entropy

MDPI

Article

Robust Variable Selection and Estimation Based on Kernel Modal Regression

Changying Guo †, Biqin Song †, Yingjie Wang †, Hong Chen * and Huijuan Xiong *

College of Science, Huazhong Agricultural University, Wuhan 430070, China; gcy2019666@163.com (C.G.); biqin.song@mail.hzau.edu.cn (B.S.); yjaywang@126.com (Y.W.)

* Correspondence: chenh@mail.hzau.edu.cn (H.C.); hj_xiong@mail.hzau.edu.cn (H.X.)

† These authors contributed equally to this work.

Received: 19 February 2019; Accepted: 9 April 2019; Published: 16 April 2019

Abstract: Model-free variable selection has attracted increasing interest recently due to its flexibility in algorithmic design and outstanding performance in real-world applications. However, most of the existing statistical methods are formulated under the mean square error (MSE) criterion, and susceptible to non-Gaussian noise and outliers. As the MSE criterion requires the data to satisfy Gaussian noise condition, it potentially hampers the effectiveness of model-free methods in complex circumstances. To circumvent this issue, we present a new model-free variable selection algorithm by integrating kernel modal regression and gradient-based variable identification together. The derived modal regression estimator is related closely to information theoretic learning under the maximum correntropy criterion, and assures algorithmic robustness to complex noise by replacing learning of the conditional mean with the conditional mode. The gradient information of estimator offers a model-free metric to screen the key variables. In theory, we investigate the theoretical foundations of our new model on generalization-bound and variable selection consistency. In applications, the effectiveness of the proposed method is verified by data experiments.

Keywords: modal regression; maximum correntropy criterion; variable selection; reproducing kernel Hilbert space; generalization error

MSC: 62J02; 68T05; 62F35

1. Introduction

Variable selection has attracted increasing attention in the machine learning community due to the massive requirements of high-dimensional data mining. Under different motivations, many variable selection methods have been constructed and shown promising performance in various applications. From the viewpoint of hypothesis function space, there are mainly two types of variable selection approaches with respect to linear assumption and nonlinear additive assumption, respectively. For the linear model assumption, variable selection algorithms are usually formulated based on the least-squares empirical risk and the sparsity-induced regularization, which include Least Absolute Shrinkage and Selection Operator (Lasso) [1], Group Lasso [2] and Elastic net [3] as special examples. For the nonlinear additive model assumption, various additive models have been developed to relax the linear restriction on regression function [4,5]. It is well known that additive models enjoy the flexibility and interpretability of their representation and can remedy the curse of dimensionality of high-dimensional nonparametric regression [6–8]. Typical examples of additive models include Sparse Additive Models (SpAM) [9], Component Selection and Smoothing Operator (COSSO) [10] and Group Sparse Additive Models (GroupSpAM) [11]. Most of the above approaches are formulated under

Tikhonov regularization scheme with special hypothesis function space (e.g., linear function space, nonlinear function space with additive structure).

More recently, some works have been made in [12–15] to alleviate the restriction on the hypothesis function space, which just require that the regression function belongs to a reproducing kernel Hilbert space (RKHS). In contrast to the traditional structure assumption on regression function, these methods identify the important variable via the gradient of kernel-based estimator. There are two strategies to improve the model flexibility through the gradient information of predictor. One follows the learning gradient methods in [13,14,16], where the functional gradient is used to construct the loss function for forming the empirical risk. Under this strategy, two model-free variable selection methods are presented by combining the error metric associated with the gradient information of estimator and the coefficient-based $\ell_{2,1}$-regularizer in [13] and $\| \cdot \|_K$-regularizer in [14], respectively. In particular, the variable selection consistency is also established based on the properties of RKHS and mild parameter conditions (e.g., the regularization parameter, the width of kernel). The other follows the structural sparsity issue in [15,17], where the functional gradient is employed to construct the sparsity-induced regularization term. Rosasco et al. in [17] proposes a least-squares regularization scheme with nonparametric sparsity, which can be solved by an iterative procedure associated with the theory of RKHS and proximal methods. Magda et al. [15] introduces a nonparametric structured sparsity by considering two regularizers based on partial derivatives and offers its optimization with the alternating direction method of multiples (ADMM) [18]. Moreover, to further improve the computation feasibility, a three-step variable selection algorithm is developed in [12] with the help of the three building blocks: kernel ridge regression, functional gradient in RKHS, and a hard threshold. Meanwhile, the effectiveness of the proposed algorithm in [12] is supported by theoretical guarantees on variable selection consistency and empirical verification on simulated data.

Despite the aforementioned methods showing promising performance for identifying the active variables, all of them rely heavily on the least-squares loss under the MSE criterion, which is sensitive to non-Gaussian noise [19,20], e.g., the heavy-tailed noise, the skewed noise, and outliers. In essence, learning methods under MSE aim to find an approximator to the conditional mean based on empirical observations. When the data are contaminated by a complex noise without zero mean, the mean-based estimator is difficult to reveal with the intrinsic regression function. This motivates us to formulate a new variable selection strategy in terms of other criterion with respect to different statistical metric (e.g., the conditional mode). Following the research line in [12,19], we consider a new robust variable selection method by integrating the issues of modal regression (for estimating the conditional mode function) and variable screening based on functional derivatives. To the best of our knowledge, this is the first paper to address robust model-free variable selection.

Statistical models for learning the conditional mode can be traced back to [21,22], which include the local modal regression in [23,24] and the global modal regression in [25–27]. Recently, the idea of modal regression has been successfully incorporated into machine learning methods from theoretical analysis [19] and application-oriented studies (e.g., cognitive impairment prediction [20] and cluster estimation [28]). Particularly, Feng et al. [19] considers a learning theory approach to modal regression and illustrates some relations between modal regression and learning under the maximum correntropy criterion [29–31]. In addition, Wang et al. [20] formulates a regularized modal regression (RMR) under modal regression criterion (MRC), and establishes its theoretical characteristics on generalization ability, robustness, and sparsity. It is natural to extend the RMR under linear regression assumption to general model-free variable selection setting.

Inspired by recent works in [12,19], we propose a new robust gradient-based variable selection method (RGVS) by integrating the RMR in RKHS and the model-free strategy for variable screening. Here, the kernel-based RMR is used to construct the robust estimator, which can reveal the truly conditional mode, even when facing data with non-Gaussian noise and outliers. Moreover, we evaluate the information quantity of each input variable by computing the corresponding gradient of estimator. Finally, a hard threshold is used to identify the truly active variables after offering the empirical norm of each gradient associated with hypothesis function. The above three steps assure the robustness and flexibility of our new approach.

To better highlight the novelty of RGVS, we present Table 1 to illustrate its relation with other related methods, e.g., linear models (Lasso [1], RMR [20]), additive models (SpAM [9], COSSO [10]), and General variable selection Method (GM) [12].

Our main contributions can be summarized as follows.

- We formulate a new RGVS method by integrating the RMR in RKHS and the model-free strategy for variable screening. This algorithm can be implemented via the half-quadratic optimization [32]. To our knowledge, this algorithm is the first one for robust model-free variable selection.
- In theory, the proposed method enjoys statistical consistency on regression estimator under much general conditions on data noise and hypothesis space. In particular, the learning rate with polynomial decay $\mathcal{O}(n^{-\frac{2}{5}})$ is obtained, which is faster than $\mathcal{O}(n^{-\frac{1}{7}})$ in [20] for linear RMR. It should be noted that our work is established under the MRC, while all previous model-free methods are formulated under the MSE criterion. In addition, variable selection consistency is obtained for our approach under a self-calibration condition.
- In application, the proposed RGVS shows the empirical effectiveness on both simulated and real-world data sets. In particular, our approach can achieve much better performance than the model-free algorithm in [12] for complex noise data, e.g., containing Chi-square noise, Exponential noise, and Student noise. Experimental results together with theoretical analysis support the effectiveness of our approach.

The rest of this paper is organized as follows. After recalling the preliminaries of modal regression, we formulate the RGVS algorithm in Section 2. Then, theoretical analysis, optimization algorithm, and empirical evaluation are provided from Section 3 to Section 5 respectively. Finally, we conclude this paper in Section 6.

Table 1. Properties of different regression algorithms.

	Lasso [1]	**RMR** [20]	**SpAM** [9]	**COSSO** [10]	**GM** [12]	**Ours**
Learning criterion	MSE	MRC	MSE	MSE	MSE	MRC
Model assumption	linear	linear	additive	additive	model-free	model-free

2. Gradient-Based Variable Selection in Modal Regression

Let $\mathcal{X} \in \mathbb{R}^p$ and $\mathcal{Y} \in \mathbb{R}$ be a compact input space and an output space, respectively. We consider the following data-generating setting

$$y = f^*(x) + \epsilon, \tag{1}$$

where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and ϵ is a random noise. For the feasibility of theoretical analysis, we denote the intrinsic distribution of $(x, y) \in \mathcal{Z} := (\mathcal{X}, \mathcal{Y})$ generated in (1) as ρ. Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n \in \mathcal{Z}^n$ be empirical observations drawn independently according to the unknown distribution ρ. Unlike sparse methods with certain model assumption (e.g., Lasso [1], SpAM [9]), the gradient-based sparse algorithms [12,13] mainly aim at screening out the informative variables according to the gradient information of intrinsic function. For input vector $u = (u_1, ..., u_p)^T \in \mathbb{R}^p$, the variable information is characterized by the gradient function $g_j^*(u) := \partial f^*(u)/\partial u_j$. Clearly, $g_j^*(u) = 0$ implies that the j-th variable is uninformative [12,17]. Considering an ℓ_2-norm measure on the partial derivatives, we denote the true active set as

$$\mathcal{S}^* = \{j : \|g_j^*\|_2^2 > 0\}, \tag{2}$$

where $\|g_j^*\|_2^2 = \int_{\mathcal{X}} (g_j^*(x))^2 d\rho_{\mathcal{X}}(x)$ and $\rho_{\mathcal{X}}$ is the marginal distribution of ρ.

Indeed, all the gradient-based variable selection algorithms [12,13,17] are constructed under Tikhonov regularization scheme in RKHS $\mathcal{H}_K$ [33,34]. The RKHS $\mathcal{H}_K$ associated with the Mercer kernel K is the closure of the linear span of $\{K_x := K(x,\cdot) : x \in \mathcal{X}\}$. Such a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive semi-definite function. Denote $< \cdot,\cdot >_K$ as the inner product in $\mathcal{H}_K$, the reproducing properties of RKHS means $< f, K_x >_K = f(x), \forall f \in \mathcal{H}_K$.

2.1. Gradient-Based Variable Selection Based on Kernel Least-Squares Regression

In this subsection, we recall the gradient-based variable selection algorithm in [12] associated with least-squares error metric. When the noise ϵ in (1) satisfies $\mathbb{E}(\epsilon|X) = 0$ (i.e., Gaussian noise), the regression function equals to the conditional mean, which can be represented by

$$f^*(x) = \mathbb{E}(Y|X = x) = \int_{\mathcal{Y}} y d\rho_{Y|X=x}(y). \tag{3}$$

Here $\rho_{Y|X=x}$ denotes the conditional distribution of Y given x. Theoretically, the regression function f^* in (3) is the minimizer of expected least-squares risk

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (y - f(x))^2 d\rho(x,y).$$

As ρ is unknown in practice, we cannot get f^* directly by minimizing $\mathcal{E}(f)$ over certain hypothesis space. Given training samples $\mathbf{z}$, the empirical risk with respect to the expected risk $\mathcal{E}(f)$ is denoted by

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2. \tag{4}$$

The gradient-based variable selection algorithm in [12] depends on the estimator defined as below:

$$\tilde{f}_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_K} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda ||f||_K^2\}, \tag{5}$$

where $\lambda > 0$ is the regularization parameter and $||f||_K$ is the kernel-norm of f. The properties of RKHS [33] assure that

$$\tilde{f}_{\mathbf{z}}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i K(x_i, x) = (\tilde{\alpha}_{\mathbf{z}})^T \mathbf{K}_n(x),$$

where $\mathbf{K}_n(x) = (K(x_1,x), \cdots, K(x_n,x))^T \in \mathbb{R}^n$ and $\tilde{\alpha}_{\mathbf{z}} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_n)^T \in \mathbb{R}^n$. Denote $\mathbf{K} = (\mathbf{K}_n(x_1), \cdots, \mathbf{K}_n(x_n)) \in \mathbb{R}^{n \times n}$ and $\mathbf{Y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$, the closed-form solution is

$$\tilde{\alpha}_{\mathbf{z}} = (\mathbf{K}^T\mathbf{K} + n\lambda\mathbf{K})^{-1}\mathbf{K}\mathbf{Y}.$$

Following Lemma 1 in [12], for any $j \in \{1, \cdots, p\}$, we have $\tilde{g}_j(x) = (\tilde{\alpha}_{\mathbf{z}})^T \partial_j \mathbf{K}_n(x)$, where

$$\partial_j \mathbf{K}_n(u) = \Big(\frac{\partial K(x_1,u)}{\partial u_j}, ..., \frac{\partial K(x_n,u)}{\partial u_j}\Big)^T, u = (u_1, ..., u_p)^T \in \mathcal{X}.$$

After imposing the empirical norm on $\tilde{g}_j$, i.e.,

$$\|\tilde{g}_j\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{g}_j(x_i))^2,$$

we get the estimated active set

$$\tilde{\mathcal{S}} = \{j : \|\tilde{g}_j\|_n^2 > v_n\},$$

where v_n is a pre-configured constant for variable selection.

The general variable selection method has shown some theoretical advantages in [12], e.g., the representation flexibility and the computation feasibility. However, the gradient-based method [12] may result in a degraded performance for real-world data without the zero-mean noise condition. Inspired by the modal regression [19,35] to learn the conditional mode, we propose a new robust gradient-based variable selection method under much general noise condition.

2.2. Robust Gradient-Based Variable Selection Based on Kernel Modal Regression

Unlike the traditional zero-mean noise assumption [12,17], the modal regression requires that the conditional mode of random noise ϵ is zero for any $x \in \mathcal{X}$, i.e.,

$$\text{mode}(\epsilon|X=x) = \arg\max_{t\in\mathbb{R}} P_{\epsilon|X}(t|X=x) = 0,$$

where $P_{\epsilon|X}$ is the conditional density of ϵ conditioned on X. In fact, this assumption imposes no restrictions on conditional mean, and can include the heavy-tailed noise, the skewed noise, and outliers. Then, we can verify that the mode-regression function

$$f^*(x) = \text{mode}(Y|X) = \arg\max_{t\in\mathbb{R}} P_{Y|X}(t|X=x), \tag{6}$$

where $P_{Y|X}(\cdot|X)$ denotes the conditional density of Y conditioned on $x \in \mathcal{X}$. It is worth noting that $P_{Y|X}(\cdot|X=x)$ is assumed to be unique and existing here. As shown in [19,20], f^* in (6) is the maximizer of the MRC over all measurable functions, which is defined as

$$\mathcal{R}(f) = \int_{\mathcal{X}} P_{Y|X}(f(x)|X=x)d\rho_{\mathcal{X}}(x).$$

The maximizer of $\mathcal{R}(f)$ is difficult to be obtained since both $P_{Y|X}$ and $\rho_{\mathcal{X}}$ are unknown. Fortunately, Theorem 5 of [19] has proved that $\mathcal{R}(f) = P_{E_f}(0)$, where $P_{E_f}(0)$ is the density function of $E_f = Y - f(x)$ at 0 and which can be easily approximated by the kernel density method [20]. With the help of modal kernel $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ for the density estimation, we can formulate an empirical kernel density estimator $\hat{P}_{E_f}$ at 0

$$\hat{P}_{E_f}(0) \quad = \quad \frac{1}{n\sigma}\sum_{i=1}^{n} K_\sigma(y_i - f(x_i), 0) = \frac{1}{n\sigma}\sum_{i=1}^{n} K_\sigma(y_i, f(x_i)) := \mathcal{R}^\sigma_{\mathbf{z}}(f). \tag{7}$$

Setting $\phi(\frac{y-f(x)}{\sigma}) := K_\sigma(y, f(x))$, we get the corresponding expected version

$$\mathcal{R}^\sigma(f) = \frac{1}{\sigma}\int_{\mathcal{X}\times\mathcal{Y}} \phi(\frac{y-f(x)}{\sigma})d\rho(x,y). \tag{8}$$

In addition, the modal regression also can be interpreted by minimizing a mode-induced error metric [19]. When $\phi(u) \le \phi(0)$ for any $u \in \mathbb{R}$, the mode-induced loss can be defined as

$$\mathcal{L}_\sigma(y - f(x)) = \sigma^{-1}(\phi(0) - \phi((y-f(x))/\sigma)),$$

which is related closely with the correntropy-induced loss in [19,36]. Given training samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, we can formulate the RMR in RKHS as

$$f_{\mathbf{z}} = \arg\max_{f\in\mathcal{H}_K} \{\mathcal{R}^\sigma_{\mathbf{z}}(f) - \lambda||f||_K^2\}, \tag{9}$$

where $\lambda > 0$ is a turning parameter that controls the complexity of the hypothesis space, and $\|f\|_K^2 = \langle f, f\rangle_K$ is the kernel-norm of $f \in \mathcal{H}_K$.

Denote $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)^T \in \mathbb{R}^n$, $\mathbf{K}_n(\mathbf{x}) = (K(x_1, x), ..., K(x_n, x))^T \in \mathbb{R}^n$ and $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$. From the representer theorem of kernel methods, we can deduce that

$$f_{\mathbf{z}}(x) = \sum_{i=1}^n \hat{\alpha}_{\mathbf{z},i} K(x_i, x) = \hat{\alpha}_{\mathbf{z}}^T \mathbf{K}_n(x),$$

with

$$\hat{\alpha}_{\mathbf{z}} = \arg\max_{\alpha \in \mathbb{R}^n} \{ \frac{1}{n\sigma} \sum_{i=1}^n \phi(\frac{y_i - \mathbf{K}_n^T(x_i)\alpha}{\sigma}) - \lambda \alpha^T \mathbf{K} \alpha \}. \tag{10}$$

From Lemma 1 in [12], we know that for any $f \in \mathcal{H}_K$ and $u = (u_1, ..., u_p)^T \in \mathbb{R}^p$,

$$\hat{g}_j(u) = \frac{\partial f(u)}{\partial u_j} = \langle f, \frac{\partial K(u, \cdot)}{\partial u_j} \rangle_K.$$

The empirical measure on gradient function $\hat{g}_j(x)$ is

$$\|\hat{g}_j\|_n^2 = \frac{1}{n} \sum_{i=1}^n (\hat{g}_j(x_i))^2 = \frac{1}{n} \sum_{i=1}^n (\hat{\alpha}_{\mathbf{z}}^T \partial_j \mathbf{K}_n(x_i))^2. \tag{11}$$

Then, the identified active set can be written as

$$\hat{\mathcal{S}} = \{j : \|\hat{g}_j\|_n^2 > v_n\}, \tag{12}$$

where v_n is a positive threshold selected under the sample-adaptive tuning framework [37].

3. Generalization-Bound and Variable Selection Consistency

This section establishes the theoretical guarantees on generalization ability and variable selection for the proposed RGVS. Firstly, we introduce some necessary assumptions.

Assumption 1. *The representing function ϕ associated with modal kernel $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ satisfies: (i) ϕ is bounded with $\int_{\mathbb{R}} \phi(u)du = 1$, $\phi(u) = \phi(-u)$ and $\phi(u) \le \phi(0), \forall u \in \mathbb{R}$; (ii) $\phi(\cdot)$ is differentiable with $\|\phi^{'}\|_\infty < \infty$ and $\int_{\mathbb{R}} u^2 \phi(u) du < \infty$.*

Observe that some smoothing kernels meet Assumption 1, such as Gaussian kernel and Logistic kernel, etc.

Assumption 2. *The conditional density function $P_{\epsilon|X}$ is second-order differentiable and $\|P_{\epsilon|X}^{''}\|_\infty$ is bounded.*

Assumption 2 has been used in [19,20], which assures upper bound on $|\mathcal{R}(f) - \mathcal{R}^\sigma(f)|$ together with Assumption 1.

Assumption 3. *Let $\mathcal{C}^s$ be a space of s-times continuous differentiable functions. Assume that $\sup_{x \in \mathcal{X}} \sqrt{K(x, x)} < \infty$ with $K \in \mathcal{C}^s$ with $s > 0$, and for a given constant M, the target function satisfies $f^* \in \mathcal{H}_K$ with $\|f^*\|_\infty \le M$.*

Assumption 3 has been used extensively in learning theory literatures, see, e.g., [38–44]. In particular, the Gaussian kernel belongs to $\mathcal{C}^\infty$.

Our error analysis begins with the following inequality in [19], where the relationship between $\mathcal{R}^\sigma(f)$ and $\mathcal{R}(f)$ is provided.

Lemma 1. *Under Assumptions 1–2, there holds*

$$\left|\mathcal{R}(f^*) - \mathcal{R}(f) - (\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f))\right| \le c_1 \sigma^2$$

for any measurable function f*:* $\mathcal{X} \to \mathbb{R}$*, where* $c_1 = ||P''_{\epsilon|x}||_\infty \int_{\mathbb{R}} u^2\phi(u)du$*.*

This indicates us to bound the excess risk $\mathcal{R}(f^*) - \mathcal{R}(f)$ via estimating $\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}})$. To be specific, we further make an error decomposition as follows.

Lemma 2. *Under Assumptions 1–3, there holds*

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \le \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})) + \lambda||f^*||_K^2 + c_1\sigma^2.$$

Proof. According to the definition of $f_{\mathbf{z}}$ in (9), we have

$$\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \lambda\|f^*\|_K^2 - (\mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \lambda\|f_{\mathbf{z}}\|_K^2) \le 0.$$

Then, we can deduce that

$$\begin{aligned} \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) &= \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f^*) + \mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \lambda\|f^*\|_K^2 + \lambda\|f^*\|_K^2 \\ &\quad -(\mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \lambda\|f_{\mathbf{z}}\|_K^2) - \lambda\|f_{\mathbf{z}}\|_K^2 + \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \\ &\le \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f^*) + \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \\ &\quad +\{\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \lambda\|f^*\|_K^2 - (\mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \lambda\|f_{\mathbf{z}}\|_K^2)\} + \lambda\|f^*\|_K^2 \\ &\le \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})) + \lambda\|f^*\|_K^2. \end{aligned}$$

This together with Lemma 1 yields the desired result. □

Observe that $\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}))$ characterizes the divergence between the data-free risk $\mathcal{R}^\sigma(f)$ and the empirical risk $\mathcal{R}^\sigma_{\mathbf{z}}(f)$. To establish its uniform estimation, we need to give the upper bound of $\|f_{\mathbf{z}}\|_K$ firstly.

According to the definition of $f_{\mathbf{z}}$, we have

$$\mathcal{R}^\sigma_{\mathbf{z}}(0) \le \mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \lambda\|f_{\mathbf{z}}\|_K^2.$$

Then,

$$\|f_{\mathbf{z}}\|_K \le \sqrt{\frac{\mathcal{R}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{R}^\sigma_{\mathbf{z}}(0)}{\lambda}} \le \sqrt{\frac{\|\phi\|_\infty}{\lambda\sigma}}.$$

Lemma 3. *For* $f_{\mathbf{z}}$ *in (9), there holds*

$$\|f_{\mathbf{z}}\|_K \le \sqrt{\frac{\|\phi\|_\infty}{\lambda\sigma}}.$$

Lemma 3 tells us that $f_{\mathbf{z}} \in \mathcal{B}_r$ with $r = \sqrt{\frac{\|\phi\|_\infty}{\lambda\sigma}}$ for any $\mathbf{z} \in \mathcal{Z}^n$, where $\mathcal{B}_r = \{f \in \mathcal{H}_K : \|f\|_K \le r\}$. This motivates us to measure the capacity of $\mathcal{B}_r$ through the empirical covering number [45].

Definition 1. *Suppose that* $\mathcal{F}$ *is a set of functions on* $\mathbf{x} = \{x_1, ..., x_n\}$ *with the* ℓ_2*-empirical metric* $d_{2,\mathbf{x}}(f,g) = \left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2\right)^{\frac{1}{2}}$*,* $\forall f, g \in \mathcal{F}$*. Then, the* ℓ_2*-empirical covering number of function set* $\mathcal{F}$ *is defined as*

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{n\in N}\sup_{\mathbf{x}} \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon), \epsilon > 0,$$

where

$$\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon) = \inf\left\{l \in N : \exists\{f_j\}_{j=1}^l \subset \mathcal{F}, s.t., \mathcal{F} \subset \cup_{j=1}^l B(f_j, \epsilon)\right\}$$

with $B(f_j, \epsilon) = \{f \in \mathcal{F} : d_{2,\mathbf{x}}(f, f_j) < \epsilon\}$

Next, we introduce a concentration inequality established in [46].

Lemma 4. *Let $\mathcal{T}$ be a function set associated with function t. Suppose that there are some constants $B, c_s, c_\theta > 0$ and $s \in [0,1]$ satisfying $\|t\|_\infty \le B$, $Et^2 \le c_s(Et)^s$ for any $t \in \mathcal{T}$. If for $0 < \theta < 2$ and $\log \mathcal{N}_2(\mathcal{T}, \epsilon) \le c_\theta \epsilon^{-\theta}, \forall \epsilon > 0$, then for any $0 < \delta < 1$ and given $\mathbf{z} = \{z_i\}_{i=1}^n \subset \mathcal{Z}$, there holds*

$$Et - \frac{1}{n}\sum_{i=1}^n t(z_i) \le \frac{1}{2}\eta^{1-s}(Et)^s + c_\theta' \eta + 2(\frac{c_s \log(1/\delta)}{n})^{\frac{1}{2-s}} + \frac{18B\log(1/\delta)}{n}, \forall t \in \mathcal{T},$$

where c_θ' is a constant only depending on θ and

$$\eta = \max\left\{ c_s^{\frac{2-\theta}{4-2s+\theta s}} (\frac{c_\theta}{n})^{\frac{2}{4-2s+\theta s}}, B^{\frac{2-\theta}{2+\theta}} (\frac{c_\theta}{n})^{\frac{2}{2+\theta}} \right\}.$$

Theorem 1. *Under Assumptions 1–3, taking $\sigma = n^{-\frac{1}{5}}$ and $\lambda = n^{-\frac{2}{5}}$, we have for any $0 < \delta < 1$*

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \le Cn^{-\zeta}\log(\frac{1}{\delta})$$

with confidence at least $1 - \delta$, where $\zeta = \min\left\{\frac{8-9\theta}{20}, \frac{8-6\theta}{5(2+\theta)}, \frac{2}{5}\right\}$ and

$$\theta = \begin{cases} \frac{2p}{p+2s} & 0 < s \le 1 \\ \frac{2p}{p+2} & 1 < s \le 1 + \frac{p}{2} \\ \frac{p}{s} & s > 1 + \frac{p}{2}. \end{cases} \tag{13}$$

Proof. Denote a function-based random variable set by

$$\mathcal{T} = \left\{ t(z) := t_f(z) = \frac{1}{\sigma}(\phi(\frac{y - f^*(x)}{\sigma}) - \phi(\frac{y - f(x)}{\sigma})) : f \in \mathcal{B}_r \right\}.$$

Under Assumption 1, for any $f_1, f_2 \in \mathcal{B}_r$, we have

$$\begin{aligned} |t_{f_1}(z) - t_{f_2}(z)| &= \frac{1}{\sigma}|\phi(\frac{y - f_1(x)}{\sigma}) - \phi(\frac{y - f_2(x)}{\sigma})| \\ &\le \frac{\|\phi'\|_\infty}{\sigma}|\frac{y - f_1(x)}{\sigma} - \frac{y - f_2(x)}{\sigma}| \\ &\le \frac{\|\phi'\|_\infty}{\sigma^2}|f_1(x) - f_2(x)|. \end{aligned}$$

Combining the above inequality and the properties of empirical covering number [40,41], we have

$$\log \mathcal{N}_2(\mathcal{T}, \epsilon) \le \log \mathcal{N}_2(\mathcal{B}_1, \frac{\epsilon\sigma^2}{r\|\phi'\|_\infty}) \le C_\theta' r^\theta \sigma^{-2\theta} \epsilon^{-\theta}, \tag{14}$$

where θ is defined in (13).

According to Assumption 1, there exists $\|t\|_\infty \le \frac{\|\phi\|_\infty}{\sigma}$. Furthermore, we get

$$\begin{aligned} Et^2 &= \frac{\|\phi\|_\infty}{\sigma} E\left[\frac{1}{\sigma}\phi(\frac{y - f^*}{\sigma}) - \frac{1}{\sigma}\phi(\frac{y - f(x)}{\sigma})\right] = \frac{\|\phi\|_\infty}{\sigma}(\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f)) \\ &\le \frac{\|\phi\|_\infty}{\sigma}(P_{E_{f^*}}(0) - P_{E_f}(0) + c_1\sigma^2) \le \frac{\|\phi\|_\infty}{\sigma}(P_{E_{f^*}}(0) - P_{E_f}(0)) + \|\phi\|_\infty c_1 \sigma \\ &\le \sigma^{-1}c_2 + \sigma c_3, \end{aligned} \tag{15}$$

where $c_2 = \|\phi\|_\infty(P_{E_{f^*}}(0) - P_{E_f}(0))$ and $c_3 = c_1\|\phi\|_\infty$ are the bounded constants.

Recalling (14) and (15), we know Lemma 4 holds true for $t \in \mathcal{T}$ with $c_\theta = c_\theta^{'} r^\theta \sigma^{-2\theta}$, $B = \frac{\|\phi\|_\infty}{\sigma}$, $s = 0$, and $c_s = c_2\sigma^{-1} + c_3\sigma$. That is to say, for any $t \in \mathcal{T}$ and $0 < \delta < 1$, with confidence $1 - \delta$

$$\begin{aligned} & \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\mathcal{R}^\sigma_{\mathbf{z}}(f^*) - \mathcal{R}^\sigma_{\mathbf{z}}(f)) \\ \leq \ & (\frac{1}{2} + c_\theta^{'}) \max\Big\{(c_2\sigma^{-1} + c_3\sigma)^{\frac{2-\theta}{4}} (\frac{c_\theta^{'} r^\theta \sigma^{-2\theta}}{n})^{\frac{1}{2}}, (\frac{\|\phi\|_\infty}{\sigma})^{\frac{2-\theta}{2+\theta}} (\frac{c_\theta^{'} r^\theta \sigma^{-2\theta}}{n})^{\frac{2}{2+\theta}}\Big\} \\ & +2\sqrt{\frac{(c_2\sigma^{-1} + c_3\sigma)\log(1/\delta)}{n}} + \frac{18\|\phi\|_\infty \log(1/\delta)}{n\sigma}. \end{aligned} \tag{16}$$

Combining Lemma 2 and (16) with $r = \sqrt{\|\phi\|_\infty / \lambda\sigma}$, we have with confidence at least $1 - \delta$

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \leq C_{n,\sigma,\lambda} \log(\frac{1}{\delta})\big(\max\{\sigma^{-\frac{2+5\theta}{4}} n^{-\frac{1}{2}}, \sigma^{-\frac{2+4\theta}{2+\theta}} n^{-\frac{2}{2+\theta}} \lambda^{-\frac{\theta}{2+\theta}}\} + n^{-\frac{1}{2}}\sigma^{-\frac{1}{2}} + \lambda + \sigma^2\big), \tag{17}$$

where $C_{n,\sigma,\lambda}$ is positive constants independently of n, σ, λ.

Setting $\sigma^2 = n^{-\frac{1}{2}}\sigma^{-\frac{1}{2}}$ and $\lambda = \sigma^2$, we have $\sigma = n^{-\frac{1}{5}}$ and $\lambda = n^{-\frac{2}{5}}$. Putting these selected parameters into (17), we get the desired estimation. □

Theorem 1 provides the upper bound to the excess risk of $f_{\mathbf{z}}$ under the MRC, which extends the previous ERM-based analysis in [19] to the regularized learning scheme. In addition, we can further bound $\|f_{\mathbf{z}} - f^*\|^2_{L^2_{\rho_\mathcal{X}}}$ after imposing Assumption 3 in [19].

Corollary 1. *Let the conditions of Theorem 1 be true. Assume that $K \in \mathcal{C}^\infty$, we have*

$$\mathcal{R}(f^*) - \mathcal{R}(f_{\mathbf{z}}) \leq \mathcal{O}(n^{-\frac{2}{5}} \log(1/\delta))$$

with confidence at least $1 - \delta$.

The learning rate derived in Corollary 1 is faster than $O(n^{-\frac{1}{7}})$ for the linear regularized modal regression [20]. Meanwhile, it should be noted that some kernel functions meet $K \in \mathcal{C}^\infty$, e.g., Gaussian kernel, Sigmoid kernel, and Logistic kernel.

Since the proposed RGVS employs the non-convex mode-induced loss, our variable selection analysis is completely different from kernel method with least-squares loss [12]. Here, we introduce the following self-calibration inequality, which addresses that a weak convergence on risk implies a strong convergence in kernel-norm under certain conditions.

Assumption 4. *For any given σ and $\mathcal{B}_r$ with $r = \|\phi\|_\infty^{\frac{1}{2}} n^{\frac{1}{4}} \sigma^{-\frac{1}{4}}$, there exists a universal constant $C_1 > 0$ such that*

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) \geq C_1\|f^* - f\|^2_K, \forall f \in \mathcal{B}_r.$$

Assumption 4 characterizes the concentration of our estimator near f^* with the kernel-norm metric. Indeed, the current restriction is related to Assumption 4 in [12], Theorem 2.7 in [47] for quartile regression, and the so-called RNI condition in [48,49] as well.

In addition, the following condition is required, which implies that the gradient function associated with truly informative variables is separated well from zero. Similar assumptions can also be found in [12,50]. For simplicity, we denote $\|g_j\|_2 := \inf_{\mathcal{X}}(g_j(x))^2 d\rho_\mathcal{X}(x)$.

Assumption 5. *There exists some constant $C_2 > 0$ such that*

$$\min_{j \in S^*} \|g_j^*\|_2^2 > C_2 n^{-\min\{\frac{1}{8}, \frac{4-\theta}{16+8\theta}\}}.$$

Theorem 2. *Let Assumptions 1–5 be true. For any given* $\sigma > \max\{n^{-\frac{4-\theta}{8+14\theta}}, n^{-\frac{2}{4+10\theta}}\}$*, set* $\lambda = n^{-\frac{1}{2}}\sigma^{-\frac{1}{2}}$ *in (9) and* $v_n = C_2 n^{-min\{\frac{1}{8},\frac{4-\theta}{16+8\theta}\}}$ *in (12). Then,* $Prob\{\hat{S} = S^*\} \to 1 \quad as \quad n \to \infty$.

Proof. As shown in [12], by direct computation, there holds

$$\Big| \|\hat{g}_j\|_n^2 - \|g_j^*\|_2^2 \Big| \leq a_1 \big(3\|f_{\mathbf{z}} - f^*\|_K + \|\hat{D}_j^* \hat{D}_j - D_j^* D_j\|_{HS}\big), \forall j, \tag{18}$$

where HS denotes the Hilbert-Schmidt operator on $\mathcal{H}_K$, $D_j^* D_j f = \int \alpha_j K_x g_j(x) d\rho_{\mathcal{X}}(x)$, $\hat{D}_j^* \hat{D}_j f = \frac{1}{n}\sum_{i=1}^{n} \alpha_j K_{x_j} g_j(x_i)$, and a_1 is a positive constant. The concentration inequality for kernel operator in [17] states that

$$\|\hat{D_j^*} \hat{D}_j - D_j^* D_j\|_{HS} \leqslant \sqrt{\frac{8\kappa^2}{n} \log\big(\frac{4p}{\delta}\big)} \tag{19}$$

with confidence $1 - \delta$.

Meanwhile, with the similar proof of Theorem 1, we can deduce that

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \leqslant a_2 \{ \log(1/\delta) \max\{\sigma^{-\frac{2+5\theta}{4}} n^{-\frac{1}{2}}, \sigma^{-\frac{2+4\theta}{2+\theta}} n^{-\frac{2+4\theta}{2+\theta}}\} + n^{-\frac{1}{2}}\sigma^{-\frac{1}{2}} + \lambda\}$$

with confidence at least $1 - \delta$, where a_2 is a positive constant. Setting $\lambda = n^{-\frac{1}{2}}\sigma^{-\frac{1}{2}}, \sigma^{-\frac{2+5\theta}{4}} n^{-\frac{1}{4}} \leqslant 1$ and $\sigma^{-\frac{4+7\theta}{4+2\theta}} n^{-\frac{4-\theta}{8+4\theta}} \leqslant 1$, we further get

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \leqslant a_3 \log(1/\delta) n^{-\min\{\frac{1}{4},\frac{4-\theta}{8+4\theta}\}}$$

with confidence $1 - \delta$. This excess risk estimation together with Assumption 4 implies that

$$\|f_{\mathbf{z}} - f^*\|_K^2 \leqslant a_3 C_1^{-1} \log(1/\delta) n^{-\min\{\frac{1}{4},\frac{4-\theta}{8+4\theta}\}} \tag{20}$$

with confidence $1 - \delta$, where a_3 is a positive constant.

Combining (18)–(20), we have with confidence $1 - \delta$

$$\forall j, \big|\|\hat{g}_j\|_n^2 - \|g_j^*\|_2^2\big| \leqslant a_4 \sqrt{log(p/\delta)} n^{-\min\{\frac{1}{8},\frac{4-\theta}{6+8\theta}\}}, \tag{21}$$

where $a_4 > 0$ is a constant independently of n, δ, λ.

Now we turn to investigate the relationship between $\hat{S}$ in (12) and S^* in (2). Firstly, we suppose there exists some $j' \in S^*$ but $j' \notin \hat{S}$. That is to say $\|\hat{g}_{j'}\|_n^2 \leqslant v_n$. By Assumption 5 with $C_2 = 2a_4\sqrt{\log(p/\delta)}$, we have

$$\big|\|\hat{g}_{j'}\|_n^2 - \|g_{j'}^*\|_2^2\big| \geq \|g_{j'}^*\|^2 - \|\hat{g}_{j'}\|_n^2 > a_4 \sqrt{\log(p/\delta)} n^{-\min\{\frac{1}{8},\frac{4-\theta}{16+8\theta}\}},$$

which contradicts with (21). This implies that $S^* \subset \hat{S}$ with confidence $1 - \delta$.

Secondly, we suppose there exists some $j' \in \hat{S}$ but $j' \notin S^*$. This means $\|g_{j'}^*\|_2^2 = 0$ and $\|\hat{g}_{j'}\|_n^2 > v_n$. Then

$$\big|\|\hat{g}_{j'}\|_n^2 - \|g_{j'}^*\|_2^2\big| = \|\hat{g}_{j'}\|_n^2 > v_n = a_4 \sqrt{\log(p/\delta)} n^{-\min\{\frac{1}{8},\frac{4-\theta}{16+8\theta}\}},$$

which contradicts with (21) with confidence $1 - \delta$. Therefore, the desired property follows by combining these two results. □

Theorem 2 demonstrates that the identified variables are consistent with truly informative variables with probability 1 as $n \to \infty$. This result guarantees the variable selection performance of our approach, provided that the active variables have enough gradient signal. In the future, it is necessary to further investigate the self-calibration assumption for RMR in RKHS.

When choosing Gaussian kernel as the modal kernel, the modal regression is consistent with regression under the maximum correntropy criterion (MCC) [36]. In terms of the breakdown point theory, Theorem 24 in [19] established the robustness characterization of kernel regression under MCC and Theorem 3 in [36] provided robust analysis for RMR. These results imply the robustness of our approach.

4. Optimization Algorithm

With the help of half-quadratic (HQ) optimization [32], the maximization problem (9) can be transformed into a weighted least-squares problem, and then get the estimator via the ADMM [18]. Indeed, the kernel-based RMR (9) can be implemented directly by the optimization strategy in [36,51] for Gaussian kernel-based modal representation, and in [20] for Epanechnikov kernel-based modal representation. For completeness, we provide the optimization steps of (9) associated with Logistic kernel-based density estimation.

Consider a convex function

$$f(a) = 1/(\exp(\sqrt{a}) + 2 + \exp(-\sqrt{a})),\ a > 0.$$

As illustrated in [52], a convex function $f(a)$ and its convex conjugate function $g(b)$ satisfy

$$f(a) = \max_b (ab - g(b)). \tag{22}$$

According to the Logistic-based representation ϕ and (22), we have

$$\phi(t) = f(t^2) = \max_b (t^2 b - g(b)), t \in \mathbb{R}. \tag{23}$$

Applying (23) into (10), we can obtain the augmented objective function

$$\max_{\alpha\in\mathbb{R}^n, b\in\mathbb{R}^n} \Big\{ \frac{1}{n\sigma} \sum_{i=1}^{n} (b_i (\frac{y_i - \alpha^T \mathbf{K}_n(x_i)}{\sigma})^2 - g(b_i)) - \lambda\alpha^T\mathbf{K}\alpha \Big\}, \tag{24}$$

where $\alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$, and $b = (b_1, ..., b_n)^T \in \mathbb{R}^n$ is the auxiliary vector. Then the maximization problem (24) can be solved by the following iterative optimization algorithm.

According to Theorem 1 in [20], we have $\arg\max_b (ab - g(b)) = f^{'}(a)$. Then, for a fixed α, b_i can be updated by $b_i = f'((\frac{y_i - \alpha^T\mathbf{K}_n(x_i)}{\sigma})^2)$. While b is settled down, update α via

$$\arg\max_{\alpha\in\mathbb{R}^n} \Big\{ \sum_{i=1}^{n} \frac{b_i}{\sigma} (y_i - \alpha^T \mathbf{K}_n(x_i))^2 - \lambda\alpha^T\mathbf{K}\alpha \Big\}. \tag{25}$$

For $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and $\mathbf{Y} = (y_1, ..., y_n) \in \mathbb{R}^n$, the problem (25) can be rewritten as

$$\arg\min_{\alpha\in\mathbb{R}^n} (\mathbf{Y} - \mathbf{K}\alpha)^T diag(-\frac{b}{\sigma})(\mathbf{Y} - \mathbf{K}\alpha) + \lambda\alpha^T\mathbf{K}\alpha$$

where $diag(\cdot)$ is an operator that transforms the vector into a diagonal matrix. By setting $\partial[(\mathbf{Y} - \mathbf{K}\alpha)^T diag(-b/\sigma)(\mathbf{Y} - \mathbf{K}\alpha) + \lambda\alpha^T\mathbf{K}\alpha]/\partial\alpha = 0$, we have

$$\alpha = 4(\mathbf{K}diag(-\frac{b}{\sigma}) + \lambda\mathbf{I})^{-1} diag(-\frac{b}{\sigma})\mathbf{Y}. \tag{26}$$

When α is obtained from (26), we can calculate the gradient-based measure $\|\hat{g}_j\|_n^2$ by (11) directly. Then we apply a pre-specified threshold v_n to identify the truly active set $\hat{\mathcal{S}}_n = \{j : \|\hat{g}_j\|_n^2 > v_n\}$. Here, the threshold v_n is selected by the stability-based criterion [37], which include two steps as below. Firstly, the training samples are randomly divided into two subsets, and the identified active variable sets $\mathcal{J}_{\mathbf{z},1k}$ and $\mathcal{J}_{\mathbf{z},2k}$ are obtained under given v_n for the k-th splitting of training samples. Then, the threshold v_n is updated by maximizing the Cohen kappa statistical measure $\frac{1}{T}\sum\limits_{k=1}^{T}\kappa(\mathcal{J}_{\mathbf{z},1k},\mathcal{J}_{\mathbf{z},2k})$.

The optimization steps of RGVS are summarized in Algorithm 1.

```
Algorithm 1: Optimization algorithm of RGVS with Logistic kernel
Input: Samples $\mathbf{z}$, the modal representation $\phi$ (Logistic kernel), Mercer kernel $K$;
Initialization: $t = 0$, $\alpha$, bandwidth $\sigma$, Max-iter$= 10^2$, $\varepsilon = 10^{-3}$;
Obtain $f_{\mathbf{z}}$ in RKHS:
    While $\alpha$ not converged and $t <$ Max-iter;
        1. Fixed $\alpha^t$, update $b_i^{t+1} = -\frac{\exp(q)-\exp(-q)}{2q(\exp(q)+2+\exp(-q))^2}, q = \frac{y_i-\mathbf{K}_n^T(x_i)\alpha^t}{\sigma}$;
        2. Fixed $b^{t+1}$, update $\alpha^{t+1} = 4(\mathbf{K}diag(-\frac{b^{t+1}}{\sigma}) + \lambda\mathbf{I})^{-1}diag(-\frac{b^{t+1}}{\sigma})\mathbf{Y}$;
        3. Check the convergence condition: $\|\alpha^{t+1} - \alpha^t\|^2 < \varepsilon$;
        4. $t \leftarrow t + 1$;
    End While
    Output: $\hat{\alpha}_{\mathbf{z}} = \alpha^{t+1}$;
Variable Selection: $\hat{\mathcal{S}}_n = \{j : \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}_{\mathbf{z}}^T\partial_j\mathbf{K}_n(x_i))^2 > v_n\}$.
Output: $\hat{\mathcal{S}}_n$
```

5. Empirical Assessments

This section assesses the empirical performance of our proposed method on simulated and real-world datasets. Three variable selection methods are introduced as the baselines, which include *Least Absolute Shrinkage and Selection Operator* (Lasso) [1], *Sparse Additive Models* (SpAM) [9], and *General Variable Selection Method* (GM) [12].

In all experiments, the RKHS $\mathcal{H}_K$ associated with Gaussian kernel $K_h(u,v) = \exp\big(-\frac{\|u-v\|_2^2}{2h^2}\big)$ is employed as the hypothesis function space. For simplicity, we denote $RGVS_{Gau}$ and $RGVS_{Log}$ as the proposed RGVS method with Gaussian modal kernel and Logistic modal kernel, respectively. In the simulated experiments, we generate three datasets (with identical sample size) independently as the training set, the validation set, and the testing set, respectively. The hyper-parameters are tuned via grid research on validation set, and the corresponding grids are displayed as follows: (i) the regularization parameter λ: $\{10^{-3}, 5\times 10^{-3}, 10^{-2}, 5\times 10^{-2}, ..., 1, 5, 10\}$; (ii) the bandwidth σ and h: $\{1 + 10^{-1}i, i = 0, 1, ..., 100\}$; (iii) the threshold v_n: $\{10^{-3+0.1t}, t = 0, ..., 60\}$.

5.1. Simulated Data

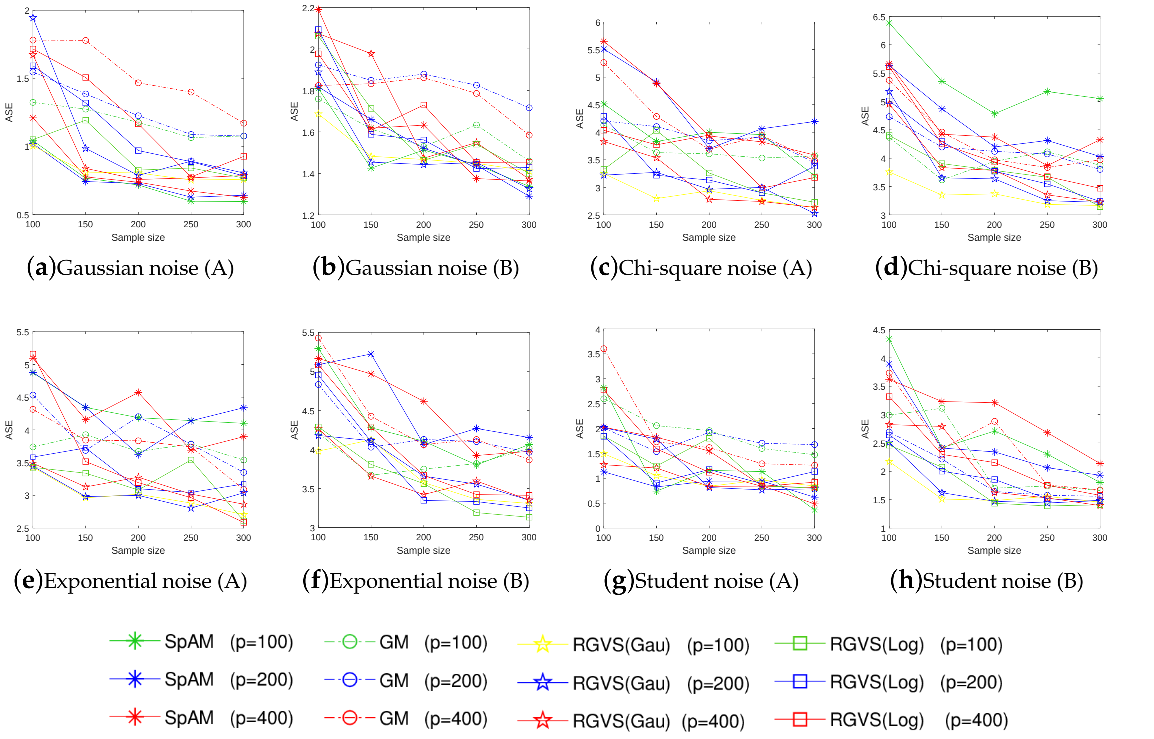
Now we evaluate our approach on two synthetic data used in [12,13]. The first example is a simple additive function and the second one is a function that includes interaction terms.

Example 1. *We generate the p-dimension input* $x_i = (x_{i1}, ..., x_{ip})$ *by* $x_{ij} = \frac{W_{ij}+\eta V_i}{1+\eta}$, *where both* W_{ij} *and* V_i *are extracted from the uniform distribution* $U(-0.5, 0.5)$ *and* $\eta = 0.2$. *The output* y_i *is generated by* $y_i = f^*(x_i) + \epsilon_i$, *where* $f^*(x_i) = 5x_{i1} + 4(x_{i2} - 1)^2 + 0.5\sin(\pi x_{i3}) + \cos(\pi x_{i3}) + 1.5(\sin(\pi x_{i3}))^2 + 2.5(\sin(\pi x_{i3}))^3 + 2(\cos(\pi x_{i3}))^3 + 6\sin(\pi x_{i4})/(2 - \sin(\pi x_{i4}))$ *and* ϵ_i *is a random noise. Here, we consider the Gaussian noise* $\mathcal{N}(0,1)$, *the Chi-square noise* $\mathcal{X}^2(2)$, *the Student noise* $t(2)$, *and the Exponential noise* $E(2)$, *respectively.*

Example 2. *This example follows the way of Example 1 to generate data. The differences are that* W_{ij} *and* V_i *are extracted from the same distribution* $U(0,1)$ *and the true function* $f^*(x_i) = 20x_{i1}x_{i2}x_{i3} + 5x_{i4}^2 + 5x_{i5}$.

For each evaluation, we consider training set with different size $n = 100, 150, 200$ and dimension $p = 150$. To make sure the results are reliable, each evaluation is repeated 50 times. Since the truly informative variables are usually unknown in practice, we evaluate the algorithmic performance according to the *average squares error*(ASE) defined as $ASE := \frac{1}{n}\sum_{i=1}^{n}(f^*(x_i) - f_{\mathbf{z}}(x_i))^2$. To better evaluate the algorithmic performance, we also adopt some metrics used in [12,13] to measure the quality of model regression, e.g., Cp (correct-fitting), SIZE (the average number of selected variables), TP (the average number of the selected true informative variables), FP (the average number of the selected uninformative variables), Up (under-fitting probability), Op (over-fitting probability). The detail result is summarized in Tables 2 and 3. To further support the competitive performance of the proposed method, we also provide the experimental results on ASE in Figure 1 and Cp in Figure 2 with $n = [100 : 50 : 300]$ and $p = 100, 200, 400$. Figures 1 and 2 show that our method has always performed well with different n.

Empirical evaluations on simulated examples verify the promising performance of RGVS on variable selection and regression estimation, even for data with non-Gaussian noises (e.g., the Chi-square noise $\mathcal{X}^2(2)$, the Student noise $t(2)$, and the Exponential noise $E(2)$). Meanwhile, GM and RGVS have similar performance under the Gaussian noise setting, which is consistent with our motivation for algorithmic design.



(**a**)Gaussian noise (A) (**b**)Gaussian noise (B) (**c**)Chi-square noise (A) (**d**)Chi-square noise (B)

(**e**)Exponential noise (A) (**f**)Exponential noise (B) (**g**)Student noise (A) (**h**)Student noise (B)

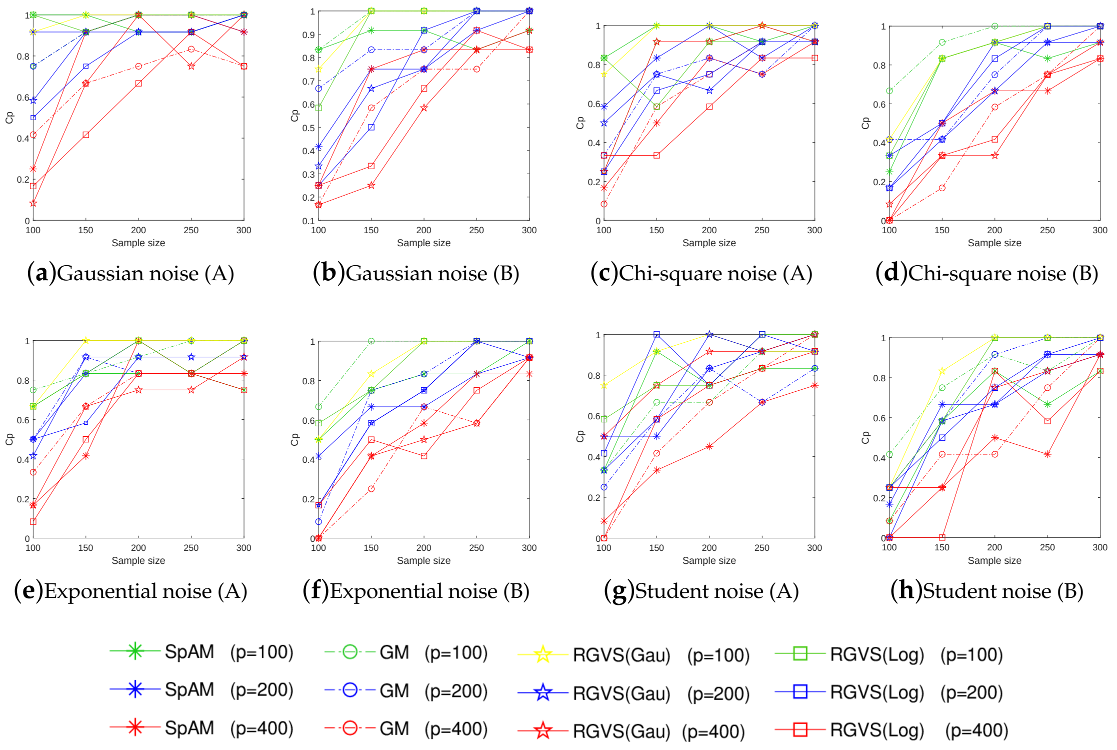
Figure 1. The average squares error (ASE) vs. the sample size n under different noise (A and B represent *Example 1.* and *Example 2* respectively).

Table 2. The averaged performance on simulated data in *Example 1* (left) and *Example 2* (right).

Noise	(n, p)	Method	SIZE	TP	FP	Up	Op	Cp	ASE	SIZE	TP	FP	Up	Op	Cp	ASE
$\mathcal{N}(0,1)$ Gaussian Noise	(100, 150) $n < p$	Lasso	3.92	3.92	0.00	0.36	0.00	0.64	1.369	4.40	4.28	0.12	0.44	0.12	0.44	5.112
		SpAM	4.12	3.92	0.20	0.08	0.16	0.76	1.075	5.02	4.98	0.04	0.04	0.04	**0.92**	1.611
		GM	4.12	3.88	0.24	0.12	0.16	0.72	1.123	5.14	4.98	0.16	0.04	0.12	0.84	1.775
		$RGVS_{Gau}$	4.00	3.92	0.08	0.08	0.04	**0.88**	**1.003**	5.12	5.00	0.12	0.00	0.08	**0.92**	**1.565**
		$RGVS_{Log}$	3.84	3.80	0.04	0.20	0.04	0.76	1.131	5.12	4.92	0.20	0.08	0.16	0.76	1.914
	(150, 150) $n = p$	Lasso	4.20	3.92	0.04	0.16	0.12	0.72	1.245	4.48	4.28	0.20	0.40	0.16	0.44	4.794
		SpAM	4.00	4.00	0.00	0.00	0.00	**1.00**	**0.804**	5.00	5.00	0.00	0.00	0.00	**1.00**	1.612
		GM	3.96	3.92	0.04	0.08	0.04	0.88	1.011	5.04	5.00	0.04	0.00	0.04	0.96	1.627
		$RGVS_{Gau}$	3.96	3.96	0.00	0.04	0.00	0.96	0.899	5.00	5.00	0.00	0.00	0.00	**1.00**	**1.500**
		$RGVS_{Log}$	3.96	3.80	0.04	0.08	0.04	0.88	1.083	5.02	5.00	0.02	0.00	0.03	0.97	1.622
	(200, 150) $n > p$	Lasso	4.00	3.92	0.08	0.20	0.00	0.80	1.252	4.52	4.52	0.00	0.40	0.00	0.60	2.507
		SpAM	4.00	4.00	0.00	0.00	0.00	**1.00**	0.829	5.04	5.00	0.04	0.00	0.02	0.98	1.497
		GM	3.96	3.96	0.00	0.04	0.00	0.96	1.012	5.06	5.00	0.06	0.00	0.04	0.96	1.528
		$RGVS_{Gau}$	4.00	4.00	0.00	0.00	0.00	**1.00**	**0.813**	5.00	5.00	0.00	0.00	0.00	**1.00**	1.485
		$RGVS_{Log}$	3.96	3.96	0.00	0.04	0.00	0.96	0.915	5.00	5.00	0.00	0.00	0.00	**1.00**	**1.453**
$\mathcal{X}^2(2)$ Chi-square Noise	(100, 150) $n < p$	Lasso	3.72	3.64	0.08	0.36	0.04	0.60	5.854	4.32	3.72	0.60	0.68	0.12	0.20	6.888
		SpAM	4.24	3.76	0.48	0.24	0.28	0.48	4.342	5.52	5.00	0.52	0.00	0.40	0.60	4.328
		GM	4.24	3.80	0.44	0.20	0.36	0.44	4.012	4.48	4.44	0.04	0.30	0.05	0.65	3.611
		$RGVS_{Gau}$	4.16	3.96	0.20	0.04	0.20	**0.76**	2.937	5.08	4.92	0.16	0.06	0.16	**0.78**	**3.486**
		$RGVS_{Log}$	4.18	3.90	0.28	0.16	0.12	0.72	**2.681**	5.08	4.84	0.24	0.08	0.18	0.74	3.968
	(150, 150) $n = p$	Lasso	5.32	3.84	1.48	0.16	0.48	0.36	5.392	5.16	4.16	1.00	0.60	0.16	0.24	4.503
		SpAM	4.04	3.96	0.08	0.04	0.08	**0.88**	2.765	5.32	5.00	0.32	0.00	0.24	0.76	3.748
		GM	4.00	3.88	0.12	0.12	0.08	0.80	2.873	4.98	4.92	0.06	0.05	0.05	0.90	4.173
		$RGVS_{Gau}$	3.96	3.92	0.04	0.08	0.04	**0.88**	2.809	5.02	5.00	0.02	0.00	0.02	**0.98**	**2.929**
		$RGVS_{Log}$	4.08	3.96	0.12	0.04	0.12	0.84	**2.097**	5.04	5.00	0.04	0.00	0.04	0.96	3.519
	(200, 150) $n > p$	Lasso	4.24	4.00	0.24	0.00	0.28	0.72	5.805	4.36	4.32	0.04	0.52	0.04	0.44	3.754
		SpAM	4.08	4.00	0.08	0.00	0.08	0.92	2.463	5.04	5.00	0.04	0.00	0.04	0.96	3.634
		GM	4.04	4.00	0.04	0.00	0.04	**0.96**	2.523	5.18	5.00	0.18	0.00	0.20	0.80	3.816
		$RGVS_{Gau}$	3.96	3.96	0.00	0.04	0.00	**0.96**	2.449	5.00	5.00	0.00	0.00	0.00	**1.00**	**2.989**
		$RGVS_{Log}$	3.96	3.96	0.00	0.04	0.00	**0.96**	**1.738**	5.00	5.00	0.00	0.00	0.00	**1.00**	3.457
$E(2)$ Exponential Noise	(100, 150) $n < p$	Lasso	3.46	3.46	0.00	0.60	0.00	0.40	4.631	4.64	4.00	0.64	0.60	0.20	0.20	4.567
		SpAM	4.28	3.88	0.40	0.12	0.28	0.60	4.599	5.84	5.00	0.84	0.00	0.44	0.56	4.224
		GM	4.20	3.64	0.56	0.36	0.36	0.28	3.941	5.36	4.68	0.68	0.32	0.28	0.40	4.528
		$RGVS_{Gau}$	4.20	3.88	0.32	0.12	0.20	**0.68**	3.274	5.06	4.82	0.24	0.14	0.14	0.72	3.907
		$RGVS_{Log}$	3.96	3.80	0.16	0.20	0.16	0.64	**2.775**	5.12	4.92	0.20	0.04	0.16	**0.80**	**3.667**
	(150, 150) $n = p$	Lasso	5.30	3.66	1.64	0.20	0.44	0.36	4.747	5.64	4.16	1.48	0.48	0.28	0.24	4.786
		SpAM	4.04	3.96	0.08	0.04	0.08	0.88	3.403	5.28	5.00	0.28	0.16	0.00	0.84	4.969
		GM	4.08	3.96	0.12	0.04	0.12	0.84	3.177	4.98	4.92	0.06	0.08	0.04	0.88	4.129
		$RGVS_{Gau}$	4.00	4.00	0.00	0.00	0.00	**1.00**	2.724	5.02	4.98	0.04	0.02	0.04	**0.94**	**2.964**
		$RGVS_{Log}$	4.00	4.00	0.00	0.00	0.00	**1.00**	**2.643**	5.00	4.96	0.04	0.04	0.04	0.92	3.918
	(200, 150) $n > p$	Lasso	3.80	3.80	0.00	0.20	0.00	0.80	4.291	4.68	4.60	0.08	0.28	0.08	0.64	3.669
		SpAM	4.00	4.00	0.00	0.00	0.00	**1.00**	2.988	5.24	5.00	0.24	0.00	0.20	0.80	4.808
		GM	3.96	3.96	0.00	0.04	0.00	0.96	3.016	4.98	4.98	0.00	0.04	0.00	0.96	3.878
		$RGVS_{Gau}$	4.00	4.00	0.00	0.00	0.00	**1.00**	2.884	5.00	5.00	0.00	0.00	0.00	**1.00**	**3.041**
		$RGVS_{Log}$	3.96	3.92	0.04	0.09	0.00	0.91	3.113	4.96	4.96	0.00	0.04	0.00	0.96	3.771
$t(2)$ Student Noise	(100, 150) $n < p$	Lasso	4.92	3.80	1.12	0.28	0.32	0.40	2.301	6.52	3.92	2.60	0.64	0.20	0.16	6.971
		SpAM	4.90	3.80	1.1	0.24	0.20	0.56	1.698	7.92	4.72	3.20	0.24	0.44	0.32	4.658
		GM	5.00	3.64	1.36	0.32	0.32	0.36	1.551	5.68	4.32	1.32	0.40	0.32	0.28	3.561
		$RGVS_{Gau}$	4.14	3.94	0.20	0.05	0.10	**0.85**	**0.822**	5.00	4.84	0.16	0.08	0.16	**0.76**	**2.308**
		$RGVS_{Log}$	4.14	3.88	0.26	0.12	0.16	0.72	1.208	4.96	4.80	0.16	0.16	0.12	0.72	2.339
	(150, 150) $n = p$	Lasso	5.08	3.72	1.36	0.24	0.40	0.36	1.793	6.32	3.80	2.52	0.68	0.20	0.12	6.020
		SpAM	4.30	4.00	0.30	0.00	0.32	0.68	0.955	5.44	5.00	0.44	0.00	0.28	0.72	2.739
		GM	4.04	3.80	0.24	0.16	0.16	0.68	1.046	5.56	4.60	0.96	0.28	0.08	0.64	2.557
		$RGVS_{Gau}$	4.00	4.00	0.00	0.00	0.00	**1.00**	**0.757**	4.98	4.98	0.00	0.08	0.00	0.92	**1.716**
		$RGVS_{Log}$	3.92	3.88	0.04	0.12	0.04	0.84	1.169	4.96	4.96	0.00	0.04	0.00	**0.96**	1.723
	(200, 150) $n > p$	Lasso	5.00	3.92	1.08	0.32	0.20	0.48	1.262	5.44	4.36	1.08	0.44	0.28	0.28	2.976
		SpAM	4.10	4.00	0.10	0.00	0.27	0.73	1.060	5.64	5.00	0.64	0.00	0.28	0.72	2.427
		GM	4.00	3.96	0.04	0.04	0.04	0.92	1.011	5.20	4.72	0.48	0.20	0.04	0.76	2.350
		$RGVS_{Gau}$	4.04	4.00	0.04	0.00	0.10	0.90	**0.681**	5.00	5.00	0.00	0.00	0.00	**1.00**	**1.517**
		$RGVS_{Log}$	4.04	4.00	0.04	0.00	0.04	**0.96**	0.884	4.96	4.96	0.00	0.04	0.00	0.96	1.672

Table 3. The averaged performance with simulated data in *Example 1*.

Noise	(n, p)	Method	SIZE	TP	FP	Up	Op	Cp	ASE
$\mathcal{N}(0, 1)$ Gaussian Noise	(300, 500) $n < p$	Lasso	1.98	1.98	0.00	1.00	0.00	0.00	1.98
		GM	4.04	4.00	0.04	0.00	0.04	**0.96**	0.80
		$RGVS_{Gau}$	4.06	4.00	0.06	0.00	0.06	0.94	**0.63**
		$RGVS_{Log}$	4.14	3.98	0.16	0.01	0.03	**0.96**	0.88
	(500, 500) $n = p$	Lasso	1.92	1.92	0.00	1.00	0.00	0.00	1.35
		GM	4.06	4.00	0.06	0.00	0.06	0.94	0.78
		$RGVS_{Gau}$	4.02	4.00	0.02	0.00	0.02	**0.98**	**0.59**
		$RGVS_{Log}$	4.04	4.00	0.04	0.00	0.02	**0.98**	0.74
	(700, 500) $n > p$	Lasso	1.88	1.88	0.00	1.00	0.00	0.00	1.55
		GM	4.04	4.00	0.04	0.00	0.04	0.96	0.77
		$RGVS_{Gau}$	4.02	4.00	0.02	0.00	0.02	0.98	**0.62**
		$RGVS_{Log}$	4.00	4.00	0.00	0.00	0.00	**1.00**	0.73
$\mathcal{X}^2(2)$ Chi-square Noise	(300, 500) $n < p$	Lasso	1.80	1.80	0.00	1.00	0.00	0.00	4.45
		GM	4.18	4.00	0.18	0.00	0.14	0.86	2.92
		$RGVS_{Gau}$	4.09	4.00	0.09	0.00	0.11	**0.89**	2.39
		$RGVS_{Log}$	4.06	3.88	0.18	0.12	0.14	0.74	**1.95**
	(500, 500) $n = p$	Lasso	1.74	1.74	0.00	1.00	0.00	0.00	4.62
		GM	4.14	4.00	0.14	0.00	0.14	0.86	3.01
		$RGVS_{Gau}$	4.08	4.00	0.08	0.00	0.06	**0.94**	2.22
		$RGVS_{Log}$	4.04	3.98	0.06	0.02	0.06	0.92	**1.82**
	(700, 500) $n > p$	Lasso	1.86	1.86	0.00	1.00	0.00	0.00	4.37
		GM	4.28	4.00	0.28	0.00	0.24	0.76	2.96
		$RGVS_{Gau}$	4.02	4.00	0.02	0.00	0.02	**0.98**	2.13
		$RGVS_{Log}$	4.02	4.00	0.02	0.00	0.02	**0.98**	**1.72**
$E(2)$ Exponential Noise	(300, 500) $n < p$	Lasso	2.04	2.04	0.00	1.00	0.00	0.00	4.25
		GM	3.94	3.87	0.07	0.13	0.05	0.82	3.14
		$RGVS_{Gau}$	4.02	4.00	0.02	0.00	0.02	**0.98**	2.36
		$RGVS_{Log}$	3.98	3.94	0.04	0.06	0.02	0.92	**1.92**
	(500, 500) $n = p$	Lasso	1.94	1.94	0.00	1.00	0.00	0.00	4.34
		GM	4.12	4.00	0.12	0.00	0.10	0.90	2.35
		$RGVS_{Gau}$	3.99	3.96	0.03	0.04	0.03	0.93	2.37
		$RGVS_{Log}$	4.02	4.00	0.02	0.00	0.02	**0.98**	**1.71**
	(700, 500) $n > p$	Lasso	1.90	1.90	0.00	1.00	0.00	0.00	4.67
		GM	4.08	4.00	0.08	0.00	0.06	0.94	2.33
		$RGVS_{Gau}$	3.99	3.99	0.00	0.01	0.00	**0.99**	**1.74**
		$RGVS_{Log}$	4.05	4.00	0.05	0.00	0.05	0.95	1.92
$t(2)$ Student Noise	(300, 500) $n < p$	Lasso	1.96	1.96	0.00	1.00	0.00	0.00	4.63
		GM	3.50	3.46	0.04	0.24	0.04	0.72	2.48
		$RGVS_{Gau}$	4.14	3.94	0.20	0.06	0.10	0.84	**0.82**
		$RGVS_{Log}$	4.00	3.98	0.02	0.02	0.00	**0.98**	0.90
	(500, 500) $n = p$	Lasso	1.76	1.76	0.00	0.98	0.00	0.02	3.83
		GM	4.30	4.00	0.30	0.00	0.16	0.84	1.96
		$RGVS_{Gau}$	4.00	4.00	0.00	0.00	0.00	**1.00**	0.76
		$RGVS_{Log}$	4.02	4.00	0.02	0.00	0.01	0.99	**0.75**
	(700, 500) $n > p$	Lasso	1.96	1.96	0.00	0.96	0.00	0.04	2.46
		GM	4.06	4.00	0.06	0.00	0.04	0.96	1.95
		$RGVS_{Gau}$	4.04	4.00	0.04	0.00	0.06	0.94	**0.68**
		$RGVS_{Log}$	4.00	4.00	0.00	0.00	0.00	**1.00**	0.74



(**a**)Gaussian noise (A) (**b**)Gaussian noise (B) (**c**)Chi-square noise (A) (**d**)Chi-square noise (B)

(**e**)Exponential noise (A) (**f**)Exponential noise (B) (**g**)Student noise (A) (**h**)Student noise (B)

Figure 2. The correct-fitting probability (Cp) vs. the sample size n under different noise (A and B represent *Example 1.* and *Example 2* respectively).

5.2. Real-World Data

We now evaluate our RGVS on Auto-Mpg and Requirements of buildings, which are all collected from UCI. Since the variable number is very limited for the current datasets, 100 irrelative variables are added, which are generated from the distribution of $U(-0.5, 0.5)$.

Auto-Mpg data describes the mile per gallon of automobile (MPG). It contains 398 samples and 7 variables, including Cylinders, Displacement, Horsepower, Weight, Acceleration, Model year, and Origin. The second real data sets is obtained to assess the heating load and cooling load requirements of buildings which contains 768 samples and 8 input variables, including Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distribution. In particular, it has two response variables (heating load and cooling load).

Now, we use the 5-fold cross validation to tune the hyper-parameters and employ the *relative sum of the squared errors* (RSSE) to measure learning performance. Here $RSSE = \sum_{x\in X_{test}}(f(x) - f_{\mathbf{z}}(x))^2 / \sum_{x\in X_{test}}(f(x) - E(f))^2$, where $f_{\mathbf{z}}$ is the estimator of f and $E(f)$ denotes the average value of f on the test set X_{test}. Experimental results are reported in Tables 4 and 5.

As shown in Table 4, our method identifies similar variables as GM, but can achieve the smaller RSSE. At same time, SpAM and Lasso tend to select less variables than GM and RGVS, which may discard the truly informative variable for regression estimation. Table 5 shows RGVS has better performance for both the *Heating Load* data and the *Cooling Load* data. All these empirical evaluations validate the effectiveness of our learning strategy consistently.

Table 4. Learning performance on Auto-Mpg.

Variable	CyL	DISP	HPOWER	WEIG	ACCELER	YEAR	ORIGN	RSSE(std)
Lasso	-	-	-	✓	-	✓	-	0.5918(0.3762)
SpAM	✓	✓	-	✓	-	-	-	0.2754(0.0191)
GM	✓	✓	✓	✓	-	✓	✓	0.2547(0.0313)
RGVS_{Gau}	✓	✓	✓	✓	-	✓	-	0.1425(0.0277)
RGVS_{Log}	✓	✓	✓	✓	-	✓	✓	**0.1379(0.0183)**

Table 5. Learning performance on Heating Load (UP) and Cooling Load (DOWN).

Variable	RC	SA	WA	RA	OH	ORIENT	GA	GAD	RSSE(std)
Lasso	-	-	-	-	✓	-	✓	-	0.1739(0.0801)
SpAM	-	-	-	✓	✓	-	-	-	0.1684(**0.0045**)
GM	✓	✓	✓	✓	✓	-	-	-	0.1244(0.0383)
RGVS_{Gau}	-	✓	✓	✓	✓	-	✓	-	**0.0935**(0.0099)
RGVS_{Log}	-	-	✓	✓	✓	-	✓	-	0.1110(0.0066)
Lasso	-	-	✓	-	✓	-	✓	-	0.2119(0.0926)
SpAM	-	-	-	✓	✓	-	-	-	0.1910(0.0131)
GM	✓	✓	✓	✓	✓	-	-	-	0.1515(0.0120)
RGVS_{Gau}	-	✓	✓	✓	✓	-	✓	-	**0.1339**(0.0116)
RGVS_{Log}	✓	✓	✓	✓	✓	-	-	-	0.1368(**0.0077**)

6. Conclusions

This paper proposes a new RGVS method rooted in kernel modal regression. The main advantages of RGVS are its flexibility on mimicking the decision function and adaptivity on screening the truly active variables. The proposed approach is evaluated by the theoretical analysis on the generalization error and variable selection, and by the empirical results on data experiments. In theory, our method can achieve the polynomial decay rate with $O(n^{-\frac{2}{5}})$. In applications, our model has shown the competitive performance for data with non-Gaussian noises.

Author Contributions: Methodology, B.S., H.C.; software, C.G., Y.W.; validation, B.S., H.X. and Y.W.; formal analysis, C.G., H.C.; writing—original draft preparation, C.G.,Y.W.; writing—review and editing, H.C. and H.X.

Funding: This research was funded in part by the National Natural Science Foundation of China (NSFC) under Grants 11601174 and 11671161.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Tibshirani, R. Regression shrinkage and delection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.
2. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.* **2006**, *68*, 49–67. [CrossRef]
3. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
4. Stone, C.J. Additive regression and other nonparametric models. *Ann. Stat.* **1985**, *13*, 689–705. [CrossRef]
5. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman and Hall: London, UK, 1990.
6. Kandasamy, K.; Yu, Y. Additive approximations in high dimensional nonparametric regression via the SALSA. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016.
7. Kohler, M.; Krzyżak, A. Nonparametric regression based on hierarchical interaction models. *IEEE Trans. Inf. Theory* **2017**, *63*, 1620–1630. [CrossRef]

8. Chen, H.; Wang, X.; Huang, H. Group sparse additive machine. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 198–208.
9. Ravikumar, P.; Liu, H.; Lafferty, J.; Wasserman, L. SpAM: Sparse additive models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 1009–1030. [CrossRef]
10. Lin, Y.; Zhang, H.H. Component selection and smoothing in multivariate nonparametric regression. *Ann. Stat.* **2007**, *34*, 2272–2297. [CrossRef]
11. Yin, J.; Chen, X.; Xing, E.P. Group sparse additive models. In Proceedings of the International Conference on Machine Learning (ICML), Edinburgh, UK, 26 June–1 July 2012.
12. He, X.; Wang, J.; Lv, S. Scalable kernel-based variable selection with sparsistency. *arXiv* **2018**, arXiv:1802.09246.
13. Yang, L.; Lv, S.; Wang, J. Model-free variable selection in reproducing kernel Hilbert space. *J. Mach. Learn. Res.* **2016**, *17*, 1–24.
14. Ye, G.; Xie, X. Learning sparse gradients for variable selection and dimension reduction. *Mach. Learn.* **2012**, *87*, 303–355. [CrossRef]
15. Gregorová, M.; Kalousis, A.; Marchand-Maillet, S. Structured nonlinear variable selection. *arXiv* **2018**, arXiv:1805.06258.
16. Mukherjee, S.; Zhou, D.X. Analysis of half-quadratic minimization methods for signal and image recovery. *J. Mach. Learn. Res.* **2006**, *7*, 519–549.
17. Rosasco, L.; Villa, S.; Mosci, S.; Santoro, M.; Verri, A. Nonparametric sparsity and regularization. *J. Mach. Learn. Res.* **2013**, *14*, 1665–1714.
18. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [CrossRef]
19. Feng, Y.; Fan, J.; Suykens, J.A.K. A statistical learning approach to modal regression. *arXiv* **2017**, arXiv:1702.05960.
20. Wang, X.; Chen, H.; Cai, W.; Shen, D.; Huang, H. Regularized modal regression with applications in cognitive impairment prediction. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1448–1458.
21. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
22. Chernoff, H. Estimation of the mode. *Ann. Inst. Stat. Math.* **1964**, *16*, 31–41. [CrossRef]
23. Yao, W.; Lindsay, B.G.; Li, R. Local modal regression. *J. Nonparametr. Stat.* **2012**, *24*, 647–663. [CrossRef]
24. Chen, Y.C.; Genovese, C.R.; Tibshirani, R.J.; Wasserman, L. Nonparametric modal regression. *Ann. Stat.* **2014**, *44*, 489–514. [CrossRef]
25. Collomb, G.; Härdle, W.; Hassani, S. A note on prediction via estimation of the conditional mode function. *J. Stat. Plan. Inference* **1986**, *15*, 227–236. [CrossRef]
26. Lee, M.J. Mode regression. *J. Econom.* **1989**, *42*, 337–349. [CrossRef]
27. Sager, T.W.; Thisted, R.A. Maximum likelihood estimation of isotonic modal regression. *Ann. Stat.* **1982**, *10*, 690–707. [CrossRef]
28. Li, J.; Ray, S.; Lindsay, B. A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **2007**, *8*, 1687–1723.
29. Liu, W.; Pokharel, P.P.; Príncipe, J.C. Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Trans. Signal Process.* **2007**, *55*, 5286–5298. [CrossRef]
30. Príncipe, J.C. *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
31. Feng, Y.; Huang, X.; Shi, L.; Yang, Y.; Suykens, J.A.K. Learning with the maximum correntropy criterion induced losses for regression. *J. Mach. Learn. Res.* **2015**, *16*, 993–1034.
32. Nikolova, M.; Ng, M.K. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.* **2005**, *27*, 937–966. [CrossRef]
33. Aronszajn, N. Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404. [CrossRef]
34. Cucker, F.; Zhou, D.X. *Learning Theory: An Approximation Theory Viewpoint*; Cambridge University Press: Cambridge, UK, 2007.
35. Yao, W.; Li, L. A new regression model: Modal linear regression. *Scand. J. Stat.* **2013**, *41*, 656–671. [CrossRef]

36. Chen, H.; Wang, Y. Kernel-based sparse regression with the correntropy-induced loss. *Appl. Comput. Harmon. Anal.* **2018**, *44*, 144–164. [CrossRef]
37. Sun, W.; Wang, J.; Fang, Y. Consistent selection of tuning parameters via variable selection stability. *J. Mach. Learn. Res.* **2012**, *14*, 3419–3440.
38. Zou, B.; Li, L.; Xu, Z. The generalization performance of ERM algorithm with strongly mixing observations. *Mach. Learn.* **2009**, *75*, 275–295. [CrossRef]
39. Guo, Z.C.; Zhou, D.X. Concentration estimates for learning with unbounded sampling. *Adv. Comput. Math.* **2013**, *38*, 207–223. [CrossRef]
40. Shi, L.; Feng, Y.; Zhou, D.X. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.* **2011**, *31*, 286–302. [CrossRef]
41. Shi, L. Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.* **2013**, *34*, 252–265. [CrossRef]
42. Chen, H.; Pan, Z.; Li, L.; Tang, Y. Error analysis of coefficient-based regularized algorithm for density-level detection. *Neural Comput.* **2013**, *25*, 1107–1121. [CrossRef]
43. Zou, B.; Xu, C.; Lu, Y.; Tang, Y.Y.; Xu, J.; You, X. k-Times markov sampling for SVMC. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 1328–1341. [CrossRef] [PubMed]
44. Li, L.; Li, W.; Zou, B.; Wang, Y.; Tang, Y.Y.; Han, H. Learning with coefficient-based regularized regression on Markov resampling. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 4166–4176.
45. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2008.
46. Wu, Q.; Ying, Y.; Zhou, D.X. Multi-kernel regularized classifiers. *J. Complex.* **2007**, *23*, 108–134. [CrossRef]
47. Steinwart, I.; Christmann, A. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **2011**, *17*, 211–225. [CrossRef]
48. Belloni, A.; Chernozhukov, V. ℓ_1-penalized quantile regression in high dimensional sparse models. *Ann. Stat.* **2009**, *39*, 82–130. [CrossRef]
49. Kato, K. Group Lasso for high dimensional sparse quantile regression models. *arXiv* **2011**, arXiv:1103.1458.
50. Lv, S.; Lin, H.; Lian, H.; Huang, J. Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *Ann. Stat.* **2018**, *46*, 781–813. [CrossRef]
51. Wang, Y.; Tang, Y.Y.; Li, L. Correntropy matching pursuit with application to robust digit and face recognition. *IEEE Trans. Cybern.* **2017**, *47*, 1354–1366. [CrossRef] [PubMed]
52. Rockafellar, R.T. *Convex Analysis*; Princeton Univ. Press: Princeton, NJ, USA, 1997.