sup_{g\in\mathcal{G}} \min_{j=1,2,\ldots,N} \|g - g_j\|_n \le \epsilon.$$

Then, $H(\mathcal{G}, \epsilon, \|\cdot\|_n) = \log \mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_n)$ is called the ϵ-empirical covering number of $\mathcal{G}$ for the empirical norm.

Note that $H(\mathcal{G}, \epsilon, \|\cdot\|_n)$ is a random quantity due to the empirical norm. We can define a deterministic entropy notation by taking expectation or supremum for $H(\mathcal{G}, \epsilon, \|\cdot\|_n)$. Without loss of generality, we assume that the following inequality holds with probability 1.

We are concerned with the subspace as

$$\mathcal{F} := \left\{ f = \sum_{j=1}^{p} f_j \,\middle|\, f_j \in \mathcal{H}_j, \quad \text{and } f_j \in \mathbb{B}_{\mathcal{H}_j} \right\},$$

where $\mathbb{B}_{\mathcal{H}_j}$ is the unit ball of $\mathcal{H}_j$ for any fixed j.

Assumption A1 The empirical covering number of $(\{f_j : \|f_j\|_{\mathcal{H}} \le 1\}, \|\cdot\|_n)$ is denoted by $H(\cdot)$. We assume that for all j

$$H(\epsilon) \le A\epsilon^{-2(1-\alpha)}, \quad \epsilon > 0, \tag{14}$$

where $0 < \alpha < 1$ and A are constants. There exist many classical positive kernels satisfying this assumption, including any space with finite VC-dimension, Sobolev/Besov classes, and Gaussian kernels. For example, one has $\alpha = 3/4$ for the Sobolev space with the second derivative.

Define the subset as

$$\mathcal{F}_S := \left\{ f = \sum_{j=1}^{p} f_j, \quad \text{satisfying } \sum_{j=1}^{p} \sqrt{\|f_j\|_n^2 + \rho_n \|f_j\|_{\mathcal{H}}^2} \le 4 \sum_{j\in S} \sqrt{\|f_j\|_n^2 + \rho_n \|f_j\|_{\mathcal{H}}^2}. \right\}.$$

Let $\widehat{\Delta} = \hat{f} - f^*$, and Proposition 3 below tells us that $\widehat{\Delta}$ belongs to $\mathcal{F}_S$ with a high probability. In fact, the similar results have been proved with respect to the least square approaches (Mernshausen and Yu 2009; Raskutti et al. 2012). Thus, it is sufficient to conduct our analysis over the restricted subset $\mathcal{F}_S$.

Assumption A2 There exist universal constants $C_1 > 0$ and $q \in (0, 2)$, such that for all $f \in \mathcal{F}_S$, one has

$$\sqrt{\mathcal{E}(f) - \mathcal{E}(f^*)} \geq C_1 \|f - f^*\|_q.$$

This assumption shows that the weak convergence induced by ρ_τ implies a more strong convergence. This has been verified under some mild conditions on the underlying distribution (Steinwart and Christmann 2011; Lv et al. 2016).

Let $\mu_n = \left(\frac{c_0 \log p}{2n}\right)^{\frac{1}{2(2-\alpha)}}$, and η is constant large sufficiently. We can establish the main results as follows.

Theorem 1 *Let $\hat{f}$ be the minimizer of the convex program (2) with regularization parameters $\lambda_n = \eta\mu_n$ and $\rho_n = \eta\mu_n^2$. When $p \geq 2\log n$ and Assumptions 1 − 2 both hold, then there exists two constants $N > 4$ and c_0, with probability at least $1 - c_0 \exp\left(-\frac{\log p}{c_0}\right) - 3p^{-N/2}$, we have*

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\sqrt{2} s \eta^{3/2} \mu_n \sqrt{1 + \mu_n^2},$$

and there also holds

$$\left\| \sum_{j=1}^{p} (\hat{f}_j - f_j^*) \right\|_q = \mathcal{O}\left(\frac{\log p}{n}\right)^{\frac{1}{4(2-\alpha)}}.$$

Note that, the first result measures the prediction ability on the penalized QR, and the second one measures estimation error accordingly. This non-asymptotic inequality is established under simpler conditions that is easy to interpret, compared with the existing literatures (Belloni and Chernozhukov 2011; Van der Geer 2008; Kato 2016). In particular, most existing oracle inequalities depend heavily on the dependency among the covariates, such as compatibility and irrepresentable conditions. It is still unclear whether these conditions hold in general.

In addition, despite its name known as "slow rate", this inequity has been shown in some cases to give faster rate of convergence than the more standard oracle rates (Van der Geer 2008). They are particularly helpful in situations, where these various assumptions imposed by the fast rate are hard to be verified, or would be quite difficult to interpret. It is also worth pointing out, since α is induced by empirical covering number, our regularization parameters are adaptive to the unknown distribution and the sparsity of the problem. This also shows that the LASSO is tuning insensitive, since the theoretical choice does not depend on the unknowns.

In this case of $p \geq \log n$, the proposed estimator can handle a nonpolynomially growing dimension of covariates as high as $\log p = o(n)$, while the dimension of the true sparse model grows as $s = o(n^{1/4})$ for all $0 < \alpha < 1$. We also notice that the existence of any moments is not required and allows for heavy-tailed distributions.

5 Simulation and real examples

In this section, we examine the effectiveness of the proposed method against some existing nonparametric methods in the literature, including Kato (2016) using a group lasso penalty, Xue (2009) assuming additive models for mean regression, Li et al. (2007) focusing on estimation of the quantile function in RKHS, and the unpenalized QR. For simplicity, we denote the aforementioned methods as KSQ, GLasso, Add, QR, and QR_0, respectively. The kernel function is set as radial basis kernel for all method, $K(s,t) = e^{-\|s-t\|^2/2\sigma^2}$, where σ^2 is set as the median of all the pairwise distances among the training sample (Jaakkola et al. 1999). Note that the performance of all methods rely on the value of tuning parameters, and thus, cross validation is used to estimate the validation error on a left-aside validation set. For all the methods, the tuning parameter is determined via a grid search for the smallest validation error. The grid for λ is set as $\lambda = 10^{-2+0.1s}$; $s = 0, \ldots, 40$. For KSQ, the grid is set as $\rho_n \in \{0.1, 0.3, 0.5, 0.7, 1\}$.

To compare the variable selection performance, a number of performance measures are used. Specifically, size, TP, and FP are used to represent the averaged number of selected informative variables, the number of truly informative variables selected and the number of truly non-informative variables selected, and C, U, and O are the times of correct-fitting, under-fitting, and over-fitting, respectively. Furthermore, the averaged test errors of each method based on the check loss are reported as well. Since QR_0 does not conduct variable selection, only its test errors are compared against other methods.

5.1 Simulated examples

The simulated examples are generated in the same fashion as in Yuan (2006) and Li et al. (2007), where different types of error distribution are examined. The true model is given as

$$y = 40\exp\left[8\left((x_1-.5)^2+(x_2-.5)^2\right)\right] \times \left(\exp\left[8\left((x_1-.2)^2+(x_2-.7)^2\right)\right]+\exp\left[8\left((x_1-.7)^2+(x_2-.2)^2\right)\right]\right)^{-1}+\epsilon,$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^p$ with $x_{ij} \sim \text{Uniform}(0,1)$ for any i and j. Three different error distributions are considered: (i) mixture normal distribution, $\epsilon \sim 0.1N(0,5^2)+0.9N(0,1)$; (ii) t-distribution, $\epsilon \sim t(3)$; and (iii) standard Laplace distribution, $\epsilon \sim Laplace(0,1)$.

Table 1 Averaged performance measures of various variable selection methods: $(n, p) = (100, 10)$

Type-error	Method	Size	TP	FP	C	U	O
Mixture-normal	KSQ	**2.020**	**2.000**	**0.020**	49	0	1
	GLasso	2.300	2.000	0.300	36	0	14
	ADD	2.260	2.000	0.260	41	0	9
	QR	10.000	2.000	8.000	0	0	50
Student-t	KSQ	**2.000**	**2.000**	**0.000**	50	0	0
	GLasso	2.420	2.000	0.420	35	0	15
	ADD	2.480	2.000	0.480	34	0	16
	QR	10.000	2.000	8.000	0	0	50
Double-exponential	KSQ	**2.040**	**2.000**	**0.040**	48	0	2
	GLasso	2.300	2.000	0.300	36	0	14
	ADD	2.240	2.000	0.240	41	0	9
	QR	10.000	2.000	8.000	0	0	50

The bold values represent the best performance

Table 2 Averaged performance measures of various variable selection methods: $(n, p) = (100, 20)$

Type-error	Method	Size	TP	FP	C	U	O
Mixture-normal	KSQ	**2.060**	**2.000**	**0.060**	47	0	3
	GLasso	2.480	2.000	0.480	36	0	14
	ADD	3.000	2.000	1.000	24	0	26
	QR	20.000	2.000	18.000	0	0	50
Student-t	KSQ	**2.060**	**2.000**	**0.060**	47	0	3
	GLasso	2.560	2.000	0.560	36	0	14
	ADD	2.640	2.000	0.640	28	0	22
	QR	20.000	2.000	18.000	0	0	50
Double-exponential	KSQ	**2.100**	**2.000**	**0.100**	46	0	4
	GLasso	2.380	2.000	0.380	36	0	14
	ADD	2.440	2.000	0.440	33	0	17
	QR	20.000	2.000	18.000	0	0	50

The bold values represent the best performance

We consider three scenarios with $(n, p) = (100, 10), (100, 20), (200, 30)$, where the validation set is of the same size as training set and the test set has 10000 observations. The quantile is fixed to be $\tau = 0.5$ in all scenarios. Each scenario is replicated 50 times, and the averaged performance measures are summarized in Tables 1, 2, and 3.

Furthermore, the test errors over 50 replications are summarized and displayed in Table 4 and a side-by-side boxplots for each scenario in Table 4 and Fig. 1, respectively.

It is evident that the proposed KSQ method has delivered superior numerical performance and outperforms other competitors in most scenarios. Specifically, KSQ yields better variable selection performance than both GLasso and Add. As Tables 1–3 showed, KSQ exactly selects the two informative variables almost every time,

Table 3 Averaged performance measures of various variable selection methods: $(n, p) = (200, 30)$

Type-error	Method	Size	TP	FP	C	U	O
Mixture-normal	KSQ	**2.000**	**2.000**	**0.000**	50	0	0
	GLasso	3.280	2.000	1.280	18	0	32
	ADD	2.880	2.000	0.880	26	0	24
	QR	30.000	2.000	28.000	0	0	50
Student-t	KSQ	**2.040**	**2.000**	**0.040**	48	0	2
	GLasso	5.240	2.000	3.240	10	0	40
	ADD	2.500	2.000	0.500	35	0	15
	QR	30.000	2.000	28.000	0	0	50
Double-exponential	KSQ	**2.040**	**2.000**	**0.040**	48	0	2
	GLasso	4.960	2.000	2.960	10	0	40
	ADD	3.160	2.000	1.160	18	0	32
	QR	30.000	2.000	28.000	0	0	50

The bold values represent the best performance

Table 4 Averaged test errors and standard errors of various methods in different scenarios

Type-error	Method	(100,10)	(100,20)	(200,30)
Mixture-normal	KSQ	7.09 (0.05)	7.07 (0.03)	7.08 (0.06)
	GLasso	7.15 (0.06)	7.11 (0.04)	7.14 (0.08)
	ADD	7.10 (0.05)	7.13 (0.05)	7.01 (0.04)
	QR	7.77 (0.05)	8.33 (0.07)	8.03 (0.07)
	QR_0	10.81 (0.43)	12.48 (0.31)	13.22 (0.52)
Student-t	KSQ	7.21 (0.05)	7.13 (0.05)	7.13 (0.05)
	GLasso	7.30 (0.05)	7.17 (0.06)	7.21 (0.06)
	ADD	7.18 (0.04)	7.17 (0.05)	7.17 (0.06)
	QR	7.84 (0.04)	8.37 (0.05)	8.13 (0.07)
	QR_0	12.53 (0.60)	11.94 (0.46)	12.90 (0.52)
Double-exponential	KSQ	7.16 (0.05)	7.13 (0.04)	7.18 (0.04)
	GLasso	7.18 (0.05)	7.14 (0.04)	7.26 (0.05)
	ADD	7.19 (0.04)	7.13 (0.07)	7.14 (0.05)
	QR	7.87 (0.05)	8.30 (0.05)	8.01 (0.05)
	QR_0	11.81 (0.46)	11.80 (0.40)	12.65 (0.48)

whereas both GLasso and Add tend to select more variables. QR and QR_0 focus on the estimation of the quantile function, and does not conduct variable selection. Furthermore, the test error of KSQ is also smaller than that of GLasso, Add, and QR, as showed in Fig. 1.

Note that the test errors of QR_0 are not plotted in Fig. 1, since its test errors are substantially larger than other competitors. In addition, as Add is developed based on

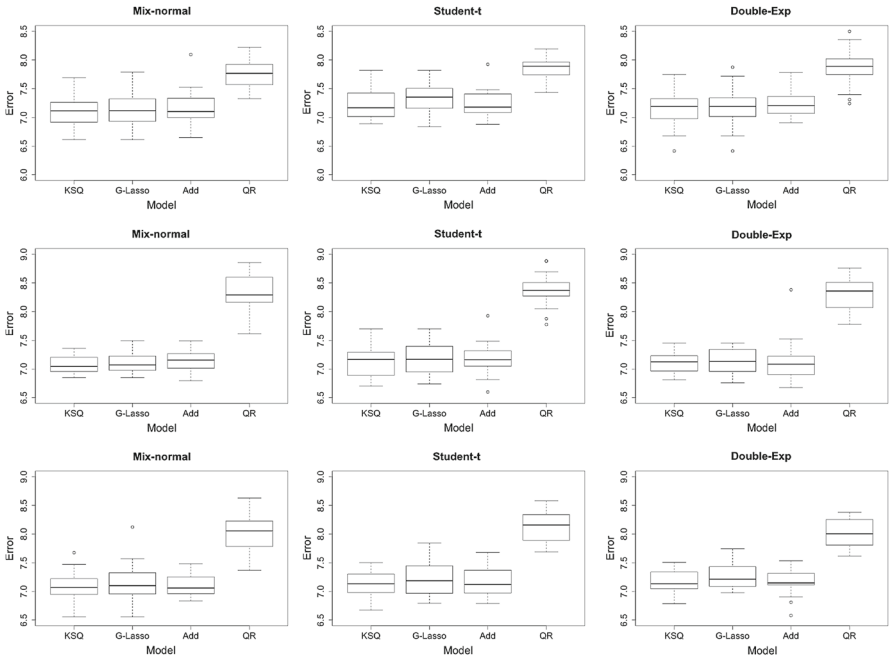

Fig. 1 Test errors of various methods in different scenarios. *Each row* represents a different scenario with $(n, p) = (100, 10)$, $(100, 20)$, or $(200, 30)$, respectively

mean regression, we refit the model with the selected informative variables by Add to calculate its corresponding test error with the check loss.

5.2 Japanese industrial chemical firm data

In this section, the proposed KSQ method is applied to analyze a real data set on Japanese industrial chemical firms (Lian 2012; Yafeh and Yosha 2003). The data set includes 186 Japanese industrial chemical firms listed on the Tokyo stock exchange, and the goal is to check whether the concentrated shareholding is associated with lower expenditure on activities with scope for managerial private benefits. The data set consists of a response variable MH5 (the general sales and administrative expenses deflated by sales), and 12 covariates: ASSETS (log(assets)), AGE (the age of the firm), LEVERAGE (ratio of debt to total assets), VARS (variance of operating profits to sales), OPERS (operating profits to sales), TOP10 (the percentage of ownership held by the 10 largest shareholders), TOP5 (the percentage of ownership held by the 5 largest shareholders), OWNIND (ownership Herfindahl index), AOLC (amount owed to largest creditor), SHARE (share of debt held by largest creditor), BDHIND (bank debt Herfindahl index), and BDA (bank debt to assets). The data set is available online through the Economic Journal at http://www.res.org.uk.

The data set is pre-processed by removing all the observations with missing values, and the response and the covariates are all standardized. The selected variables by

Table 5 Selected variables as well as the corresponding prediction errors by various selection methods in the Japanese industrial chemical firm data set

Variables	KSQ	GLasso	Add	QR
ASSETS	–	–	–	√
AGE	–	–	–	√
LEVERAGE	√	√	√	√
VARS	√	√	√	√
OPERS	√	√	√	√
TOP10	–	–	–	√
TOP5	–	–	–	√
OWNIND	–	–	√	√
AOLC	–	–	–	√
SHARE	√	√	√	√
BDHIND	√	√	√	√
BDA	–	–	–	√
Pred. err.	**0.283 (0.007)**	**0.283 (0.007)**	0.288 (0.007)	0.296 (0.008)

The bold values represent the best performance

various methods are reported in Table 5. To check the validity of the selected variables, we then randomly split the data set, with 20 observations for testing and 20 observations for validation, and the remaining are for training. The splitting is replicated 100 times, and the averaged prediction errors are reported in Table 5.

Both KSQ and GLasso select five informative variables, LEVERAGE, VARS, OPERS, SHARE, and BDHIND, whereas Add include one more variable OWNIND. The average prediction error of KSQ and GLasso is smaller than that of Add and QR, indicating that the prediction performance is improved with more non-informative variables screened. The prediction error of QR_0 is the same as that of QR and thus omitted here. Furthermore, a deviance test for the model with the five selected variables against the saturated model yields a p value 0.386, which concurs with the conclusion that the screened variables by KSQ are indeed statistically non-significant.

6 Conclusion

This article proposes a new sparsity-smoothness penalty over an RKHS in the high-dimensional additive quantile model, which allows for both the dimension p and the sparsity s to diverge with the sample size n. The proposed method provides a flexible modeling framework and includes many existing methods as its special cases. The resultant optimization task is tackled by an efficient computing algorithm, which couples the MM algorithm and the proximal gradient descent algorithm. More importantly, the oracle inequalities for the proposed estimators are provided under weak conditions. Numerical experiments on simulated and real examples are also supportive of the effectiveness of the proposed method. One potential future direction is to further reduce the computational cost. As the proposed method allows for a more

flexible modeling framework in RKHS and involves with a non-convex penalty term, its computational cost can be expensive. It is also of interest to extend the framework to a completely model-free scenario by relaxing the additive model assumption.

Acknowledgements SL's research is supported partially by NSFC-11301421, JBK141111, JBK14TD0046, JBK140210, and KLAS-130026507, and JW's research is supported partially by HK GRF-11302615 and CityU SRG-7004244. The authors also acknowledge Professor Fukumizu for providing an immensely hospitable and fruitful environment when SL visited ISM of Japan, and this work is partially supported by MEXT Grant-in-Aid for Scientific Research on Innovative Areas of Japan (25120012).

Appendix: Main Proofs

To simplify the proof, we only consider the special case where $\mu_\tau = 0$ in our model (1). Lemma 1 presents the behavior of weight empirical process (see Lemma 8.4 of Van der Geer (2000)).

Lemma 1 *Let $\mathcal{G}$ be a collection of functions $g : \{z_1, \ldots, z_n\} \to \mathbb{R}$, endowed with a metric induced by the norm $\|g\|_n$. Let $H(\cdot)$ be the entropy of $\mathcal{G}$. Suppose that*

$$H(\varepsilon) \leq A\varepsilon^{-2(1-\alpha)}, \quad \forall\, \varepsilon > 0,$$

where A is some constant and $\alpha \in (0, 1)$. In addition, let $\epsilon_1, \ldots, \epsilon_n$ be independent centered random variables, satisfying

$$\max_i \mathbb{E}[exp(\epsilon_i^2/L)] \leq M. \tag{15}$$

Denote $\langle \epsilon, g\rangle_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i g(z_i)$ with any given $g \in \mathcal{G}$, then for a constant c_0 depending on α, A, L, and M, we have for all $T \geq c_0$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{2\langle \epsilon, g\rangle_n}{\|g\|_n^\alpha} > \frac{T}{\sqrt{n}}\right) \leq c_0 \exp\left(-\frac{T^2}{c_0^2}\right).$$

According to Lemma 1, we can establish the following technical lemma, which tells us that the key quantity involved in empirical process can be bounded by the proposed regularization term. It turns out that the corresponding oracle rates are improved.

Lemma 2 *Under the same conditions of Lemma 1. Define the following event as*

$$\Theta := \left\{\forall j = 1, 2, \ldots, p \ \left|\langle \epsilon, f_j\rangle_n\right| \leq \mu_n\sqrt{\|f_j\|_n^2 + \mu_n^2\|f_j\|_{\mathcal{H}}}, \quad \text{for all } f_j \in \mathcal{H}\right\},$$

where c_0 is some universal constant, which may differ from that of Lemma 1. When $2\log p \geq c_0$, we have

$$\mathbb{P}(\Theta) \geq 1 - c_0 \exp\left(-\frac{\log p}{c_0}\right).$$

Proof Let $\mathcal{G} = \{g_j : \|g_j\|_{\mathcal{H}} = 1\}$ involved in Lemma 1. Then, applying Lemma 1, it follows that

$$\sup_{f_j} \frac{2\langle \epsilon, f_j \rangle_n}{\|f_j\|_n^{\alpha} \|f_j\|_{\mathcal{H}}^{1-\alpha}} = \sup_{f_j} \frac{2\langle \epsilon, f_j/\|f_j\|_{\mathcal{H}} \rangle_n}{\|f_j/\|f_j\|_{\mathcal{H}}\|_n^{\alpha}} \leq \frac{T}{\sqrt{n}}$$

with probability at least $1 - c_0 \exp(-T^2/c_0^2)$. Let $T = \sqrt{2c_0 \log p}$, and the assumption $2 \log p \geq c_0$ implies that $T \geq c_0$. Then, we have

$$\mathbb{P}\left(\max_j \sup_{f_j} \frac{2\langle \epsilon, f_j \rangle_n}{\|f_j\|_n^{\alpha} \|f_j\|_{\mathcal{H}}^{1-\alpha}} > \sqrt{\frac{2c_0 \log p}{n}} \right) \leq c_0 p \exp\left(-\frac{T^2}{c_0^2} \right) \leq c_0 \exp\left(-\frac{\log p}{c_0} \right).$$

In other words, with probability at least $1 - c_0 \exp\left(-\frac{\log p}{c_0}\right)$, there holds

$$\sup_{f_j \in \mathcal{H}} \frac{\langle \epsilon, f_j \rangle_n}{\|f_j\|_n^{\alpha} \|f_j\|_{\mathcal{H}}^{1-\alpha}} \leq \sqrt{\frac{c_0 \log p}{2n}}, \quad \text{for all } j \in \{1, 2, \ldots, p\}.$$

Thus, we derive our first desired conclusion for Θ based on the basic inequality:

$$x^{\alpha} y^{1-\alpha} \leq \sqrt{x^2 + y^2}, \quad \text{for any } \alpha \in (0, 1) \text{ and } x, y > 0.$$

□

Similar results on the Rademacher complexity and Gaussian complexity have been established in Koltchinskii and Yuan (2010) and Raskutti et al. (2012), respectively.

The next lemma shows that the quantities $\sum_{j=1}^{p} \sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n \|\widehat{\Delta}_j\|_{\mathcal{H}}^2}$ can be controlled by the corresponding one as applied to the active set S. They provide a way to prove sparsity oracle inequalities for the estimator (2).

Proposition 3 *Conditioned on the events Θ, with the choices of $\lambda_n \geq 2\mu_n$ and $\rho_n \geq \mu_n^2$, we have*

$$\sum_{j=1}^{p} \sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n \|\widehat{\Delta}_j\|_{\mathcal{H}}^2} \leq 4 \sum_{j \in S} \sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n \|\widehat{\Delta}_j\|_{\mathcal{H}}^2}.$$

Proof Define the functional

$$\widetilde{\mathcal{L}}(\Delta) = \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau} (\epsilon_i - \Delta(X_i)) + \lambda_n \sum_{j=1}^{p} \sqrt{\|f_j^* + \Delta_j\|_n^2 + \rho_n \|f_j^* + \Delta_j\|_{\mathcal{H}}^2}$$

and note that by definition of our M estimator, the error function $\widehat{\Delta} := \widehat{f} - f^*$ minimizes $\widetilde{\mathcal{L}}$. From the inequality $\widetilde{\mathcal{L}}(\widehat{\Delta}) \leq \widetilde{\mathcal{L}}(0)$, that is

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(\epsilon_i - \widehat{\Delta}(X_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(\epsilon_i\right)\\ &\quad\leq \lambda_n \sum_{j=1}^{p}\sqrt{\|f_j^*\|_n^2 + \rho_n\|f_j^*\|_{\mathcal{H}}^2} - \lambda_n\sum_{j=1}^{p}\sqrt{\|f_j^* + \widehat{\Delta}_j\|_n^2 + \rho_n\|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}^2}.\end{aligned} \tag{16}$$

Denote $a(t) = \tau - 1_{\{t\leq 0\}}(t)$. Recall that ρ_τ is a convex function and $a(t) \in \partial\rho_\tau(t)$, where $\partial\rho_\tau(t)$ is denoted to be the sub-gradient of ρ_τ at point t. By the definition of sub-gradient, we have

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(\epsilon_i - \widehat{\Delta}(X_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(\epsilon_i\right) \geq -\frac{1}{n}\sum_{i=1}^{n}a(\epsilon_i)\widehat{\Delta}(X_i). \tag{17}$$

This in connection with (16) shows that

$$\begin{aligned}&-\frac{1}{n}\sum_{i=1}^{n}a(\epsilon_i)\widehat{\Delta}(X_i)\\ &\quad\leq \lambda_n\sum_{j=1}^{p}\left(\sqrt{\|f_j^*\|_n^2 + \rho_n\|f_j^*\|_{\mathcal{H}}^2}. - \sqrt{\|f_j^* + \widehat{\Delta}_j\|_n^2 + \rho_n\|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}^2}\right).\end{aligned} \tag{18}$$

It is easy to check that $J_n(f_j) := \sqrt{\|f_j\|_n^2 + \rho_n\|f_j\|_{\mathcal{H}}^2}$ forms a standard mixed norm with any $f_j \in \mathcal{H}$. For any $j \in S$, by the triangle inequality with respect to the norm, we have

$$\sqrt{\|f_j^*\|_n^2 + \rho_n\|f_j^*\|_{\mathcal{H}}^2} - \sqrt{\|f_j^* + \widehat{\Delta}_j\|_n^2 + \rho_n\|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}^2} \leq \sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n\|\widehat{\Delta}_j\|_{\mathcal{H}}^2}.$$

On the other hand, for any $j \in S^c$, we have

$$\sqrt{\|f_j^*\|_n^2 + \rho_n\|f_j^*\|_{\mathcal{H}}^2} - \sqrt{\|f_j^* + \widehat{\Delta}_j\|_n^2 + \rho_n\|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}^2} = -\sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n\|\widehat{\Delta}_j\|_{\mathcal{H}}^2}.$$

This in connection with (18) implies that

$$-\frac{1}{n}\sum_{i=1}^{n}a(\epsilon_i)\widehat{\Delta}(X_i) \leq \lambda_n\sum_{j\in S}\sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n\|\widehat{\Delta}_j\|_{\mathcal{H}}^2} - \lambda_n\sum_{j\in S^c}\sqrt{\|\widehat{\Delta}_j\|_n^2 + \rho_n\|\widehat{\Delta}_j\|_{\mathcal{H}}^2}. \tag{19}$$

In addition, it is clear that $\{a(\epsilon_i)\}_{i=1}^n$ are bounded and independent variables with zero-mean, so the condition of (15) is satisfied. Thus, by Lemma 2 on Θ, one gets

$$\frac{1}{n}\sum_{i=1}^{n} a(\epsilon_i)\widehat{\Delta}(X_i) \le \mu_n \sum_{j=1}^{p} \sqrt{\|\widehat{\Delta}_j\|_n^2 + \mu_n^2 \|\widehat{\Delta}_j\|_{\mathcal{H}}^2},$$

with the choices of $\lambda_n \ge 2\mu_n$ and $\rho_n \ge \mu_n^2$, the above quantity is plugged into (19) to yield our desired result immediately. □

Now, we introduce the local Rademacher complexity, which is critical to our derived results. Given the bounded function class $\mathcal{G}$ with the star-shaped property [see Bartlett et al. (2005)], satisfying $\|g\|_\infty \le b (b \ge 1)$ for all $g \in \mathcal{G}$. Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of variables from X, drawn according to some distribution $\mathbb{Q}$. For each $a > 0$, we define the local Rademacher complexity:

$$\mathcal{R}_n(\mathcal{G}; a) := \mathbb{E}_{x,\sigma}\left[\sup_{g\in\mathcal{G}, \|g\|_2 \le a} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i g(x_i)\right|\right],$$

where $\{\sigma_i\}_{i=1}^n$ is an i.i.d. sequence of Rademcher variables, taking values $\{\pm 1\}$ with probability $1/2$. Denote v_n to be the smallest solution to the inequality:

$$\mathcal{R}_n(\{f_j : \|f_j\| \le 1\}; v_n) = \frac{v_n^2}{40}. \tag{20}$$

Note that such an v_n exists, since the star-shape property ensures that the function $\mathcal{R}_n(\mathcal{G}; a)/a$ is non-increasing in a.

Lemma 3 *For any $j \in \{1, 2, \ldots, p\}$, suppose that $\|f_j\|_\infty \le b$ for all $f_j \in \mathcal{H}$. For any $t \ge v_n$, define*

$$\mathrm{E}_j(t) := \left\{\frac{1}{2}\|f_j\|_2 \le \|f_j\|_n \le \frac{3}{2}\|f_j\|_2, \quad \text{for all } f_j \in \mathcal{H} \text{ with } \|f_j\|_2 \ge bt\right\}. \tag{21}$$

Denote $\mathrm{E}(t) := \bigcap_{j=1}^p \mathrm{E}_j(t)$. If $t \ge \sqrt{\frac{\log p}{n}}$ also holds, then there exist universal constants (c_1, c_2), such that

$$\mathbb{P}[\mathrm{E}(t)] \ge 1 - c_1 \exp(-c_2 n t^2).$$

To establish the relationship between α of empirical covering number and v_n of local Rademacher complexity, we need the following conclusion, showing that local Rademacher averages can be estimated by empirical covering numbers.

Lemma 4 *Let $\mathcal{G}$ be a class of measurable functions from X to $[-1, 1]$. Suppose that Assumption A1 holds for some $\alpha \in (0, 1)$. Then, there exists a constant c_α depending only on α, such that*

$$\mathcal{R}_n(\mathcal{H}; r) \le c_\alpha \max\left\{ r^\alpha \left(\frac{A}{n}\right)^{1/2}, \left(\frac{A}{n}\right)^{1(2-\alpha)} \right\}.$$

Furthermore, for the case of a single RKHS $\mathcal{H}$, we need the relationship between the empirical and $\|\cdot\|_2$ norms for function in $\mathcal{H}$. The following conclusion is derived immediately combining Theorem 4 of Koltchinskii and Yuan (2010) and Lemma 3 above.

Lemma 5 *Suppose that $N \ge 4$ and $p \ge 2\log n$. Then, there exists a universal constant $c > 0$, such that with probability at least $1 - p^{-N}$, for all $f \in \mathcal{H}$*

$$\|f\|_2 \le c(\|f\|_n + \mu_n \|f\|_{\mathcal{H}}),$$
$$\|f\|_n \le c(\|f\|_2 + \mu_n \|f\|_{\mathcal{H}}).$$

For any given $\Delta_-,\ \Delta_+ > 0$, we define the function subset of $\mathcal{F}$ as

$$\mathcal{F}(\Delta_-, \Delta_+) := \{f : \mu_n \|f - f^*\|_{2,1} \le \Delta_-, \mu_n^2 \|f - f^*\|_{\mathcal{H},1} \le \Delta_+\},$$

where $\|f\|_{2,1} = \sum_{j=1}^p \|f_j\|_2$ and $\|f\|_{\mathcal{H},1} = \sum_{j=1}^p \|f_j\|_{\mathcal{H}}$ for any $f = \sum_{j=1}^p f_j$. Equipped with this result, we can then prove a refined uniform convergence rate.

Proposition 4 *Let $\mathcal{F}(\Delta_-, \Delta_+)$ be a measurable function subset defined as above. Suppose that assumption (14) holds for each univariate $\mathcal{H}$. For some $N > 4$ involved in c_0, with confidence at least $1 - c_0 \exp\left(-\frac{\log p}{c_0}\right) - 2p^{-N/2}$, the following bound holds uniformly on $\Delta_- \le e^p$ and $\Delta_- \le e^p$:*

$$[\mathcal{E}(f) - \mathcal{E}(f^*)] - [\mathcal{E}_n(f) - \mathcal{E}_n(f^*)] \le c_1(\Delta_- + \Delta_+) + e^{-p}, \quad \forall\, f \in \mathcal{F}(\Delta_-, \Delta_+).$$

Proof of Theorem 1 By the definition of $\hat{f}$, it follows that

$$\mathcal{E}_n(\hat{f}) + \lambda_n \sum_{j=1}^p \sqrt{\|\hat{f}_j\|_n^2 + \rho_n \|\hat{f}_j\|_{\mathcal{H}}^2} \le \mathcal{E}_n(f^*) + \lambda_n \sum_{j=1}^p \sqrt{\|f_j^*\|_n^2 + \rho_n \|f_j^*\|_{\mathcal{H}}^2}.$$

This can be rewritten as

$$\begin{aligned}
&\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) + \lambda_n \sum_{j=1}^p \sqrt{\|\hat{f}_j\|_n^2 + \rho_n \|\hat{f}_j\|_{\mathcal{H}}^2} \\
&\le [\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)] - [\mathcal{E}_n(\hat{f}) - \mathcal{E}_n(f^*)] + \lambda_n \sum_{j=1}^p \sqrt{\|f_j^*\|_n^2 + \rho_n \|f_j^*\|_{\mathcal{H}}^2}.
\end{aligned}$$

By the triangle inequality, we get

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) + \lambda_n \sum_{j \in S^c} \sqrt{\|\hat{f}_j\|_n^2 + \rho_n \|\hat{f}_j\|_{\mathcal{H}}^2}$$

$$\leq [\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)] - [\mathcal{E}_n(\hat{f}) - \mathcal{E}_n(f^*)] + \lambda_n \sum_{j \in S} \sqrt{\|\hat{f}_j - f_j^*\|_n^2 + \rho_n \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2}. \tag{22}$$

Note that on $j \in S^c$, we have $\|\hat{f}_j\|_n = \|\hat{f}_j - f_j^*\|_n$ and $\|\hat{f}_j\|_{\mathcal{H}} = \|\hat{f}_j - f_j^*\|_{\mathcal{H}}$. $\sum_{j \in S} \sqrt{\|\hat{f}_j\|_n^2 + \rho_n \|\hat{f}_j\|_{\mathcal{H}}^2}$ is added to both the sides of (22), this implies that

$$\begin{aligned} &\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) + \lambda_n \sum_{j=1}^{p} \sqrt{\|\hat{f}_j - f_j^*\|_n^2 + \rho_n \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2} \\ &\leq [\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)] - [\mathcal{E}_n(\hat{f}) - \mathcal{E}_n(f^*)] \\ &\quad + 2\lambda_n \sum_{j \in S} \sqrt{\|\hat{f}_j - f_j^*\|_n^2 + \rho_n \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2}. \end{aligned} \tag{23}$$

Applying Lemma 5 for $\|\hat{f}_j - f_j^*\|_n$, $j = 1, \ldots, p$, with probability at least $1 - p^{-N}$, we have

$$\|\hat{f}_j - f_j^*\|_n^2 \geq c^{-2}/2 \|\hat{f}_j - f_j^*\|_2^2 - \mu_n^2 \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2.$$

When $\zeta > 2$ is satisfied, the quantity (23) can be further formulated as

$$\begin{aligned} &\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) + \lambda_n \sum_{j=1}^{p} \sqrt{c^{-2}/2 \|\hat{f}_j - f_j^*\|_2^2 + \rho_n/2 \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2} \\ &\leq [\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)] - [\mathcal{E}_n(\hat{f}) - \mathcal{E}_n(f^*)] \\ &\quad + 2\lambda_n \sum_{j \in S} \sqrt{\|\hat{f}_j - f_j^*\|_n^2 + \rho_n \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2}. \end{aligned}$$

We can claim that

$$\mu_n \|\hat{f} - f^*\|_{2,1} \leq e^p, \quad \mu_n^2 \|\hat{f} - f^*\|_{\mathcal{H},1} \leq e^p,$$

with probability 1. For simplicity, we only verify the first term. Note that $\|f_j\|_n \leq \|f_j\|_{\mathcal{H}} \leq 1$ for any $f_j \in \mathcal{H}$, and we see that

$$\mu_n \|\hat{f} - f^*\|_{2,1} \leq 2p \left(\frac{\log p}{n}\right)^{\frac{1}{2(2-\alpha)}} \leq 2p \left(\frac{\log p}{n}\right)^{\frac{1}{4}} \leq e^p, \quad \text{for all } n \geq 1,\ \alpha \in (0, 1).$$

This together Proposition 4 implies that, with probability at least $1-c_0\exp\left(-\frac{\log p}{c_0}\right)-3p^{-N/2}$

$$\begin{aligned}&\mathcal{E}\big(\hat{f}\big)-\mathcal{E}(f^*)+\lambda_n/\sqrt{2}\sum_{j=1}^{p}\sqrt{c^{-2}\|\hat{f}_j-f_j^*\|_2^2+\rho_n\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}\\&\le c_1\mu_n\sum_{j=1}^{p}\sqrt{\|\hat{f}_j-f_j^*\|_2^2+\mu_n^2\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}\\&\quad+e^{-p}+2\lambda_n\sum_{j\in S}\sqrt{\|\hat{f}_j-f_j^*\|_n^2+\rho_n\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}.\end{aligned}$$

Let η be large sufficiently, such that $\max\{2\sqrt{2}cc_1,1\}\le\eta$, then with the same probability as above, we have

$$\begin{aligned}&\mathcal{E}\big(\hat{f}\big)-\mathcal{E}(f^*)+\lambda_n/4\sum_{j=1}^{p}\sqrt{c^{-2}\|\hat{f}_j-f_j^*\|_2^2+\rho_n\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}\\&\le e^{-p}+2\lambda_n\sum_{j\in S}\sqrt{\|\hat{f}_j-f_j^*\|_n^2+\rho_n\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}.\end{aligned}\tag{24}$$

On the other hand, with the choices $\rho_n=\eta\mu_n$ and $\lambda_n^2=\eta\mu_n^2$, it follows that

$$\lambda_n\sum_{j\in S}\sqrt{\|\hat{f}_j-f_j^*\|_2^2+\rho_n\|\hat{f}_j-f_j^*\|_{\mathcal{H}}^2}\le 4\sqrt{2}s\eta^{3/2}\mu_n\sqrt{1+\mu_n^2},$$

where we used the fact $\|f_j\|_n\le\|f_j\|_{\mathcal{H}}\le 1$ for any $f_j\in\mathcal{H}$, $j=1,\ldots,p$. Plugging the above quantity into the right side of (24) yields

$$\mathcal{E}\big(\hat{f}\big)-\mathcal{E}(f^*)\le 4\sqrt{2}s\eta^{3/2}\mu_n\sqrt{1+\mu_n^2}+e^{-p}.$$

It is verified easily that $p\ge\log n$ implies that $e^{-p}\le 4\sqrt{2}s\eta^{3/2}\mu_n\sqrt{1+\mu_n^2}$; then, we have

$$\mathcal{E}\big(\hat{f}\big)-\mathcal{E}(f^*)\le 8\sqrt{2}s\eta^{3/2}\mu_n\sqrt{1+\mu_n^2}.$$

□

References

Bartlett, P. L., Bousquet, O., Mendelson, S. (2005). Local Rademacher complexities. *Annals of Statistics*, *33*, 1497–1537.

Beck, A., Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*, 183–202.

Belloni, A., Chernozhukov, V. (2011). ℓ_1 penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, *39*, 83–130.
Combettes, P., Wajs, V. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, *4*, 1168–1200.
Donoho, D. L., Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*, 1200–1224.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Hastie, T., Tibshirani, R. (1990). *Monographs on Statistics and Applied Probability, Generalized Additive Models (1st ed.)*, London: Chapman and Hall.
He, X. M., Wang, L., Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, *41*, 324–369.
Hunter, D. R., Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, *11*, 60–77.
Jaakkola, T., Diekhans, M., Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *In Proceedings of Seventh International Conference on Intelligent Systems for Molecular Biology*, 149–158.
Kato, K. (2016). Group Lasso for high dimensional sparse quantile regression models. Manuscript.
Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, *25*, 239–262.
Koenker, R., Basset, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.
Koltchinskii, V., Yuan, M. (2008). Sparse recovery in large ensembles of kerenl machines. In: 21st Annual Conference on Learning Theory, Helsinki, 229–238.
Koltchinskii, V., Yuan, M. (2010). Sparsity in multiple kernel learning. *Annals of Statistics*, *38*, 3660–3695.
Li, Y., Zhu, J. (2008). l^1-norm quantile regressions. *Journal of Computational and Graphical Statistics*, *17*, 163–185.
Li, Y., Liu, Y., Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, *102*, 255–268.
Lian, H. (2012). Estimation of additive quantile regression models by two-fold penalty. *Journal of Business and Economic Statistics*, *30*, 337–350.
Lv, S. G., Lin, H. Z., Lian, H., Huang, J. (2016). Oracle inequalities for sparse additive quantile regression models in reproducing kernel Hilbert space. Manuscript.
Meier, L., Van der Geer, S., Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, *37*, 3779–3821.
Mernshausen, N., Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, *37*, 246–270.
Moreau, J. J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *Reports of the Paris Academy of Sciences, Series A*, *255*, 2897–2899.
Mosci, S., Rosasco, L., Santoro, M., Verri, A., Villa, S. (2010). Solving structured sparsity regularization with proximal methods. *Machine Learning and Knowledge Discovery in Databases*, *6322*, 418–433.
Negahban, S., Ravikumar, P., Wainwright, M. J., Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, *27*, 538–557.
Pearce, N. D., Wand, M. P. (2006). Penalized splines and reproducing kernel methods. *The American Statistician*, *60*, 233–240.
Raskutti, G., Wainwright, M., Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, *13*, 389–427.
Ravikumar, P., Liu, H., Lafferty, J., Wasserman, L. (2009). SpAM: Sparse additive models. *Journal of the Royal Statistical Society: Series B*, *71*, 1009–1030.
Rosasco, L., Villa, S., Mosci, S., Santoro, M., Verri, A. (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, *14*, 1665–1714.
Steinwart, I., Christmann, A. (2011). Estimating conditional quantiles with the help of pinball loss. *Bernoulli*, *17*, 211–225.
Tseng, P. (2010). Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, *125*, 263–295.
Van der Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge: Cambridge University Press.
Van der Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, *36*, 614–645.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. *Advances in kernel methods: support vector learning*, pp. 69–88.

Wang, L., Wu, Y. C., Li, R. Z. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, *107*, 214–222.

Xue, L. (2009). Consistent variable selection in additive models. *Statistical Science*, *19*, 1281–1296.

Yafeh, Y., Yosha, O. (2003). Large Shareholders and banks: Who monitors and how? *The Economic Journal*, *113*, 128–146.

Yuan, M. (2006). GACV for Quantile Smoothing Splines. *Computational Statistics and Data Analysis*, *5*, 813–829.