Gradient-induced Model-free Variable Selection Based on Composite Quantile Regression in Reproducing Kernel Hilbert Space

Xin Tan¹, Yan Kong¹ and Mei Yan²

¹ University of Electronic Science of Technology of China, Chengdu, China
² Yunnan Normal University, Yunnan, China
E-mail: tanxin_happy@qq.com

Abstract. Variable selection plays an important role to identify truly informative variables in high-dimensional data analysis. In his paper, we propose a variable selection method with composite quantile regression in reproducing kernel Hilbert space (RKHS), which has two main advantages. The first is that our method requires no special model structure assumption and no independence of error term. It is suitable for general non-parametric models and even heteroscedastic models. The second is that the calculation is simple and fast. So, it can also work in high-dimensional situations. Finally, the numerical experiments and real data analysis demonstrate its superior performance in variable selection.

1. Introduction
When analyzing high-dimensional data, we often believed that only a few variables really provide information, which is called the sparsity of data. Therefore, identifying the truly informative variables is regarded as one of the main goals of high-dimensional data analysis and many practical applications (such as health research). In literature, a wide spectrum of variable selection methods had been proposed. For example, the least absolute shrinkage and selection operator (LASSO, [1]), the smoothly clipped absolute deviation (SCAD, [2]), the adaptive LASSO ([3]), the group LASSO ([4]) for the linear model assumption; [5] , [6] , the component selection and smoothing operator (COSSO, [7]), [8] for the additive structure assumption. The above methods all have requirements for the model structure. In recent years, a type of model-free variable selection method has been proposed, which do not require any special structure of the model. [9] introduced a new variable selection method based on the measurement error model, which can be applied to non-parametric kernel regression. The model-free variable selection method is based on the idea that a variable is truly informative with respect to the regression function if the gradient of the regression function is away from zero. Thus, this type of variable selection method focuses on learning the gradient function. In this paper, we focus on estimating the gradient by directly estimating the quantile function in a reproducing kernel Hilbert space (RKHS, [10]). Based on this idea, [11] and [12] both proposed a model-free variable selection method which identified the unimportant variables by whether the gradient of the regression function is zero. However, their methods are not applicable to the case that the error term is independent of variables X, which can be shown later. To overcome this problem, [13] utilized whether the gradient of the composite quantile function is zero to distinguish important variables. However, their calculations are complex and need iteratively optimize O(p) vectors, which is not suitable for high-dimensional setting. Assuming that the conditional quantiles belong to some smoothing RKHS, we propose an efficient sparse learning procedure, where we first estimate the conditional quantile functions at multiple quantile levels, then
integrate their gradients and finally select informative variables via hard thresholding. We just need to solve $m$ (shown later and independent of $p$) optimization problems separately, and thus it is computationally efficient for high-dimensional data.

The rest of this article is organized as follows. Section 2 introduces our method and three-step procedure. In section 3, we provide a method to choose parameter $\lambda$ and $\nu_n$. The simulation studies and real applications are contained in Section 5 and 6. Finally, a brief discussion is given to conclude this work.

2. Method

In this paper, we consider the non-parametric model

$$Y = f(X) + \varepsilon,$$

where $X \in \mathbb{R}^p$ and $\varepsilon$ may not independent of $X$. Note that $X^l$ is regarded as uninformative if and only if

$$g^l_t(x) = \partial_t Q_t(X) = \frac{\partial Q_t(X)}{\partial x^l} \equiv 0, \text{ for any } x \text{ and } \tau \in (0,1),$$

where $Q_t(X)$ is the $\tau$-th conditional quantile function of $Y$ given $X$. Thus, based on this property, we get the following procedure to select informative variables:

**Procedure of our method**

**Step 1:** Obtain an estimation $\hat{Q}_{tk}(x)$ in RKHS for any given $\tau_k \in (0,1)$, $k = 1,2,\cdots, m$.

**Step 2:** Compute $\hat{g}^l_{tk} = \frac{\partial_t q_{tk}(x)}{\partial x^l}$ for $l = 1,2,\cdots,p$ and $k = 1,2,\cdots, m$.

**Step 3:** Identify the informative variables by checking the norm of $\hat{g}^l = (\hat{g}^l_1,\cdots, \hat{g}^l_m)$ and the norm is defined as $\|\hat{g}^l\|_m^2 = \frac{1}{m} \sum_{k=1}^m \|\hat{g}^l_{tk}\|_n^2$.

For the step 1, we employ the KQR (kernel quantile regression, [14]) method,

$$\hat{Q}_{tk}(\cdot) = \sum_{i=1}^n \alpha^{(k)}_i \mathcal{K}(X_i,\cdot) = K_n(\cdot)\alpha^{(k)},$$

where $\alpha^{(k)} = (\alpha^{(k)}_1, \alpha^{(k)}_2, \cdots, \alpha^{(k)}_n)^T$ and $K_n(\cdot) = (\mathcal{K}(X_1,\cdot), \mathcal{K}(X_2,\cdot), \cdots, \mathcal{K}(X_n,\cdot))^T$.

Next, for the step 2, it follows from Lemma 1 of [13]. This lemma implies that if we want to estimate the gradient of $Q_t$ within the smooth RKHS, it suffices to estimate $Q_t$ without loss of information. Thus, if $\hat{Q}_{tk}$ is obtained in step 1, $g^l_{tk}(x)$ can be estimated as $\hat{g}^l_{tk}(x) = (\partial_t K_n(x))\alpha^{(k)}$ where $\partial_t K_n(x) = (\partial_t \mathcal{K}(X_1,x), \partial_t \mathcal{K}(X_2,x), \cdots, \partial_t \mathcal{K}(X_n,x))$.

For the step 3, note that $\mathcal{A}^* = \{l : \int_0^1 E[|g^l_t(X)|^2] \, dt \geq 0\}$. Since we don't know the distribution of $X$, in practice, we apply the empirical norm, i.e., $\|\hat{g}^l_{tk}\|_n^2 = \frac{1}{n} \sum_{i=1}^n (\hat{g}^l_{tk}(X_i))^2$. Thus, we can estimate the active set as $\hat{A} = \{l : \|\hat{g}^l\|_m^2 > \nu_n\}$ where $\nu_n$ is a parameter.
3. Choosing parameter $\lambda$ and $\nu_n$

Based on some other literatures, for example, [13] and [15], we set $\lambda = 0.001$ in our experiments. Then we mainly focus on the choice of $\nu_n$. Note that the $\nu_n$ control the variable selection performance, thus we employ the Bayes information criterion (BIC, [16] and [16]) to choose it. Firstly, for a given $\nu_n$, we can get $\hat{A}$ by our proposed method. Secondly, we can get the quantile estimation by step 1 based on sample $\{(X_{\hat{A},i},Y_i)\}_{i=1}^n$. Thus, we can get the value of $\nu_n$ of correct times of correct-fitting, over-fitting and under-fitting (C, O and U) to evaluate the variable selection performance. The following two examples are considered, which were also used in [12].

Example 4.1. Let $X = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ with $X^{(k)} = (X^{(k)}_1, X^{(k)}_2, \ldots, X^{(k)}_n)^T$ and $X^{(k)}_i = (W^{(k)}_i + rU_i)/(1 + r)$ for $r \in [0,1]$, $k = 1,2,\ldots,p$, $i = 1,2,\ldots,n$ where $W^{(k)}_i$ and $U_i$ are independently from $U(-0.5,0.5)$. When $r = 0$, all variables are independent. The true regression function is $g(X_i) = g_2(1)X^{(1)}_i + g_2(2)X^{(2)}_i + g_3(X^{(3)}_i) + g_4(X^{(4)}_i) + g_5(X^{(5)}_i)$ with $g_2(u) = 2u - 1$, $g_2(u) = 2u + 1$, $g_4(u) = 0.1 \sin(\pi u) + 0.2 \cos(\pi u) + 0.3 \sin^2(\pi u) + 0.4 \cos^2(\pi u) + 0.5 \sin^3(\pi u)$ and $g_5(u) = \sin(\pi u)/(2 - \sin(\pi u))$. Generate the response variable by $Y_i = g(X_i) + \epsilon_i$ where $\epsilon_i$ is distributed as $N(0,1)$. Obviously, the first five variables are the active variables in this example.

Example 4.2. Let $X^{(k)}_i$, $i = 1,2,\ldots,n$, $k = 1,2,\ldots,p$ be generate similarly to Example 4.1 except that $W^{(k)}_i$ and $U_i$ are independently drawn from $U(0,1)$. The model is $Y_i = 4X^{(1)}_iX^{(3)}_i + 3X^{(3)}_i\epsilon_i$ with $\epsilon_i \sim N(0,1)$. Clearly, the first three variables are the informative variables in this example.

For each example, we consider 12 settings with $r = 0$, $0.1$, $n = 200$, 400 and $p = 20$, 50, 100. For each setting, we replicate the simulation 50 times. In this paper, we focus on the times of correct-fitting to compare the variable selection performance and use bold-font to highlight the methods that has the best performance.

As a summary of the simulation results, our proposed method has much outperformance over the other competitors in most scenarios. The results of examples are given in table 1 and table 2. Note that example 4.1 is an additive model and the error term is independent of $X$, thus all six methods are applicable and thus are expected to do well. However, our proposed method still has superior performance yet RF and COSSO get worse when the dimension becomes higher.

| Table 1. The averaged performance measures of variable selection in Example 4.1. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| $p$  | $n$  | Method | $r = 0$ | $r = 0.1$ |
| 20 | 200 | GM_cqr | C | O | U | C | O | U |
| 5.00 | 50 | 0 | 0 | 0 | 5.00 | 50 | 0 | 0 |
| MF | 5.12 | 40 | 3 | 7 | 5.06 | 43 | 2 | 5 |
| RF | 5.42 | 31 | 19 | 0 | 5.74 | 26 | 24 | 0 |
| Median | 5.00 | 0 | 0 | 0 | 5.00 | 50 | 0 | 0 |
MF can also identify the third informative variable. However, its performance is not as good as GM_cqr over other competitors. Our proposed method can find all the informative variables in the regression function. The averaged performance measures of variable selection in Example 4.1.

| p  | n  | Method  | Size | C   | O   | U   | Size | C   | O   | U   |
|----|----|---------|------|-----|-----|-----|------|-----|-----|-----|
| 20 | 200| GM_cqr  | 2.82 | 43  | 0   | 7   | 2.70 | 37  | 0   | 13  |
|    |    | MF      | 3.24 | 24  | 9   | 17  | 3.22 | 20  | 8   | 22  |
|    |    | RF      | 9.08 | 0   | 27  | 23  | 8.80 | 0   | 29  | 21  |
|    |    | Median  | 2.42 | 5   | 16  | 29  | 2.88 | 1   | 10  | 39  |
|    |    | GM      | 2.12 | 0   | 2   | 48  | 2.16 | 0   | 2   | 48  |
|    |    | COSSO   | 2.48 | 0   | 2   | 48  | 2.50 | 2   | 1   | 47  |
5. Real data analysis

In this section, we apply our proposed method to the residential building data. The residential building data at https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set, studies the construction costs and sale price corresponding to real estate single-family residential apartments in Tehran, Iran. There are 105 covariates with 372 observations and two response variables. We focus on the cost price in this paper. In our analysis, we take log transformation to the response variable. We randomly take 2/3 of the data as training set and the rest as testing set, and then apply the six methods in the simulations. We repeat the simulations 200 times and use their averages to assess the performance of the methods. The results are given in table 3. GM_cqr and Median selected the same less variables and provides smaller prediction error.

Table 3. The number of selected variables and the prediction errors by various selection methods in the residential building dataset.

| Method | GM_cqr | MF | Median | GM | COSSO | RF |
|--------|--------|----|--------|----|-------|----|
| Number of selected variables | 2 | 9 | 2 | 1 | 105 | 49 |
| Prediction error ($10^{-2}$) | 5.46 | 17.13 | **4.74** | 56.08 | 6.85 | 112.92 |

(1.41) (2.36) (1.30) (12.86) (12.72) (7.38)
6. Conclusion
In this paper, we focus on estimating the gradient by estimating the quantile function in RKHS. Our propose method benefits the following two points: (i) Our proposed method is model-free and also applicable to heteroscedastic models which means that there may be informative variables in the error term; (ii) The calculation procedure of our method is provided. And it is also applicable to high-dimensional settings.

7. References
[1] Tibshirani, R 1996 Regression shrinkage and selection via the lasso Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–88
[2] Fan, J and Li, R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties Journal of the American statistical Association, 96(456):1348–60
[3] Zou, H 2006 The adaptive lasso and its oracle properties Journal of the American statistical association, 101(476):1418–29
[4] Yuan, M and Lin, Y 2006 Model selection and estimation in regression with grouped variables Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67
[5] Shively, T S, Kohn, R, and Wood, S 1999 Variable selection and function estimation in additive nonparametric regression using a data-based prior Journal of the American Statistical Association, 94(447):777–94
[6] Huang, J Z and Yang, L 2004 Identification of non-linear additive autoregressive models Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(2):463–77
[7] Lin, Y and Zhang, H H 2006 Component selection and smoothing in multivariate nonparametric regression The Annals of Statistics, 34(5):2272–97
[8] Xue, L (2009) Consistent variable selection in additive models Statistica Sinica, 19:1281–96
[9] Stefanski, L, Wu, Y, and White, K 2014 Variable selection in nonparametric classification via measurement error model selection likelihoods Journal of the American Statistical Association, 109(506):574–89
[10] Wahba, G 1999 Support vector machines, reproducing kernel Hilbert spaces, and randomized gacv Advances in kernel methods, 6:69–88
[11] Yang, L, Lv, S, and Wang, J 2016 Model-free variable selection in reproducing kernel Hilbert space The Journal of Machine Learning Research, 17(1):2885–2908
[12] He, X, Wang, J, and Lv, S 2018b Gradient-induced model-free variable selection with composite quantile regression Statistica Sinica, 28(3):1521–38
[13] He, X, Wang, J, and Lv, S 2018a Efficient kernel-based variable selection with sparsistency arXiv preprint arXiv:1802.09246
[14] Huang, J, Horowitz, J L and Wei, F 2010 Variable selection in nonparametric additive models Annals of statistics, 38(4):2282-313
[15] Wang, X and Leng, C 2016 High dimensional ordinary least squares projection for screening variables Journal of The Royal Statistical Society Series B-statistical Methodology, 78(3):589–611
[16] Schwarz, G 1978 Estimating the dimension of a model Annals of statistics, 6(2): 461–64
[17] Watanabe, S 2013 A widely applicable bayesian information criterion Journal of Machine Learning Research, 14(Mar):867–97
[18] Breiman, L 2001 Random forests Machine learning, 45(1):5–32

Acknowledgment
This work was supported by the Doctoral Foundation of Yunnan Normal University (No. 00800205020503127) and the Yunnan Provincial Education Department (No. 2021J0434).