Statistica Sinica Preprint No: SS-2016-0222R2	
Title	Gradient-induced Model-free Variable Selection with Composite Quantile Regression
Manuscript ID	SS-2016.0222
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0222
Complete List of Authors	Shaogao Lv Xin He and Junhui Wang
Corresponding Author	Shaogao Lv
E-mail	lvsg716@swufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Gradient-induced Model-free Variable Selection with Composite Quantile Regression*

Xin He†, Junhui Wang† and Shaogao Lv‡

† Department of Mathematics
City University of Hong Kong

‡ College of Statistics and Mathematics
Zhejiang Gongshang University, China.

Abstract

Variable selection is central to sparse modeling, and many methods have been proposed under various model assumptions. Most existing methods are based on an explicit functional relationship, while we are concerned with a model-free variable selection method that attempts to identify informative variables that are related to the response by simultaneously examining the sparsity in multiple conditional quantile functions. It does not require specification of the underlying model for the response. The proposed method is implemented via an efficient computing algorithm that couples the majorize-minimization algorithm and the proximal gradient descent algorithm. Its asymptotic estimation and variable selection consistencies are established, without explicit model assumptions, that assure the truly informative variables are correctly identified with high probability. The effectiveness of the proposed method is supported by a variety of simulated and real-life examples.

Key Words and Phrases: Lasso, learning gradients, quantile regression, reproducing kernel Hilbert space (RKHS), sparsity, variable selection

1 Introduction

With the rapid development of modern technology, it is much easier to collect a large number of observations and variables at a relatively low cost. Among the collected variables, it is generally

*Correspondence to: Shaogao Lv (lvsg716@swufe.edu.cn)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

believed that only a small number of them are truly informative for the analysis. Thus, sparse modeling that identifies the truly informative variables is critical for subsequent data analysis.

In the literature, one popular framework of sparse modeling is the regularization method, where sparsity-induced regularization terms are used so that the resultant sparse models keep only the informative variables. For linear models, a number of regularization terms have been proposed, including the least absolute shrinkage and selection operator (Lasso; Tibshirani (1996)), the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), the adaptive Lasso (Zou (2006)), the truncated ℓ_1 penalty (TLP; Shen et al. (2012)), and so on. These methods mainly focus on the conditional mean regression, and the informative variables are defined based on the corresponding regression coefficients. Similar regularization terms have also been applied to conditional quantile regression (QR)(Zhu et al. (2007); Li and Zhu (2008); Wu and Liu (2009); and Kato (2016)). To extend variable selection to a more general nonparametric context, additive models are popularly used (Shively et al. (1999); Xue (2009); Huang et al. (2010)). A further extension is the component selection and smoothing operator method (Cosso; Lin and Zhang (2006)), where the number of functional components may increase exponentially with the dimension. Recently, Ye and Xie (2012) and Yang et al. (2016) proposed a model-free variable selection method in the framework of gradient learning, where a variable is regarded as truly informative if the corresponding gradient of the mean function is significantly non-zero. All the aforementioned variable selection methods focus on a single conditional mean or quantile regression function, and their performance largely relies on the validity of the functional relationship.

Another popular framework of sparse modeling is variable screening (Fan and Lv (2008)), which examines each individual variable separately to attain the sure screening properties. More recently, a number of model-free screening schemes (Zhu et al. (2011); He et al. (2013)) have been developed under general model settings. Yet as pointed out in He et al. (2013), a potential weakness of the marginal screening methods is the ignorance of the marginally unimportant but jointly important variables. To overcome this difficulty, a higher-order screening method was

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

developed (Hao and Zhang (2014)).

We propose a new model-free variable selection method in a regularized gradient learning framework. The proposed method attempts to identify the informative variables that are related to the response by fully exploiting the underlying distribution. To fully characterize dependence between the variables and the response, multiple conditional quantile functions are simultaneously examined, and a variable is deemed informative if it contributes to any of the conditional quantile functions. Thus the proposed method is formulated as a gradient learning framework associated with the composite quantile functions (Zou and Yuan (2008)) on a flexible reproducing kernel Hilbert space (RKHS; Wahba (1999)). Gradient learning can be traced to Härdle and Gasser (1985) and Müller et al. (1987), and some of its recent developments include Jarrow et al. (2004) and Brabanter et al. (2013). The proposed method equips the gradient learning framework with a group lasso penalty that can be viewed as an extension of the classical finite-dimensional Lasso penalty in a functional space. An efficient computing algorithm is developed, which combines the MM algorithm (Hunter and Lange (2000)) and the proximal gradient descent algorithm (Rockafellarl (1970)). The performance of the proposed method is supported by a variety of simulations and data examples, as well as its asymptotic estimation and variable selection consistencies. In particular, our results assure that the proposed method recovers the truly informative variables with probability tending to one, and converges to the true gradient function.

The proposed method aims to finding variables that may contribute to not only the conditional mean or quantile function, but the conditional distribution of the response. Thus the identified non-informative variables by the proposed method can be regarded as independent of the response given other variables. Asymptotic variable selection consistency is obtained without assuming any explicit model, a contrast to most existing theoretical results based on model assumptions. Yet, as in many nonparametric variable selection methods (Xue (2009); Huang et al. (2010); Yang et al. (2016)), the asymptotic results for the proposed method are established in the scenario of a fixed dimension.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

The rest of the article is organized as follows. Section 2 presents the general framework of the proposed model-free variable selection method, as well as its computing algorithm. Section 3 establishes asymptotic estimation and variable selection consistencies. Section 4 contains the numerical results on simulations and data examples, followed by a concluding summary. The computational details are provided in the appendix, and proofs are contained in the online supplemental materials.

2 Methodology

2.1 Variable selection and conditional independence

Suppose that a training set consists of $\mathcal{Z} = (\mathbf{x}_i, y_i)$; $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathcal{X} \subset \mathcal{R}^p$ and $y_i \in \mathcal{R}$ are independently sampled as $(\mathbf{X}, Y)$ with $\mathbf{X} = (X^1, \ldots, X^p)^T$ supported on a compact metric space $\mathcal{X}$. Most variable selection methods are based on an additive model $y = \mu + \sum_{j=1}^{p} f_j^*(x_j) + \epsilon$, and define the uninformative variables as those with corresponding $f_j^* \equiv 0$. In a model-free fashion, we regard X^l as uninformative if

$$Y \perp\!\!\!\perp X^l \mid \mathbf{X}^{-l},$$

where $\mathbf{X}^{-l}$ denotes all variables except for X^l. To characterize the conditional dependence, we note that Y and X^l are conditional independent if and only if

$$\nabla Q_{\tau,l}^*(\mathbf{x}) = \partial Q_\tau^*(\mathbf{x})/\partial x^l \equiv 0, \quad \text{for any } \mathbf{x} \text{ and } \tau \in (0,1), \tag{1}$$

where $Q_\tau^*(\mathbf{X})$ is the τ-th conditional quantile function of Y given $\mathbf{X}$. This motivates the proposed variable selection method in a framework of learning sparse gradient functions. Let $\mathbf{g}_\tau^*(\mathbf{x}) = (g_{\tau,1}^*(\mathbf{x}), \ldots, g_{\tau,p}^*(\mathbf{x}))^T$ with $g_{\tau,l}^*(\mathbf{x}) = \nabla Q_{\tau,l}^*(\mathbf{x})$ as the true gradient function, and $Q_\tau(\mathbf{x})$ and $\mathbf{g}_\tau(x)$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

as estimates of $Q^*_\tau(\mathbf{x})$ and $\mathbf{g}^*_\tau(\mathbf{x})$, respectively. In this paper we restrict Q_τ to be contained in a RKHS $\mathcal{H}_K$ with a pre-specified kernel function $K(\cdot,\cdot)$. Due to the reproducing properties of the gradient functions, it can be shown under some smoothness conditions that $\mathbf{g}_\tau = \nabla Q_\tau$ is contained in $\mathcal{H}_K^p$ with $\mathcal{H}_K^p$ being a p-fold of $\mathcal{H}_K$ (Zhou (2007)).

2.2 Proposed formulation

At a given quantile level τ, the proposed method is formulated as

$$\underset{Q_\tau\in\mathcal{H}_K,\mathbf{g}_\tau\in\mathcal{H}_K^p}{\operatorname{argmin}} \frac{1}{n(n-1)}\sum_{i,j=1}^{n} w_{ij}L_\tau\Big(y_i - Q_\tau(\mathbf{x}_j) - \mathbf{g}_\tau(\mathbf{x}_i)^T(\mathbf{x}_i-\mathbf{x}_j)\Big) + J(Q_\tau,\mathbf{g}_\tau), \qquad (2)$$

where $L_\tau(u) = u(\tau - I(u<0))$ is known as the check loss for the τ-th quantile, $w_{ij} = w(\mathbf{x}_i,\mathbf{x}_j)$ is a weight function, and $J(Q_\tau,\mathbf{g}_\tau)$ is a penalty term. The first term in (2) is an empirical version of

$$\mathcal{E}(Q_\tau,\mathbf{g}_\tau) = \int\int w(\mathbf{x},\mathbf{u})L_\tau\Big(y - Q_\tau(\mathbf{u}) - \mathbf{g}_\tau(\mathbf{x})^T(\mathbf{x}-\mathbf{u})\Big)d\rho_X(\mathbf{u})d\rho(\mathbf{x},y),$$

where ρ and ρ_X are the joint distribution function of $(\mathbf{x},y)$ and the marginal distribution function of $\mathbf{x}$, respectively. Here, $Q_\tau(\mathbf{u}) + \mathbf{g}_\tau(\mathbf{x})^T(\mathbf{x}-\mathbf{u})$ can be regarded as an approximation of $Q_\tau(\mathbf{x})$ at a neighboring point $\mathbf{u}$, and $w(\mathbf{x},\mathbf{u})$ is used to ensure the local neighborhood of $\mathbf{x}$ contributing more to the estimation of $Q_\tau(\mathbf{x})$ and $\mathbf{g}_\tau(\mathbf{x})$. Typically, we set $w(\mathbf{x},\mathbf{u}) = e^{-\|\mathbf{x}-\mathbf{u}\|^2/\sigma_n^2}$, where σ_n^2 is a pre-specified scale parameter.

To make use of (1) for variable selection, we consider multiple quantile functions simultaneously, in order to identify the variables that may contain information about the conditional distribution at any quantile level. Let $0<\tau_1<\cdots<\tau_m<1$ be a pre-specified sequence of quantile levels. Let $\mathbf{Q} = (Q_{\tau_1},...,Q_{\tau_m})$ and $\mathbf{g} = (\mathbf{g}^1,...,\mathbf{g}^p)$ with $\mathbf{g}^l = (g^l_{\tau_1},...,g^l_{\tau_m})$, with

$$\mathcal{E}(\mathbf{Q},\mathbf{g}) = \frac{1}{m}\sum_{k=1}^{m}\int\int w(\mathbf{x},\mathbf{u})L_{\tau_k}\Big(y - Q_{\tau_k}(\mathbf{u}) - \mathbf{g}_{\tau_k}(\mathbf{x})^T(\mathbf{x}-\mathbf{u})\Big)d\rho_X(\mathbf{u})d\rho(\mathbf{x},y),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

and its empirical version as

$$\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g})=\frac{1}{mn(n-1)}\sum_{k=1}^{m}\sum_{i,j=1}^{n}w_{ij}L_{\tau_k}\Big(y_i-Q_{\tau_k}(\mathbf{x}_j)-\mathbf{g}_{\tau_k}(\mathbf{x}_i)^T(\mathbf{x}_i-\mathbf{x}_j)\Big).$$

The proposed method is then formulated as

$$\underset{\mathbf{Q},\mathbf{g}}{\operatorname{argmin}}\ \mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g})+\frac{\lambda_0}{m}\sum_{k=1}^{m}\|Q_{\tau_k}\|_{\mathcal{H}_K}^2+\lambda_1\sum_{l=1}^{p}\pi_l\|\mathbf{g}^l\|_{\mathcal{H}_K^m}. \tag{3}$$

Here $\|\mathbf{g}^l\|_{\mathcal{H}_K^m}=\sqrt{\frac{1}{m}\sum_{k=1}^{m}\|g_{\tau_k}^l\|_{\mathcal{H}_K}^2}$ is a group Lasso penalty (Yuan and Lin (2006)) that has the effect of pushing all or none of elements in $\|\mathbf{g}^l\|_{\mathcal{H}_K^m}$ to be exactly 0, thus achieving the purpose of variable selection. The weight π_l is adaptively assigned to different $\|\mathbf{g}^l\|_{\mathcal{H}_K^m}$ to achieve better selection performance following the suggestion of Zou and Yuan (2008), the penalty term $\|Q_{\tau_k}\|_{\mathcal{H}_k}^2$ is a standard RKHS-norm penalty, and λ_0 and λ_1 are two tuning parameters.

2.3 Computing algorithm

In this section, we develop an efficient computing algorithm to solve (3) , which couples the MM algorithm and the proximal gradient descent algorithm. The algorithm proceeds in an iterative fashion. Given the current estimate $(\widetilde{\mathbf{Q}},\widetilde{\mathbf{g}})$ and $\tilde{o}_{ij}=y_i-\widetilde{Q}_{\tau_k}(\mathbf{x}_j)-\widetilde{\mathbf{g}}_{\tau_k}(\mathbf{x}_i)^T(\mathbf{x}_i-\mathbf{x}_j)$, we first approximate the check loss $L_{\tau_k}(o_{ij})$ with a smooth loss function $L_{\tau_k}^{\epsilon}(o_{ij})=L_{\tau_k}(o_{ij})-\frac{\epsilon}{2}\ln(\epsilon+|o_{ij}|)$ and then majorize it with

$$\widetilde{L}_{\tau_k}^{\epsilon}(o_{ij}|\tilde{o}_{ij})=\frac{1}{4}\left(\frac{o_{ij}^2}{\epsilon+|\tilde{o}_{ij}|}+(4\tau_k-2)o_{ij}+c\right),$$

where c is a constant such that $\widetilde{L}_{\tau_k}^{\epsilon}(\tilde{o}_{ij}|\tilde{o}_{ij})=L_{\tau_k}^{\epsilon}(\tilde{o}_{ij})$. The minimization step is then to solve

$$\underset{\mathbf{Q},\mathbf{g}}{\operatorname{argmin}}\ R(\mathbf{Q},\mathbf{g})+\frac{\lambda_0}{m}\sum_{k=1}^{m}\|Q_{\tau_k}\|_{\mathcal{H}_K}^2+\Omega(\mathbf{g}), \tag{4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

where $\Omega(\mathbf{g}) = \lambda_1 \sum_{l=1}^{p} \pi_l \|\mathbf{g}^l\|_{\mathcal{H}_K^m}$ and

$$R(\mathbf{Q}, \mathbf{g}) = \frac{1}{mn(n-1)} \sum_{k=1}^{m} \sum_{i,j=1}^{n} w_{ij} \widetilde{L}_{\tau_k}^{\epsilon} \Big(y_i - Q_{\tau_k}(\mathbf{x}_j) - \mathbf{g}_{\tau_k}(\mathbf{x}_i)^T (\mathbf{x}_i - \mathbf{x}_j) \Big| \tilde{o}_{ij} \Big).$$

The obtained solution of (4) is used to update $\tilde{o}_{ij}$, and the iteration is stopped when some termination condition is met.

To solve the sub-optimization in (4), we employ a proximal gradient descent algorithm. Specifically, at the t-th iteration with solution $(\mathbf{Q}^t, \mathbf{g}^t)$,

$$\mathbf{Q}^{t+1} = \arg\min_{\mathbf{Q}} \left\{ R(\mathbf{Q}, \mathbf{g}^t) + \tfrac{\lambda_0}{m} \textstyle\sum_{k=1}^{m} \|Q_{\tau_k}\|_{\mathcal{H}_K}^2 \right\}, \tag{5}$$

$$\mathbf{g}^{t+1} = \text{prox}_{\frac{1}{D}\Omega} \left(\mathbf{g}^t - \tfrac{1}{D} \nabla_{\mathbf{g}} R(\mathbf{Q}^{t+1}, \mathbf{g}^t) \right), \tag{6}$$

here $\text{prox}_{\frac{1}{D}\Omega}$ is a proximal operator (Moreau (1962)), defined as

$$\text{prox}_{\frac{1}{D}\Omega}(\mathbf{g}) = \underset{\mathbf{f} \in \mathcal{H}_K^{mp}}{\text{argmin}} \left\{ \frac{1}{2} \|\mathbf{f} - \mathbf{g}\|_{\mathcal{H}_K^{mp}}^2 + \frac{1}{D} \Omega(\mathbf{f}) \right\},$$

where D is an upper bound of the maximum eigenvalues of $\nabla_{\mathbf{g}}^2 R(\mathbf{Q}, \mathbf{g})$.

To solve (5), we can solve for each Q_{τ_k} separately. By the representer theorem of RKHS, the solution of (5) must be of the form $Q_{\tau_k}(\mathbf{x}) = \sum_{i=1}^{n} c_i^k K(\mathbf{x}_i, \mathbf{x})$ with $\mathbf{c}^k = (c_1^k, ..., c_n^k) \in \mathbb{R}^n$. Then $\mathbf{c}^k$ can be obtained by solving the equation system

$$\mathbf{c}^k \Big\{ (e_{k,1} \mathbf{K}_{\mathbf{x}_1}^T, ..., e_{k,n} \mathbf{K}_{\mathbf{x}_n}^T) + 2\lambda_0 I_n \Big\} = (z_{k,1}, ..., z_{k,n}), \tag{7}$$

where $\mathbf{K}_{\mathbf{x}} = (K(\mathbf{x}_1, \mathbf{x}), ..., K(\mathbf{x}_n, \mathbf{x}))^T$, and $e_{k,i}$ and $z_{k,i}$ are defined as in the appendix.

The representer theorem of RKHS also implies that the solution of (6) must be of the form $g_{\tau_k}^l(\mathbf{x}) = \mathbf{K}_{\mathbf{x}}^T \boldsymbol{\alpha}^{kl} = \sum_{i=1}^{n} \alpha_i^{kl} K(\mathbf{x}_i, \mathbf{x})$. If $\bar{\mathbf{g}}^t = \mathbf{g}^t - \frac{1}{D} \nabla_{\mathbf{g}} R(\mathbf{Q}^{t+1}, \mathbf{g}^t)$ and $\widetilde{\mathbf{g}}^{t+1} = ([\widetilde{\mathbf{g}}^{t+1}]_1, ..., [\widetilde{\mathbf{g}}^{t+1}]_p)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

then

$$[\widetilde{\mathbf{g}}^{t+1}]_l = \frac{[\bar{\mathbf{g}}^t]_l}{\|[\bar{\mathbf{g}}^t]_l\|_{\mathcal{H}_K^m}} \Big(\|[\bar{\mathbf{g}}^t]_l\|_{\mathcal{H}_K^m} - \frac{\lambda_l}{D} \Big)_+,$$

with $\lambda_l = \lambda_1 \pi_l$ and $\nabla_{\mathbf{g}} R(\mathbf{Q}^{t+1}, \mathbf{g}^t)$ defined in (10) of the Appendix.

Algorithm 1

given parameters λ_0, λ_1, π_l; $l = 1, 2, \ldots, p$, ϵ, c, and quantile vector $\boldsymbol{\tau} > 0$
initialize $\mathbf{g}^0 = \mathbf{Q}^0 = \mathbf{0}$, $t = 1$
repeat

$$\mathbf{Q}^t = \arg\min_{\mathbf{Q}} \left\{ R(\mathbf{Q}, \mathbf{g}^{t-1}) + \frac{\lambda_0}{m} \sum_{k=1}^{m} \|Q_{\tau_k}\|_{\mathcal{H}_K}^2 \right\}$$

$$\mathbf{g}^t = \text{prox}_{\frac{1}{D}\Omega} \left(\mathbf{g}^{t-1} - \frac{1}{D} \nabla R(\mathbf{Q}^t, \mathbf{g}^{t-1}) \right)$$

$$\tilde{o} \leftarrow o(\mathbf{Q}^t, \mathbf{g}^t)$$

$$t \leftarrow t + 1$$

until $(\mathbf{Q}^t, \mathbf{g}^t)$ converges.

The complexity of the proposed variable selection method is linear in m, which is not much different than conducting variable selection for m separate quantile regression models. A possible drawbacks is that the Lipschitz constant D in the proximal gradient algorithm is not always computable. For large-scale problem, this quantity is intractable computationally, and a backtracking scheme (Beck and Teboulle (2009)) can be used to approximate the value of D.

3 Asymptotic consistencies

This section establishes the asymptotic estimation and variable selection consistencies of the proposed method. Let $(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}})$ be the minimizer of (3), $(\mathbf{Q}^*, \mathbf{g}^*)$ be the true quantile and true gradient functions, and $\mathcal{A}^* = \{X^1, X^2, ..., X^{p_0}\}$ be, with $p_0 < p$, the true active set.

Assumption 1. The support $\mathcal{X}$ is a non-degenerate compact subset of $\mathcal{R}^p$. For any $\tau \in (0, 1)$, there exists a positive constant c_1 such that $\sup_{\mathbf{x} \in \mathcal{X}} \| \mathbf{H}_\tau^*(\mathbf{x}) \|_2 \leq c_1$, where $\mathbf{H}_\tau^*(\mathbf{x}) = \nabla^2 Q_\tau^*(\mathbf{x})$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

a Hessian matrix for any given $\mathbf{x}$ and $\|\cdot\|_2$ is the matrix 2-norm.

Assumption 2. For some positive constants c_2 and θ, the marginal density $p(\mathbf{x})$ exists and satisfies $|p(\mathbf{x}) - p(\mathbf{u})| \leq c_2 d_X(\mathbf{x}, \mathbf{u})^\theta$, for any $\mathbf{x}, \mathbf{u} \in \mathcal{X}$, where $d_{\mathbf{X}}(\cdot, \cdot)$ is the Euclidean distance.

Assumption 3. There exist positive constants c_3 and c_4 such that $c_3 \leq \lim_{n\to\infty} \min_{1\leq l\leq p_0} \pi_l \leq \lim_{n\to\infty} \max_{1\leq l\leq p_0} \pi_l \leq c_4$.

Assumption 1 gives regularity conditions on the support $\mathcal{X}$. The boundedness assumption on the largest eigenvalues of $\mathbf{H}^*_\tau(\mathbf{x})$ is necessary to prevent the loss function from diverging (Ye and Xie (2012)). Assumption 2 characterizes the smoothness of underlying distribution of $\mathbf{x}$ by introducing a Lipschitz condition on its density. It follows from Assumptions 1 and 2 that there exists some constant c_5 such that $\sup_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \leq c_5$. Assumption 3 restricts the behavior of the adaptive weights when n diverges.

Theorem 1. *Suppose Assumptions 1-3 are hold and* $\mathbf{Q}^* \in \mathcal{H}_K^m$. *If* $\lambda_0 = n^{-\frac{1}{4}}$, $\lambda_1 = n^{-\frac{\theta}{2(p+2+2\theta)}}$, *and* $\sigma_n = n^{-\frac{\theta}{2(p+2+2\theta)}}$, *then there exists some constant* c_6 *such that with probability at least* $1-\delta$,

$$\left|\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)\right| \leq c_6 \big(\log(4/\delta)\big)^{1/2} n^{-\Theta},$$

with $\Theta = \min\{\frac{p+2}{4(p+2+2\theta)}, \frac{\theta}{2(p+2+2\theta)}\}$.

Theorem 1 establishes the weak convergence of $(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}})$ with the convergence rate that depends on the choice of λ_0, λ_1 and σ_n. This might be improved with a more involved derivation. The assumption $\mathbf{Q}^* \in \mathcal{H}_K^m$ can be relaxed by considering the approximation error between $\mathbf{Q}^*$ and $\mathcal{H}_K^m$ (Ye and Xie (2012)). The proposed method can also be improved by considering the weighted average of the selected quantile levels, which might lead to a smaller constant in the upper bound of $\left|\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)\right|$.

Let the estimated active set be $\widehat{\mathcal{A}} = \{X^l : \sum_{k=1}^m \|\widehat{g}^l_{\tau_k}\|_1 \neq 0\}$, where $\|\widehat{g}^l_{\tau_k}\|_1 = \int_{\mathcal{X}} |\widehat{g}^l_{\tau_k}(\mathbf{x})| d\rho_{\mathbf{X}}(\mathbf{x})$. We need some assumptions.

Assumption 4. As n diverges, $n^{-\frac{1}{2}} \psi_{max}^{-\frac{1}{2}} \psi_{min} \lambda_1 \min_{l>p_0} \pi_l \to \infty$, where ψ_{max} and ψ_{min} are

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

the largest and smallest eigenvalues of $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$, respectively.

Assumption 5. There exist positive constants c_7, c_8, and $q \in (0, 2)$, such that

$$\inf_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}(\mathbf{Q}, \mathbf{g}) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)| \geq \frac{c_7}{m}\sum_{k=1}^m \|Q_{\tau_k} - Q^*_{\tau_k}\|_q^2 + \frac{c_8}{m}\sum_{k=1}^m \|\mathbf{g}_{\tau_k} - \mathbf{g}^*_{\tau_k}\|_q^2,$$

where $\mathcal{F}_{r_n} = \{(\mathbf{Q}, \mathbf{g}) \in \mathcal{H}_K^{m(p+1)} : \frac{\lambda_0}{m}\sum_{k=1}^m \|Q_{\tau_k}\|_{\mathcal{H}_K}^2 \leq r_n \text{ and } \lambda_1 \sum_{l=1}^p \pi_l \|g_\tau^l\|_{\mathcal{H}_K} \leq r_n\}$, with $r_n \geq \frac{1}{mn(n-1)}\sum_{k=1}^m \sum_{i,j=1}^n w_{ij}|y_i|$ and $\|g_\tau\|_q = \big(\int_{\mathcal{X}} \sum_{l=1}^p |g_\tau^l(\mathbf{x})|^q d\rho_{\mathbf{X}}(\mathbf{x})\big)^{1/q}$ is the norm induced by $L^q_{\rho_{\mathbf{X}}}$.

Assumption 6. For some positive constants c_9 and ζ, the true gradient function satisfies

$$\sup_{\mathbf{x},l} |g^{*l}_{\tau'}(\mathbf{x}) - g^{*l}_{\tau}(\mathbf{x})| \leq c_9|\tau' - \tau|^\zeta, \text{ for any } \tau', \tau \in (0, 1). \tag{8}$$

When $l > p_0$, $\mathbf{g}^{*l}_\tau(\mathbf{x}) \equiv 0$ for any $\tau \in (0, 1)$ and $\mathbf{x} \in \mathcal{X}$ almost surely, and when $l \leq p_0$, there exist $t > 0$ and τ_0 such that $\int_{\mathcal{X}\setminus\mathcal{X}_t} (g^{*l}_{\tau_0}(\mathbf{x}))^2 d\rho_{\mathbf{X}}(\mathbf{x}) > 0$, where $\mathcal{X}_t = \{\mathbf{x} \in \mathcal{X} : d_{\mathbf{X}}(\mathbf{x}, \partial\mathcal{X}) < t\}$, $d_{\mathbf{X}}(\mathbf{x}, \partial\mathcal{X}) = \inf_{\mathbf{u}\in\partial\mathcal{X}} d_{\mathbf{X}}(\mathbf{x}, \mathbf{u})$, and $\partial\mathcal{X}$ is the boundary of $\mathcal{X}$.

Assumption 4 further quantifies the asymptotic behavior of the adaptive weights. When the second-order Sobolev kernel is used, ψ_{max} and ψ_{min} are of order $O_p(n)$ and $O_p(n^{-1})$ (Braun (2006); Wainwright et al. (2012)). Then Assumption 4 is satisfied with $\pi_l = \|\widetilde{\mathbf{g}}^l\|_2^{-\gamma}$, where γ is determined by the given $\psi_{max}, \psi_{min}, \lambda_1$, and $\widetilde{\mathbf{g}}^l$ is the solution of (3) with $\lambda_0 = \lambda_1 = 0$. The verification can be done similar to the proof of Theorem 1. Assumption 5 connects the strong convergence $\sum_{k=1}^m \|\mathbf{g}_{\tau_k} - \mathbf{g}^*_{\tau_k}\|_q$ with the weak convergence measured by the difference of $\mathcal{E}(\mathbf{Q}, \mathbf{g})$ and $\mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)$. It is similar to that used in Steinwart and Christmann (2011) and Lv et al. (2016) in proving the strong convergence of nonparametric function estimation. Assumption 6 quantifies the smoothness of the true gradient functions. Similar Lipschitz conditions are also used in Belloni and Chernozhukov (2011) for parametric cases. Assumption 6 requires the gradient functions with respect to the truly informative variables be significantly away from 0, and those with respect to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

the non-informative variables be 0. This discriminates between informative and non-informative variables without imposing an explicit model assumption.

Theorem 2. *If the assumptions in Theorem 1 and Assumptions 4-6 hold, then* $P(\widehat{\mathcal{A}} = \mathcal{A}^*) \to 1$ *as* $m, n \to \infty$.

4 Numerical experiments

In this section, the effectiveness of the proposed method is compared against some existing non-parametric variable selection methods. Specifically, the random forest (Breiman (2001)) can be adjusted to conduct nonparametic variable selection; Xue (2009) assumes an additive model for the conditional mean function to conduct variable selection; Lin and Zhang (2006) conduct component selection and a smoothing operator for the nonparametric mean regression; a modified formulation of (3) with $\tau = 0.5$ conducts variable selection for the conditional median regression function; He et al. (2013) consider the independent sure screening method. We denote these methods as MF, RF, Add, Cosso, Median, and QaSIS, respectively. The quantile levels were set as $\tau = (0.01, 0.25, 0.75, 0.99)$ for the proposed MF, and $\tau = 0.75$ for QaSIS.

For all methods, the kernel function was set as the radial basis kernel for computational convenience, $K(s,t) = e^{-\|s-t\|^2/\sigma^2}$, where σ^2 was set as the median of all pairwise distances among the training sample (Jaakkola et al. (1999)). The performance of all methods was tuned through a stability-based selection criterion (Sun et al. (2013)). One randomly splits the training set into two subsets, and applies any variable selection method to them to generate two estimated active sets. A measure of the agreement of these two estimated active sets is defined as the variable selection stability, and the selection criterion looks for the tuning parameter corresponding the largest stability measure. The search was conducted via a grid search, where the grid was set as $\{10^{-2+0.1s} : s = 0, \ldots, 40\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

4.1 Simulated examples

Two simulations were done. In the first, only the mean function relied on the variables, whereas in the second both the mean function and the error term relied on the variables.

Example 1: We generated $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbf{R}^p$ with x_{ij}= $\frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently generated from $U(-0.5, 0.5)$. We took $f^*(\mathbf{x}_i) = 6f_1(x_{i1}) + 4f_2(x_{i2})f_3(x_{i3}) + 6f_3(x_{i4}) + 5f_4(x_{i5})$, with $f_1(u) = u, f_2(u) = 2u + 1, f_3(u) = 2u - 1, f_4(u) = 0.1\sin(\pi u) + 0.2\cos(\pi u) + 0.3(\sin(\pi u))^2 + 0.4(\cos(\pi u))^3 + 0.5(\sin(\pi u))^3$ and $f_5(u) = \sin(\pi u)/(2 - \sin(\pi u))$. The response y_i was generated as $y_i = f^*(\mathbf{x}_i) + \epsilon_i$, with ϵ_i's independently $N(0, 1)$. Here the true regression function is additive and contains a interaction term. The first five variables are the informative variables.

Example 2: The generating scheme was similar to that of Example 1, except that W_{ij} and U_i were independently $U(0, 1)$ and the response y_i was generated as $y_i = 4x_{i1}x_{i2} + 3|x_{i3}|\epsilon_i$. Here, (X^1, X^2, X^3) were the all informative variables.

For each example, we considered scenarios with $(n, p) = (200, 10), (200, 20)$ and $(400, 100)$. In each scenario, $\eta = 0$ and $\eta = 0.1$ were examined. When $\eta = 0$, the data was completely independent, whereas when $\eta = 0.1$, correlation structure had been added among the variables. Each scenario was replicated 50 times. The averaged performance measures are summarized in Tables 1 and 2, where Size is the averaged number of selected informative variables, TP is the number of truly informative variables selected, FP is the number of truly non-informative variables selected, and C, U, O are the times of correct-fitting, under-fitting, and over-fitting, respectively.

Tables 1 and 2 about here

It is evident that MF outperforms the other competitors in most scenarios. In Example 1, MF yields similar performance as Median and ADD, but outperforms the other methods. In Example 2, MF delivers a much larger advantage against the other five methods. Those methods focus on a

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

single mean or quantile function, and miss the X^3 that affects the response through the variance, while, MF is able to identify X^3 in most replications. In both examples with $\eta = 0.1$, the correlation structure increases the difficulty of identifying the informative variables, and here MF outperforms its competitors in most scenarios.

4.2 Japanese industrial chemical firm data

This section reports on the application of MF to a dataset on Japanese industrial chemical firms (Yafeh et al.,2003). The dataset includes 186 Japanese industrial chemical firms listed on the Tokyo stock exchange, and the goal is to check whether concentrated shareholding is associated with lower expenditure on activities with scope for managerial private benefits. The dataset 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 dataset is available online through the Economic Journal at http://www.res.org.uk.

The dataset was pre-processed by removing all the missing values, and the response and the covariates were all standardized. We then randomly split the dataset, with 20 observations for testing and the remainder for training. The splitting was replicated 100 times, and the variable selection performance and the averaged prediction errors are summarized in Table 3.

Table 3 about here

As Table 3 shows, MF selects four informative variables, including LEVERAGE, VARS, OPERS, and BDA, whereas Median and Add select five variables, Cosso and RF select seven vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

ables, and QaSIS selects six variables. The average prediction error of MF is smaller than that of the other five methods, suggesting that these methods may include some noise variables that deteriorate their prediction performance. Figure 1 displays scatter plots of MH5 against the variables selected by MF. Among all the variables selected by MF, BDA is ignored by Median, Cosso and ADD. However, it's clear from the scatter plot that the variance of MH5 appears to shrink as BDA increases, even though its mean does not change much with BDA. The modified Levene test yields a significant p-value, providing strong evidence against the constant variance of MH5 given BDA. This supports the advantage of MF in identifying informative variables might influence the conditional distribution of the response.

Figure 1 about here

5 Summary

This article proposes a gradient-induced model-free variable selection method to identify the informative variables that are dependent of the response in a general sense. The proposed method formulates the variable selection task in a flexible framework of learning gradients of multiple quantile regression functions. The proposed method works under the classical setting with fixed dimension; it would be of interest to extend it to the case of diverging dimensions, One possible route is to first implement a model-free sure screening algorithm (Fan and Lv (2008)) to screen out most uninformative variables, and then the proposed method can be applied to identify the truly informative variables within the reduced candidate variable set.

Supplementary Materials

The proofs of Theorems 1 and 2 and the related lemmas and propositions are provided in the online supplementary materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Acknowledgment

JW's research is partially supported by HK GRF-11302615 and HK GRF-11331016, and SL's research is partially supported by NSFC-11301421.

Appendix: updating Q and g

I.1 Updating $\mathbf{Q}$: For any fixed k, let

$$\tilde{w}_{ij} = \frac{w_{ij}}{2(\epsilon + |\tilde{o}_{ij}|)}, \quad h_{ijk} = \frac{w_{ij}(2\tau_k - 1)}{2}, \; g_{ijk} = y_i - \mathbf{g}_{\tau_k}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j).$$

Summing up the derivative of (5) with respect to c_j^k yields that, for any $\mathbf{x} \in \mathcal{X}$,

$$\frac{1}{mn(n-1)} \sum_{i,j=1}^{n} \tilde{w}_{ij}\big(g_{ijk} - Q_{\tau_k}(\mathbf{x}_j)\big)K(\mathbf{x}_j, \mathbf{x}) + \frac{1}{mn(n-1)} \sum_{i,j=1}^{n} h_{ijk}K(\mathbf{x}_j, \mathbf{x}) - 2\frac{\lambda_0}{m}Q_{\tau_k}(\mathbf{x}) = 0.$$

Let $z_{k,j} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\tilde{w}_{ij} g_{ijk} + h_{ijk})$ and $e_{k,j} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \tilde{w}_{ij}$. Since the above equality holds for any $\mathbf{x}$, we have

$$2\lambda_0 \mathbf{c}^k = (z_{k,1}, ..., z_{k,n}) - \mathbf{c}^k(e_{k,1}\mathbf{K}_{\mathbf{x}_1}^T, ..., e_{k,n}\mathbf{K}_{\mathbf{x}_n}^T),$$

I.2 Updating $\mathbf{g}$: Since $\Omega(\cdot)$ is one-homogeneous, $\Omega(\theta f) = \theta\Omega(f)$ for $\theta > 0$, the Moreau identity (Combettes and Wajs (2005)) gives an equivalent relationship between the proximal operator and the projection operator,

$$\text{prox}_{\mu\Omega} = I - \pi_{\mu\mathcal{C}_n}, \tag{9}$$

where $\mathcal{C}_n = (\partial\Omega(0))$ is the subdifferential of Ω at the origin, and $\pi_{\mu\mathcal{C}_n} : \mathcal{H}_K^{mp} \to \mathcal{H}_K^{mp}$ is the projection on $\mu\mathcal{C}_n$, which is well defined since $\mathcal{C}_n$ is a closed subset of $\mathcal{H}_K^{mp}$. We can efficiently compute the projection of $\pi_{\mu\mathcal{C}_n}$ from the following lemma (Rosasco et al. (2009)).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Lemma 1. *For all $l = 1, ..., p$, let $\mathcal{G}_l$ be a Hilbert space with norm $\|\cdot\|_l$ and $\mathcal{J}_l : \mathcal{F} \to \mathcal{G}_l$ be a bounded linear operator. Let $J(\mathbf{f}) = \sum_{l=1}^{p} \|\mathcal{J}_l(\mathbf{f})\|_l$ and $\mathcal{G} = \Pi_{l=1}^{p}\mathcal{G}_l$, so that $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_p)$ with $\mathbf{v}_l \in \mathcal{G}_l$ and $\| \mathbf{v} \| = \sum_{l=1}^{p} \| \mathbf{v}_l \|_l$. If $\mathcal{J} : \mathcal{F} \to \mathcal{G}$ is such that $\mathcal{J}(\mathbf{f}) = (\mathcal{J}_1(\mathbf{f}), ..., \mathcal{J}_p(\mathbf{f}))$ and $Ker\mathcal{J} = \{0\}$, then*

$$\partial J(0) = \{\mathcal{J}^T \mathbf{v} : \mathbf{v} \in \mathcal{G}, \|\mathbf{v}_l\|_l \leq 1, \textit{ for any } l\},$$

where $\mathcal{J}^T : \mathcal{G} \to \mathcal{F}$ is the adjoint of $\mathcal{J}$ that can be written as $\mathcal{J}^T\mathbf{v} = \sum_{l=1}^{p} \mathcal{J}_l^T\mathbf{v}_l$. The projection of an element $g \in \mathcal{F}$ on the set $\mu\partial J(0)$ is given by $\mu\mathcal{J}^T\bar{\mathbf{v}}$, with

$$\bar{\mathbf{v}} = \arg \min_{\mathbf{v}\in\mathcal{G}, \| \mathbf{v}_l \|_l \leq 1} \|\mu\mathcal{J}^T \mathbf{v} - g\|_{\mathcal{F}}^2,$$

Let $\mathcal{J} : \mathcal{H}_K^{mp} \to \mathcal{H}_K^{mp}$ to be the weighted group operator, $\mathcal{J}(\mathbf{f}) = (\mathcal{J}_1(\mathbf{f}), ..., \mathcal{J}_p(\mathbf{f}))$, where $\mathcal{J}_l(\mathbf{f}) = \sqrt{\lambda_l}[\mathbf{f}]_l$ with $[\mathbf{f}]_l \in \mathcal{H}_K^m$. We reformulate the penalty term as $\Omega(\mathbf{f}) = \sum_{l=1}^{p} \|\mathcal{J}_l(\mathbf{f})\|_{\mathcal{H}_K^m}$. Proposition 2 of Rosasco et al. (2009) shows that the projection can be defined as $\pi_{\mu\mathcal{C}_n}(\mathbf{g}) = \mu\bar{\mathbf{v}}$ with $\bar{\mathbf{v}} = (\lambda_1\bar{\mathbf{v}}_1, ..., \lambda_p\bar{\mathbf{v}}_p)$, where

$$\bar{\mathbf{v}}_l = \underset{\|\mathbf{v}_l\|_{\mathcal{H}_K^m} \leq 1}{\operatorname{argmin}} \|\mu\lambda_l\mathbf{v}_l - [\mathbf{g}]_l\|_{\mathcal{H}_K^m}^2, \; l = 1, ..., p.$$

It can be computed block-wise as

$$\bar{\mathbf{v}}_l = \min\left\{1, \frac{\|[\mathbf{g}]_l\|_{\mathcal{H}_K^m}}{\mu\lambda_l}\right\} \frac{[\mathbf{g}]_l}{\|[\mathbf{g}]_l\|_{\mathcal{H}_K^m}},$$

which implies that

$$[I - \pi_{\mu\mathcal{C}_n}(\mathbf{g})]_l = [\mathbf{g}]_l - \min\left\{\mu\lambda_l, \|[\mathbf{g}]_l\|_{\mathcal{H}_K^m}\right\} \frac{[\mathbf{g}]_l}{\|[\mathbf{g}]_l\|_{\mathcal{H}_K^m}} = \frac{[\mathbf{g}]_l}{\|[\mathbf{g}]_l\|_{\mathcal{H}_K^m}}(\|[\mathbf{g}]_l\|_{\mathcal{H}_K^m} - \mu\lambda_l)_+, \quad l = 1, ..., p.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

In our case $\mu = \frac{1}{D}$ and, by (9), the proximal operator of (6) can be expressed explicitly as

$$[\widetilde{\mathbf{g}}^{t+1}]_l = \frac{[\bar{\mathbf{g}}^t]_l}{\|[\bar{\mathbf{g}}^t]_l\|_{\mathcal{H}_K^m}}(\|[\bar{\mathbf{g}}^t]_l\|_{\mathcal{H}_K^m} - \frac{\lambda_l}{D})_+,$$

where $\bar{\mathbf{g}}^t = \mathbf{g}^t - \frac{1}{D}\nabla_{\mathbf{g}} R(\mathbf{Q}^{t+1}, \mathbf{g}^t)$. A direct computation yields that $\nabla_{\mathbf{g}} R(\mathbf{Q}^{t+1}, \mathbf{g}^t) = (\mathbf{V}_1^t, ..., \mathbf{V}_m^t)^T$, where $\mathbf{V}_k^t$ is a p-dimensional vector, whose l-th element is

$$[\mathbf{V}_k^t]_l(x) = \frac{1}{mn(n-1)} \sum_{i,j=1}^{n} w_{ij} \left(\frac{y_i - Q_{\tau_k}^{t+1}(\mathbf{x}_j) - \mathbf{g}_{\tau_k}^t(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j)}{2(\epsilon + |\tilde{o}_{ij}|)} + (\tau_k - 0.5) \right) K(\mathbf{x}_i, \mathbf{x})(x_{j,l} - x_{i,l}). \tag{10}$$

References

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

Belloni, A. and Chernozhukov, V. (2011). $l1$-penalty quantile regression in high-dimensional spares models. *Annals of Statistics* **39**, 82-130.

Brabanter, K., Brabanter, J. and Moor, B. (2013). Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research* **14**, 281-301.

Braun, M. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research* **7**, 2303-2328.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5-32.

Combettes, P. and WAJS, V. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation* **4**, 1168–1200.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space

(with discussion). *Journal of the Royal Statistical Society Series B* **70**, 849-911.

Hao, N. and Zhang, H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109**, 1285-1301.

Härdle, W. and Gasser, T. (1985). On robust kernel estimation of derivatives of regression functions. *Scandinavian Journal of Statistics* **12**, 233-240.

He, X., Wang, L. and Hong, H. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics* **41**, 342-369.

Huang, J., Horowitz, J. and Wei, F. (2010). Variable selection in nonparameteric additive models. *Annals of Statistics* **38**, 2282-2313.

Hunter, D. and Lange, K. (2000). Quantile regression via a MM algorithm. *Journal of Computational and Graphical Statistics* **9**, 60-77.

Jaakkola, T., Diekhans, m. and 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.

Jarrow, R., Ruppert, D. and Yu, Y. (2004). Estimating the term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association* **99**, 57-66.

Kato, K. (2017). Group Lasso for high dimensional sparse quantile regression models. Manuscript.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association* **102**, 255-268.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics* **17**, 163-185.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annal of Statistics* **34**, 2272-2297.

Lv, S. He, X. and Wang, J. (2017). A unified penalized method for sparse additive quantile models:

a RKHS approach. In press.

Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien, *Reports of the Paris Academy of Sciences, Series A* **255**, 2897-2899.

Müller, H., Standtmuller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74**, 743-749.

Raskutti, G., Wainwright, M. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **13**, 389-427.

Rockafellar, R. T. (1970). Convex analysis.*Princeton University Press*.

Rosasco, L., Mosci, S., Santoro, M., Verri, A., and Villa, S.(2009). Iterative projection methods for structured sparsity regularization. *Computer Science and Artificial Intelligence Laboratory Technical Report*, MIT-CSAIL-TR-2009-050.

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107**, 223-232.

Shively, T., Kohn, R. and Wood, S. (1999). Variable selection and function estimation in additive non-parametric regression using a data-based prior. *Journal of the American Statistical Association* **94**, 777-794.

Steinwart, I and Christmann, A. (2011). Estimating conditional quantiles with the help of pinball loss. *Bernoulli* **17**, 211-225.

Sun, W., Wang, J. and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research* **14**, 3419-3440.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B* **58**, 267-288.

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

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. it Statistica Sinica **19**, 801-817.

Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica* **19**, 1281-1296.

Yafeh, Y and Yosha, O. (2003). Large shareholders and banks: who monitors and how? *The Economic Journal* **113**, 128-146.

Yang, L., Lv, S. and Wang, J. (2016). Model-free variable selection in reproducing kernel Hilbert space. *Journal of Machine Learning Research* **17**, 1-24.

Ye, G. and Xie, X. (2012). Learning sparse gradients for variable selection and dimension reduction. *Journal of Machine Learning Research* **87**, 303-355.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with group variables. *Journal of the Royal Statistical Society Series B* **68**, 49-67.

Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics* **220**, 456-463.

Zhu, L., Li, L., Li, R. and Zhu, L. (2011). Model-free feature screening for ultra-high dimensional data. *Journal of the American Statistical Association* **106**, 1464-1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108–1126.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Table 1: The averaged performance measures of various variable selection methods in Example 1.

(n, p, η)	Method	Size	TP	FP	C	U	O
(200,10,0)	MF	5.06	4.98	0.08	45	1	4
	Median	5.02	4.98	0.04	47	1	2
	Cosso	3.88	3.88	0.00	36	14	0
	Add	4.96	4.96	0.00	48	2	0
	RF	5.48	5.00	0.48	31	0	19
	QaSIS	4.68	4.54	0.14	27	19	4
(200,20,0)	MF	5.12	4.94	0.18	40	3	7
	Median	5.22	4.96	0.26	41	2	7
	Cosso	4.22	4.22	0.00	40	10	0
	Add	4.96	4.96	0.00	48	2	0
	RF	5.62	5.00	0.62	24	0	26
	QaSIS	4.82	4.42	0.40	20	22	8
(400,100,0)	MF	5.16	5.00	0.16	44	0	6
	Median	5.10	5.00	5.00	45	0	5
	Cosso	5.10	4.46	0.64	27	13	10
	Add	5.00	5.00	0.00	50	0	0
	RF	5.92	5.00	0.92	30	0	20
	QaSIS	4.82	4.80	0.02	39	10	1
(200,10,0.1)	MF	5.10	4.98	0.12	43	1	6
	Median	5.06	4.98	0.08	45	1	4
	Cosso	3.18	3.18	0.00	27	23	0
	Add	5.00	5.00	0.00	50	0	0
	RF	5.44	5.00	0.44	30	0	20
	QaSIS	4.24	4.20	0.04	19	30	1
(200,20,0.1)	MF	5.06	4.96	0.10	43	2	5
	Median	5.16	4.94	0.22	36	3	11
	Cosso	4.18	4.18	0.00	39	11	0
	Add	4.96	4.96	0.00	48	2	0
	RF	5.70	5.00	0.70	22	0	28
	QaSIS	4.36	4.24	0.08	22	28	3
(400,100,0.1)	MF	5.14	5.00	0.14	44	0	6
	Median	5.14	5.00	0.14	45	0	5
	Cosso	4.62	4.38	0.24	31	14	5
	Add	5.00	5.00	0.00	50	0	0
	RF	6.36	5.00	1.36	14	0	36
	QaSIS	4.56	4.54	0.02	30	19	1

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Table 2: The averaged performance measures of various variable selection methods in Example 2.

(n, p, η)	Method	Size	TP	FP	C	U	O
(200,10,0)	MF	3.14	2.92	0.22	38	4	8
	Median	2.26	2.02	0.24	2	47	1
	Cosso	1.82	1.72	0.10	2	48	0
	Add	2.06	1.72	0.34	1	47	2
	RF	3.46	2.34	1.12	5	33	12
	QaSIS	2.68	2.44	0.22	20	23	7
(200,20,0)	MF	3.24	2.76	0.48	24	9	17
	Median	2.26	2.00	0.26	0	49	1
	Cosso	1.64	1.60	0.04	0	50	0
	Add	1.90	1.38	0.52	1	49	0
	RF	4.24	2.36	1.88	1	32	17
	QaSIS	3.22	2.38	0.84	9	26	15
(400,100,0)	MF	3.16	2.96	0.20	38	2	10
	Median	2.12	2.04	0.08	2	48	0
	Cosso	1.60	1.60	0.00	0	50	0
	Add	1.56	1.56	0.00	0	48	2
	RF	4.46	2.28	2.18	2	36	12
	QaSIS	2.64	2.48	0.16	22	24	4
(200,10,0.1)	MF	3.22	2.76	0.46	27	9	14
	Median	2.38	2.02	0.36	2	46	2
	Cosso	1.80	1.70	0.10	3	46	1
	Add	1.98	1.60	0.38	2	44	4
	RF	3.16	2.28	0.88	5	34	11
	QaSIS	2.10	1.94	0.16	9	36	5
(200,20,0.1)	MF	3.22	2.72	0.50	20	8	22
	Median	2.30	1.92	0.38	1	49	0
	Cosso	1.44	1.42	0.02	0	50	0
	Add	3.60	1.62	1.98	0	42	8
	RF	3.84	2.20	1.64	0	40	10
	QaSIS	2.74	2.30	0.44	14	27	9
(400,100,0.1)	MF	3.32	2.80	0.54	28	8	14
	Median	2.08	1.96	0.12	0	50	0
	Cosso	1.64	1.60	0.04	0	50	0
	Add	4.72	1.72	3.00	0	48	2
	RF	4.38	2.12	1.26	1	44	5
	QaSIS	2.26	2.16	0.10	14	34	2

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Table 3: The selected variables, as well as the corresponding averaged prediction errors, by various selection methods in the Japanese industrial chemical firm dataset.

Variables	MF	Median	Cosso	Add	RF	QaSIS
ASSETS	-	-	-	-	-	-
AGE	-	-	√	-	-	-
LEVERAGE	√	√	√	√	√	√
VARS	√	√	√	√	√	√
OPERS	√	√	√	√	√	-
TOP10	-	-	-	-	√	-
TOP5	-	-	-	-	-	-
OWNIND	-	-	√	-	√	-
AOLC	-	-	√	-	√	√
SHARE	-	√	√	√	-	√
BDHIND	-	√	-	√	-	√
BDA	√	-	-	-	√	√
Pred. Err.	0.273	0.316	0.286	0.316	0.276	0.296
S.D.	(0.006)	(0.008)	(0.006)	(0.008)	(0.007)	(0.007)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0222

Figure 1: The scatter plot of MH5 against the selected variables by MF in the Japanese industrial chemical firm dataset. The solid lines are the fitted curve by local smoothing, and the dashed lines display the fitted mean plus or minus one standard deviation.