Gradient Learning With the Mode-Induced Loss: Consistency Analysis and Applications

Hong Chen, Youcheng Fu, Xue Jiang, Yanhong Chen, Weifu Li, Yicong Zhou, *Senior Member, IEEE*, and Feng Zheng, *Member, IEEE*

***Abstract*— Variable selection methods aim to select the key covariates related to the response variable for learning problems with high-dimensional data. Typical methods of variable selection are formulated in terms of sparse mean regression with a parametric hypothesis class, such as linear functions or additive functions. Despite rapid progress, the existing methods depend heavily on the chosen parametric function class and are incapable of handling variable selection for problems where the data noise is heavy-tailed or skewed. To circumvent these drawbacks, we propose sparse gradient learning with the mode-induced loss (SGLML) for robust model-free (MF) variable selection. The theoretical analysis is established for SGLML on the upper bound of excess risk and the consistency of variable selection, which guarantees its ability for gradient estimation from the lens of gradient risk and informative variable identification under mild conditions. Experimental analysis on the simulated and real data demonstrates the competitive performance of our method over the previous gradient learning (GL) methods.**

***Index Terms*— Gradient learning (GL), learning theory, mode-induced loss, Rademacher complexity, variable selection.**

I. Introduction

DUE to the demand of computation feasibility and result interpretability, variable selection associated with high-dimensional data has attracted increasing attentions in the statistics and machine learning communities [1], [2], [3], [4]. There is a wide spectrum of variable selection methods, which can be divided mainly into linear models, nonlinear additive models, and partial linear models (PLMs). Under linear model assumption, active variables are selected directly by the information metric of covariates (e.g., the Bayesian information criterion (BIC) [5] and Akaike information criterion (AIC) [6]), or by Tikhonov regularization schemes with sparse penalty on regression coefficients (e.g., least absolute shrinkage and selection operator (LASSO) [1], smoothly clipped absolute deviation (SCAD) [7], and least angle regression (LARS) [8]). As a natural extension of linear models, additive models are proposed for nonlinear approximation and variable selection [9], [10], [11], where popular algorithms include component selection and smoothing operator (COSSO) [12], nonparametric independence screening (NIS) [13], sparse additive models (SpAM) [14], GroupSpAM [15], and sparse modal additive model (SpMAM) [16], [17]. As a tradeoff between the linear and nonlinear models, PLMs assume some covariates are linearly related to the response while the others are nonlinear [18]. Some efforts have been made for the PLM-based variable selection and function estimation, such as linear and nonlinear discoverer (LAND) [19], the model pursuit approach [20], and the sparse PLMs [21].

Although the methods mentioned above have shown promising performance in some applications, their success is limited by the assumption of parametric function class and its interplay with the intrinsic target function. In real-world applications, it may be difficult to know the prior structure of the target function. Naturally, the previous models may deteriorate seriously in settings of model misspecification.

Unlike the previous strategy depending on parametric hypothesis, the gradient learning (GL) method can be considered as a model-free framework [26], [27], which learns the gradients of target function for variable selection [22], [28], [29], [30], [31]. Several GL algorithms have been proposed including sparse gradient learning (SGL) [23], model-free (MF) variable selection [24], and robust GL (RGL) [25]. The sparse regularization in reproducing kernel Hilbert space (RKHS) is integrated into GL [23] for selecting active variables. Different from data-independent RKHS used in [22], [23] and [26], MF [24] is associated with data-dependent hypothesis space, where the gradients belong to the coefficient-based representation. Moreover, the variable selection strategy of GL is applied to the correntropy regression [25] and the multitask learning [30].

Although the GL algorithms enjoy the model-free property, most of them are sensitive to non-Gaussian noises due to using the square loss [11], [32], [33], which is associated closely with the conditional mean. Different from that, the conditional mode is usually robust to the heavy-tailed noise,

Manuscript received 11 May 2021; revised 15 January 2022 and 31 May 2022; accepted 8 January 2023. Date of publication 18 January 2023; date of current version 9 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 12071166, Grant 62122035 and Grant 61972188; in part by the Fundamental Research Funds for the Central Universities of China under Grant 2662020LXQD002 and Grant 2662021JC008; in part by the Science and Technology Development Fund, Macau SAR, under Grant 0049/2022/A1; and in part by the University of Macau under Grant MYRG2022-00072-FST. *(Corresponding author: Weifu Li.)*

Hong Chen, Youcheng Fu, and Weifu Li are with the College of Science, Huazhong Agricultural University, Wuhan 430070, China, and also with the Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn; moshi@webmail.hzau.edu.cn; liweifu@mail.hzau.edu.cn).

Xue Jiang and Feng Zheng are with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: jiangx2022@mail.sustech.edu.cn; f.zheng@ieee.org).

Yanhong Chen is with the National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chenyh@nssc.ac.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2023.3236345, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2023.3236345

TABLE I
PROPERTIES OF DIFFERENT GL MODELS

Methods	Hypothesis space	Loss function	Regularizer	Robustness	Sparsity	Error bound	Selection consistency
GL[22]	Data-independent $\mathcal{H}_K^p$	the square loss	2-norm	×	×	✓	×
SGL[23]	Data-independent $\mathcal{H}_K^p$	the square loss	1-norm	×	✓	✓	×
MF[24]	Data-dependent $\mathcal{H}_{K,D}^p$	the square loss	$2,1$-norm	×	✓	✓	✓
RGL[25]	Data-dependent $\mathcal{H}_{K,D}^p$	the correntropy-induced loss	q, q'-norm	✓	✓	×	×
SGLML(ours)	Data-dependent $\mathcal{H}_{K,D}^p$	the mode-induced loss	$2,1$-norm	✓	✓	✓	✓

the skewed noise, and outliers [34], [35]. Therefore, in this article, we consider an RGL scheme motivated by the mode-induced loss (refer to (12)).

The mode-induced loss has been used for modal regression, which aims to estimate the conditional mode of response with given covariates [35]. The empirical results verify its effectiveness and robustness for sparse linear regression [36], the atomic representation-based classification [37], the multi-view learning [38], and the multitask additive models [16].

Inspired from [16], [24], we propose SGL with the mode-induced loss (SGLML) for variable selection, where the regularization penalty is incorporated into the GL scheme with data-dependent hypothesis space. The mode-induced loss is used for improving robustness, and the sparse regularization helps conduct variable selection for addressing interpretability. As a robust extension of MF [24], the proposed SGLML enjoys the model-free flexibility and the robustness to non-Gaussian noises simultaneously. Indeed, RGL [25] can be considered as a special case of our SGLML using Gaussian kernel for density estimation [16], [35].

The main contributions of this article are summarized as below.

1) A new GL scheme, called SGLML, is proposed to mitigate the drawbacks of previous variable selection methods, e.g., relying on specific model assumption [1], [14], [16] and lacking the robustness to non-Gaussian noises [22], [23], [24]. To the best of our knowledge, the GL working together with the mode-induced loss has not been investigated before.
2) Learning theory analysis for SGLML is established on the excess risk bound and variable selection consistency by developing the error decomposition [16], [36] and the concentration estimation techniques [23], [24]. Theoretical results demonstrate that our estimator achieves the consistency in terms of generalization ability when the sample size goes to infinity, as well as identifies the active variable under proper parameter conditions. In particular, the current analysis fills partly the theoretical gap for RGL in [25].
3) Empirical evaluations on the simulated and real-word data show satisfactory performance of our approach over the existing GL methods under non-Gaussian noises' setting.

To better highlight our contribution, we present Table I to summarize the properties of the gradient algorithms from both the model design (hypothesis space, loss function, and regularizer) and theoretical foundations (error bound and variable selection consistency).

The rest of this article is organized as follows. Section II recalls the background and related works of GL. Section III formulates our SGLML and presents its computing algorithm. Section IV states the main theoretical results on the asymptotic estimation and variable selection consistency. Section V reports the experimental analysis of our approach. Finally, Section VI closes this article.

II. Preliminaries and Related Works

Let $\mathcal{X} \subset \mathbb{R}^p$ be a compact input space and $\mathcal{Y} \subset \mathbb{R}$ be an output space. Denote (X, Y) as the pair of explanatory and response variables taking values in $\mathcal{X} \times \mathcal{Y}$. Assume that we are given a training set of i.i.d. observations $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ that are generated by

$$Y = f^*(X) + \epsilon \tag{1}$$

where $f^* : \mathcal{X} \to \mathcal{Y}$ is an intrinsic target function, and ϵ is a random noise satisfying some certain conditions, e.g., the zero-mean noise assumption $\mathbb{E}(\epsilon|X) = 0$ or the zero-mode noise condition in (9). For the feasibility of the theoretical analysis, we denote ρ over $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ as an intrinsic probability distribution with respect to the sampling process (1), and $\rho_{\mathcal{X}}$ as the corresponding marginal distribution of $\mathcal{X}$. Let $L^2_{\rho_{\mathcal{X}}}$ be the function space of the square-integrable functions with respect to $\rho_{\mathcal{X}}$.

A. Gradient Learning

The regularized GL model is proposed in [22] for variable selection, which removes parametric assumption on f^*. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T \in \mathcal{X}$. If the partial derivatives of f^* exist, we define its gradient ∇f^* as the vector of functions

$$\nabla f^* = \left(\frac{\partial f^*}{\partial x_1}, \ldots, \frac{\partial f^*}{\partial x_p}\right)^T.$$

The relevance between each coordinate element x_l and f^* can be evaluated by the norm of a partial derivative $\|(\partial f^*/\partial x_l)\|$ (e.g., $\|(\partial f^*/\partial x_l)\|_{L^2_{\rho_{\mathcal{X}}}} = (\int_{\mathcal{X}} |((\partial f^*(\mathbf{x}))/\partial x_l)|^2 d\rho_{\mathcal{X}}(\mathbf{x}))^{1/2}$ as $(\partial f^*/\partial x_l) \in L^2_{\rho_{\mathcal{X}}}$), where a large norm implies a large change in the function f^* with respect to the lth coordinate [22]. This fact gives an intuitive motivation for GL.

According to the first-order Taylor series expansion, we know $f^*(\mathbf{x}) \approx f^*(\mathbf{u}) + \nabla f^*(\mathbf{u})^T(\mathbf{x} - \mathbf{u})$ for $\mathbf{u}$ in the neighborhood of $\mathbf{x}$, and $\mathbf{x}and\mathbf{u}$ are the interior points of $\mathcal{X}$. The empirical risk $\widetilde{\mathcal{E}}_D(\mathbf{g})$ for an estimator $\mathbf{g}$ of ∇f^* is defined as

$$\widetilde{\mathcal{E}}_D(\mathbf{g}) = \frac{1}{n^2}\sum_{i,j=1}^{n} w_{ij}\Big(y_i - y_j - \mathbf{g}\big(\mathbf{x}_j\big)^T\big(\mathbf{x}_i - \mathbf{x}_j\big)\Big)^2 \tag{2}$$

where $w_{ij} := w(\mathbf{x}_i, \mathbf{x}_j)$ serves as a weight in local regression. A typical choice for w_{ij} is given by $\exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2s^2\}$ with $s > 0$, e.g., [23], [24], and [25].

Denote the RKHS associated with a Mercer kernel K as $\mathcal{H}_K$, its norm as $\|\cdot\|_K$, and $\mathcal{H}_K^p = \mathcal{H}_K \times \cdots \times \mathcal{H}_K$ as the $p$fold product of $\mathcal{H}_K$. Let $\mathbf{g} = (g_1, \ldots, g_p)^T$ be a vector function in $\mathcal{H}_K^p$ with $g_l \in \mathcal{H}_K$ for $l = 1, \ldots, p$. GL [22] aims to get the estimator $\hat{\mathbf{g}}$ by the following regularized optimization problem

$$\hat{\mathbf{g}} = \mathop{\arg\min}_{\mathbf{g}\in\mathcal{H}_K^p}\left\{\widetilde{\mathcal{E}}_D(\mathbf{g}) + \lambda\sum_{l=1}^{p}\|g_l\|_K^2\right\} \tag{3}$$

where $\lambda > 0$ is a regularization parameter, and the regularization term $\sum_{l=1}^{p}\|g_l\|_K^2$ is the square of the 2-norm of the vector $(\|g_1\|_K, \ldots, \|g_p\|_K)^T \in \mathbb{R}^p$. Besides the experimental validation, the error analysis demonstrates the convergence of the empirical estimator (3) to ∇f^* [22].

B. Sparse GL

Following the research line of GL [22], Ye and Xie [23] propose the SGL to further address the sparsity for high-dimensional variable selection. The SGL [23] is formulated as

$$\hat{\mathbf{g}} = \mathop{\arg\min}_{\mathbf{g}\in\mathcal{H}_K^p}\left\{\widetilde{\mathcal{E}}_D(\mathbf{g}) + \lambda\sum_{l=1}^{p}\|g_l\|_K\right\} \tag{4}$$

where the sparse regularization is the 1-norm of $(\|g_1\|_K, \ldots, \|g_p\|_K)^T \in \mathbb{R}^p$. The theoretical analysis assures the convergence of the estimator (4) to ∇f^* in both the Euclidean and the manifold setting. Empirical examples verify its utility for variable selection [23].

C. Coefficient-Based SGL

According to the representer theorem in RKHS (see e.g., [24], [39], and [40]), the minimizer $\hat{\mathbf{g}} = (\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_p)^T$ of (3) or (4) satisfies

$$\hat{g}_l(\mathbf{x}) = \sum_{t=1}^{n}\hat{\alpha}_{lt}K(\mathbf{x}, \mathbf{x}_t), \quad \hat{\alpha}_{lt} \in \mathbb{R}, \quad l = 1, 2, \ldots, p.$$

Inspired from the above representing property, Yang et al. [24] use the data-dependent hypothesis space

$$\mathcal{H}_{K,D}^p = \left\{\mathbf{g} = \left(g_1, g_2, \ldots, g_p\right)^T : \right.$$
$$\left. g_l(\mathbf{x}) = \sum_{t=1}^{n}\alpha_{lt}K(\mathbf{x}, \mathbf{x}_t), 1 \le l \le p\right\} \tag{5}$$

for GL, where $\alpha_{lt} \in \mathbb{R}$. It should be noted that $\mathcal{H}_{K,D}^p$ depends on observations D and is a subspace of $\mathcal{H}_K^p$.

Let $\Omega(g_l) = \inf\{\|\boldsymbol{\alpha}_l\|_2 : g_l(\mathbf{x}) = \sum_{t=1}^{n}\alpha_{lt}K(\mathbf{x}, \mathbf{x}_t) \in \mathcal{H}_{K,D}\}$, where $\|\boldsymbol{\alpha}_l\|_2 = (\sum_{t=1}^{n}|\alpha_{lt}|^2)^{1/2}$ is the standard 2-norm of the vector $\boldsymbol{\alpha}_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{ln})^T \in \mathbb{R}^n$.

Observe that $g_l \equiv 0$ is equivalent to $\|\boldsymbol{\alpha}_l\|_2 = 0$. Then, it is natural to impose a coefficient-based sparse penalty on $\mathcal{H}_{K,D}^p$ as

$$\Omega(\mathbf{g}) = \sum_{l=1}^{p}\pi_l\Omega(g_l) \tag{6}$$

where $\pi_l, l = 1, \ldots, p$, are the adaptive parameters. Let

$$A = \left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p\right) \in \mathbb{R}^{n\times p} \tag{7}$$

and let $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$. The jth column of $\mathbf{K}$ is denoted by $\mathbf{K}_j \in \mathbb{R}^n$. For $\mathbf{g} \in \mathcal{H}_{K,D}^p$, the empirical risk $\widetilde{\mathcal{E}}_D(\mathbf{g})$ can be rewritten as

$$\widetilde{\mathcal{E}}_D(\mathbf{g}) = \frac{1}{n^2}\sum_{i,j=1}^{n}w_{ij}\left(y_i - y_j - \mathbf{K}_j^T A\left(\mathbf{x}_i - \mathbf{x}_j\right)\right)^2.$$

The optimization scheme of MF [24] is formulated as

$$\hat{\mathbf{g}} = \mathop{\arg\min}_{\mathbf{g}\in\mathcal{H}_{K,D}^p}\left\{\widetilde{\mathcal{E}}_D(\mathbf{g}) + \lambda\Omega(\mathbf{g})\right\}. \tag{8}$$

Different from $\mathcal{H}_K^p$ used in GL [22] and SGL [23], MF searches the gradient function $\mathbf{g}$ in the data-dependent hypothesis space $\mathcal{H}_{K,D}^p$. The statistical analysis is provided in [24] to guarantee the effectiveness of MF on regression estimation and active variable discovery under Gaussian noise setting.

D. Robust GL

The maximum correntropy criterion (MCC) [41], [42], [43] is incorporated into the GL in [25] to improve its robustness to complex noises. The MCC-based regression estimates the target function by

$$f_{D,\sigma} = \mathop{\arg\max}_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}K_\sigma(y_i, f(\mathbf{x}_i)), \quad \sigma > 0$$

where K_σ is the Gaussian kernel, and $\mathcal{H}$ is a hypothesis space. Correspondingly, the correntropy-induced loss ℓ_σ is defined as

$$\ell_\sigma(y_i, f(\mathbf{x}_i)) = \sigma^2(1 - K_\sigma(y_i, f(\mathbf{x}_i))), \quad \sigma > 0$$

and the correntropy regression scheme can be rewritten as

$$f_{D,\sigma} = \mathop{\arg\min}_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\ell_\sigma(y_i, f(\mathbf{x}_i)).$$

Naturally, the empirical GL risk with the correntropy-induced loss ℓ_σ can be denoted by

$$\bar{\mathcal{E}}_D(\mathbf{g}) = \frac{1}{n^2}\sum_{i,j=1}^{n}w_{ij}\ell_\sigma\Big(y_i, y_j + \mathbf{g}\left(\mathbf{x}_j\right)^T\left(\mathbf{x}_i - \mathbf{x}_j\right)\Big).$$

Based on the hypothesis space $\mathcal{H}_{K,D}^p$ in (5), the RGL [25] is formulated as

$$\hat{\mathbf{g}} = \mathop{\arg\min}_{\mathbf{g}\in\mathcal{H}_{K,D}^p}\left\{\bar{\mathcal{E}}_D(\mathbf{g}) + \lambda\sum_{l=1}^{p}\left(\sum_{t=1}^{n}|\alpha_{lt}|^q\right)^{\frac{q'}{q}}\right\}, \quad q, q' \ge 1.$$

Naturally, we can set $(q, q') \in \{(2, 1), (2, 2), (1, 1)\}$ with respect to the 2, 1-norm, the 2-norm, and the 1-norm regularization, respectively. Although the computing algorithm and empirical evaluations are provided in [25], there is no learning theory analysis for RGL that we are aware of.

III. Sparse GL With the Mode-Induced Loss

A. Proposed SGLML

Under the zero-mean noise assumption $\mathbb{E}(\epsilon|X) = 0$, the intrinsic regression function $f^*(\mathbf{x})$ in (1) can be rewritten as the conditional mean $\mathbb{E}(Y|X = \mathbf{x})$, that is,

$$f^*(\mathbf{x}) = \mathbb{E}(Y|X = \mathbf{x}) = \int_{\mathcal{Y}} y d\rho(y|X = \mathbf{x})$$

where $\rho(y|X = \mathbf{x})$ is the conditional distribution of y given $X = \mathbf{x}$. Under this setting, it is well-known that f^* is the minimizer of expected risk $\int_{\mathcal{Z}} (y - f(\mathbf{x}))^2 d\rho(\mathbf{x}, y)$ over the measurable function space [44], [45].

However, when data contain non-Gaussian noises, the target function f^* becomes inconsistent with the conditional mean $\mathbb{E}(Y|X = \cdot)$. Accordingly, the existing GL methods [22], [23], [24] may suffer from the degraded performance on variable selection. Different from the zero-mean noise assumption, the zero-mode noise condition assumes the mode of the conditional density of the noise to be zero, i.e.,

$$\text{mode}(\epsilon|X = \mathbf{x}) := \arg\max_{t\in\mathbb{R}} p_{\epsilon|X}(t|X = \mathbf{x}) = 0 \quad \forall \mathbf{x} \in \mathcal{X} \tag{9}$$

where $p_{\epsilon|X}$ is the conditional density of ϵ conditioned on X. Indeed, the zero-mode condition is satisfied for many non-Gaussian noises [35]. Performing the mode operator on both sides of (1), we observe that the *modal regression function* f_M is equivalent to f^* [35], i.e.,

$$f_M(\mathbf{x}) := \arg\max_{t\in\mathbb{R}} p_{Y|X}(t|X = \mathbf{x}) = f^*(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X} \tag{10}$$

where $p_{Y|X}$ denotes the conditional density of Y given X. Throughout this article, we assume the existence and uniqueness of f^* [16], [17], [34], [35]. Obviously, this assumption holds if the global mode of $p_{\epsilon|X}$ exists and is unique.

Under the zero-mode noise condition, the modal regression metric $\mathcal{R}(f)$ is defined in [35] as

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

It can be verified that f^* in (10) is the maximizer of $\mathcal{R}(f)$ over all the measurable functions. However, it is difficult to maximize $\mathcal{R}(f)$ directly due to the unknown conditional density $p_{Y|X}$. A surrogate modal regression criterion is introduced by converting the estimation of $p_{Y|X}$ into the density estimation of 1-D random variable $E_f := Y - f(X)$ [35]. Theorem 5.1 in [35] states

$$p_{E_f}(0) = \mathcal{R}(f)$$

where p_{E_f} is the density function of E_f. Therefore, f^* is also a maximizer of $p_{E_f}(0)$, which can be approximated by the kernel density estimation (KDE) technique. Particularly, a kernel for estimating $p_{E_f}(0)$ is called a modal kernel $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ [35], and the estimator of $p_{E_f}(0)$ is

$$\hat{p}_{E_f}(0) = \frac{1}{n\sigma} \sum_{i=1}^{n} K_\sigma(y_i, f(\mathbf{x}_i)) := \mathcal{R}_D^\sigma(f).$$

For feasibility, we set

$$\phi\left(\frac{y_i - f(\mathbf{x}_i)}{\sigma}\right) := K_\sigma(y_i, f(\mathbf{x}_i)) \tag{11}$$

and denote the expectation form of $\mathcal{R}_D^\sigma(f)$ as

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

Theorem 9 in [35] assures that $\mathcal{R}^\sigma(f)$ will converge to $\mathcal{R}(f)$ when $\sigma \to 0$. By imposing certain conditions on the density of ϵ conditioned on X (see Assumption 3 in [35]), Theorem 19 in [35] states that $\|f - f^*\|_{L^2_{\rho_\mathcal{X}}}$ can be bounded by $\mathcal{R}(f^*) - \mathcal{R}(f)$ with a constant multiplier.

Following this line, [16] further introduces the mode-induced loss $\psi_\sigma : \mathbb{R} \to [0, +\infty)$ for robust variable selection, which is defined as

$$\psi_\sigma(y - f(x)) = \frac{1}{\sigma}\left(\phi(0) - \phi\left(\frac{y - f(x)}{\sigma}\right)\right). \tag{12}$$

Based on ψ_σ, we introduce the mode-induced gradient loss

$$\ell(\mathbf{g}, (\mathbf{x}, y), (\mathbf{u}, v)) = w(\mathbf{x}, \mathbf{u}) \psi_\sigma\left(y - v - \mathbf{g}(\mathbf{u})^T(\mathbf{x} - \mathbf{u})\right) \tag{13}$$

where $(\mathbf{x}, y), (\mathbf{u}, v) \in \mathcal{Z}$ and $w(\mathbf{x}, \mathbf{u}) = \exp\{-\|\mathbf{x}-\mathbf{u}\|^2/2s^2\}$ is an adaptive weight. Subsequently, we define the corresponding expectation and empirical risks, respectively, as

$$\mathcal{E}(\mathbf{g}) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(\mathbf{g}, (\mathbf{x}, y), (\mathbf{u}, v)) d\rho(\mathbf{x}, y) d\rho(\mathbf{u}, v)$$

and

$$\mathcal{E}_D(\mathbf{g}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \ell\left(\mathbf{g}, (\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\right).$$

Therefore, an empirical estimator of ∇f^* can be obtained via the SGLML formulated by

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}\in\mathcal{H}_{K,D}^p} \{\mathcal{E}_D(\mathbf{g}) + \lambda\Omega(\mathbf{g})\} \tag{14}$$

where $\lambda > 0$ is a parameter measuring the tradeoff between the empirical risk $\mathcal{E}_D(\mathbf{g})$ and the sparsity penalty $\mathbf{\Omega}(\mathbf{g})$ defined in (6).

Remark: The differences between our method and RGL [25] are threefold. 1) Our approach depends on the mode-induced loss [16], [35], while the one in [25] is associated with the correntropy-induced loss l_σ under the MCC [43]. 2) The RGL [25] can be considered as a special case of SGLML when Gaussian kernel is used for KDE. Indeed, the error metric (12) used in our method generalizes the correntropy-induced loss to general setting, since the candidate kernels for KDE also include Sigmoid kernel, Logistic kernel, etc. Refer to Remark 2 in [16] for the plots of loss functions with different modal kernels. 3) Our work establishes the analysis on excess risk bound and variable selection consistency, which fills the theoretical gap in part for RGL [25].

B. Computing Algorithm

For A defined in (7), denote $\|A\|_{2,1} = \sum_{l=1}^{p}(\sum_{t=1}^{n} |\alpha_{lt}|^2)^{1/2}$ and

$$\mathcal{E}(A) = \frac{1}{n^2\sigma} \sum_{i,j=1}^{n} w_{ij}\left(\phi(0) - \phi\left(Z_{ij}/\sigma\right)\right)$$

with

$$Z_{ij} = y_i - y_j - \mathbf{K}_j^T A\big(\mathbf{x}_i - \mathbf{x}_j\big). \tag{15}$$

Then, the minimization problem (14) can be rewritten as

$$\min_{A \in \mathbb{R}^{n\times p}} \big\{\mathcal{E}(A) + \lambda \|A\|_{2,1}\big\}. \tag{16}$$

Following [25], we can estimate A iteratively by

$$A^{t+1} = \arg\min_{A\in\mathbb{R}^{n\times p}} \frac{1}{2}\big\|\zeta^{t+1} - A\big\|_F^2 + \lambda\gamma\,\|A\|_{2,1} \tag{17}$$

where $\zeta^{t+1} = A^t - \gamma\,\nabla\mathcal{E}(A^t)$, t is the iterative number, and γ is the step size.

A standard approach to solve (17) is using a soft thresholding operator $S_{\lambda\gamma}$ [46] such that

$$A^{t+1} = S_{\lambda\gamma}\left(\zeta^{t+1}\right) := \left(S_{\lambda\gamma}\left(\boldsymbol{d}_1^{t+1}\right), S_{\lambda\gamma}\left(\boldsymbol{d}_2^{t+1}\right), \dots, S_{\lambda\gamma}\left(\boldsymbol{d}_p^{t+1}\right)\right)$$

where $\zeta^{t+1} = [\boldsymbol{d}_1^{t+1}, \dots, \boldsymbol{d}_p^{t+1}] \in \mathbb{R}^{n\times p}$ and

$$S_{\lambda\gamma}\left(\boldsymbol{d}_l^{t+1}\right) = \begin{cases} 0 & \text{if } \left\|\boldsymbol{d}_l^{t+1}\right\|_F \le \lambda\gamma \\ \dfrac{\left\|\boldsymbol{d}_l^{t+1}\right\|_F - \lambda\gamma}{\left\|\boldsymbol{d}_l^{t+1}\right\|_F}\boldsymbol{d}_l^{t+1} & \text{otherwise.} \end{cases}$$

The computing steps of SGLML are summarized as below.

Algorithm 1 Iterative Optimization for SGLML

```
Require: Data $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$, weight matrix $W = (w(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$, kernel matrix $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$, regularization parameter $\lambda > 0$, step size $\gamma > 0$.
    Initial $A^0 = [\boldsymbol{\alpha}_1^0, \dots, \boldsymbol{\alpha}_p^0]$ with $\boldsymbol{\alpha}_l^0 \in \mathbb{R}^n$ for $l = 1, \dots, p$, representing width $\sigma > 0$, stopping threshold $\epsilon_0 > 0$.
Ensure: Coefficient matrix $A^{t+1} = [\boldsymbol{\alpha}_1^{t+1}, \dots, \boldsymbol{\alpha}_p^{t+1}]$ and the SGLML estimator
    $\hat{\mathbf{g}}(\mathbf{x}) = \left(\sum_{i=1}^n \alpha_{1i}^{t+1} K(\mathbf{x}, \mathbf{x}_i), \dots, \sum_{i=1}^n \alpha_{pi}^{t+1} K(\mathbf{x}, \mathbf{x}_i)\right)^T.$
  while $\|A^{t+1} - A^t\|_F \ge \epsilon_0$ do
    1) Compute $Z_{ij} = y_i - y_j - \mathbf{K}_j^T A^t(\mathbf{x}_i - \mathbf{x}_j)$.
    2) Compute gradient $\nabla\mathcal{E}(A^t)$ and the descent step
         $\zeta^{t+1} = A^t - \gamma\,\nabla\mathcal{E}\big(A^t\big).$
    3) Perform the soft threshold $S_{\lambda\gamma}$ on $\zeta^{t+1}$ to obtain $A^{t+1}$ by $A^{t+1} = S_{\lambda\gamma}(\zeta^{t+1})$.
  end while
```

Without loss of generality, we state the convergence analysis of Algorithm 1 for the mode-induced gradient loss (13) associated with the Gaussian kernel. Under this setting, $\mathcal{E}(A)$ can be further written as

$$\mathcal{E}(A) = \frac{1}{n^2}\sum_{i,j=1}^n w_{ij}\left(1 - \exp\left(-\frac{Z_{ij}^2}{2\sigma^2}\right)\right)$$

and its gradient

$$\nabla\mathcal{E}(A) = \frac{1}{n^2\sigma^2}\sum_{ij=1}^n w_{ij}\exp\left(-\frac{Z_{ij}^2}{2\sigma^2}\right) Z_{ij}\big(\mathbf{x}_j - \mathbf{x}_i\big)\mathbf{K}_j^T.$$

It is easy to check that $\nabla\mathcal{E}(A)$ is Lipschitz continuous with constant L (also see Section III in [25]). Let A^t be the sequence generated by Algorithm 1 with $(1/\gamma) > L$. According to Theorem 1 in [25], we know that the limit of $\{A^t\}_{t\ge 1}$ is a stationary point of (16).

IV. Learning Theory Analysis

This section establishes our main theoretical results on the asymptotic estimation and variable selection consistency of the proposed SGLML.

The following assumptions are required for our analysis, which have been used extensively in machine learning literatures, e.g., [24] and [35].

Assumption 1: Assume that $\mathcal{Y} \subset [-M, M]$, $\sup_{\mathbf{x},\mathbf{x}'} w(\mathbf{x}, \mathbf{x}') \le 1$, and there exists a constant $\tilde{C}$ such that $\sup_{\mathbf{x}} \|\mathbf{x}\| \le \tilde{C}$. Also, the kernel function involving in RKHS satisfies $\sup_{\mathbf{x}} |K(\mathbf{x}, \mathbf{x})| \le 1$ and the largest eigenvalue of kernel matrix $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ is of order $O(n^\mu)$ with $\mu \in [0, 1)$.

Assumption 2: The function ϕ defined in (11) satisfies the following conditions: 1) $\forall u, \phi(u) = \phi(-u)$, $\phi(u) \le \phi(0)$, and $\int_{\mathbb{R}} \phi(u)du = 1$; (ii) ϕ is bounded and differentiable with $\|\phi'\|_\infty < \infty$; and (iii) $\int_{\mathbb{R}} u^2\phi(u)du < \infty$.

Assumption 2 holds true for the Gaussian kernel, the logistic kernel, and the Sigmoid kernel [16], [35], [36].

Assumption 3: The probability density $p(\mathbf{x})$ exists and satisfies

$$|p(\mathbf{x}) - p(\mathbf{u})| \le C d_{\mathcal{X}}(\mathbf{x}, \mathbf{u})^\theta \quad \forall \mathbf{x}, \mathbf{u} \in \mathcal{X}$$

where $d_{\mathcal{X}}$ is the Euclidean distance on $\mathcal{X}$, and $Cand\theta$ are positive constants.

Assumption 3 introduces a Lipschitz condition on $p(\mathbf{x})$ to assure the smoothness of the marginal distribution $\rho_{\mathcal{X}}$, which is a natural condition for learning gradient [22], [23], [24].

Assumption 4: Assume that the target gradient $\mathbf{g}^* \in \mathcal{H}_K^p$.

Since SGLML depends on a subspace of RKHS, it is natural to require $\mathbf{g}^* = \nabla f^* \in \mathcal{H}_K^p$, which is consistent with [22], [23], [24].

Inspired from the error decomposition in [16] and [24], we introduce the following stepping-stone function

$$\bar{\mathbf{g}} = \arg\min_{\mathbf{g}\in\mathcal{H}_K^p}\left\{\mathcal{E}_D(\mathbf{g}) + \lambda\sum_{l=1}^p \pi_l \|g_l\|_K^2\right\}. \tag{18}$$

The representer theorem of kernel methods yields $\bar{\mathbf{g}} = (\bar{g}_1, \bar{g}_2, \dots, \bar{g}_p)^T$ with $\bar{g}_l(\cdot) = \sum_{t=1}^n \bar{\alpha}_{lt} K(\cdot, \mathbf{x}_t), \bar{\alpha}_{lt} \in \mathbb{R}$ for each $l \in \{1, \dots, p\}$.

The regularization risk of SGLML can be decomposed as below.

Proposition 1: Let Assumption 4 be true. For the SGLML-based estimator $\hat{\mathbf{g}}$ in (14), there holds

$$\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) \le \{\mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}})\} + \big\{\mathcal{E}_D\big(\mathbf{g}^*\big) - \mathcal{E}\big(\mathbf{g}^*\big)\big\} + \lambda\Omega(\bar{\mathbf{g}}) + \left\{\mathcal{E}\big(\mathbf{g}^*\big) + \lambda\sum_{l=1}^p \pi_l \big\|g_l^*\big\|_K^2\right\}.$$

Proof: Based on the definitions of $\hat{\mathbf{g}}$ in (14) and $\bar{\mathbf{g}}$ in (18), we have

$$\begin{aligned}
&\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) \\
&\quad= \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}}) + \mathcal{E}_D(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) \\
&\quad\le \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}}) + \mathcal{E}_D(\bar{\mathbf{g}}) + \lambda\Omega(\bar{\mathbf{g}}) \\
&\quad\le \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}}) + \mathcal{E}_D(\bar{\mathbf{g}}) + \lambda\Omega(\bar{\mathbf{g}}) + \lambda\sum_{l=1}^{p}\pi_l\|\bar{g}_l\|_K^2 \\
&\quad\le \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}}) + \mathcal{E}_D\big(\mathbf{g}^*\big) + \lambda\sum_{l=1}^{p}\pi_l\big\|g_l^*\big\|_K^2 + \lambda\Omega(\bar{\mathbf{g}}) \\
&\quad= \mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}}) + \mathcal{E}_D\big(\mathbf{g}^*\big) - \mathcal{E}\big(\mathbf{g}^*\big) + \lambda\Omega(\bar{\mathbf{g}}) + \mathcal{E}\big(\mathbf{g}^*\big) \\
&\qquad + \lambda\sum_{l=1}^{p}\pi_l\big\|g_l^*\big\|_K^2,
\end{aligned}$$

where the last inequality follows from Assumption 4. This completes the proof. ■

In the sequel, we focus on bounding $\lambda\Omega(\bar{\mathbf{g}})$, $\mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}})$, and $\mathcal{E}_D(\mathbf{g}^*) - \mathcal{E}(\mathbf{g}^*)$, respectively.

Proposition 2: Under Assumptions 1–3, there holds

$$\lambda\Omega(\bar{\mathbf{g}}) \le \frac{p\tilde{C}\|\phi'\|_\infty}{\sigma^2\sqrt{n}}.$$

Proof: Denote $\bar{\boldsymbol{\alpha}}_l = (\bar{\alpha}_{l1}, \bar{\alpha}_{l2}, \ldots, \bar{\alpha}_{ln})^T$. Since $\bar{\mathbf{g}}$ is the minimizer in (18) involving the loss function (12), we deduce that

$$\bar{\mathbf{g}}\big(\mathbf{x}_j\big) = \big(\mathbf{K}_j^T\bar{\boldsymbol{\alpha}}_1, \mathbf{K}_j^T\bar{\boldsymbol{\alpha}}_2, \ldots, \mathbf{K}_j^T\bar{\boldsymbol{\alpha}}_p\big)^T$$

with

$$\lambda\pi_l\mathbf{K}\bar{\boldsymbol{\alpha}}_l - \frac{1}{n^2}\sum_{i,j=1}^{n} w_{ij}\psi'_\sigma\big(Z_{ij}\big)\mathbf{K}_j\big(x_{il} - x_{jl}\big) = 0 \quad \forall l \in \{1, \ldots, p\}.$$

Observe that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j=1}^{n} w_{ij}\psi'_\sigma\big(Z_{ij}\big)\mathbf{K}_j\big(x_{il} - x_{jl}\big) \\
&\quad= \frac{1}{n^2}\mathbf{K}\Bigg(\sum_{i=1}^{n} w_{1i}\psi'_\sigma(Z_{i1})\big(x_{il} - x_{jl}\big), \ldots, \\
&\qquad\qquad \sum_{i=1}^{n} w_{1i}\psi'_\sigma(Z_{in})\big(x_{il} - x_{jl}\big)\Bigg)^T.
\end{aligned}$$

Based on the positive definiteness of $\mathbf{K}$, we have

$$\lambda\pi_l\bar{\alpha}_{tl} = \frac{1}{n^2}\sum_{i=1}^{n} w_{it}\psi'_\sigma(Z_{it})(x_{il} - x_{tl}) \quad \forall t \in \{1, \ldots, n\}.$$

Then, by direct computation, we have

$$\begin{aligned}
\lambda\Omega(\bar{\mathbf{g}}) &= \sum_{l=1}^{p}\pi_l\sqrt{\sum_{t=1}^{n}\left|\frac{1}{n^2\pi_l}\sum_{i=1}^{n} w_{it}\psi'_\sigma(Z_{it})(x_{il} - x_{tl})\right|^2} \\
&\le \sum_{l=1}^{p}\sqrt{\sum_{t=1}^{n}\left(\sum_{i=1}^{n}\frac{\tilde{C}\|\phi'\|_\infty}{n^2\sigma^2}\right)^2} = \frac{p\tilde{C}\|\phi'\|_\infty}{\sqrt{n}\sigma^2}
\end{aligned}$$

where the inequality follows from Assumption 2. ■

Now we recall McDiarmid's inequality [47].

Lemma 1: Let $z_1, \ldots, z_n$ and z_i' be independent random variables with values in $\mathcal{Z}$. For any $i \in \{1, 2, \ldots, n\}$, if $f : \mathcal{Z}^n \to \mathbb{R}$ satisfies $\sup_{z_1,\ldots,z_n,z_i'\in\mathcal{Z}}|f(z_1, \ldots, z_n) - f(z_1, \ldots, z_i', \ldots, z_n)| \le C_i$, then

$$\mathrm{Prob}\{f(z_1, \ldots, z_n) - \mathbb{E}f(z_1, \ldots, z_n) > t\} \le \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}{C_i}^2}\right\}.$$

Define the sphere with radius r as

$$\mathcal{G}_r = \left\{\mathbf{g} \in \mathcal{H}_{K,D}^p : \Omega(\mathbf{g}) \le r\right\}$$

and

$$S(D, r) := \sup_{\mathbf{g}\in\mathcal{G}_r}|\mathcal{E}(\mathbf{g}) - \mathcal{E}_D(\mathbf{g})|.$$

Lemma 2: Let Assumption 1 be true. For any given $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and $\hat{\mathbf{g}}$ in (14), there holds $\hat{\mathbf{g}} \in \mathcal{G}_r$ with $r = (\phi(0)/(\lambda\sigma))$.

Proof: Following the definition of $\hat{\mathbf{g}}$, we get

$$\begin{aligned}
\mathcal{E}_D(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) \le \mathcal{E}_D(0) + \lambda\Omega(0) &\le \frac{1}{\sigma}\left(\phi(0) - \phi\left(\frac{2M}{\sigma}\right)\right) \\
&\le \frac{\phi(0)}{\sigma}.
\end{aligned}$$

It yields $\Omega(\hat{\mathbf{g}}) \le ((\phi(0))/(\lambda\sigma))$. ■

Proposition 3: Under Assumptions 1–4, for any $\delta \in (0, 1)$, we have

$$\mathcal{E}_D\big(\mathbf{g}^*\big) - \mathcal{E}\big(\mathbf{g}^*\big) \le \frac{4\phi(0)}{\sigma}\sqrt{\frac{\ln(1/\delta)}{n}}$$

with confidence at least $1 - \delta$.

Proof: Let $z_i' = (\mathbf{x}_i', y_i')$ be an observation drawn from the distribution ρ and independent of $z_i = (\mathbf{x}_i, y_i), 1 \le i \le n$. Recall that $D = \{z_i\}_{i=1}^n$ and denote $D^i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$.

Observe that

$$\begin{aligned}
\mathcal{E}_D\big(\mathbf{g}^*\big) = \frac{1}{n^2}\Bigg\{&\sum_{k\neq i, j\neq i}^{n}\ell\big(\mathbf{g}^*, z_k, z_j\big) + \sum_{j=1}^{n}\ell\big(\mathbf{g}^*, z_i, z_j\big) \\
&+ \sum_{k=1}^{n}\ell\big(\mathbf{g}^*, z_k, z_i\big)\Bigg\}.
\end{aligned}$$

By direct computation, we get

$$\begin{aligned}
\mathcal{E}_D\big(\mathbf{g}^*\big) - \mathcal{E}_{D^i}\big(\mathbf{g}^*\big) = \frac{1}{n^2}\Bigg\{&\sum_{j=1}^{n}\big(\ell\big(\mathbf{g}^*, z_i, z_j\big) - \ell\big(\mathbf{g}^*, z_i', z_j\big)\big) \\
&+ \sum_{k=1}^{n}\big(\ell\big(\mathbf{g}^*, z_k, z_i\big) - \ell\big(\mathbf{g}^*, z_k, z_i'\big)\big)\Bigg\}.
\end{aligned}$$

Moreover,

$$\begin{aligned}\left|\mathcal{E}_D(\mathbf{g}^*) - \mathcal{E}_{D^i}(\mathbf{g}^*)\right| &\le \frac{1}{n^2}\left\{\sum_{j=1}^n \left|\ell(\mathbf{g}^*, z_i, z_j) - \ell(\mathbf{g}^*, z_i', z_j)\right|\right.\\ &\qquad \left.+\sum_{k=1}^n \left|\ell(\mathbf{g}^*, z_k, z_i) - \ell(\mathbf{g}^*, z_k, z_i')\right|\right\}\\ &\le \frac{2}{n^2}\sum_{j=1}^n \|\ell\|_\infty \le \frac{2\phi(0)}{n\sigma} := \widetilde{C}_i.\end{aligned}$$

This verifies the bounded difference property of $f(z_1,\ldots,z_n) := \mathcal{E}_D(\mathbf{g}^*)$. Then, according to Lemma 1, we have

$$\mathrm{Prob}\left\{\mathcal{E}_D(\mathbf{g}^*) - \mathcal{E}(\mathbf{g}^*) \ge \epsilon\right\} \le \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^n \widetilde{C}_i^2}\right\}.$$

Setting $\delta = \exp\{-((2\epsilon^2)/(\sum_{i=1}^n \widetilde{C}_i^2))\}$, we get the desired result. ■

Now we recall a basic result for Rademacher complexity, which is a natural extension of Lemma 22 [48].

Lemma 3: Let the kernel K be defined on $\mathcal{X}\times\mathcal{X}$ satisfying $\sup_{\mathbf{x}} K(\mathbf{x},\mathbf{x}) \le 1$. Let Φ be a feature map from input $\mathcal{X}$ to Hilbert space $\mathcal{H}$ with inner product $< \cdot,\cdot >$ such that $K(\mathbf{x},\mathbf{x}') =< \Phi(\mathbf{x}), \Phi(\mathbf{x}') >$ for all $\mathbf{x},\mathbf{x}' \in \mathcal{X}$. Denote

$$\mathcal{F} = \left\{h : \mathcal{X} \to \mathbb{R} : h(\cdot) = \sum_{t=1}^n \beta_t K(\cdot, \mathbf{x}_t),\ \|h\|_K \le B\right\}.$$

Then, $\mathcal{F} \subset \{\mathbf{x} \to< w, \Phi(\mathbf{x}) >: \|w\| \le B\}$ and the Rademacher complexity $R_n(\mathcal{F}) \le (B/\sqrt{n})$.

Now we state the upper bound of $\mathcal{E}(\hat{\mathbf{g}}) - \mathcal{E}_D(\hat{\mathbf{g}})$.

Proposition 4: Under Assumptions 1–3, for any $\delta \in (0,1)$, we have

$$\mathcal{E}(\hat{\mathbf{g}})-\mathcal{E}_D(\hat{\mathbf{g}})\le\frac{4\phi(0)}{\sigma}\sqrt{\frac{\ln(1/\delta)}{n}}+\frac{2\phi(0)}{\sigma\sqrt{n}}+\frac{4\|\phi'\|_\infty\phi(0)\tilde{C}n^{\frac{\mu}{2}}}{\sqrt{n}\lambda\sigma^3\min_j\pi_j}$$

with confidence at least $1-\delta$.

Proof: Lemma 2 assures that $\hat{\mathbf{g}} \in \mathcal{G}_r$ with $r := ((\phi(0))/(\lambda\sigma))$. Similar to the proof of Proposition 3, we can obtain that

$$|\mathcal{E}_D(\mathbf{g}) - \mathcal{E}_{D^i}(\mathbf{g})| \le \frac{2\phi(0)}{\sigma n}. \tag{19}$$

Considering $S(D,r) = \sup_{\mathbf{g}\in\mathcal{G}_r} |\mathcal{E}(\mathbf{g}) - \mathcal{E}_D(\mathbf{g})|$, we can deduce that

$$\begin{aligned}\left|S(D,r)-S(D^i,r)\right| &\le \sup_{\mathbf{g}\in\mathcal{G}_r}\big||\mathcal{E}(\mathbf{g}) - \mathcal{E}_D(\mathbf{g})| - |\mathcal{E}(\mathbf{g}) - \mathcal{E}_{D^i}(\mathbf{g})|\big|\\ &\le \sup_{\mathbf{g}\in\mathcal{G}_r}|\mathcal{E}_D(\mathbf{g}) - \mathcal{E}_{D^i}(\mathbf{g})| \le \frac{2\phi(0)}{\sigma n}.\end{aligned}$$

Applying McDiarmid's inequality in Lemma 1 to $f(z_1, z_2,\ldots, z_n) = S(D,r)$, we have

$$\mathrm{Prob}\{S(D,r) - \mathbb{E}S(D,r) \ge \epsilon\} \le \exp\left\{-\frac{n\epsilon^2\sigma^2}{4\phi^2(0)}\right\}.$$

Setting $\delta = \exp\{-((n\epsilon^2\sigma^2)/(4\phi^2(0)))\}$, we get $\epsilon = ((2\phi(0))/\sigma)(((\ln(1/\delta))/n))^{1/2}$. Then,

$$S(D,r) \le \mathbb{E}S(D,r) + \frac{2\phi(0)}{\sigma}\sqrt{\frac{\ln(1/\delta)}{n}} \tag{20}$$

with confidence at least $1-\delta$.

The rest part is to bound $\mathbb{E}S(D,r)$ with analysis techniques in [48] and [23]. We can verify that

$$\begin{aligned}S(D,r) &\le \sup_{\mathbf{g}\in\mathcal{G}_r}\left|\mathcal{E}(\mathbf{g}) - \frac{1}{n}\sum_{j=1}^n \mathbb{E}_{(\mathbf{x},y)}\ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{x}_j, y_j))\right|\\ &\quad+\sup_{\mathbf{g}\in\mathcal{G}_r}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{E}_{(\mathbf{x},y)}\ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{x}_j, y_j)) - \mathcal{E}_D(\mathbf{g})\right|\\ &\le S_1 + S_2\end{aligned} \tag{21}$$

where

$$\begin{aligned}S_1 &= \sup_{\mathbf{g}\in\mathcal{G}_r}\mathbb{E}_{(\mathbf{x},y)}\left|\mathbb{E}_{(\mathbf{u},\upsilon)}\ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{u},\upsilon))\right.\\ &\qquad\left.-\frac{1}{n}\sum_{j=1}^n \ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{x}_j, y_j))\right|\\ S_2 &= \frac{1}{n}\sum_{j=1}^n \sup_{\mathbf{g}\in\mathcal{G}_r}\sup_{(\mathbf{u},\upsilon)\in Z}\left|\mathbb{E}_{(\mathbf{x},y)}\ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{u},\upsilon))\right.\\ &\qquad\left.-\frac{1}{n}\sum_{i=1,i\neq j}^n \ell(\mathbf{g}, (\mathbf{x}_i,y_i), (\mathbf{u},\upsilon))\right|.\end{aligned}$$

Let ϵ_i, $1 \le i \le n$, be independent Rademacher variables. In terms of the properties of Rademacher complexities in [23], [48], we have

$$\begin{aligned}\mathbb{E}S_1 &\le 2\sup_{(\mathbf{x},y)}\mathbb{E}\sup_{\mathbf{g}\in\mathcal{G}_r}\left|\frac{1}{n}\sum_{j=1}^n \epsilon_j\ell(\mathbf{g}, (\mathbf{x},y), (\mathbf{x}_j, y_j))\right|\\ &= 2\sup_{(\mathbf{x},y)}\mathbb{E}\sup_{\mathbf{g}\in\mathcal{G}_r}\left|\frac{1}{n}\sum_{j=1}^n \epsilon_j w(\mathbf{x},\mathbf{x}_j)\right.\\ &\qquad\times\Big[\psi_\sigma\Big(y - y_j - \mathbf{g}(\mathbf{x}_j)^T(\mathbf{x} - \mathbf{x}_j)\Big)\\ &\qquad\left.- \psi_\sigma(y - y_j) + \psi_\sigma(y - y_j)\Big]\right|\\ &\le 2\sup_{(x,y)}\mathbb{E}\sup_{\mathbf{g}\in\mathcal{G}_r}\left|\frac{1}{n\sigma}\sum_{j=1}^n \epsilon_j w(\mathbf{x},\mathbf{x}_j)\right.\\ &\qquad\times\Big[\psi_\sigma\Big(y - y_j - \mathbf{g}(\mathbf{x}_j)^T(\mathbf{x} - \mathbf{x}_j)\Big)\\ &\qquad\left.- \psi_\sigma(y - y_j)\Big]\right| + \frac{2\phi(0)}{\sigma\sqrt{n}}\\ &\le \frac{2\|\phi'\|_\infty}{\sigma^2}\sup_{(\mathbf{x},y)}\mathbb{E}\sup_{\mathbf{g}\in\mathcal{G}_r}\left|\frac{1}{n}\sum_{j=1}^n \epsilon_j\mathbf{g}(\mathbf{x}_j)^T(\mathbf{x} - \mathbf{x}_j)\right| + \frac{2\phi(0)}{\sigma\sqrt{n}}.\end{aligned}$$

Based on Assumption 1, for any $\mathbf{g} \in \mathcal{G}_r$

$$\|\mathbf{g}\|_K \leq \sum_{l=1}^{p} \|g_l\|_K \leq n^{\frac{\mu}{2}} \sum_{l=1}^{p} \|\boldsymbol{\alpha}_l\|_2 \leq \frac{n^{\frac{\mu}{2}} r}{\min_j \pi_j}.$$

Then, from Lemma 3 and Theorem 12 in [48], we get

$$\sup_{(\mathbf{x},y)} \mathbb{E} \sup_{\mathbf{g}\in\mathcal{G}_r} \left| \frac{1}{n} \sum_{j=1}^{n} \epsilon_j \mathbf{g}(\mathbf{x}_j)^T (\mathbf{x} - \mathbf{x}_j) \right| \leq \frac{n^{\frac{\mu}{2}} \phi(0) \tilde{C}}{\min_j \pi_j \sqrt{n} \lambda \sigma}.$$

Integrating the above estimations, we have

$$\mathbb{E} S_1 \leq \frac{2\|\phi'\|_\infty n^{\frac{\mu}{2}} \phi(0) \tilde{C}}{\min_j \pi_j \sqrt{n} \lambda \sigma^3} + \frac{2\phi(0)}{\sigma\sqrt{n}}.$$

Similarly, we can get the statement

$$\mathbb{E} S_2 \leq \frac{2\|\phi'\|_\infty n^{\frac{\mu}{2}} \phi(0) \tilde{C}}{\min_j \pi_j \sqrt{n} \lambda \sigma^3} + \frac{2\phi(0)}{\sigma\sqrt{n}}.$$

The desired result follows by combining the above upper bounds of $\mathbb{E}S_1, \mathbb{E}S_2$ with (20) and (21). ■

It is a position to state the upper bound on $\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}})$.

Theorem 1: Let Assumptions 1–4 be true. For any $0 < \delta \leq 1$, with confidence at least $1 - \delta$, there holds

$$\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) - \left\{ \mathcal{E}(\mathbf{g}^*) + \lambda \sum_{l=1}^{p} \pi_l \|g_l^*\|_K^2 \right\} \leq \frac{2p\|\phi'\|_\infty \tilde{C}}{\sigma^2\sqrt{n}} + \frac{8\phi(0)}{\sigma} \sqrt{\frac{\ln(2/\delta)}{n}} + \frac{n^{\frac{\mu-1}{2}} \phi(0) \tilde{C}}{\lambda \sigma^3 \min_j \pi_j}.$$

Moreover, setting $\lambda = n^{-\nu}$ and $\sigma = n^{-\zeta}$ with positive η, ζ, and $3\zeta + \nu < 1 - \mu$, we have

$$\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) - \left\{ \mathcal{E}(\mathbf{g}^*) + \lambda \sum_{l=1}^{p} \pi_l \|g_l^*\|_K^2 \right\} \leq C n^{-\frac{1-\mu-3\zeta-\nu}{2}} \ln(2/\delta)$$

with confidence at least $1 - \delta$, where C is a positive constant independently of $nand\delta$.

Proof: Combining Propositions 1-4, we have

$$\mathcal{E}(\hat{\mathbf{g}}) + \lambda\Omega(\hat{\mathbf{g}}) \leq \frac{2p\|\phi'\|_\infty \tilde{C}}{\sigma\sqrt{n}} + \frac{10\phi(0)}{\sigma} \sqrt{\frac{\ln(2/\delta)}{n}} + \frac{n^{\frac{\mu-1}{2}} \phi(0) \tilde{C}}{\lambda \sigma^3 \min_j \pi_j} + \mathcal{E}(\mathbf{g}^*) + \lambda \sum_{l=1}^{p} \pi_l \|g_l^*\|_K^2$$

with confidence $1 - \delta$. Putting the selected parameters into the above inequality, we get the desired result. ■

Theorem 1 establishes the concentration estimation of the data-dependent regularization risk of our estimator (14) to the data-free regularization risk of $\mathbf{g}^*$. When $\mathcal{E}(\mathbf{g}^*) + \lambda \sum_{l=1}^{p} \pi_l \|g_l^*\|_K^2 \leq O(n^{-((1-\mu-3\zeta-\nu)/2)})$, the mode-induced gradient risk $\mathcal{E}(\hat{\mathbf{g}})$ tends to zero with polynomial decay rate $O(n^{-((1-\mu-3\zeta-\nu)/2)})$ under proper parameters. As a byproduct, our result also fills the gap of learning theory analysis to RGL [25] partly.

Under the Gaussian noise condition, Lemma 1 in [24] illustrated the relationship between the minimizer of risk functionals and the true gradient of regression function in probability. However, the convergence stated in Theorem 1 does not imply the consistency of the learned function to the true conditional mode function. It may be a challenge to get the function approximation guarantee directly under the zero-mode assumption. We leave it for future work.

The following theorem characterizes the properties of nonzero $\hat{\boldsymbol{\alpha}}_l$ associated with SGLML (14).

Theorem 2: Let $\{\hat{\boldsymbol{\alpha}}_l\}_{l=1}^{p}$ be the coefficients associated with $\hat{\mathbf{g}}$ in (14). For $l \in \{1, \ldots, p\}$ satisfying $\|\hat{\boldsymbol{\alpha}}_l\|_2 \neq 0$, there holds

$$\left\| \frac{1}{\sigma^2 n^2} \sum_{i,j=1}^{n} w(\mathbf{x}_i, \mathbf{x}_j) \phi'(Z_{ij}) \mathbf{K}_j (x_{il} - x_{jl}) \right\|_2 = \lambda \pi_l$$

where Z_{ij} is defined in (15).

Proof: According to (14), we know that $\{\hat{\boldsymbol{\alpha}}_l\}_{l=1}^{p}$ are with respect to the minimizer of

$$C(\boldsymbol{\alpha}) = \frac{1}{n^2} w_{ij} \psi_\sigma \Big(y_i - y_j - \mathbf{g}(\mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) \Big) + \lambda \sum_{l=1}^{p} \pi_l \|\boldsymbol{\alpha}_l\|_2$$

where ψ_σ is defined in (12), and $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$.

For each $\|\boldsymbol{\alpha}_l\|_2 \neq 0$, we take partial derivative of $C(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}_l$ and get

$$\frac{1}{\sigma^2 n^2} \sum_{i,j=1}^{n} w_{ij} \phi'(Z_{ij}) \mathbf{K}_j (x_{il} - x_{jl}) = \frac{\lambda \pi_l \boldsymbol{\alpha}_l}{\|\boldsymbol{\alpha}_l\|_2}. \tag{22}$$

The desired result follows by taking 2-norm on the both sides of (22). ■

Theorem 2 provides the necessary condition for the nonzero coefficient of SGLML, which can also be used as the stepping stone to our selection consistency analysis.

Without loss of generality, denote

$$J^* = \{1, 2, \ldots, p^*\}$$

as the index set of truly informative variables and let the active set identified by SGLML be

$$\hat{J} = \{l : \|\hat{\boldsymbol{\alpha}}_l\|_2 \geq \upsilon_n\}$$

for a threshold $\upsilon_n \geq 0$. In applications, υ_n can be obtained in terms of the stability-based selection strategy [49].

Theorem 3: Let Assumptions 1 and 2 be true. If $\lambda \pi_l \sigma^2 n^{-(1/2)} > \tilde{C}\|\phi'\|_\infty$ for $j > p*$, there holds $\hat{J} \subset J^*$ for any $D \in \mathcal{Z}^n$.

Proof: Suppose that $\|\hat{\boldsymbol{\alpha}}_l\|_2 \neq 0$ for some $l > p^*$. It is easy to check that

$$\left\| \frac{1}{n^2 \sigma^2} \sum_{i,j=1}^{n} w_{ij} \phi'(Z_{ij}) \mathbf{K}_j (x_{il} - x_{jl}) \right\|_2 \leq \frac{\tilde{C}\|\phi'\|_\infty}{\sigma^2 n} \sum_{j=1}^{n} \|\mathbf{K}_j\|_2 \leq \frac{\tilde{C}\|\phi'\|_\infty \sqrt{n}}{\sigma^2}.$$

Combining this inequality with Theorem 2, we obtain that $\lambda \pi_l \leq ((\tilde{C}\|\phi'\|_\infty \sqrt{n})/\sigma^2)$. This contradicts with the parameter condition $\lambda \pi_l \sigma^2 n^{-(1/2)} > \tilde{C}\|\phi'\|_\infty$. Hence, we know $\|\hat{\boldsymbol{\alpha}}_l\|_2 = 0$ for any $l > p^*$. This completes the proof. ■

Theorem 3 extends the analysis of variable selection consistency for the existing GL (8) (e.g., Theorem 3 in [24]) to the RGL setting.

TABLE II
FIVE NOISES USED IN THE SYNTHETIC DATA

Noise Type	Expression
I: No noise, no outliers	0
II: Gaussian noise	$0.1 \times \mathcal{N}(0, 1)$
III: Cauchy noise	$0.05 \times \text{Caunchy}(0, 1)$
IV: 30% Outliers	$0.1 \times (30\%\mathcal{N}(0, 10) + 70\%\mathcal{N}(0, 1))$
V: 30% Outliers + 70% Cauchy noise	$0.1 \times 30\%\mathcal{N}(0, 10)+$ $0.05 \times 70\%\text{Cauchy}(0, 1)$

V. EXPERIMENTAL ANALYSIS

This section evaluates the robustness of the proposed SGLML to non-Gaussian noises.

For feasibility, we denote SGLML associated with the Gaussian kernel, Logistic kernel, and Sigmoid kernel in KDE as $\text{SGLML}_{\text{Gau}}$, $\text{SGLML}_{\text{Log}}$, and $\text{SGLML}_{\text{Sig}}$, respectively. The Gaussian kernel $K_h(\mathbf{x}, \mathbf{u}) = \exp\{-\|\mathbf{x} - \mathbf{u}\|^2/2h^2\}$ is used to construct the data-dependent hypothesis space $\mathcal{H}_{K,D}^p$ and h is set as the median pairwise distance of $\{\mathbf{x}_i\}_{i=1}^n$ [25]. We choose each weight $w(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2s^2\}$ with $s = h$ and consider each π_l as 1 for simplicity. The baseline models are SGL [23], MF [24], and RGL [25].

For all $l \in \{1, \ldots, p\}$, the variable importance is measured by $r_l = ((\|g_l\|_K^2)/(\sum_{l=1}^p \|g_l\|_K^2))$ for SGL [23] and by $r_l' = ((\sum_{t=1}^n |\alpha_{lt}|^2)/(\sum_{l=1}^p \sum_{t=1}^n |\alpha_{lt}|^2))$ for MF [24], RGL [25] and our SGLML in (14).

A. Parameter Tuning and Evaluation Metrics

The SGLML and RGL contain two tuning parameters (the regularization parameter λ and the bandwidth σ in the mode-induced loss), while SGL and MF just involve λ. The bandwidth parameter σ plays a key role in KDE, which controls the smoothness of the estimated curve. As stated in [25], the variable selection results are usually not very sensitive to the choice of the regularization parameter λ when $\lambda \in [10^{-8}, 10^{-2}]$.

Following [16] and [24], we select the optimal parameters for each method using the stability-based selection criterion [49]. The stability criterion $s_{\lambda,\sigma}$ measures the stability of variable selection results and is estimated by

$$\hat{s}_{\lambda,\sigma} = \frac{1}{T} \sum_{t=1}^{T} \kappa\left(\widehat{\mathcal{A}}_{1t}, \widehat{\mathcal{A}}_{2t}\right)$$

where $\widehat{\mathcal{A}}_{1t}$ and $\widehat{\mathcal{A}}_{2t}$ are two selected variable sets, κ is the Cohen kappa coefficient [50] measuring the similarity between two selected variable sets, and T is the repeated times. The optimal parameters are selected by maximizing $\hat{s}_{\lambda,\sigma}$. Considering the probability of underfitting, we choose the one producing both the maximum number of iterations and the maximum kappa coefficient from the above selected parameters.

For the simulated data, seven metrics are adopted to measure the performance of all the methods, including size (the average number of selected variables), TP (the average number of selected truly relevant variables), FP (the average number of selected truly irrelevant variables), and C (the times of correct-fitting), U (the times of under-fitting), O (the times of over-fitting), and stability criterion $\hat{s}_{\lambda,\sigma}$. For the real-word data, we only consider the stability criterion $\hat{s}_{\lambda,\sigma}$ as the performance measurement because truly informative variables are unknown.

B. Experiments on Simulated Data

Inspired from the simulated experiments in [25] and [51], we consider the regression model

$$y = f^*(\mathbf{u}) + \epsilon, \quad \mathbf{u} = \left(u_1, \ldots, u_p\right) \in \mathbb{R}^p$$

under the following two settings.

Example 1: The additive regression function

$$f^*(\mathbf{u}) = -2\tan(0.5u_1) + u_2 + u_3 + \exp(-u_4).$$

Example 2: The nonadditive regression function

$$f^*(\mathbf{u}) = (2u_1 - 1)(2u_2 - 1).$$

We refer to [24] for the process of generating variables. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and $x_{ij} = ((W_{ij} + \eta U_i)/(1 + \eta))$, where W_{ij} and U_i are independent of $U(-0.5, 0.5)$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. For each example, we consider $\eta = 0, 1$ for noncorrelated features and correlated features. We generate data with $(n, p) = (100, 50), (100, 100), (100, 150)$ corresponding to $n > p, n = p$, and $n < p$, respectively. To estimate $\hat{s}_{\lambda,\sigma}$, we sample data S_{1t} and S_{2t} with the same size, and then apply the variable selection methods on each data to get active sets $\widehat{\mathcal{A}}_{1t}$ and $\widehat{\mathcal{A}}_{2t}$. The average stability $\hat{s}_{\lambda,\sigma}$ is obtained with $T = 10$. The evaluated results are obtained by applying each method with the selected parameters on the testing data. After 50 repetitions, we state the average results on size, TP, and FP and report the happening times of C, U, and O.

Table II summarizes the noises explored in the simulated experiments. To better evaluate the robustness of learning models, we report the variable selection results in Tables III and IV for data with the non-Gaussian noises, where SGLML usually can achieve the competitive performance than SGL and MF. As a special case of the SGLML associated with Gaussian kernel for KDE, the RGL also enjoys a similar performance.

For completeness, additional evaluations for no noise and Gaussian noise settings are stated in Appendix, where our SGLML shows comparable performance with other methods.

C. Experiments on Real-World Data

We apply the proposed SGLML to select the active variables associated with the arrival time of coronal mass ejections (CMEs). CME data (https://cdaw.gsfc.nasa.gov/CME_list/) contain 193 observations with 21 variables, including center projection angle (CPA), angle width (AW), linear speed (LS), SND speed final (SSF), SND speed 20RS (SSRS), ACCEL, MASS, kinetic energy (KE), measurement position angle (MPA), field magnitude average (FMA), BX, BY, BZ, speed (S), VX, VY, VZ, proton density (PD), temperature (T), flow pressure (FP), and plasma beta (PB). Considering the missing values in some covariates (especially for ACCEL), we remove

TABLE III
AVERAGED PERFORMANCE OF VARIABLE SELECTION UNDER NON-GAUSSIAN NOISES (EXAMPLE 1)

			Uncorrelated Variables ($\eta = 0$)							Correlated Variables ($\eta = 1$)						
Case	(n, p)	Models	Size	TP	FP	C	U	O	$\hat{s}_{\lambda,\sigma}$	Size	TP	FP	C	U	O	$\hat{s}_{\lambda,\sigma}$
III	$(100, 50)$ $n > p$	SGL	3.540	2.980	0.560	28	19	3	0.718	2.960	1.840	1.120	12	38	0	0.293
		MF	3.100	2.820	0.280	24	24	2	0.731	3.220	2.000	1.220	13	36	1	0.359
		RGL	4.000	4.000	0.000	50	0	0	1.000	4.000	**4.000**	0.000	**50**	0	0	0.988
		SGLML$_{Gau}$	4.000	4.000	0.000	50	0	0	1.000	4.020	**4.000**	0.020	49	0	1	**1.000**
		SGLML$_{Log}$	4.000	4.000	0.000	50	0	0	1.000	3.900	3.900	0.000	48	2	0	**1.000**
		SGLML$_{Sig}$	4.000	**4.000**	0.000	**50**	0	0	**1.000**	3.920	3.920	0.000	49	1	0	**1.000**
	$(100, 100)$ $n = p$	SGL	3.460	2.920	0.540	26	20	4	0.510	1.080	1.020	0.060	4	46	0	0.270
		MF	3.420	2.900	0.520	29	19	2	0.528	1.480	1.100	0.380	2	48	0	0.277
		RGL	4.000	4.000	0.000	50	0	0	1.000	4.000	**4.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	4.000	4.000	0.000	50	0	0	1.000	4.000	**4.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Log}$	4.000	4.000	0.000	50	0	0	1.000	3.800	3.800	0.000	44	6	0	0.966
		SGLML$_{Sig}$	4.000	**4.000**	0.000	**50**	0	0	**1.000**	3.920	3.880	0.040	46	2	2	**1.000**
	$(100, 150)$ $n < p$	SGL	2.880	2.420	0.460	19	27	4	0.347	0.860	0.860	0.000	0	50	0	0.270
		MF	3.320	2.500	0.820	22	27	1	0.388	2.220	1.600	0.620	3	46	1	0.096
		RGL	4.000	4.000	0.000	50	0	0	1.000	4.000	**4.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	4.000	4.000	0.000	50	0	0	1.000	3.960	3.960	0.000	49	1	0	**1.000**
		SGLML$_{Log}$	4.000	4.000	0.000	50	0	0	1.000	4.000	**4.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Sig}$	4.000	**4.000**	0.000	**50**	0	0	**1.000**	4.020	**4.000**	0.020	49	0	1	0.988
IV	($(100, 50)$ $n > p$)	SGL	3.860	3.760	0.100	35	11	4	0.972	4.080	2.460	1.620	2	47	1	0.307
		MF	4.040	3.860	0.180	35	7	8	0.988	3.080	2.020	1.060	1	49	0	0.322
		RGL	4.000	4.000	0.000	50	0	0	1.000	4.080	**4.000**	0.080	47	0	3	**1.000**
		SGLML$_{Gau}$	4.000	4.000	0.000	50	0	0	1.000	4.020	**4.000**	0.020	**49**	0	1	**1.000**
		SGLML$_{Log}$	4.000	4.000	0.000	50	0	0	1.000	4.080	3.980	0.100	44	1	5	**1.000**
		SGLML$_{Sig}$	4.000	**4.000**	0.000	**50**	0	0	**1.000**	4.040	**4.000**	0.040	48	0	2	**1.000**
	$(100, 100)$ $n = p$	SGL	4.300	3.860	0.440	33	7	10	0.909	1.960	1.460	0.500	1	49	0	0.182
		MF	4.200	3.860	0.340	34	7	9	0.933	2.340	1.640	0.700	1	48	1	0.256
		RGL	4.000	4.000	0.000	50	0	0	1.000	4.020	**3.940**	0.080	45	3	2	**1.000**
		SGLML$_{Gau}$	4.000	4.000	0.000	50	0	0	1.000	3.940	**3.940**	0.000	**47**	3	0	**1.000**
		SGLML$_{Log}$	4.000	4.000	0.000	50	0	0	1.000	3.860	3.700	0.160	32	13	5	0.852
		SGLML$_{Sig}$	4.000	**4.000**	0.000	**50**	0	0	**1.000**	4.120	3.900	0.220	41	4	5	**1.000**
	$(100, 150)$ $n < p$	SGL	3.660	3.420	0.240	27	22	1	0.885	1.320	0.940	0.380	0	50	0	0.178
		MF	3.920	3.660	0.260	33	14	3	0.925	1.640	1.120	0.520	0	50	0	0.215
		RGL	3.860	3.860	0.000	43	7	0	**1.000**	3.680	3.580	0.100	34	15	1	0.922
		SGLML$_{Gau}$	3.840	3.840	0.000	42	8	0	**1.000**	3.980	**3.820**	0.160	**42**	5	3	**1.000**
		SGLML$_{Log}$	4.020	**4.000**	0.020	**49**	0	1	**1.000**	4.080	3.500	0.580	17	20	13	0.750
		SGLML$_{Sig}$	4.160	3.960	0.200	41	2	7	0.989	3.700	3.580	0.120	33	15	2	0.916
V	$(100, 50)$ $n > p$	SGL	3.340	2.660	0.680	11	34	5	0.377	2.820	1.620	1.200	1	48	1	0.258
		MF	2.540	2.340	0.200	10	40	0	0.374	4.940	2.040	2.900	0	47	3	0.306
		RGL	4.000	**4.000**	0.000	**50**	0	0	0.988	4.000	3.860	0.140	42	5	3	0.965
		SGLML$_{Gau}$	3.980	3.980	0.000	49	1	0	**1.000**	3.960	3.800	0.160	39	7	4	0.965
		SGLML$_{Log}$	3.920	3.920	0.000	49	1	0	**1.000**	3.800	3.640	0.160	33	15	2	0.907
		SGLML$_{Sig}$	3.900	3.900	0.000	48	2	0	**1.000**	3.980	**3.880**	0.100	**43**	4	3	**0.985**
	$(100, 100)$ $n = p$	SGL	3.080	2.540	0.540	15	32	3	0.442	1.900	0.900	1.000	0	50	0	0.185
		MF	2.700	2.300	0.400	9	39	2	0.469	2.240	0.980	1.260	0	50	0	0.133
		RGL	4.020	3.980	0.040	47	1	2	0.956	4.080	**3.740**	0.320	32	12	6	0.878
		SGLML$_{Gau}$	4.000	3.980	0.020	48	1	1	**1.000**	3.440	3.200	0.240	28	21	1	0.810
		SGLML$_{Log}$	4.020	**4.000**	0.020	**49**	0	1	**1.000**	4.080	3.060	1.020	5	36	9	0.552
		SGLML$_{Sig}$	4.000	3.960	0.040	46	2	2	0.951	3.900	**3.740**	0.160	**39**	9	2	**0.921**
	$(100, 150)$ $n < p$	SGL	1.460	1.440	0.020	2	48	0	0.362	0.960	0.600	0.360	0	50	0	0.185
		MF	1.220	0.800	0.420	0	50	0	0.397	1.040	0.680	0.360	0	50	0	0.133
		RGL	3.800	3.680	0.120	32	15	3	0.908	4.060	**3.440**	0.620	**27**	20	3	**0.860**
		SGLML$_{Gau}$	3.920	3.900	0.020	45	4	1	0.971	3.900	3.040	0.860	8	36	6	0.666
		SGLML$_{Log}$	4.040	**3.960**	0.080	**47**	2	1	**1.000**	3.580	2.860	0.720	10	38	2	0.672
		SGLML$_{Sig}$	3.680	3.660	0.020	36	14	0	0.922	3.620	3.240	0.380	20	27	3	0.784

TABLE IV
AVERAGED PERFORMANCE OF VARIABLE SELECTION UNDER NON-GAUSSIAN NOISES (EXAMPLE 2)

			Uncorrelated Variables ($\eta = 0$)							Correlated Variables ($\eta = 1$)						
Case	(n, p)	Models	Size	TP	FP	C	U	O	$\hat{s}_{\lambda,\sigma}$	Size	TP	FP	C	U	O	$\hat{s}_{\lambda,\sigma}$
III	$(100, 50)$ $n > p$	SGL	1.760	1.760	0.000	43	7	0	0.637	3.040	1.500	1.540	25	17	8	0.507
		MF	1.760	1.760	0.000	43	7	0	0.608	2.440	1.500	0.940	24	17	9	0.570
		RGL	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Log}$	2.000	2.000	0.000	50	0	0	1.000	1.960	1.960	0.000	49	0	1	1.000
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	1.960	1.960	0.000	49	0	1	1.000
	$(100, 100)$ $n = p$	SGL	1.580	1.580	0.000	37	13	0	0.731	1.680	1.360	0.320	24	19	7	0.490
		MF	1.680	1.600	0.080	37	12	1	0.731	1.660	1.340	0.320	24	19	7	0.512
		RGL	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Log}$	1.960	1.960	0.000	49	1	0	1.000	1.980	1.960	0.020	48	1	1	1.000
		SGLML$_{Sig}$	1.960	1.960	0.000	49	1	0	1.000	1.980	1.960	0.020	48	1	1	1.000
	$(100, 150)$ $n < p$	SGL	1.520	1.460	0.060	34	16	0	0.733	1.460	1.400	0.060	29	20	1	0.282
		MF	1.660	1.500	0.160	34	16	0	0.733	2.700	1.640	1.060	14	13	23	0.285
		RGL	2.000	2.000	0.000	50	0	0	1.000	2.040	2.000	0.040	49	0	1	1.000
		SGLML$_{Gau}$	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Log}$	2.000	2.000	0.000	50	0	0	1.000	2.080	2.000	0.080	46	0	4	1.000
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.000	**2.000**	0.000	**50**	0	0	**1.000**
IV	$(100, 50)$ $n > p$	SGL	2.000	2.000	0.000	50	0	0	1.000	1.980	1.920	0.060	44	4	2	0.958
		MF	2.000	2.000	0.000	50	0	0	1.000	2.200	1.940	0.260	37	3	10	0.923
		RGL	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Log}$	2.000	2.000	0.000	50	0	0	1.000	2.040	2.000	0.040	48	0	2	1.000
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.000	**2.000**	0.000	**50**	0	0	**1.000**
	$(100, 100)$ $n = p$	SGL	2.020	2.000	0.020	49	0	1	1.000	2.240	1.980	0.260	36	1	13	0.884
		MF	2.020	2.000	0.020	49	0	1	1.000	2.180	1.980	0.200	39	1	10	0.884
		RGL	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Gau}$	2.000	2.000	0.000	50	0	0	1.000	2.060	1.980	0.080	46	1	3	1.000
		SGLML$_{Log}$	2.000	2.000	0.000	50	0	0	1.000	2.000	**2.000**	0.000	**50**	0	0	**1.000**
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.120	2.000	0.120	44	0	6	1.000
	$(100, 150)$ $n < p$	SGL	2.000	2.000	0.000	50	0	0	1.000	2.240	1.920	0.320	35	4	11	0.933
		MF	2.000	2.000	0.000	50	0	0	1.000	2.280	1.920	0.360	33	4	13	0.933
		RGL	2.020	2.000	0.020	49	0	1	1.000	2.080	2.000	0.080	46	0	4	1.000
		SGLML$_{Gau}$	2.040	2.000	0.040	48	0	2	1.000	2.060	**2.000**	0.060	**47**	0	3	**1.000**
		SGLML$_{Log}$	2.020	2.000	0.020	49	0	1	1.000	2.080	**2.000**	0.060	**47**	0	3	**1.000**
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.060	**2.000**	0.060	**47**	0	3	**1.000**
V	$(100, 50)$ $n > p$	SGL	1.940	1.840	0.100	43	4	3	0.829	1.920	1.620	0.340	30	14	6	0.645
		MF	1.900	1.860	0.040	46	4	0	0.829	1.940	1.620	0.320	30	13	7	0.676
		RGL	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.060	2.000	0.060	49	0	1	1.000
		SGLML$_{Gau}$	2.020	2.000	0.020	49	0	1	1.000	2.040	2.000	0.040	49	0	1	1.000
		SGLML$_{Log}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.040	2.000	0.040	48	0	2	1.000
		SGLML$_{Sig}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.040	**2.000**	0.040	**49**	0	1	**1.000**
	$(100, 100)$ $n = p$	SGL	2.240	1.800	0.440	38	9	3	0.846	2.020	1.500	0.520	21	21	8	0.700
		MF	1.860	1.800	0.060	41	6	3	0.815	1.680	1.420	0.260	22	22	6	0.687
		RGL	2.020	2.000	0.020	49	0	1	1.000	2.060	1.920	0.140	42	3	5	1.000
		SGLML$_{Gau}$	2.040	2.000	0.040	48	0	2	1.000	2.000	1.940	0.060	46	3	1	1.000
		SGLML$_{Log}$	2.000	**2.000**	0.000	**50**	0	0	**1.000**	2.020	**2.000**	0.020	**49**	0	1	**1.000**
		SGLML$_{Sig}$	2.020	2.000	0.020	49	0	1	1.000	2.040	**2.000**	0.040	**49**	0	1	**1.000**
	$(100, 150)$ $n < p$	SGL	1.680	1.580	0.100	34	12	4	0.966	3.500	1.580	1.920	8	14	28	0.126
		MF	1.680	1.540	0.140	33	14	3	0.966	3.920	1.640	2.280	4	12	34	0.195
		RGL	2.040	1.960	0.080	44	2	4	1.000	2.100	1.980	0.120	43	1	6	0.980
		SGLML$_{Gau}$	2.040	1.960	0.080	44	2	4	1.000	2.180	1.980	0.200	40	1	9	0.980
		SGLML$_{Log}$	2.020	2.000	0.020	49	0	1	1.000	2.040	**2.000**	0.040	**48**	0	2	**1.000**
		SGLML$_{Sig}$	2.000	**2.000**	**0.000**	**50**	0	0	**1.000**	2.100	1.980	0.120	43	1	6	1.000

TABLE V

Variable Selection Results and Stability Criterion on Real Data

Method	$\varrho = 0$	$\varrho = 0.1$	$\varrho = 0.5$
SGL	LS, SSF, SSRS (0.922)	AW, LS, SSF, SSRS, irre20 (0.727)	U/{KE, FP} (0.361)
MF	LS, SSF, SSRS (0.906)	AW, LS, SSF, SSRS, irre17, irre20 (0.643)	U/{LS, SSF, FP, irre1, irre3, irre18, irre19} (0.352)
RGL	LS, SSF, SSRS (0.938)	AW, LS, SSF, SSRS (0.593)	CPA, AW, LS, SSF, SSRS, MASS (0.353)
SGLML$_{Gau}$	LS, SSF, SSRS (**0.969**)	AW, LS, SSF, SSRS (0.618)	CPA, AW, LS, SSF, SSRS (0.435)
SGLML$_{Log}$	LS, SSF, SSRS (0.922)	AW, LS, SSF, SSRS (**0.828**)	CPA, AW, LS, SSF, SSRS (**0.494**)
SGLML$_{Sig}$	LS, SSF, SSRS (**0.969**)	AW, LS, SSF, SSRS (0.770)	CPA, AW, LS, SSF, SSRS (0.491)

[1] U is denoted as the set of all 40 variables and $U/\text{set1}$ is denoted as the set $\{x : x \in U$ and $x \notin \text{set1}\}$.

the feature of ACCEL first, and then delete the instances with other missing feature. Twenty irrelevant variables are generated from the uniform distribution $U(-0.5, 0.5)$ to enlarge the features, which are denoted as irre1, ..., irre20. The remaining 134 observations with 40 variables are considered for our analysis.

To estimate $\hat{s}_{\lambda,\sigma}$, we randomly divide the data equally into two parts with ten repetitions. The noise from $(\varrho \times \mathcal{N}(0, 3) + (1 - \varrho) \times \mathcal{N}(0, 1))$ is added to the data, where its strength is controlled by $\varrho \in [0, 1]$.

The variable selection results are presented in Table V. When $\varrho = 0$, all the methods have similar performance, while SGLML behaves more stable as it has the maximum $\hat{s}_{\lambda,\sigma}$. For SGL and MF, some irrelevant variables are selected when $\varrho = 0.1$ and fail to select active variables when $\varrho = 0.5$. However, the SGLML can achieve satisfactory variable selection performance under different noise levels.

VI. Conclusion

This article formulated a class of model-free variable selection models by integrating SGL and mode-induced loss into the Tikhonov regularization scheme. The effectiveness of the proposed approach is validated by the theoretical analysis on error bound and variable selection consistency, as well as experimental analysis on the simulated and real data.

Appendix

The experimental results for the simulated data under no noise and Gaussian noise settings.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[2] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, Jun. 2019.

[3] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.

[4] M. Yuan and D.-X. Zhou, "Minimax optimal rates of estimation in high dimensional additive models," *Ann. Statist.*, vol. 44, no. 6, pp. 2564–2593, Dec. 2016.

[5] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Jan. 1978.

[6] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Cham, Switzerland: Springer, 1998, pp. 199–213.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its Oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[9] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Statist.*, vol. 13, no. 2, pp. 689–705, 1985.

[10] T. Hastie and R. Tibshirani, "Generalized additive models," *Stat. Sci.*, vol. 1, no. 3, pp. 297–318, 1986.

[11] S. Lv, H. Lin, H. Lian, and J. Huang, "Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space," *Ann. Statist.*, vol. 46, no. 2, pp. 781–813, Apr. 2018.

[12] Y. Lin and H. H. Zhang, "Component selection and smoothing in multivariate nonparametric regression," *Ann. Statist.*, vol. 34, no. 5, pp. 2272–2297, Oct. 2006.

[13] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *J. Amer. Stat. Assoc.*, vol. 106, no. 494, pp. 544–557, Jun. 2011.

[14] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, "Sparse additive models," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 71, no. 5, pp. 1009–1030, 2009.

[15] J. Yin, X. Chen, and E. P. Xing, "Group sparse additive models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1–15.

[16] H. Chen, Y. Wang, F. Zheng, C. Deng, and H. Huang, "Sparse modal additive model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2373–2387, Jun. 2021.

[17] Y. Wang, H. Chen, F. Zheng, C. Xu, T. Gong, and Y. Chen, "Multi-task additive models for robust estimation and automatic structure discovery," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1–13.

[18] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, "Semiparametric estimates of the relation between weather and electricity sales," *J. Amer. Stat. Assoc.*, vol. 81, no. 394, pp. 310–320, Jun. 1986.

[19] H. H. Zhang, G. Cheng, and Y. Liu, "Linear or nonlinear? Automatic structure discovery for partially linear models," *J. Amer. Stat. Assoc.*, vol. 106, no. 495, pp. 1099–1112, 2011.

[20] J. Huang, F. Wei, and S. Ma, "Semiparametric regression pursuit," *Statistica Sinica*, vol. 22, no. 4, pp. 1403–1426, 2012.

[21] Y. Lou, J. Bien, R. Caruana, and J. Gehrke, "Sparse partially linear additive models," *J. Comput. Graph. Statist.*, vol. 25, no. 4, pp. 1026–1040, 2018.

[22] S. Mukherjee and D.-X. Zhou, "Learning coordinate covariances via gradients," *J. Mach. Learn. Res.*, vol. 7, pp. 519–549, Mar. 2006.

[23] G.-B. Ye and X. Xie, "Learning sparse gradients for variable selection and dimension reduction," *Mach. Learn.*, vol. 87, no. 3, pp. 303–355, 2012.

[24] L. Yang, S. Lv, and J. Wang, "Model-free variable selection in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2885–2908, 2016.

[25] Y. Feng, Y. Yang, and J. A. K. Suykens, "Robust gradient learning with applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 822–835, Apr. 2016.

[26] S. Mukherjee and Q. Wu, "Estimation of gradients and coordinate covariation in classification," *J. Mach. Learn. Res.*, vol. 7, pp. 2481–2514, Nov. 2006.

[27] S.-G. Lv, "Refined generalization bounds of gradient learning over reproducing kernel Hilbert spaces," *Neural Comput.*, vol. 27, no. 6, pp. 1294–1320, Jun. 2015.

[28] S. Mukherjee, Q. Wu, and D.-X. Zhou, "Learning gradients on manifolds," *Bernoulli*, vol. 16, no. 1, pp. 181–207, 2010.

[29] Q. Wu, J. Guinney, M. Maggioni, and S. Mukherjee, "Learning gradients: Predictive models that infer geometry and statistical dependence," *J. Mach. Learn. Res.*, vol. 11, pp. 2175–2198, Aug. 2010.

[30] J. Guinney, Q. Wu, and S. Mukherjee, "Estimating variable structure and dependence in multitask learning via gradients," *Mach. Learn.*, vol. 83, no. 3, pp. 265–287, 2011.

[31] Y. Ying, Q. Wu, and C. Campbell, "Learning the coordinate gradients," *Adv. Comput. Math.*, vol. 37, no. 3, pp. 355–378, 2012.

[32] H. Chen and Y. Wang, "Kernel-based sparse regression with the correntropy-induced loss," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 1, pp. 144–164, 2018.

[33] Y. Feng and Y. Ying, "Learning with correntropy-induced losses for regression with mixture of symmetric stable noise," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 795–810, Mar. 2020.

[34] W. Yao and L. Li, "A new regression model: Modal linear regression," *Scandin. J. Statist.*, vol. 41, no. 3, pp. 656–671, Sep. 2014.

[35] Y. Feng, J. Fan, and J. A. Suykens, "A statistical learning approach to modal regression," *J. Mach. Learn. Res.*, vol. 21, no. 2, pp. 1–35, 2020.

[36] X. Wang, H. Chen, W. Cai, D. Shen, and H. Huang, "Regularized modal regression with applications in cognitive impairment prediction," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.

[37] Y. Wang, Y. Y. Tang, L. Li, and H. Chen, "Modal regression-based atomic representation for robust face recognition and reconstruction," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4393–4405, Oct. 2020.

[38] J. Xu, F. Wang, Q. Peng, X. Y. S. Wang, X.-Y. Jing, and C. L. P. Chen, "Modal regression based structured low-rank matrix recovery for multiview learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1204–1216, Mar. 2021.

[39] H. Chen, X. Wang, C. Deng, and H. Huang, "Group sparse additive machine," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 1–15.

[40] G. Liu, H. Chen, and H. Huang, "Sparse shrunk additive models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–12.

[41] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[42] J. C. Príncipe, *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer, 2010.

[43] Y. Feng, X. Huang, L. Shi, Y. Yang, and J. A. K. Suykens, "Learning with the maximum correntropy criterion induced losses for regression," *J. Mach. Learn. Res.*, vol. 16, pp. 993–1034, Jan. 2015.

[44] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[45] I. Steinwart and A. Christmann, *Support Vector Machines*. Cham, Switzerland: Springer, 2008.

[46] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.

[47] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*. Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 148–188.

[48] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Jan. 2002.

[49] W. Sun, J. Wang, and Y. Fang, "Consistent selection of tuning parameters via variable selection stability," *J. Mach. Learn. Res.*, vol. 14, no. 9, pp. 3419–3440, 2013.

[50] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[51] L. Yang, S. Sperlich, and W. Härdle, "Derivative estimation and testing in generalized additive models," *J. Stat. Planning Inference*, vol. 115, no. 2, pp. 521–542, Aug. 2003.

Hong Chen received the B.S., M.S., and Ph.D. degrees from Hubei University, Wuhan, China, in 2003, 2006, and 2009, respectively.

From February 2016 to August 2017, he worked as a Post-Doctoral Researcher with the Department of Computer Science and Engineering, University of Texas, Arlington, TX, USA. He is currently a Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan, China. His relevant works have been published at the prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the *Neural Computation*, the Advances in Neural Information Processing Systems, and the International Conference on Machine Learning. His research interests include machine learning, statistical learning theory, and approximation theory.

Youcheng Fu recieved the B.S. degree from Huazhong Agricultural University, Wuhan China, in 2020, where he is currently pursuing the master's degree with the College of Science.

Xue Jiang received the B.Sc. degree in electronic information engineering and the M.Sc. degree in computer science and technology from Wuhan University, Wuhan, China, in 2019 and 2022, respectively. She is currently pursuing the joint Ph.D. degree with the Southern University of Science and Technology (SUSTech) and Hong Kong Baptist University (HKBU).

Her research interests lie in the area of machine learning and computer vision.

Yanhong Chen received the Ph.D. degree in 2011 from National Space Science Center, Chinese Academy of Sciences, Beijing, China.

Since 2011, she has been working as an Assistant Research Fellow with the National Space Science Center, mainly on space environment research and modeling. Her current research interests include integration research of space weather and machine learning and its application in space weather prediction.

Weifu Li received the B.S., M.S. and Ph.D. degrees from Hubei University, Wuhan, China, in 2014, 2017, and 2020, respectively.

From July 2015 to August 2019, he was a Joint Student with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. From September 2019 to July 2020, he was a Joint Student with the Research Center for Smart Vehicles, Toyota Technological Institute, Nagoya, Japan. He is currently an Associate Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His relevant works have been published at the prestigious journals, such as *Science*, *Nature Genetics*, *Molecular Plant*, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. His research interests include machine learning and statistical learning theory and its applications.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University in Changsha, China, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020 and 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Feng Zheng (Member, IEEE) received the Ph.D. degree from The University of Sheffield, Sheffield, U.K., in 2017.

He is currently an Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include machine learning, computer vision, and human–computer interaction.