Electronic Journal of Statistics
Vol. 15 (2021) 1610–1635
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1824

Learning sparse conditional distribution: An efficient kernel-based approach*

Fang Chen and Xin He†

School of Statistics and Management
Shanghai University of Finance and Economics
e-mail: fchen@163.sufe.edu.cn; he.xin17@mail.shufe.edu.cn

Junhui Wang

School of Data Science
City University of Hong Kong
e-mail: j.h.wang@cityu.edu.hk

Abstract: This paper proposes a novel method to recover the sparse structure of the conditional distribution, which plays a crucial role in subsequent statistical analysis such as prediction, forecasting, conditional distribution estimation and others. Unlike most existing methods that often require explicit model assumption or suffer from computational burden, the proposed method shows great advantage by making use of some desirable properties of reproducing kernel Hilbert space (RKHS). It can be efficiently implemented by optimizing its dual form and is particularly attractive in dealing with large-scale dataset. The asymptotic consistencies of the proposed method are established under mild conditions. Its effectiveness is also supported by a variety of simulated examples and a real-life supermarket dataset from Northern China.

MSC2020 subject classifications: Primary 68Q32, 62G08; secondary 62J07.
Keywords and phrases: Conditional distribution, consistency, parallel computing, RKHS, sparse learning.

Received April 2020.

Contents

*Xin He's research is supported in part by NSFC-11901375, Shanghai Pujiang Program 2019PJC051 and the Fundamental Research Funds for the Central Universities. Junhui Wang's research is supported in part by HK RGC Grants GRF-11303918, GRF-11300919 and GRF-11304520.
†Xin He is the corresponding author.

1. Introduction

In modern statistical analysis, the study on the conditional distribution and the relevant issues have attracted more and more attention [12, 13, 17, 6, 42] with many applications, including in economics and finance [5, 46], in prediction and forecasting for time series analysis [10] and the inference on conditional distribution estimation [42]. Yet, as pointed out by [13], a conventional nonparametric estimator of the conditional distribution will suffer poor accuracy, even the number of covariates collected is relatively small. Hence, in high dimensional data analysis, the sparse modelling is demanded in the sense that it is generally believed that only a few covariates truly affect the conditional distribution. And thus it is crucial to identify truly informative covariates acting on the conditional distribution as the first step for subsequent statistical analysis.

This paper proposes a novel and efficient method to exactly identify all the covariates acting on the conditional distribution. The proposed method is motivated by the key observations that joint quantile regressions can be used as an efficient tool to exploit the conditional distribution and the gradient functions provide an appropriate definition of the informative covariates without explicit model specification. More importantly, unlike most existing learning gradients methods, which estimate the gradients under the regularization framework [28, 48, 14] and thus suffer computational burden, we notice that the derivative reproducing property in RKHS [52] provides an efficient alternative for fast computation of the gradient functions. Specifically, the proposed method first fits joint kernel-based quantile regressions, followed by the fast computation of corresponding gradients by using the derivative reproducing property in RKHS and a hard-thresholding procedure. The proposed method can be efficiently implemented by dual optimizations and is particularly attractive in dealing with large-scale cases. The asymptotic consistency of the proposed method is established without requiring any explicit model specification under mild conditions and the numerical experiments illustrate the superior performance of the proposed method against some state-of-the-art competitors.

The major contributions of the proposed method are four-fold: (i) It avoids directly estimating the gradient functions but efficiently computes the estimated gradient functions by using the derivative property in RKHS, which significantly reduces the computational cost; (ii) It only requires standard quadratic optimization with linear constraints, which can be efficiently solved by some existing packages in many statistical softwares. Besides, it can be applied to deal with big data with slight modification by using some distributed platform such as Apache Spark with Map-reduce steps. (iii) It is as efficient as the screening procedure but the strong marginal correlation condition is not required. Specifically, it can be regarded as a nonparametric joint screening method in the sense that each gradient function is computed given all the other covariates and computation procedures can be done in a parallel fashion. (iv) With the help of functional operators in learning theory, the asymptotic selection consistency is established under mild condition, which ensures that all the covariates acting on the conditional distribution can be exactly identified with probability tending to one.

1.1. Related works

In literature, sparse learning methods have been proposed to recovery the dependence relationship between the covariates and the conditional distribution under certain assumptions. One popularly used assumption assumes that the covariates act on the conditional distribution only through the conditional mean function. Various strategies have been developed to propose sparse learning methods under the linear modelling, including sparse-induced regularization [40, 7, 53, 31, 32], sure independence screening [8, 45], and knockoff filter [2]. Extended methods have also been developed under the nonparametric additive model [23, 16, 9], or in the reproducing kernel Hilbert space (RKHS) [28, 48]. Yet these methods require explicit model assumptions that are difficult to check in practice, or expensive computational cost that undermines its scalability. It is also interesting to point out that the aforementioned methods only focus on modelling the mean of the conditional distribution.

Beyond conditional mean regression, many methods have also been developed to detect a more general dependence between the covariates and the conditional distribution [47, 25, 15, 14] through individual or composite quantile regression [20]. Specifically, [15] proposes a nonparametric screening method to retain all the covariates acting on the conditional quantile function at some given quantile level. It can be extended to retain all the covariates acting on the conditional distribution by estimating the marginal quantile utility at multiple quantile levels. Yet, the screening-based method aims to retain a large number of covariates to guarantee the sure screening property, which is often much larger than the number of truly informative covariates in sparse learning. [14] proposes a learning-gradient-based method to identify all the covariates acting on the conditional distribution, and establishes the asymptotic selection consistency. It learns the gradient of conditional quantile functions at multiple quantile levels in a RKHS, and employs a functional group lasso penalty in the formulated

regularization framework. It is also interesting to notice that [28] proposes a novel learning gradients method, which adds an empirical functional penalty on the gradients to a standard kernel ridge regression in a RKHS, and it can be further extended to recover the sparse structure of the conditional distribution.

1.2. Paper organization

The rest of this paper is organized as follows. Section 2 provides the background and some key motivations and Section 3 introduces the proposed method. The computational algorithm and detail of tuning procedure are provided in Section 4. The asymptotic consistencies of the proposed method are established under mild conditions in Section 5. Section 6 reports the numerical experiments on the simulated and real examples. A brief summary is provided in Section 7, and all the technical proofs are given in Appendix.

2. Motivation and background

2.1. Motivation

Given a random pair $\mathcal{Z} = (\mathbf{x}, y)$ drawn from some unknown distribution $\rho_{\mathbf{x},y}$ with covariates $\mathbf{x} = (x_1, ..., x_p)^T \in \mathcal{X}$, where $\mathcal{X} \subset \mathcal{R}^p$ is assumed to be a compact set, and response $y \in \mathcal{R}$. It is of great interest to estimate the conditional distribution function

$$F_y(\mathbf{x}) = P(Y \leq y|\mathbf{x}),$$

which has been widely studied in literature [12, 13, 17, 6, 42]. Yet, many existing methods are developed for the low-dimensional case, and yield suboptimal performance when the number of covariates becomes relatively large [12]. In the high-dimensional case, it is generally believed that only a small number of covariates have effects on $F_y(\mathbf{x})$, while the others are noise. Thus, it is crucial to first identify all the truly informative covariates acting on $F_y(\mathbf{x})$ for subsequent statistical analysis. To this end, we regard a covariate x_l as noise if

$$F_y(\mathbf{x}) = F_y(\mathbf{x}_{-l}), \tag{2.1}$$

where $\mathbf{x}_{-l}$ denotes all the covariates except x_l. This criterion implies that a noise covariate x_l does not have any functional relationship with the conditional distribution function $F_y(\mathbf{x})$.

As pointed out by [21], $F_y(\mathbf{x})$ and the conditional quantile function $Q^*_\tau(\mathbf{x})$ are equivalent characterizations of the conditional distribution of y given $\mathbf{x}$. Precisely, $F_y(\mathbf{x})$ can be uniquely quantified by $Q^*_\tau(\mathbf{x})$ as $Q^*_\tau(\mathbf{x}) = \inf_{y\in\mathcal{R}}\{y : F(y|\,\mathbf{x}) \geq \tau\}$, and vice versa. In the rest of this paper, we require $Q^*_\tau \in \mathcal{H}_K$ for any $\tau \in (0, 1)$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) induced by some pre-specified two-times differentiable kernel $K(\cdot, \cdot)$. Consequently, the criterion in (2.1) is equivalent to $Q^*_\tau(\mathbf{x}) = Q^*_\tau(\mathbf{x}_{-l})$ for any $\tau \in (0, 1)$, and thus we only need to check whether the covariate x_l has any functional relationship

with $Q^*_\tau(\mathbf{x})$ for any $\tau \in (0,1)$. Particularly, it suffices to check whether the corresponding gradient functions

$$g^*_{l,\tau}(\mathbf{x}) = \partial Q^*_\tau(\mathbf{x})/\partial x_l = 0$$

almost surely for any $\mathbf{x} \in \mathcal{X}$ and $\tau \in (0,1)$. Then, the importance of each covariate can be measured via its $\mathcal{L}^2$-norm that

$$\|g^*_{l,\tau}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_\mathbf{x})} = \int_\mathcal{X} (g^*_{l,\tau}(\mathbf{x}))^2 \mathrm{d}\rho_\mathbf{x}(\mathbf{x}), \quad \text{for any } \tau \in (0,1),$$

where $\rho_\mathbf{x}$ denotes the marginal distribution of $\mathbf{x}$. Then, the true active set $\mathcal{A}^*$, which contains all the covariates acting on the conditional distribution, can be defined as

$$\mathcal{A}^* = \left\{ l : \|g^*_{l,\tau}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_\mathbf{x})} > 0, \text{ for any } \tau \in (0,1) \right\}.$$

More importantly, we notice that the derivative reproducing property [52] in RKHS that

$$g^*_{l,\tau}(\mathbf{x}) = \partial Q^*_\tau(\mathbf{x})/\partial x_l = \langle Q^*_\tau, \partial_l K_\mathbf{x} \rangle_K, \tag{2.2}$$

where $K_\mathbf{x}(\cdot) := K(\mathbf{x}, \cdot)$, ensures that once a consistent estimator of Q^*_τ is obtained, its gradient function $g^*_{l,\tau}$ can be efficiently obtained by simply matrix multiplication.

These key facts motivate us to recover the sparse structure of the conditional distribution by employing a two-step learning algorithm that we first obtain some consistent estimator of Q^*_τ, and then apply the derivative reproducing property in (2.2) to compute the empirical norm of the estimated gradient function g^*_τ and check whether it is substantially different from 0.

2.2. Reproducing kernel

We briefly introduce some basic knowledge of RKHS and interested readers may refer to [30] for more details about RKHS. Precisely, let $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ be a bounded, symmetric and positive semidefinite function in that $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} K(\mathbf{x},\mathbf{x}') < \infty$. Then, the RKHS $\mathcal{H}_K$ associated with the kernel $K(\cdot,\cdot)$ is the completion of the linear span of functions $\{K_\mathbf{x}(\cdot) := K(\mathbf{x},\cdot) : \mathbf{x} \in \mathcal{X}\}$ with the inner product given by $\langle K_\mathbf{x}, K_\mathbf{u} \rangle_K = K(\mathbf{x},\mathbf{u})$ for any $\mathbf{x},\mathbf{u} \in \mathcal{X}$. Note that the RKHS $\mathcal{H}_K$ is uniquely determined by the kernel $K(\cdot,\cdot)$. By Mercer's theorem [26], under some regularity conditions, the kernel function has an eigen-expansion that $K(\mathbf{x},\mathbf{u}) = \sum_{k=1}^\infty \mu_k \phi_k(\mathbf{x}) \phi_k(\mathbf{u})$, where $\mu_1 \geq \mu_2 \geq ... \geq 0$ are a non-negative sequence of eigenvalues and $\{\phi_k\}_{k=1}^\infty$ are the associated eigenfunctions, taken to be orthonormal in $\mathcal{L}^2(\mathcal{X},\rho_\mathbf{x}) = \left\{Q : \int_\mathcal{X} Q^2(\mathbf{x}) \mathrm{d}\rho_\mathbf{x}(\mathbf{x}) < \infty\right\}$. Moreover, for any $Q \in \mathcal{H}_K$, we have $Q(\mathbf{x}) = \sum_{k=1}^\infty a_k \phi_k(\mathbf{x})$, where $a_k = \langle Q, \phi_k \rangle_{\mathcal{L}^2(\mathcal{X},\rho_\mathbf{x})} = \int_\mathcal{X} Q(\mathbf{x}) \phi_k(\mathbf{x}) \mathrm{d}\rho_\mathbf{x}(\mathbf{x})$ are the Fourier coefficients, $\|Q\|_K = \sum_{j \geq 1} \frac{\langle Q, \phi_j \rangle^2_2}{\mu_j}$. Note that the above results require that $\mathcal{H}_K \subset \mathcal{L}^2(\mathcal{X},\rho_\mathbf{x})$ and it

is directly satisfied if $\sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x},\mathbf{x})$ is assumed to be finite. It is interesting to notice that the finiteness requirement is equivalent to the boundness of the kernel.

More importantly, the reproducing property of RKHS is critical in both theoretical analysis and computation, which states that $\langle Q, K_{\mathbf{x}}\rangle_K = Q(\mathbf{x})$ for any $Q \in \mathcal{H}_K$. It is worth pointing out that the RKHS induced by some universal kernel, such as the Gaussian kernel, is a fairly large functional space in the sense that any continuous function can be arbitrarily well approximated by an intermediate function in its induced RKHS under the infinity norm [36].

3. Methodology

3.1. Sparsity learning

Given the random sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which are copies of $\mathcal{Z} = (\mathbf{x}, y)$, we first solve the following optimization task that

$$\underset{Q_{\tau_1},\ldots,Q_{\tau_m}\in\mathcal{H}_K}{\operatorname{argmin}} \frac{1}{nm}\sum_{k=1}^{m}\sum_{i=1}^{n} L_{\tau_k}(y_i - Q_{\tau_k}(\mathbf{x}_i)) + \frac{\lambda}{m}\sum_{k=1}^{m}\|Q_{\tau_k}\|_K^2, \tag{3.1}$$

where the first term is the sample version of $\frac{1}{m}\sum_{k=1}^m \mathcal{E}(Q_{\tau_k}) = \frac{1}{m}\sum_{k=1}^m EL_{\tau_k}(y - Q_{\tau_k}(\mathbf{x}))$ with $L_\tau(u) = u(\tau - I_{\{u\le 0\}})$ denoting the check loss function, $\tau_1,\ldots,\tau_m$ denote m quantile levels, and λ denotes the parameter controlling the model complexity. In practice, $\tau_1,\ldots,\tau_m$ can be chosen equidistantly from the interval $(0,1)$ to exploit the skeleton of the conditional distribution. Moreover, the number of quantile levels m and the covariate dimension p are both allowed to diverge with n. For simplicity, we suppress their dependence on n and still use m and p, and denote the cardinality of $\mathcal{A}^*$ as $|\mathcal{A}^*| = p_0 \ll p$. Note that the optimization task (3.1) resembles the joint quantile regression (JQR) [29] without imposing the (shape) non-crossing constraints to facilitate the computation.

By the representer theorem [43], the solution of (3.1) must have a finite form that

$$\widehat{Q}_{\tau_k}(\mathbf{x}) = \sum_{i=1}^{n}\widehat{\alpha}_i^k K(\mathbf{x}_i,\mathbf{x}) = \widehat{\boldsymbol{\alpha}}_{\tau_k}^T\mathbf{K}_n(\mathbf{x}), \quad \text{for } k = 1,\ldots,m, \tag{3.2}$$

where $\widehat{\boldsymbol{\alpha}}_{\tau_k} = (\widehat{\alpha}_1^k,\ldots,\widehat{\alpha}_n^k)^T \in \mathcal{R}^n$ denotes the representer coefficients and $\mathbf{K}_n(\cdot) = (K(\mathbf{x}_1,\cdot),\ldots,K(\mathbf{x}_n,\cdot))^T$ denotes the kernel vector with transposition $(\cdot)^T$. It is worth pointing out that the representer theorem [43] converts the original optimization problem (3.1) over an infinite functional space $\mathcal{H}_K$ into an optimization problem over a finite n-dimensional vector space of $\boldsymbol{\alpha}_{\tau_k}$. Thus, solving the optimization task (3.1) is equivalent to solving

$$\widehat{\boldsymbol{\alpha}}_{\tau_1},\ldots,\widehat{\boldsymbol{\alpha}}_{\tau_m} = \underset{\substack{\boldsymbol{\alpha}_{\tau_1},\ldots,\boldsymbol{\alpha}_{\tau_m}\\ \in\mathcal{R}^n}}{\operatorname{argmin}} \frac{1}{nm}\sum_{k=1}^{m}\sum_{i=1}^{n} L_{\tau_k}(y_i - \boldsymbol{\alpha}_{\tau_k}^T\mathbf{K}_n(\mathbf{x}_i)) + \frac{\lambda}{m}\sum_{k=1}^{m}\boldsymbol{\alpha}_{\tau_k}^T\mathbf{K}\boldsymbol{\alpha}_{\tau_k}, \tag{3.3}$$

where $\mathbf{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathcal{R}^{n\times n}$ denotes the kernel matrix.

Once $\widehat{\boldsymbol{\alpha}}_{\tau_1}, ..., \widehat{\boldsymbol{\alpha}}_{\tau_m}$ are obtained, the estimated gradient function $\widehat{g}_{l,\tau_k}$, $l = 1, ..., p$, can be directly computed as

$$\widehat{g}_{l,\tau_k}(\mathbf{x}) = \frac{\partial \widehat{Q}_{\tau_k}(\mathbf{x})}{\partial x_l} = \widehat{\boldsymbol{\alpha}}_{\tau_k}^T \partial_l \mathbf{K}_n(\mathbf{x}), \tag{3.4}$$

where $\partial_l \mathbf{K}_n(\mathbf{x}) = (\frac{\partial K(\mathbf{x}_1,\mathbf{x})}{\partial x_l}, \cdots, \frac{\partial K(\mathbf{x}_n,\mathbf{x})}{\partial x_l})^T$ is given once the kernel $K(\cdot,\cdot)$ is pre-specified. It is interesting to point out that we only need to solve (3.3) for $\widehat{\boldsymbol{\alpha}}_{\tau_k}, k = 1, ..., m$ one time, and the estimated gradient functions can be directly computed by using the derivative reproducing property in (3.4) without incurring any additional computational cost.

In practice, since the marginal distribution $\rho_{\mathbf{x}}$ is usually unknown, we adopt the empirical norm as a reasonable substitute for the $\mathcal{L}^2$-norm. Thus, the estimated informative set is defined as

$$\widehat{\mathcal{A}}_{v_n} = \left\{ l : \frac{1}{m} \sum_{k=1}^{m} \|\widehat{g}_{l,\tau_k}\|_n^2 > v_n \right\},$$

where $\|\widehat{g}_{l,\tau_k}\|_n^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{g}_{l,\tau_k}(\mathbf{x}_i))^2$ denotes the empirical norm and v_n represents some pre-specified thresholding value.

Note that the proposed method is computationally efficient in the sense that it only requires to solve the convex optimization task (3.3) for $\widehat{Q}_{\tau_k}(\mathbf{x})$, and its gradients can be analytically computed as in (3.4). Once the gradients are obtained, sparse learning can be done by comparing its empirical norm with some given thresholding values. It is clear that the selection performance of the proposed method relies on the choice of the thresholding value v_n, which can be appropriately determined through a stability-based selection criterion [38]. More details of the employed criterion are provided in Section 4.2.

3.2. Parallel and distributed computing

In this section, we illustrate that the proposed method is particularly attractive and useful in dealing with the large-scale cases, where the dimension p or the sample size n is extremely large. When the dimension p is extremely large that $p > n$, the proposed method only needs to estimate the n-dimensional representer coefficients $\boldsymbol{\alpha}_{\tau_k}, k = 1, ..., m$, and the estimated gradient function can be computed in a parallel fashion. Specifically, the proposed method only needs to estimate $\boldsymbol{\alpha}_{\tau_k}, k = 1, ..., m$ by solving the optimization task (3.3), where the computational complexity is only related with n and m. Once the estimated coefficients $\widehat{\boldsymbol{\alpha}}_{\tau_k}, k = 1, ..., m$ are obtained, the empirical norm of the estimated gradient functions can be directly computed as

$$\|\widehat{g}_{l,\tau_k}\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\boldsymbol{\alpha}}_{\tau_k}^T \partial_l \mathbf{K}_n(\mathbf{x}_i) \right)^2,$$

for each $l = 1, ..., p$ simultaneously. Finally, given a pre-specified v_n, the empirical norm of the estimated gradient functions, $\frac{1}{m}\sum_{k=1}^{m} \|\widehat{g}_{l,\tau_k}\|_n^2$, can be truncated by a hard-thresholding step, and thus the estimated informative set $\widehat{\mathcal{A}}_{v_n}$ is obtained.

When the sample size n is extremely large, the proposed method can be applied to deal with the large n case by using the idea of the divide-and-conquer method [51] with slight modification. Specifically, the divide-and-conquer method usually assumes that the sample can be partitioned into several disjoint subsets and each subset is stored in a local machine. Therefore, a local estimator can be obtained in each local machine, and then these local estimators are communicated to a central processor and a global estimator is synthesized by taking an average. Figure 1 illustrates the idea of the divide-and-conquer method. It is worth pointing out that the proposed method only requires to communicate the p-dimensional vectors of the empirical norms to the central processor, which is quite computationally efficient with the assistance of some distributed platform, such as Apache Spark with Map-reduce steps. The readers may refer to [19] for detailed introduction of Apache Spark.

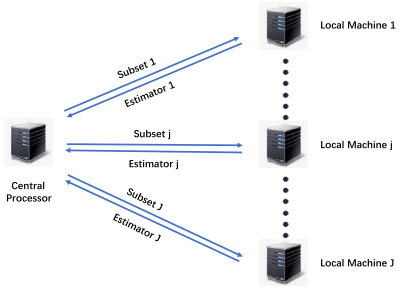

FIG 1. *Illustration of the divide-and-conquer method.*

Without loss of generality, we assume that the sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ can be exactly divided into the J disjoint subsets $\mathcal{Z}_1^n, \ldots, \mathcal{Z}_J^n$ at random that $J \mid n$, and each subset $\mathcal{Z}_j^n$ contains n/J observations stored in the distributed local machines. For the local machine j, the joint kernel-based quantile regressions (3.3) are fitted with the subset $\mathcal{Z}_j^n$ to estimate the local representer coefficients $(\widehat{\boldsymbol{\alpha}}_{\tau_1}^{(j)}, ..., \widehat{\boldsymbol{\alpha}}_{\tau_m}^{(j)})$ by solving

$$\min_{\boldsymbol{\alpha}_{\tau_1}^{(j)},\ldots,\boldsymbol{\alpha}_{\tau_m}^{(j)}} \frac{1}{m|\mathcal{Z}_j^n|} \sum_{k=1}^{m} \sum_{\mathbf{x}_i \in \mathcal{Z}_j^n} L_{\tau_k}(y_i - (\boldsymbol{\alpha}_{\tau_k}^{(j)})^T \mathbf{K}_n(\mathbf{x}_i)) + \frac{\lambda}{m} \sum_{k=1}^{m} (\boldsymbol{\alpha}_{\tau_k}^{(j)})^T \mathbf{K} \boldsymbol{\alpha}_{\tau_k}^{(j)},$$

where $|\mathcal{Z}_j^n|$ denotes the cardinality of the subset $\mathcal{Z}_j^n$. Then, the empirical norm of the estimated gradient functions based on $\mathcal{Z}_j^n$ can be computed as

$$\|\widehat{g}_l^{(j)}\|_n^2 = \frac{1}{m|\mathcal{Z}_j^n|} \sum_{k=1}^{m} \sum_{\mathbf{x}_i \in \mathcal{Z}_j^n} \left((\widehat{\boldsymbol{\alpha}}_{\tau_k}^{(j)})^T \partial_l \mathbf{K}_n(\mathbf{x}_i) \right)^2,$$

for each $l = 1, ..., p$. Note that the above procedure can be done simultaneously for each distributed local machine. Once the estimated gradients in each local machine are obtained, we only need to communicate $\left(\|\widehat{g}_1^{(j)}\|_n^2, ..., \|\widehat{g}_p^{(j)}\|_n^2\right)^T \in \mathcal{R}^p$ to the central processor, and then the empirical norm with respect to the whole sample $\mathcal{Z}^n$ can be computed as

$$\|\widehat{g}_l\|_n^2 = \frac{1}{J}\sum_{j=1}^{J} \|\widehat{g}_l^{(j)}\|_n^2,$$

for each $l = 1, ..., p$. Finally, given a pre-specified thresholding value v_n, the estimated informative set is $\widehat{\mathcal{A}}_{v_n} = \left\{l : \|\widehat{g}_l\|_n^2 > v_n\right\}$.

As illustrated above, we employ the standard divide-and-conquer scheme to facilitate the estimation of the quantile functions in (3.1) and their gradient functions when the sample size is very large. It is interesting to notice that the numerical efficiency and performance can be further improved by employing the distributed scheme in [18], where the nondifferentiable objective function is replaced by a smoothing approximation, leading to an analytic solution, and then the corresponding statistics can be computed locally and finally aggregated together to complete the calculation of the derived estimator. Note that the emphasis of this paper is on developing an efficient method to learn the sparsity structure of the conditional distribution, and thus we leave the detailed investigation on more sophisticated schemes for the distributed computation as the future work.

4. Computational issues

The computational detail for updating $\boldsymbol{\alpha}_{\tau_k}, k = 1, ..., m$, in (3.3) and the stability-based tuning procedure for the optimal choice of the thresholding value v_n are provided in this section.

4.1. The dual optimization

To solve (3.3), it is equivalent to solving its dual problem at each quantile level. Precisely, the dual optimization task can be computed straightforwardly by using Lagrange multipliers, and thus solving (3.3) is equivalent to solving a quadratic optimization with linear constraints that for each $k = 1, ..., m$,

$$\begin{aligned} &\min_{\boldsymbol{\alpha}_{\tau_k} \in \mathcal{R}^n} \frac{1}{2}\boldsymbol{\alpha}_{\tau_k}^T \mathbf{K}\, \boldsymbol{\alpha}_{\tau_k} - \boldsymbol{\alpha}_{\tau_k}^T \mathbf{y}, \\ &\text{s.t. } C(\tau_k - 1) \le \alpha_i \le C\tau_k, \quad \forall \ 1 \le i \le n, \text{ and } \quad \mathbf{1}_n^T \boldsymbol{\alpha}_{\tau_k} = 0, \end{aligned} \tag{4.1}$$

where C denotes some tuning parameter with $C = \frac{1}{n\lambda}$ and $\mathbf{1}_n = (1, ..., 1)^T \in \mathcal{R}^n$. Moreover, the optimization task (4.1) can be rewritten as

$$\min_{\boldsymbol{\alpha}_{\tau_k} \in \mathcal{R}^n} \frac{1}{2}\boldsymbol{\alpha}_{\tau_k}^T \mathbf{K}\, \boldsymbol{\alpha}_{\tau_k} - \boldsymbol{\alpha}_{\tau_k}^T \mathbf{y}, \quad \text{s.t. } \mathbf{A}^T \boldsymbol{\alpha}_{\tau_k} \geq_1 \mathbf{B}_k, \tag{4.2}$$

where $\mathbf{A} = (\mathbf{1}_n, \mathbf{I}_n, -\mathbf{I}_n) \in \mathcal{R}^{n\times(2n+1)}$, $\mathbf{I}_n = \text{diag}(1, ..., 1) \in \mathcal{R}^{n\times n}$, $\geq_1$ means that the first constraint is equality and the others are inequalities, and $\mathbf{B}_k = (0, C(\tau_k - 1)\mathbf{1}_n^T, -C\tau_k\mathbf{1}_n^T)^T \in \mathcal{R}^{2n+1}$. Note that the optimization task (4.2) can be efficiently solved by some commonly used package in many statistical softwares, such as the "quadprog" package in R or "qpsolvers" package in Python.

Note that the optimization task (4.2) can be done in a parallel fashion at each quantile level, and thus the estimation procedure is computationally efficient and scalable. This computational efficiency is largely due to the fact that the (shape) non-crossing constraints [29, 1] are not enforced in (3.1). With the non-crossing constraints enforced, one may except the numerical performance can be further improved, at the cost the increased computational burden. In the rest of this paper, we illustrate the proposed method by fitting JQR as introduced in Section 3.1 without enforcing the (shape) non-crossing constraints, and it yields satisfactory numerical performance in all the numerical examples in Section 6.

4.2. Tuning procedure

To determine the thresholding parameter v_n, we employ the stability-based criterion [38] to search the optimal value of v_n. Its key idea is to measure the stability of the proposed method by randomly splitting the training sample into two parts, and comparing the disagreement between the two estimated active sets.

Specifically, given a value v_n, we randomly split the training sample $\mathcal{Z}^n$ into two parts $\mathcal{Z}_1^n$ and $\mathcal{Z}_2^n$ and apply the proposed method to $\mathcal{Z}_1^n$ and $\mathcal{Z}_2^n$, to obtain two estimated active sets $\widehat{\mathcal{A}}_{1,v_n}$ and $\widehat{\mathcal{A}}_{2,v_n}$, respectively. Then, the disagreement between $\widehat{\mathcal{A}}_{1,v_n}$ and $\widehat{\mathcal{A}}_{2,v_n}$ is measured by Cohen's kappa coefficient

$$\kappa(\widehat{\mathcal{A}}_{1,v_n}, \widehat{\mathcal{A}}_{2,v_n}) = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a) = \frac{n_{11}+n_{22}}{p}$ and $Pr(e) = \frac{(n_{11}+n_{12})(n_{11}+n_{21})}{p^2} + \frac{(n_{12}+n_{22})(n_{21}+n_{22})}{p^2}$ with $n_{11} = |\widehat{\mathcal{A}}_{1,v_n} \cap \widehat{\mathcal{A}}_{2,v_n}|, n_{12} = |\widehat{\mathcal{A}}_{1,v_n} \cap \widehat{\mathcal{A}}_{2,v_n}^C|, n_{21} = |\widehat{\mathcal{A}}_{1,v_n}^C \cap \widehat{\mathcal{A}}_{2,v_n}|, n_{22} = |\widehat{\mathcal{A}}_{1,v_n}^C \cap \widehat{\mathcal{A}}_{2,v_n}^C|$ and $|\cdot|$ denotes the set cardinality. The procedure is repeated B times and the estimated sparse learning stability is measured as

$$\hat{s}(\Psi_{v_n}) = \frac{1}{B}\sum_{b=1}^{B} \kappa(\widehat{\mathcal{A}}_{1,v_n}^b, \widehat{\mathcal{A}}_{2,v_n}^b).$$

Finally, the parameter is set as $\widehat{v}_n = \max\left\{v_n \in \mathcal{R}^{\geq 0} : \frac{\hat{s}(\Psi_{v_n})}{\max_{v_n} \hat{s}(\Psi_{v_n})} \geq q\right\}$, where $q \in (0, 1)$ is some given percentage and $\mathcal{R}^{\geq 0}$ denotes the set of non-negative real numbers.

5. Asymptotic results

In this section, we establish the asymptotic results for the proposed method and for simplicity, we define $Q_\tau^* = \text{argmin}_{Q_\tau \in \mathcal{B}_\tau} \|Q_\tau\|_K^2$ with $\mathcal{B}_\tau = \{Q_\tau : Q_\tau =$

$\text{argmin}_{h_\tau \in \mathcal{H}_K} \mathcal{E}_{\mathcal{Z}^n}(h_\tau)\}$ to ensure the uniqueness of the minimizer Q_τ^* and denote $\widetilde{Q}_\tau = \text{argmin}_{Q_\tau \in \mathcal{H}_K} \mathcal{E}(Q_\tau) + \lambda \|Q_\tau\|_K^2$. The following technical conditions are needed to establish the estimation consistency.

Condition 1. There exist some positive constants κ_1 and κ_2 such that for any $l = 1, ..., p$, $\sup_{\mathbf{x} \in \mathcal{X}} \|K_\mathbf{x}\|_K \leq \kappa_1$ and $\sup_{\mathbf{x} \in \mathcal{X}} \|\partial_l K_\mathbf{x}\|_K \leq \kappa_2$.

Condition 2. There exist some positive constants c_1 and θ_1 such that there holds $\max_{k=1,...,m} \|\widetilde{Q}_{\tau_k} - Q^*_{\tau_k}\|_K = c_1 \lambda^{\theta_1}$.

Condition 1 imposes a boundedness condition on the kernel function as well as its gradient functions, which is commonly used in statistical learning literature [28, 50, 14] and satisfied by many popular kernels, including the Gaussian kernel, Sobolev kernel and the scaled linear kernel with the compact support condition on $\mathcal{X}$. Note that the compact support condition is usually assumed in machine learning literature [28, 14, 24] and also is regularly imposed for universality and the Mercer's theorem. Recently, many efforts have been made to extend them to the non-compact setting and interested readers may refer to [35, 37, 33]. Condition 2 controls the approximation error in the sense that $\lim_{\lambda \to 0} \|\widetilde{Q}_{\tau_k} - Q^*_{\tau_k}\|_K = 0$. Note that similar condition quantifying the approximation error is commonly used in statistical learning literature [28, 49].

Now we turn to establish the asymptotic estimation consistency of the proposed method.

Theorem 1. *Suppose Conditions 1–2 are satisfied. For any* $\delta > 4(\log n)^{-2} E(y^2)$, *with probability at least* $1 - \frac{\delta}{2}$, *there holds*

$$\frac{1}{m} \sum_{k=1}^{m} \|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K \leq c_3 \left(\log \frac{8}{\delta}\right)^{\frac{1}{4}} (\log n)^{1/2} \frac{(1 + \kappa_1 \lambda^{-\frac{1}{2}})^{\frac{1}{2}}}{\lambda^{\frac{1}{2}} n^{\frac{1}{4}}},$$

where $c_3 = 2\sqrt{2} \max\{1, \kappa_1\}$.

Theorem 1 provides the strong convergence result of the difference between $\widehat{Q}_{\tau_k}$ and $\widetilde{Q}_{\tau_k}$, which plays a crucial role in establishing the estimation consistency of the estimated gradient functions. Note that in machine learning literature [34, 28, 22], it is usually assumed that $|y| < M$, where M denotes some positive constant, for mathematical simplicity. When this bounded response condition is imposed, the upper bound in Theorem 1 reduces to $c_3 M \left(\log \frac{8}{\delta}\right)^{\frac{1}{4}} \frac{(1+\kappa_1 \lambda^{-\frac{1}{2}})^{\frac{1}{2}}}{\lambda^{\frac{1}{2}} n^{\frac{1}{4}}}$ for any $\delta \in (0, 1)$.

Theorem 2. *Suppose that all the conditions in Theorem 1 are satisfied. Let* $\lambda = n^{-\frac{1}{4}}$, *for some positive constant* c_4, *with probability at least* $1 - \delta$, *we have*

$$\max_{l=1,...,p} \frac{1}{m} \sum_{k=1}^{m} \left| \|\widehat{g}_{l,\tau_k}\|_n^2 - \|g^*_{l,\tau_k}\|^2_{\mathcal{L}^2(\mathcal{X}, \rho_\mathbf{x})} \right| \leq c_4 \left(\log \frac{4p}{\delta}\right)^{\frac{1}{2}} (\log n)^{1/2} n^{-\Theta},$$

where $c_4 = \max\left\{2c_3 \kappa_1^{1/2}, c_2, 2\sqrt{2}\kappa_2^2\right\} \max\left\{\kappa_2^2, \kappa_2^2 \max_k \|Q^*_{\tau_k}\|_K, \max_k \|Q^*_{\tau_k}\|_K^2\right\}$ *and* $\Theta = \min\{\frac{1}{16}, \frac{\theta_1}{4}\}$.

Theorem 2 establishes the convergence rate of the difference between the estimated gradient functions and the true gradient functions and this desired result is crucial to establish the asymptotic selection consistency of the proposed method. Note that the convergence result allows p diverging but as pointed out by [11] that the dependency is generally difficult to quantify explicitly. Further conditions are assumed to establish the selection consistency for the proposed method.

Condition 3. For some positive constants c_5 and ζ_1, the true gradient functions satisfy that for any $\tau, \tau' \in (0,1)$ and $l = 1, ..., p$,

$$\sup_{\mathbf{x}\in\mathcal{X}} |g^*_{l,\tau}(\mathbf{x}) - g^*_{l,\tau'}(\mathbf{x})| \leq c_5|\tau - \tau'|^{\zeta_1}.$$

Condition 4. For any $l \in \mathcal{A}^*$, there exists some quantile level $\tau_0^l \in (0,1)$ such that $\|g^*_{l,\tau_0^l}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} \geq c_6 \left(\log \frac{4p}{\delta}\right)^{\xi} (\log n)^{1/2} n^{-\Theta}$ for some $c_6 > 0$ and $\xi > 1/2$, and $\min_{k=1,...,m} |\tau_k - \tau_0^l| \xrightarrow{m\to\infty} 0$.

Condition 3 is a Lipschitz continuity condition, which quantifies the smoothness of the true gradient functions. Similar conditions are imposed in [4] for parametric case and [14] for nonparametric case. The first part of Condition 4 requires the true gradient function contains sufficient information about the truly informative covariates at some quantile level, which is commonly used in nonparametric modelling and is much tighter than many existing nonparametric methods [16, 48]. The second part of Condition 4 imposes the condition on the choice of quantile levels and is naturally satisfied for the equidistant quantile levels, as well as some other more sophisticated designed quantile levels. Now we establish the asymptotic consistency of the proposed method.

Theorem 3. *Suppose that all the conditions in Theorem 2 as well as Conditions 3 and 4 are satisfied. Let* $v_n = \frac{c_6}{2} \left(\log \frac{4p}{\delta}\right)^{\xi} (\log n)^{1/2} n^{-\Theta}$, *then we have*

$$P(\widehat{\mathcal{A}} = \mathcal{A}^*) \xrightarrow{n,m\to\infty} 1. \tag{5.1}$$

Theorem 3 shows that with the diverging of sample size n and the number of quantile levels m, the selected informative set can exactly recover the true active set with probability tending to 1. This result is particularly interesting given the fact that it is established without any explicit model assumption and exactly identifies all the truly informative covariates which act on the conditional distribution in any pattern.

6. Numerical experiments

In this section, we study the numerical performance of the proposed method, denoted as MF, and compare it against some state-of-the-art methods, including the distance correlation learning (DC, [39]) and the quantile-adaptive screening (QaSIS, [15]). Note that the original DC and QaSIS methods are designed to keep the first $[n/\log n]$ covariates to achieve the sure screening property, and thus we further truncate them by using some thresholding value to conduct

sparse learning and report the truncated results. It is worth pointing out that the methods considered in [14] are computationally demanding and can only work when the dimension is relatively small. Thus, we omit the numerical comparison of the proposed method with all the methods considered in [14].

In all the simulated scenarios, we apply the Gaussian kernel, $K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\| \mathbf{u}-\mathbf{v} \|_2^2}{2\sigma_n^2}\right)$, for MF and σ_n is set as the median of all the pairwise distances among the training sample. As the choice of the thresholding value highly affects the performance of all the compared methods, we apply the stability-based selection criterion introduced in Section 4.2 to determine it. The maximization of the stability criterion is conducted via a grid search, where the grid is set as $\{10^{-3+0.1s} : s = 0, ..., 60\}$.

6.1. Simulated examples

The numerical performance of MF and its competitors is evaluated under two simulated examples, which are also considered in [14]. In the both simulated examples, the quantile levels are set as $\boldsymbol{\tau} = (0.1, 0.25, 0.75, 0.9)$ for MF and $\tau = 0.75$ for QaSIS.

Example 1: (Nonlinear model) We first generate $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently drawn from $U(-0.5, 0.5)$. The data are generated as $f^*(\mathbf{x}_i) = 6f_1(x_{i1}) + 4f_2(x_{i2})f_3(x_{i3}) + 6f_4(x_{i4}) + 5f_5(x_{i5})$ with $f_1(u) = u, f_2(u) = (2u + 1), f_3(u) = (2u - 1), f_4(u) = 0.1\sin(2\pi u) + 0.2\cos(2\pi u) + 0.3(2\sin(\pi u))^2 + 0.4(\cos(2\pi u))^3 + 0.5(2\sin(\pi u))^3$ and $f_5(u) = \frac{\sin(\pi u)}{(2-\sin(\pi u))}$, and ϵ_i's are independently drawn from $N(0, 1)$.

Example 2: (Heterogeneous model) The generating scheme is the same as in Example 1 except that W_{ij} and U_i are independently drawn from $U(0, 1)$ and the response y_i is generated as $y_i = 4x_{i1}x_{i2} + 3|x_{i3}|\epsilon_i$.

Clearly, only the mean of the conditional distribution relies on the covariates in Example 1, whereas in Example 2, the covariates act on the conditional distribution through the mean function and the error term. For the both examples, we consider the scenarios that $(n, p) = (200, 100), (400, 100), (400, 1000), (400, 4000)$ and the correlation structure among each covariates is considered by setting $\eta = 0, 0.3$, and 0.8 for each scenario. Specifically, when $\eta = 0$, the covariates are completely independent, whereas when $\eta = 0.3$ and 0.8, correlation structure among the covariates are added. Each scenario is replicated 200 times and the averaged performance measures are summarized in Tables 1–2, where Size is the averaged number of selected informative covariates, TP is the number of truly informative covariates selected, FP is the number of truly non-informative covariates selected, and C, U, O are the times of correct-fitting, under-fitting, and over-fitting, respectively.

As shown in Tables 1 and 2, MF outperforms the both competitors in most scenarios. In Example 1, MF is able to identify all the five truly informative covariates in most replications. However, DC and QaSIS tend to miss some truly informative covariates. In Example 2, MF shows a much larger advantage

TABLE 1
The averaged performance measures of MF and its competitors in Example 1.

(n, p, η)	Method	Size	TP	FP	C	U	O
(200, 100,0)	MF	5.07	4.97	0.10	174	7	19
	QaSIS	2.80	2.79	0.01	18	182	0
	DC	2.36	2.35	0.01	95	104	1
(400,100,0)	MF	5.22	5.00	0.22	159	0	41
	QaSIS	4.25	4.25	0.00	101	99	0
	DC	4.84	4.84	0.00	173	27	0
(400,1000,0)	MF	5.01	4.99	0.02	196	1	3
	QaSIS	3.70	3.70	0.00	52	148	0
	DC	4.71	4.71	0.00	153	47	0
(400, 4000,0)	MF	5.04	5.00	0.04	192	0	8
	QaSIS	3.47	3.47	0.00	37	163	0
	DC	4.50	4.50	0.00	125	75	0
(200, 100,0.3)	MF	5.39	4.99	0.40	136	1	63
	QaSIS	2.43	2.41	0.02	0	200	0
	DC	3.49	3.48	0.01	40	159	1
(400,100,0.3)	MF	5.16	5.00	0.16	172	0	28
	QaSIS	3.42	3.41	0.01	6	194	0
	DC	4.40	4.40	0.00	122	78	0
(400,1000,0.3)	MF	5.00	4.99	0.01	196	2	2
	QaSIS	2.93	2.93	0.00	2	198	0
	DC	3.85	3.85	0.00	68	132	0
(400, 4000,0.3)	MF	5.11	5.00	0.11	179	0	21
	QaSIS	2.72	2.72	0.00	0	200	0
	DC	3.46	3.46	0.00	44	156	0
(200, 100,0.8)	MF	6.34	4.99	1.35	52	3	145
	QaSIS	2.44	2.25	0.19	0	200	0
	DC	2.92	2.90	0.02	0	200	0
(400, 100,0.8)	MF	4.93	4.59	0.34	118	80	2
	QaSIS	3.37	3.34	0.03	0	200	0
	DC	3.39	3.39	0.00	0	200	0
(400,1000,0.8)	MF	5.03	4.99	0.04	190	2	8
	QaSIS	2.55	2.55	0.00	0	200	0
	DC	3.03	3.03	0.00	0	200	0
(400, 4000,0.8)	MF	5.07	4.98	0.09	179	5	16
	QaSIS	2.19	2.18	0.01	0	200	0
	DC	2.78	2.78	0.00	0	200	0

against the both competitors. The two competitors tend to miss x_3 which affects the response through the variance, while MF is still able to identify x_3 in most replications and tends to overfit by including some noise covariates, which is much less severe than missing the important ones. In the both simulated scenarios with $\eta = 0.3$ and 0.8, the added correlation structure increases the difficulty of identifying the informative covariates, and here MF also outperforms its competitors in most scenarios.

We also report the computational cost of MF under different scenarios in Table 3 to illustrate the remarkable computational efficiency of MF, which supports our claim that MF is particularly useful to deal with large-scale data. Note that all the simulations are done by a computing machine with CPU Intel Xeon 5117.

TABLE 2
The averaged performance measures of MF and its competitors in Example 2.

(n, p, η)	Method	Size	TP	FP	C	U	O
(200, 100,0)	MF	3.08	2.90	0.18	148	20	32
	QaSIS	1.55	1.26	0.29	8	188	4
	DC	2.06	2.04	0.02	59	140	1
(400,100,0)	MF	2.99	2.99	0.00	199	1	0
	QaSIS	1.97	1.96	0.01	32	167	1
	DC	2.71	2.71	0.00	152	48	0
(400,1000,0)	MF	3.32	3.00	0.32	149	0	51
	QaSIS	1.56	1.51	0.01	9	191	0
	DC	2.36	2.35	0.01	95	104	1
(400, 4000,0)	MF	3.93	2.99	0.94	80	1	119
	QaSIS	1.39	1.39	0.00	6	194	0
	DC	2.08	2.08	0.00	65	135	0
(200, 100,0.3)	MF	3.52	2.79	0.73	81	40	79
	QaSIS	1.72	1.08	0.64	6	183	11
	DC	1.70	1.65	0.05	22	175	3
(400,100,0.3)	MF	2.99	2.99	0.00	197	3	0
	QaSIS	2.01	1.84	0.17	36	153	11
	DC	2.46	2.46	0.00	114	86	0
(400,1000,0.3)	MF	3.61	2.93	0.69	100	15	85
	QaSIS	1.15	1.05	0.10	3	197	0
	DC	1.91	1.90	0.01	42	157	1
(400, 4000,0.3)	MF	5.01	2.90	2.11	39	21	140
	QaSIS	1.09	0.70	0.39	1	198	1
	DC	1.61	1.59	0.02	27	172	1
(200, 100,0.8)	MF	7.11	2.62	4.49	11	72	117
	QaSIS	5.67	0.96	4.71	0	178	22
	DC	3.39	1.05	2.34	3	182	15
(400, 100,0.8)	MF	4.13	2.82	1.31	93	36	71
	QaSIS	3.69	1.40	2.29	10	171	19
	DC	2.22	1.72	0.50	19	168	13
(400,1000,0.8)	MF	4.86	2.63	2.23	31	66	103
	QaSIS	14.55	0.78	13.77	0	174	26
	DC	8.17	1.15	7.02	0	184	16
(400, 4000,0.8)	MF	18.31	2.59	15.72	4	71	125
	QaSIS	13.68	0.43	13.25	0	193	7
	DC	13.41	1.01	12.40	2	172	26

TABLE 3
The computational cost of MF under different settings.

(n, p)	(200, 100)	(400, 100)	(400, 1000)	(400, 4000)
Example 1	2.52s	20.73s	1.07mins	5.18mins
Example 2	3.03s	18.01s	1.10mins	4.88mins

6.2. Real-data analysis

In this section, MF and its competitors are applied to a supermarket dataset [44], which is collected from a major supermarket located in northern China,

TABLE 4
The number of selected covariates as well as the corresponding averaged prediction errors by all the competitors in the supermarket dataset.

Method	Size	Pred. error (std.)
MF	22	0.378 (0.018)
DC	20	0.379 (0.019)
QaSIS	15	0.386 (0.021)

TABLE 5
Selected products by MF in the supermarket dataset (∗ means the product's name is missing in the original dataset).

Serial Number	Name	Serial Number	Name
79060095	Totole	90042221	∗
90050213	Korean side dishes	90048026	∗
90110009	∗	90130031	Staple food
79030026	XianDao fructose	79020028	XianDao soy sauce
70050035	Coco Cola	76080380	Chewing gum
90052222	LiuPinXiang dessert	73050024	Refined white sugar
73050002	Rice	76010761	∗
79050120	Sweet noodles	90040001	Eggs
78010066	Powdered milk	90059374	∗
90100009	Trotters with sauce	79050615	Hand-pulled Noodle
73090019	Red date without stone		

consisting of daily sale records of $p = 6,398$ products on $n = 464$ days. This data include almost all kinds of daily necessities and the response is the number of customers on each day, and the covariates are the daily sale volumes of each product. The supermarket wants to know which product's sale volumes are highly related with the number of customers, and then some specific sale strategies based on those products can be designed to attract more customers.

Both the response and covariates are pre-processed and thus have zero mean and unit variance. We apply MF with $\boldsymbol{\tau} = (0.1, 0.25, 0.5, 0.75, 0.9)$, QaSIS with $\tau = 0.75$ and DC to the supermarket data. Note that the truly informative covariates are unknown in the real-data analysis, and thus we report the prediction performance of each method by randomly splitting the dataset into two parts, with 300 observations for training and the remaining for testing. We refit kernel-based quantile regressions in (3.3) with the selected covariates for each method on the training set, and measure the prediction performance on the testing set. The splitting procedure is repeated 1000 times and the numerical performance is summarized in Table 4.

From Table 4, MF selects 22 products, whereas DC selects 20 products and QaSIS selects 15 products. The averaged prediction error of MF is smaller than those of DC and QaSIS, implying that these two methods may miss some important products. We also report the products selected by MF in Table 5.

Clearly, MF selects the products that "Totole", "Coco Cola", "Eggs" , "Rice", "Sweet noodles", "Trotters with sauce" and others. This result suggests that

many customers are more willing to buy these necessities, and thus some special sale strategies based on the selected products can attract more customers. It is interesting to point out that the product "Korean side dishes" is also selected by MF, which is not surprising since the supermarket is located in northern China, and its food culture is more or less similar as South Korea.

7. Summary

This paper proposes an efficient kernel-based method to recovery the sparse structure of the conditional distribution, which can be regarded as a crucial step for the subsequent statistical analysis. The proposed method focuses on identifying all the covariates acting on the conditional distribution by taking advantages of the nice properties of RKHS. The implementation of the proposed method is computationally efficient by using dual optimization, and is particularly useful to deal with large-scale cases. The asymptotic consistency of the proposed method is established without requiring any explicit model conditions and the numerical experiments illustrate the superior performance of the proposed method against some state-of-the-art competitors.

Appendix: Technical proofs

Given the condition $|y| \leq M_n$, we consider the functional space

$$\mathcal{F}_{M_n} = \left\{ \mathbf{Q} = (Q_{\tau_1}, ..., Q_{\tau_m}) \ \text{ with } Q_{\tau_k} \in \mathcal{H}_K : \frac{1}{m} \sum_{k=1}^{m} \|Q_{\tau_k}\|_K^2 \leq \lambda^{-1} M_n \right\}.$$

Note that $\mathcal{F}_{M_n}$ is fairly large in the sense that the minimizer of (3.1), denoted as $\widehat{\mathbf{Q}} = (\widehat{Q}_{\tau_1}, ..., \widehat{Q}_{\tau_m})$, is contained in $\mathcal{F}_{M_n}$ by the fact that $\frac{1}{m}\sum_{k=1}^{m}(\mathcal{E}_{\mathcal{Z}^n}(\widehat{Q}_{\tau_k}) + \lambda\|\widehat{Q}_{\tau_k}\|_K^2) \leq \frac{1}{m}\sum_{k=1}^{m}(\mathcal{E}_{\mathcal{Z}^n}(0) + \lambda\|0\|_K^2) \leq \max_{i=1,...,n}|y_i| \leq M_n$. We also denote

$$\mathcal{S}(\mathcal{Z}^n, M_n) = \sup_{\mathbf{Q} \in \mathcal{F}_{M_n}} \frac{1}{m} \sum_{k=1}^{m} |\mathcal{E}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k})|.$$

Now we bound $\mathcal{S}(\mathcal{Z}^n, M_n)$ by applying McDiarmid's inequality.

Lemma 1. *(McDiarmid's Inequality) Let $z_1, ..., z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $f : \mathcal{Z}^n \to \mathcal{R}$ satisfies*

$$\sup_{z_1,...,z_n,z_i' \in \mathcal{Z}} |f(z_1, ..., z_n) - f(z_1, ..., z_i', ..., z_n)| \leq C_i,$$

for every $i \in \{1, 2, ..., n\}$. Then, for every $t > 0$,

$$P\left(|f(z_1, ..., z_n) - E\left(f(z_1, ..., z_n)\right)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} C_i^2}\right).$$

Directly by Lemma 1, we can establish the following lemma.

Lemma 2. *Suppose that Condition 1 is met. If* $|y| \leq M_n$*, for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, it holds*

$$\mathcal{S}(\mathcal{Z}^n, M_n) \leq \left(2\log\frac{2}{\delta}\right)^{1/2} \frac{M_n + \kappa_1 \lambda^{-1/2} M_n^{1/2}}{n^{1/2}} + E(\mathcal{S}(\mathcal{Z}^n, M_n)).$$

Proof of Lemma 2: Denote $\mathcal{Z}_i^n$ as a modified sample which is exactly the same as $\mathcal{Z}^n$ except that the i-th entry $(\mathbf{x}_i, y_i)$ is replaced by its copy $(\mathbf{x}_i^{'}, y_i^{'})$. By the triangle inequality, we have

$$\begin{aligned}
&\mathcal{S}(\mathcal{Z}^n, M_n) - \mathcal{S}(\mathcal{Z}_i^n, M_n) \\
&= \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{m}\sum_{k=1}^{m} |\mathcal{E}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k})| - \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{m}\sum_{k=1}^{m} \left|\mathcal{E}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}_i^n}(Q_{\tau_k})\right| \\
&\leq \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{m}\sum_{k=1}^{m} \left|\mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}_i^n}(Q_{\tau_k})\right| \leq \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{2}{nm}\sum_{k=1}^{m}\sum_{i=1}^{n} |y_i - Q_{\tau_k}(\mathbf{x}_i)|,
\end{aligned}$$

where the inequalities are trivial. Note that by the condition that $|y| \leq M_n$, the Cauchy-Schwartz inequality in RKHS and Condition 1, we have

$$\max_{i=1,\ldots,n} |y_i - Q_{\tau_k}(\mathbf{x}_i)| \leq M_n + \|Q_{\tau_k}\|_\infty \leq M_n + \kappa_1 \|Q_{\tau_k}\|_K. \tag{7.1}$$

Therefore, there holds

$$\begin{aligned}
\mathcal{S}(\mathcal{Z}^n, M_n) - \mathcal{S}(\mathcal{Z}_i^n, M_n) &\leq \frac{2}{n}\left(M_n + \kappa_1 \left(\frac{1}{m}\sum_{k=1}^{m} \|Q_{\tau_k}\|_K^2\right)^{1/2}\right) \\
&\leq \frac{2}{n}\left(M_n + \kappa_1 \lambda^{-1/2} M_n^{1/2}\right),
\end{aligned}$$

where the first inequality follows from the relationship between root mean square and arithmetic mean, the definition of $\mathcal{F}_{M_n}$ and (7.1).

Finally, by Lemma 1, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$|\mathcal{S}(\mathcal{Z}^n, M_n) - E\mathcal{S}(\mathcal{Z}^n, M_n)| \leq \left(2\log\frac{2}{\delta}\right)^{1/2} \frac{M_n + \kappa_1 \lambda^{-1/2} M_n^{1/2}}{n^{1/2}}.$$

Then the desired result follows immediately. ■

Lemma 3. *Suppose that Condition 1 is satisfied. If* $|y| \leq M_n$*, it holds*

$$E\left(\mathcal{S}(\mathcal{Z}^n, M_n)\right) \leq \frac{4(M_n + \kappa_1 \lambda^{-1/2} M_n^{1/2})}{n^{1/2}}.$$

Proof of Lemma 3: Let $\{\sigma_i\}_{i=1}^n$ be the Rademacher random variables taking values in $\{-1, 1\}$ with equal probabilities. By the properties of Rademacher complexity [3], we have

$$\begin{aligned} E(\mathcal{S}(\mathcal{Z}^n, M_n)) &\le 2E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \left| \sum_{i=1}^{n} \sigma_i L_{\tau_k}(y_i - Q_{\tau_k}(\mathbf{x}_i)) \right| \\ &\le 4E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \left| \sum_{i=1}^{n} \sigma_i (y_i - Q_{\tau_k}(\mathbf{x}_i)) \right| \\ &\le 4 \left(E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \left| \sum_{i=1}^{n} \sigma_i Q_{\tau_k}(\mathbf{x}_i) \right| + \frac{M_n}{n^{1/2}} \right), \end{aligned}$$

where the first inequality follows from the Rademacher random variable is symmetric, the second inequality follows from that the check loss is Lipschitz continuous, and the last inequality follows from the property of Rademacher complexity and the condition that $|y| \le M_n$. Note that by the reproducing property that $Q_{\tau_k}(\mathbf{x}) = \langle Q_{\tau_k}, K_{\mathbf{x}} \rangle_K$, we have

$$\begin{aligned} & E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \left| \sum_{i=1}^{n} \sigma_i Q_{\tau_k}(\mathbf{x}_i) \right| \\ &= E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \left| \langle Q_{\tau_k}, \sum_{i=1}^{n} \sigma_i K_{\mathbf{x}_i} \rangle_K \right| \\ &\le E \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \frac{1}{nm} \sum_{k=1}^{m} \|Q_{\tau_k}\|_K \left(\sum_{i,j=1}^{n} \sigma_i \sigma_j K(\mathbf{x}_i, \mathbf{x}_j) \right)^{1/2} \\ &\le \frac{\kappa_1 M_n^{1/2}}{n\lambda^{1/2}} \left(\sum_{i,j=1}^{n} E\sigma_i\sigma_j \right)^{1/2} \le \frac{\kappa_1 M_n^{1/2}}{n^{1/2}\lambda^{1/2}}, \end{aligned}$$

where the first inequality directly follows from the Cauchy-Schwartz inequality, the second inequality follows from the definition of $\mathcal{F}_{M_n}$ and Condition 1, and the last inequality follows from the property of the Rademacher random variables. Then the desired result follows immediately. ■

Proof of Theorem 1: We first denote the event $\mathcal{C}_1$ as

$$\left\{ \frac{1}{m} \sum_{k=1}^{m} \|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K \ge 4 \max\{1, \kappa_1\} \left(\log \frac{8}{\delta} \right)^{1/4} \left(\frac{(1 + \kappa_1 \lambda^{-1/2}) \log n}{\lambda n^{1/2}} \right)^{1/2} \right\}.$$

Clearly, $P(\mathcal{C}_1)$ can be decomposed as

$$\begin{aligned} P(\mathcal{C}_1) &= P\left(\mathcal{C}_1 \cap \{\ |y| > \log n\}\right) + P\left(\mathcal{C}_1 \cap \{\ |y| \le \log n\}\right) \\ &\le P\left(|y| > \log n\right) + P\left(\mathcal{C}_1 \mid |y| \le \log n\right) =: P_1 + P_2. \end{aligned}$$

To bound P_1, by Markov's inequality, it holds $P(|y| > \log n) \leq (\log n)^{-2} E(y^2)$, where $E(y^2)$ is a bounded quantity. To bound P_2, denote $\mathcal{E}^\lambda(Q_{\tau_k}) = \mathcal{E}(Q_{\tau_k}) + \lambda\|Q_\tau\|_K^2$ and $\mathcal{E}^\lambda_{\mathcal{Z}^n}(Q_{\tau_k}) = \mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k}) + \lambda\|Q_{\tau_k}\|_K^2$. Then, directly by Proposition 2 and Theorems 2.6 and 2.7 in [41], there holds

$$\begin{aligned}\Psi^\diamond_\lambda\left(\|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K\right) &\leq 4 \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \left|t_{\mathcal{E}^\lambda}\mathcal{E}^\lambda(Q_{\tau_k}) - t_{\mathcal{E}^\lambda}\mathcal{E}^\lambda_{\mathcal{Z}^n}(Q_{\tau_k})\right| \\ &\leq 4 \sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \left|\mathcal{E}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k})\right|, \end{aligned} \tag{7.2}$$

where $\Psi^\diamond_\lambda(t) = \inf\{\frac{\lambda s^2}{2} + |t-s| : s \in [0,\infty)\}$ and $t_{\mathcal{E}^\lambda}$ is the translation map defined as $t_{\mathcal{E}^{\lambda_n}} G(Q_{\tau_k}) = G(Q_{\tau_k} + \widetilde{Q}_{\tau_k}) - \mathcal{E}^\lambda(\widetilde{Q}_{\tau_k})$ for all $G : \mathcal{H}_K \to \mathcal{R}$. Since $\Psi^\diamond_\lambda$ is invertible and increasing, we can write its inverse explicitly as $(\Psi^\diamond_\lambda)^{-1}$ as

$$(\Psi^\diamond_\lambda)^{-1}(t) = \begin{cases} \sqrt{2t/\lambda}, & \text{if } t < 1/(2\lambda), \\ t + 1/(2\lambda), & \text{otherwise.} \end{cases}$$

When the upper bound of (7.2) is sufficiently small, we have

$$\begin{aligned}\frac{1}{m}\sum_{k=1}^m \|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K &\leq \frac{2\sqrt{2}}{\lambda^{1/2} m} \sum_{k=1}^m \left(\sup_{\mathbf{Q}\in\mathcal{F}_{M_n}} \left|\mathcal{E}(Q_{\tau_k}) - \mathcal{E}_{\mathcal{Z}^n}(Q_{\tau_k})\right|\right)^{1/2} \\ &\leq \frac{2\sqrt{2}}{\lambda^{1/2}} \mathcal{S}(\mathcal{Z}^n, M_n)^{1/2}, \end{aligned}$$

where the first inequality follows from (7.2), the second inequality follows from the definition of $\mathcal{S}(\mathcal{Z}^n, M_n)$. Moreover, by taking $M_n = \log n$, for any $\delta \in (0,1)$, with probability at least $1 - \delta/4$, there holds

$$\frac{1}{m}\sum_{k=1}^m \|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K \leq c_3 \left(\log\frac{8}{\delta}\right)^{1/4} (\log n)^{1/2} \frac{(1 + \kappa_1 \lambda^{-1/2})^{1/2}}{\lambda^{1/2} n^{1/4}},$$

where $c_3 = 4\max\{1, \kappa_1\}$ and the inequality follows from Lemmas 2 and 3. Therefore, we have $P_2 \leq \delta/4$, and thus for any $\delta > 4(\log n)^{-2} E(y^2)$, we have $P(\mathcal{C}_1) \leq P_1 + P_2 \leq \delta/2$. The desired results follow immediately. ■

Proof of Theorem 2: Define the sample operators for gradients, $\widehat{D}_l : \mathcal{H}_K \to \mathcal{R}^n$ and its adjoint operator $\widehat{D}_l^* : \mathcal{R}^n \to \mathcal{H}_K$ as

$$(\widehat{D}_l Q_{\tau_k})_i = \langle Q_{\tau_k}, \partial_l K_{\mathbf{x}_i}\rangle_K \text{ and } \widehat{D}_l^* \mathbf{c} = \frac{1}{n}\sum_{i=1}^n \partial_l K_{\mathbf{x}_i} c_i,$$

respectively. Similarly, the integral operators for gradients, $D_l : \mathcal{H}_K \to \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ and $D_l^* : \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) \to \mathcal{H}_K$ are defined as

$$D_l Q_{\tau_k} = \langle Q_{\tau_k}, \partial_l K_{\mathbf{x}}\rangle_K \text{ and } D_l^* Q_{\tau_k} = \int_{\mathcal{X}} \partial_l K_{\mathbf{x}} Q_{\tau_k}(\mathbf{x}) \mathrm{d}\rho_{\mathbf{x}}(\mathbf{x}).$$

Note that D_l and $\widehat{D}_l$ are Hilbert-Schimdt operators by Propositions 12 and 13 in [28], and we have

$$D_l^* D_l Q_{\tau_k} = \int_{\mathcal{X}} \partial_l K_{\mathbf{x}} g_{l,\tau_k}(\mathbf{x}) \mathrm{d}\rho_{\mathbf{x}}(\mathbf{x}) \text{ and } \widehat{D}_l^* \widehat{D}_l Q_{\tau_k} = \frac{1}{n} \sum_{i=1}^{n} \partial_l K_{\mathbf{x}_i} g_{l,\tau_k}(\mathbf{x}_i).$$

Furthermore, we denote $HS(K)$ as a Hilbert space with all Hilbert-Schmidt operators on $\mathcal{H}_K$, endowed with the norm $\|\cdot\|_{HS}$. Note that denote $T : \mathcal{H} \to \mathcal{H}$ as a (linear) bounded operator and $\mathcal{H}$ as a complex (separable) Hilbert space endowed with the operator norm $\|\cdot\|_{op}$, we have $\|T\|_{op} \leq \|T\|_{HS}$.

Then, with these functional operators, there holds

$$\begin{aligned}
&\|\widehat{g}_{l,\tau_k}\|_n^2 - \|g_{l,\tau_k}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 \\
&= \langle \widehat{Q}_{\tau_k}, \frac{1}{n} \sum_{i=1}^{n} \widehat{g}_{l,\tau_k}(\mathbf{x}_i) \partial_l K_{\mathbf{x}_i} \rangle_K - \langle Q_{\tau_k}^*, \int_{\mathcal{X}} g_{l,\tau_k}^*(\mathbf{x}_i) \partial_l K_{\mathbf{x}} \mathrm{d}\rho_{\mathbf{x}}(\mathbf{x}) \rangle_K \\
&= \left\langle \widehat{Q}_{\tau_k} - Q_{\tau_k}^*, \widehat{D}_l^* \widehat{D}_l (\widehat{Q}_{\tau_k} - Q_{\tau_k}^*) \right\rangle_K + \left\langle \widehat{D}_l^* \widehat{D}_l Q_{\tau_k}^*, \widehat{Q}_{\tau_k} - Q_{\tau_k}^* \right\rangle_K + \\
&\qquad \left\langle Q_{\tau_k}^*, \widehat{D}_l^* \widehat{D}_l (\widehat{Q}_{\tau_k} - Q_{\tau_k}^*) \right\rangle_K + \left\langle Q_{\tau_k}^*, (\widehat{D}_l^* \widehat{D}_l - D_l^* D_l) Q_{\tau_k}^* \right\rangle_K \\
&\leq \|\widehat{D}_l^* \widehat{D}_l\|_{HS} \|\widehat{Q}_{\tau_k} - Q_{\tau_k}^*\|_K^2 + 2\|Q_{\tau_k}^*\|_K \|\widehat{D}_l^* \widehat{D}_l\|_{HS} \|\widehat{Q}_{\tau_k} - Q_{\tau_k}^*\|_K + \\
&\qquad \|Q_{\tau_k}^*\|_K^2 \|\widehat{D}_l^* \widehat{D}_l - D_l^* D_l\|_{HS},
\end{aligned}$$

where $\|Q_{\tau_k}^*\|_K^2$ is a bounded quantity and the inequality follows from the Cauchy-Schwartz inequality. Note that $\|\widehat{Q}_{\tau_k} - Q_{\tau_k}^*\|_K \leq \|\widehat{Q}_{\tau_k} - \widetilde{Q}_{\tau_k}\|_K + \|\widetilde{Q}_{\tau_k} - Q_{\tau_k}^*\|_K$, and thus an upper bound for $\frac{1}{m} \sum_{k=1}^{m} \|\widehat{Q}_{\tau_k} - Q_{\tau_k}^*\|_K$ is provided by the combination of Condition 2 and Theorem 1. Now we turn to bound $\|\widehat{D}_l^* \widehat{D}_l\|_{HS}$ and $\|\widehat{D}_l^* \widehat{D}_l - D_l^* D_l\|_{HS}$.

By Condition 1 and direct calculation as that in Theorem 7 of [27], there holds

$$\left\| \widehat{D}_l^* \widehat{D}_l \right\|_{HS}^2 = \|\partial_l K\|_K^4 \leq \kappa_2^4. \tag{7.3}$$

On the other hand, by the concentration inequalities in Hilbert-Schmidt space $HS(K)$ on $\mathcal{H}_K$ [28], for any $\epsilon_n \in (0,1)$, we have

$$P\left(\left\| \widehat{D}_l^* \widehat{D}_l - D_l^* D_l \right\|_{HS} \geq \epsilon_n \right) \leq 2 \exp\left(-\frac{n\epsilon_n^2}{8\kappa_2^4} \right).$$

This implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, there holds

$$\left\| \widehat{D}_l^* \widehat{D}_l - D_l^* D_l \right\|_{HS} \leq \left(\frac{8\kappa_2^4}{n} \log \frac{4p}{\delta} \right)^{1/2}, \tag{7.4}$$

for any $l = 1, ..., p$.

Hence, when $\frac{1}{m}\sum_{k=1}^{m}\|\widehat{Q}_{\tau_k}-Q^*_{\tau_k}\|_K$ is sufficiently small, by (7.3) and (7.4), with probability at least $1-\delta$, there holds

$$\begin{aligned}
&\max_{l=1,\ldots,p}\frac{1}{m}\sum_{k=1}^{m}\left|\|\widehat{g}_{l,\tau_k}\|_n^2-\|g^*_{l,\tau_k}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}\right|\\
&\le a_1\left(\frac{3}{m}\sum_{k=1}^{m}\|\widehat{Q}_{\tau_k}-Q^*_{\tau_k}\|_K+\max_{l=1,\ldots,p}\|\widehat{D}_l^*\widehat{D}_l-D_l^*D_l\|_{HS}\right)\\
&\le 8a_1\left(c_3\left(\log\frac{8}{\delta}\right)^{1/4}\left(\frac{(1+\kappa_1\lambda^{-1/2})\log n}{\lambda n^{1/2}}\right)^{1/2}+c_2\lambda_n^{\theta_1}+\left(\frac{8\kappa_2^4}{n}\log\frac{4p}{\delta}\right)^{1/2}\right),
\end{aligned}$$

where $a_1=\max\{\kappa_2^2,\kappa_2^2\max_k\|Q^*_{\tau_k}\|_K,\max_k\|Q^*_{\tau_k}\|_K^2\}$. Then, by setting $\lambda=n^{-1/4}$, the desired result follows immediately. ■

Proof of Theorem 3. Firstly, we show that $\widehat{\mathcal{A}}\subset\mathcal{A}^*$ in probability. If not, suppose that there exists some $l'\in\widehat{\mathcal{A}}$ but $l'\notin\mathcal{A}^*$, which implies $\frac{1}{m}\sum_{k=1}^{m}\|\widehat{g}_{l',\tau_k}\|_n^2>v_n$ but for any $\tau\in(0,1)$, $\|g^*_{l',\tau}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}=0$. If the thresholding value is chosen as $v_n=\frac{c_6}{2}\left(\log\frac{4p}{\delta}\right)^{\xi}(\log n)^{1/2}n^{-\Theta}$, then with probability at least $1-\delta$, there holds

$$\frac{1}{m}\sum_{k=1}^{m}\left|\|\widehat{g}_{l',\tau_k}\|_n^2-\|g^*_{l',\tau_k}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}\right|=\frac{1}{m}\sum_{k=1}^{m}\|\widehat{g}_{l',\tau_k}\|_n^2>v_n,$$

which contradicts with Theorem 1, and thus $\widehat{\mathcal{A}}\subset\mathcal{A}^*$ with probability at least $1-\delta$.

Next, we show that $\mathcal{A}^*\subset\widehat{\mathcal{A}}$ in probability. If not, suppose there exists some $l'\in\mathcal{A}^*$ but $l'\notin\widehat{\mathcal{A}}$, which implies that for some $\tau_0^{l'}\in(0,1)$, $\|g^*_{l',\tau_0^{l'}}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}>c_6\left(\log\frac{4p}{\delta}\right)^{\xi}(\log n)^{1/2}n^{-\Theta}$ but $\frac{1}{m}\sum_{k=1}^{m}\|\widehat{g}_{l',\tau_k}\|_n^2\le v_n$. By Conditions 3 and 4, there must exist some τ_{k_0} with $k_0\in\{1,...,m\}$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}}|g^*_{l',\tau_0^l}(\mathbf{x})-g^*_{l',\tau_{k_0}}(\mathbf{x})|\le c_5|\tau_0^{l'}-\tau_{k_0}|^{\zeta_1}\to 0,\quad\text{as } m\to\infty.$$

Hence, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{k=1}^{m}\left|\|\widehat{g}_{l',\tau_k}\|_n^2-\|g^*_{l',\tau_k}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}\right|\\
&\ge\frac{1}{m}\sum_{k=1}^{m}\|g^*_{l',\tau_k}-g^*_{l',\tau_0^{l'}}+g^*_{l',\tau_0^{l'}}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}-v_n\\
&\ge\|g^*_{l',\tau_0^{l'}}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}-v_n+\frac{1}{m}\|g^*_{l',\tau_{k_0}}-g^*_{l',\tau_0^{l'}}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}+\\
&\qquad\qquad\frac{2}{m}\int_{\mathcal{X}}(g^*_{l',\tau_{k_0}}(\mathbf{x})-g^*_{l',\tau_0^{l'}}(\mathbf{x}))g^*_{l',\tau_0^{l'}}(\mathbf{x})\mathrm{d}\rho_{\mathbf{x}}(\mathbf{x}).
\end{aligned}$$

If we choose $v_n=\frac{c_6}{2}\left(\log\frac{4p}{\delta}\right)^{\xi}(\log n)^{1/2}n^{-\Theta}$, there holds

$$\frac{1}{m}\sum_{k=1}^{m}\left|\|\widehat{g}_{l',\tau_k}\|_n^2-\|g^*_{l',\tau_k}\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}\right|$$
$$>\frac{c_6}{2}\left(\log\frac{4p}{\delta}\right)^{\xi}(\log n)^{1/2}n^{-\Theta}+\frac{2}{m}\int_{\mathcal{X}}(g^*_{l',\tau_{k_0}}(\mathbf{x})-g^*_{l',\tau_0^{l'}}(\mathbf{x}))g^*_{l',\tau_0^{l'}}(\mathbf{x})\mathrm{d}\rho_{\mathbf{x}}(\mathbf{x}),$$

which contradicts with Theorem 1, and thus we have $\mathcal{A}^*\subset\widehat{\mathcal{A}}$ with probability at least $1-\delta$. Combining these two results yield the desired theoretical results. ■

Acknowledgments

The authors thank the editor, the associate editor and the two anonymous referees for their constructive suggestions, which significantly improve this paper.

References

[1] P. Aubin-Frankowski and Z. Szabó. Hard shape-constrained kernel machines. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–17, 2020.

[2] R. Barber and E. Candès. A knockoff filter for high-dimensional selective inference. *Annals of Statistics*, 47(5):2504–2537, 2019. MR3988764

[3] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. MR1984026

[4] A. Belloni and V. Chernozhukov. l^1-penalty quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011. MR2797841

[5] S. Bond and K. Patel. The conditional distribution of real estate returns: relating time variation in higher moments to downside risk measurement. *Journal of Real Estate Finance and Economics*, 26(2):319–339, 2003.

[6] S. Das and D. Politis. Nonparametric estimation of the conditional distribution at regression boundary points. *The American Statistician*, 74(3):233–242, 2020. MR4133499

[7] J.Q. Fan and R.Z Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(45):1348–1360, 2001. MR1946581

[8] J.Q. Fan and J.C. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, 70(5):849–911, 2008. MR2530322

[9] J.Q. Fan, F. Yang, and R. Song. Nonparametric independence screening in sparse ultrahigh dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. MR2847969

[10] J.Q. Fan and Q.W. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York, 1st edition, 2003. MR1964455

[11] K. Fukumizu and C.L. Leng. Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370, 2014. MR3180569

[12] P. Hall, R.C.L. Wolff, and Q.W. Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999. MR1689221

[13] P. Hall and Q.W. Yao. Approximating conditional distribution functions using dimension reduction. *Annals of Statistics*, 33(3):1404–1421, 2005. MR2195640

[14] X. He, J.H. Wang, and S.G. Lv. Gradient-induced model-free variable selection with composite quantile regression. *Statistica Sinica*, 28(3):1521–1538, 2018. MR3821016

[15] X.M. He, L. Wang, and H.G. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41(1):342–369, 2013. MR3059421

[16] J. Huang, J.W. Horowitz, and F.R. Wei. Variable selection in nonparameteric additive models. *Annals of Statistics*, 38(4):2282–2313, 2010. MR2676890

[17] R. Izbicki and A.B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017. MR3679910

[18] J. Jin, C. Ying, and Z. Yu. Distributed estimation of principal support vector machines for sufficient dimension reduction. *Technical Report (https://arxiv.org/abs/1911.12732)*, pages 1–46, 2021.

[19] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia. *Learning Spark*. O'Reilly Media, Sebastopol, CA, 2015.

[20] R. Koenker and G.J. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978. MR0474644

[21] R. Koenker, S. Leorato, and F. Peracchi. Distribution vs quantile regression. *Technical Report (https://ideas.repec.org/p/eie/wpaper/1329.html)*, pages 1–34, 2021.

[22] S.B. Lin, X. Guo, and D.X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18:1–31, 2017. MR3714255

[23] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annal of Statistics*, 34(5):2272–2297, 2006. MR2291500

[24] S.G. Lv, H.Z. Lin, H. Lian, and J. Huang. Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *Annal of Statistics*, 46(2):781–813, 2018. MR3782384

[25] S.J. Ma, R.Z. Li, and C.L. Tsai. Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data. *Journal of the American Statistical Association*, 112(518):650–663, 2017. MR3953442

[26] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the*

Royal Society A, 209:415–446, 1909.

[27] L. Rosasco, M. Belkin, and E. De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010. MR2600634

[28] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013. MR3104492

[29] M. Sangnier, O. Fercoq, and F. d'Alché-Buc. Joint quantile regression in vector valued RKHSs. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3693–3701, 2016.

[30] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machine, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[31] X.T. Shen, W. Pan, and Y.Z. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012. MR2949354

[32] X.T. Shen, W. Pan, Y.Z. Zhu, and H. Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65:807–832, 2013. MR3105798

[33] C. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018. MR3874152

[34] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007. MR2327597

[35] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011. MR2825431

[36] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005. MR2234577

[37] I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2011. MR2914365

[38] W.W. Sun, J.H. Wang, and Y.X. Fang. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14:3419–3440, 2013. MR3144467

[39] G. J. Székely, M.L. Rizzo, and N.K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007. MR2382665

[40] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**:267–288, 1996. MR1379242

[41] S. Villa, L. Rosasco, S. Mosci, and A. Verri. Consistency of learning algorithms using Attouch-Wets convergence. *Optimization*, 61:287—-305, 2012. MR2879337

[42] S. Volgushev, S.K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *Annals of Statistics*, 47(3):1634–1662, 2019.

MR3911125

[43] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. *Advances in kernel methods: support vector learning*, MIT Press:69–88, 1998.

[44] H. S. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2016. MR2750576

[45] X. Y. Wang and C.L. Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B*, 78(3):589–611, 2016. MR3506794

[46] T. Watanabe. Excess kurtosis of conditional distribution for daily stock returns: The case of Japan. *Applied Economics Letters*, 7(6):353–355, 2000.

[47] Y.C. Wu and Y.F. Liu. Variable selection in quantile regression. *Statistica Sinica*, 19:801–817, 2009. MR2514189

[48] L. Yang, S.G. Lv, and J.H. Wang. Model-free variable selection in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, 17:1–24, 2016. MR3517105

[49] C. Zhang, Y.F. Liu, and Y.C. Wu. On quantile regression in reproducing kernel Hilbert spaces with data sparsity constraint. *Journal of Machine Learning Research*, 17:1–45, 2016. MR3491134

[50] C.H Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010. MR2604701

[51] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015. MR3450540

[52] D.X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Annals of Statistics*, 220(1):456–463, 2007. MR2444183

[53] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. MR2279469