Statistica Sinica **30** (2020), 2075-2103
doi:https://doi.org/10.5705/ss.202018.0086

VARIABLE SELECTION FOR CLASSIFICATION WITH DERIVATIVE-INDUCED REGULARIZATION

Xin He, Shaogao Lv and Junhui Wang

Shanghai University of Finance and Economics, Nanjing Audit University and City University of Hong Kong

Abstract: Despite extensive research on variable selection over the past two decades, few studies exist on variable selection for classification, particularly when no assumptions are made about the model. In this paper, we propose a general variable selection framework for classification by examining the conditional probability. The proposed framework is illustrated by means of support vector machine (SVM) with derivative-induced sparsity, which makes no explicit model assumption, and takes full advantage of the mathematical properties of the reproducing kernel Hilbert space (RKHS). In contrast to many existing methods, our proposed method leads to a convex optimization task, and fully exploits gradient information by using the reproducing property of gradients in smooth RKHSs. The proposed method can also be viewed as a generalization of the classical SVM, and achieves superior empirical performance in sparse classification. Importantly, the estimation consistency and subset selection properties of the proposed method are established. Lastly, the effectiveness of the method is demonstrated using simulated and real-life examples.

Key words and phrases: Classification, gradient learning, reproducing kernel Hilbert space (RKHS), sparsity, support vector machine (SVM).

1. Introduction

In a binary classification, a random pair $\mathcal{Z} = (\mathbf{x}, y)$ is drawn from some unknown distribution $\rho_{\mathbf{x},y}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathbf{x} = (x^1, \ldots, x^p)^T \in \mathcal{X} \subset \mathcal{R}^p$ and $y \in \mathcal{Y} = \{-1, 1\}$. The main purpose of the classification is to determine a classification rule f that minimizes the misclassification error $EI(y \neq f(\mathbf{x}))$, where I is the indicator function. It is well known that the optimal classification rule that minimizes this error is the so-called Bayes rule, $f_{Bayes}(\mathbf{x}) = \text{sign}(P(y = 1|\, \mathbf{x}) - 1/2)$. In the literature, various convex surrogate functions have been proposed as alternative to the indicator function to facilitate the estimation of $f_{Bayes}(\mathbf{x})$ and improve numerical performance. Yet, as pointed out in Fan and Fan (2008), in high-dimensional classifications, the performance can be as poor as that of a random guess if all the variables are used, accounting for the accumulation of noise. Hence, it is generally believed that, among all the collected variables, only a small number are truly informative for classification, making it crucial that they can be identified correctly.

In the literature, the support vector machine (SVM) is a popular classification method, with numerous variable selection methods having been developed under its framework. For a linear SVM, Zhou et al. (2004) impose the L^1-penalty to push the coefficient of noninformative variables to be exactly zero and thus achieve variable selection. Zou (2007) proposes an adaptively weighted L^1-norm penalty to further exploit the sparsity. Zou and Yuan (2008) consider the F_∞-norm SVM to select groups of features. Zhang et al. (2016) propose a variable selection method for a linear SVM in a diverging dimension setting. For a nonlinear SVM, Bi et al. (2003) propose first identifying a subset of informative variables using a series of sparse linear SVMs, and then fitting a nonlinear model based on the identified subset. Zhang (2006) proposes a variable selection method for a nonlinear SVM and uses the smoothing spline ANOVA model. Zhao and Liu (2012) consider a combination of a functional SVM and a sparse additive model to achieve variable selection. Note that the performance of the aforementioned methods largely relies on the validity of their prespecified model assumptions.

In this paper, we propose a general variable selection framework for classification that aims to identify the informative variables that contribute to the Bayes rule. Specifically, we illustrate the proposed framework using an SVM; however, it can be extended to other classification problems. The proposed framework is motivated by the fact that the Bayes rule depends on truly informative variables only. Thus, the contribution of the noninformative variables to the Bayes rule, measured by the corresponding derivative function, shall necessarily be zero, almost surely. The proposed framework is then expressed in regularization form, with a derivative-induced penalty in a smooth reproducing kernel Hilbert space (RKHS). This penalty fully utilizes the property that derivatives are bounded linear functionals in a RKHS with suitable representation. This property makes the computation feasible and efficient, and also ensures the reliability and validity of the estimation. An efficient algorithm is developed based on Nesterov's method (Nesterov (2005)) to solve the resultant optimization task. The asymptotic estimation and subset selection results are established under mild conditions. The theoretical results ensure that the proposed method is able to recover the truly informative variables with probability tending to one.

There are several salient advantages of our proposed framework. First, a general variable selection framework is established that aims to discover all truly informative variables contributing to the Bayes rule, and can be extended to various popular loss functions. Second, unlike most existing gradient-based methods in the literature (Yang, Lv and Wang (2016)), we combine the L^2-norm of the corresponding partial derivatives with the standard RKHS-norm to avoid heavy computation in local pairwise learning. The novel regularization term not only ensures the existence and uniqueness of the minimizer in the regularization problem (Ekeland and Temam (1976)), but also ensures the explicit form of the minimizer, from the properties of the RKHS and the representer theorem (Rosasco et al. (2013)). Third, the asymptotic estimation and subset selection results are established without imposing explicit model assumptions, in contrast to most existing theoretical results that make specific model assumptions.

The rest of the paper is organized as follows. Section 2 presents the general framework of the variable selection for classification, and introduces the proposed variable selection method by means of an SVM, as well as its computing algorithm. Section 3 establishes the statistical properties of the proposed method. Section 4 contains numerical results for both simulated and real-life examples. Section 5 concludes the summary. The Appendix contains all the computational details and technical proofs.

2. Proposed Methodology

2.1. Variable selection for classification

Let $\eta(\mathbf{x}) = P(y = 1|\,\mathbf{x})$. Then, the Bayes rule $f_{Bayes}(\mathbf{x}) = \text{sign}(P(y = 1|\,\mathbf{x}) - 1/2) = \text{sign}(\eta(\mathbf{x}) - 1/2)$. In a sense, the information related to classification can be fully captured by the conditional probability $\eta(\mathbf{x})$. A variable x^l is deemed noninformative if and only if it does not contribute to $\eta(\mathbf{x})$, given all other variables $\mathbf{x}^{-l}$, or more precisely,

$$\eta(\mathbf{x}^{-l}, x^l) = \eta(\mathbf{x}^{-l}). \tag{2.1}$$

This definition differs from that of many existing methods, which often rely on various specific model assumptions (Zhang (2006); Zou (2007); Zhang et al. (2016)), and was only recently considered by Barber and Candès (2015); Lee, Li and Zhao (2016); Li and Liu (2018). In contrast to existing methods, we quantify the contribution of x^l in $\eta(\mathbf{x})$ by examining its corresponding derivative function,

$$\partial_l \eta(\mathbf{x}) = \frac{\partial \eta(\mathbf{x})}{\partial x^l}, \tag{2.2}$$

and a noninformative variable can be implied by $\partial_l \eta(\mathbf{x}) = 0$ almost surely. Hence, the true active set $\mathcal{A}^*$ is defined as $\mathcal{A}^* = \{l : \|\partial_l \eta(\mathbf{x})\|^2_{\rho_{\mathbf{x}}} > 0\}$, where $\|\partial_l \eta(\mathbf{x})\|^2_{\rho_{\mathbf{x}}} = \int_{\mathcal{X}} \big(\partial_l \eta(\mathbf{x})\big)^2 d\rho_{\mathbf{x}}$ is the L^2-norm induced by the marginal probability measure $\rho_{\mathbf{x}}$ on $\mathcal{X}$.

Let $L(yf(\mathbf{x}))$ be a margin-based surrogate loss function, and $f^* = \operatorname{argmin}_{f \in \mathcal{H}_K} EL(yf(\mathbf{x}))$, where $\mathcal{H}_K$ is an RKHS associated with a specified kernel K and endowed with the norm $\|\cdot\|_K$. Note that $f^*(\mathbf{x})$ is a function of $\eta(\mathbf{x})$ for many surrogate losses. For example, for the square error loss $L(t) = (1-t)^2$, $f^*_S(\mathbf{x}) = 2\eta(\mathbf{x}) - 1$; for the logistic loss $L(t) = (1/\ln 2)\ln(1+e^{-t})$, $f^*_L(\mathbf{x}) = \ln(\eta(\mathbf{x})/(1-\eta(\mathbf{x})))$; and for the hinge loss $L(t) = (1-t)_+$,

$$\begin{aligned} f^*_H(\mathbf{x}) &= \operatorname*{argmin}_{f \in \mathcal{H}_K} E(1 - yf(\mathbf{x}))_+ \\ &= \operatorname*{argmin}_{f \in \mathcal{H}_K} (1 - f(\mathbf{x}))_+ \eta(\mathbf{x}) + (1 + f(\mathbf{x}))_+ (1 - \eta(\mathbf{x})). \end{aligned}$$

In such cases, it follows from the chain rule of derivatives that

$$g^*_l(\mathbf{x}) = \partial_l f^*(\mathbf{x}) = \frac{\partial f^*(\mathbf{x})}{\partial \eta} \partial_l \eta(\mathbf{x}). \tag{2.3}$$

When $f^*(\mathbf{x})$ is not a degenerate function of $\eta(\mathbf{x})$, $\partial f^*(\mathbf{x})/\partial \eta$ is nonzero almost surely. With this assumption, it suffices to estimate the corresponding derivative functions g^*_l instead of $\eta(\mathbf{x})$ for the purpose of variable selection. For example, in Li and Liu (2018), the logistic loss is used and $\partial f^*(\mathbf{x})/\partial \eta = (\eta(\mathbf{x})(1-\eta(\mathbf{x})))^{-1}$, which is nonzero as long as $\eta(\mathbf{x})$ is not exactly zero or one.

We illustrate the proposed framework of the derivative-induced sparsity method using the popular hinge loss; however, the framework can be extended readily to other loss functions. Furthermore, we propose a general variable selection framework for classification by examining the sparse structure of the conditional probability, an area that is under-researched in the statistics literature, but is gaining attention (Li and Liu (2018)). In contrast to Rosasco et al. (2013), who examine the properties of the derivative-induced regularizer under a squared loss, the proposed method uses the derivative-induced regularizer as a tool to identify the variables that contribute to the Bayes rule.

2.2. Kernel SVM

Throughout this paper, suppose that a random sample $\mathcal{Z}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is an independent copy of $\mathcal{Z} = (\mathbf{x}, y)$, and that $\eta(\mathbf{x}) \in \mathcal{H}_K$. A standard Kernel

SVM problem is formulated as

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|f\|_K^2,$$

where the first term is the sample version of $\mathcal{E}(f) = E(1 - yf(\mathbf{x}))_+$, denoted as $\mathcal{E}_{\mathcal{Z}_n}(f)$, and $\|f\|_K^2$ is the RKHS-norm controlling the smoothness of f.

Any given symmetric and positive semi-definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ induces an RKHS $\mathcal{H}_K$. For each $\mathbf{x} \in \mathcal{X}$, the function $\mathbf{z} \to K(\mathbf{z}, \mathbf{x})$ is contained in $\mathcal{H}_K$. In addition, $\mathcal{H}_K$ is endowed with an inner product $\| \cdot \|_K$, such that the following reproducing property is satisfied:

$$f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot) \rangle_K, \quad \text{for any } f \in \mathcal{H}_K.$$

More importantly, under suitable smoothness conditions on K, such as twice-continuous differentiability, there holds $\partial_l K_{\mathbf{x}} = \partial_l K(\mathbf{x}, \cdot) = (\partial K(\mathbf{s}, \cdot)/\partial s^l)\big|_{\mathbf{s}=\mathbf{x}} \in \mathcal{H}_K$, by Theorem 1 of Zhou (2007). The reproducing property of derivative functions in $\mathcal{H}_K$ is also satisfied:

$$g_l(\mathbf{x}) = \partial_l f(\mathbf{x}) = \langle f, \partial_l K(\mathbf{x}, \cdot) \rangle_K, \quad \text{for } f \in \mathcal{H}_K, \tag{2.4}$$

where $g_l(\mathbf{x})$ is the partial derivative of $f(\mathbf{x})$ with respect to x^l. It can be shown that both polynomial kernels and Gaussian kernels satisfy these conditions. In particular, the Gaussian kernels are also universal on every compact subset of the Euclidean space, implying that they can closely approximate any continuous true function.

The reproducing property of derivative functions given in (2.4) is interesting, and somewhat surprising. In general, estimating a derivative function is more difficult than estimating the function itself, and the former exhibits a slower convergence rate and is more sensitive to noise . However, in a smooth RKHS, we can see from (2.4) that the difficulty of estimating a derivative function is almost the same as that of estimating the function itself. Moreover, the reproducing property in (2.4) ensures that any infinite-dimensional optimization with derivative-induced regularization within a smooth RKHS can be reduced to a finite-dimensional optimization problem. More importantly, as shown in Subsection 2.3, the number of parameters in our proposed kernel-induced optimization is at most $(n+1)p$. In contrast, existing methods that capture higher-order nonlinear effects are often expressed in an enumerate way. The resulting function is often formulated as $f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x^j) + \sum_{k<j} f_{k,j}(x^k, x^j) + \cdots$ by a group of authors, including Bach (2009) and Lin and Zhang (2006). Clearly, the number of parameters in these models is more than an exponential order of the data dimension, which incurs a tremendous computational cost in high-dimensional cases.

2.3. Proposed formulation

Our proposed method is formulated to solve the following optimization problem:

$$\underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; \mathcal{E}_{\mathcal{Z}_n}(f) + \lambda_0 \|f\|_K^2 + \lambda_1 \Omega_{[p]}(f), \tag{2.5}$$

where $\Omega_{[p]}(f) = \sum_{l=1}^{p} \pi_l \|g_l\|_{\rho_{\mathbf{x}}}$ is a lasso-type penalty to induce sparsity in f, and λ_0 and λ_1 are both tuning parameters. Note that π_l is often chosen adaptively to assign different weights to the derivative functions in order to attain asymptotic selection consistency; see Section 3. Note that our proposed penalty terms can be viewed as a functional version of the elastic net (Zou and Trevor (2005)), which deals efficiently with highly correlated data.

The true marginal measure $\rho_{\mathbf{x}}$ is unknown in practice. Thus, the sample version of $\|g_l\|_{\rho_{\mathbf{x}}}^2$ is defined as

$$\|g_l\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \big(g_l(\mathbf{x}_i)\big)^2 = \frac{1}{n} \sum_{i=1}^{n} \big(\partial_l f(\mathbf{x}_i)\big)^2.$$

Hence, the objective function in (2.5) becomes

$$\underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \big(1 - y_i f(\mathbf{x}_i)\big)_+ + \lambda_0 \|f\|_K^2 + \lambda_1 \sum_{l=1}^{p} \pi_l \|g_l\|_n, \tag{2.6}$$

where the last term is an empirical version of $\Omega_{[p]}(f)$, denoted as $\widehat{\Omega}_{[p]}(f)$. Denote the minimizer of (2.6) as $\widehat{f}$ and the selected informative variable set as $\widehat{\mathcal{A}} = \{l : \|\widehat{g}_l\|_n > 0\}$, with $\widehat{g}_l(\mathbf{x}) = \langle \widehat{f}, \partial_l K_{\mathbf{x}} \rangle_K$. Note that by an extended version of the representer theorem (Wahba (1998)), in an RKHS, the minimizer of (2.6) must have the following form (Rosasco et al. (2013)):

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^{n} \sum_{l=1}^{p} \beta_i^l \partial_l K_{\mathbf{x}_i}(\mathbf{x}) = \mathbf{c}^T \mathbf{M}_n(\mathbf{x}), \tag{2.7}$$

where $\mathbf{c} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$, for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$, with $\boldsymbol{\beta}_l = (\beta_1^l, \ldots, \beta_n^l)$; and $\mathbf{M}_n(\mathbf{x}) = (\mathbf{K}_n(\mathbf{x})^T, \partial_1 \mathbf{K}_n(\mathbf{x})^T, \ldots, \partial_p \mathbf{K}_n(\mathbf{x})^T)^T \in \mathcal{R}^{n(p+1)}$, with $\mathbf{K}_n(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x}))^T$ and $\partial_l \mathbf{K}_n(\mathbf{x}) = (\partial_l K_{\mathbf{x}_1}(\mathbf{x}), \ldots, \partial_l K_{\mathbf{x}_n} (\mathbf{x}))^T$. Moreover, $\|f\|_K^2 = \mathbf{c}^T \widetilde{\mathbf{M}} \mathbf{c}$ with $\widetilde{\mathbf{M}} = \langle \mathbf{M}_n(\cdot), \mathbf{M}_n(\cdot) \rangle_K$ and $\|g_l\|_n = ((1/n) \sum_{j=1}^n (\mathbf{c}^T \mathbf{U}_{\mathbf{x}_j}^l)^2)^{1/2}$ with $\mathbf{U}_{\mathbf{x}_j}^l = (\partial_l \mathbf{K}_{\mathbf{x}_j}^T, \partial_{1l} \mathbf{K}_n(\mathbf{x}_j)^T, \ldots, \partial_{pl} \mathbf{K}_n(\mathbf{x}_j)^T)^T$. Additional computational details for $\mathbf{M}_n(\mathbf{x}), \widetilde{\mathbf{M}}$, and $\mathbf{U}_{\mathbf{x}}^l$ are provided in Appendix A.1.

The optimization problem in (2.6) can be reformulated as

$$\operatorname*{argmin}_{\mathbf{c}} \frac{1}{n} \sum_{i=1}^n (1 - y_i \mathbf{c}^T \mathbf{M}_n(\mathbf{x}_i))_+ + \lambda_0 \mathbf{c}^T \widetilde{\mathbf{M}} \mathbf{c} + \lambda_1 \sum_{l=1}^p \pi_l \left(\frac{1}{n} \sum_{j=1}^n (\mathbf{c}^T \mathbf{U}_{\mathbf{x}_j}^l)^2 \right)^{1/2}, \tag{2.8}$$

where the last term can be regarded as a group lasso penalty (Yuan and Lin (2006)), pushing all or no elements of $\{\mathbf{c}^T \mathbf{U}_{\mathbf{x}_j}^l\}_{j=1}^n$ to be exactly zero to achieve sparsity in f.

Note that the proposed method requires that the true target function $f^* \in \mathcal{H}_K$ be induced by some predefined kernel function K, which can be set differently based on the prior information obtained about f^*. For example, if f^* is known to be linear in advance, we can set K as the linear kernel, and then (2.7) reduces to $f(\mathbf{x}) = \sum_{l=1}^p \varphi^l x^l$, with $\varphi^l = \sum_{i=1}^n (\alpha_i x_i^l + \beta_i^l)$. Direct calculation yields that the derivative-induced regularizer $\|g_l\|_n = ((1/n) \sum_{j=1}^n (\sum_{i=1}^n (\alpha_i x_i^l + \beta_i^l))^2)^{1/2} = |\varphi^l|$, reducing to the standard lasso penalty. Further computational details of the derivative-induced regularizer under other popular kernels are contained in Appendix A.1. If prior information is unavailable, we can set K as the Gaussian kernel. In this case, the resultant RKHS is fairly large, because the Gaussian kernel is known to be universal in the sense that any continuous function can be well approximated by some function in the induced RKHS under the infinity norm (Steinwart (2005)).

2.4. Computing algorithm

In this subsection, we develop an efficient algorithm for (2.8) using Nesterov's method (Nesterov (2005)). Essentially, we replace nonsmooth terms with smooth approximations, and then solve the approximated optimization problem using a gradient descent algorithm.

Clearly, the objective function in (2.8) has an equivalent form that

$$\max_{\mathbf{u}\in\mathcal{P}_1} \frac{1}{n}\sum_{i=1}^{n}(1-y_i\mathbf{c}^T\mathbf{M}_n(\mathbf{x}_i))u_i+\lambda_0\mathbf{c}^T\widetilde{\mathbf{M}}\,\mathbf{c}+\frac{1}{n^{1/2}}\max_{\mathbf{v}\in\mathcal{P}_2}\sum_{l=1}^{p}\pi_l(\mathbf{v}^l)^T\left(\mathbf{U}^l\right)^T\mathbf{c}, \tag{2.9}$$

where $\mathcal{P}_1=\{\mathbf{u}\in\mathcal{R}^n:0\le u_i\le 1\}$, $\mathcal{P}_2=\left\{\mathbf{v}^l:\|\,\mathbf{v}^l\,\|_2\le 1,\mathbf{v}^l\in\mathcal{R}^n\right\}$, and $\mathbf{U}^l=(\mathbf{U}^l_{\mathbf{x}_1},\ldots,\mathbf{U}^l_{\mathbf{x}_n})^T$. Note that the first and third terms of (2.9) can be approximated by

$$\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})\equiv\max_{\mathbf{u}\in\mathcal{P}_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(1-y_i\mathbf{c}^T\mathbf{M}_n(\mathbf{x}_i)\right)u_i-d_1^{\mu_1}(\mathbf{u})\right\}, \tag{2.10}$$

$$\widehat{\Omega}^*_{[p],\mu_2}(\mathbf{c})\equiv\max_{\mathbf{v}\in\mathcal{P}_2}\left\{\frac{1}{n^{1/2}}\sum_{l=1}^{p}\left(\pi_l(\mathbf{v}^l)^T\left(\mathbf{U}^l\right)^T\mathbf{c}-d_2^{\mu_2}(\mathbf{v}^l)\right)\right\}, \tag{2.11}$$

respectively, where $d_1^{\mu_1}(\mathbf{u})=(\mu_1/2)\|\,\mathbf{u}\,\|_2^2$ and $d_2^{\mu_2}(\mathbf{v}^l)=(\mu_2/2)\|\mathbf{v}^l\|_2^2$ are prox-functions, and μ_1 and μ_2 are some positive constants. Note that $\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})$ is convex and continuously differentiable, and can be regarded as a uniformly bounded smooth approximation of $\mathcal{E}_{\mathcal{Z}^n}(\mathbf{c})$, with $\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})\le\mathcal{E}_{\mathcal{Z}^n}(\mathbf{c})\le\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})+\mu_1$. A similar result holds for $\widehat{\Omega}^*_{\mathcal{Z}^n,\mu_2}(\mathbf{c})$. More importantly, because $d_1^{\mu_1}(\mathbf{u})$ and $d_2^{\mu_2}(\mathbf{v}^l)$ are strongly convex, the minimizers of (2.10) and (2.11) are unique and have explicit forms:

$$u_i^*=\text{median}\left(0,\frac{1-y_i\mathbf{c}^T\mathbf{M}_n(\mathbf{x}_i)}{n\mu_1},1\right)\text{ and}$$

$$\mathbf{v}_l^*=\frac{\pi_l\left(\mathbf{U}^l\right)^T\mathbf{c}}{n^{1/2}\mu_2\max\left(1,(\pi_l\|(\mathbf{U}^l)^T\,\mathbf{c}\,\|_2/\mu_2n^{1/2})\right)}.$$

Denote $\boldsymbol{\mu}=(\mu_1,\mu_2)^T$. Then, the objective function in (2.8) is approximated by the smoothed functional $F_{\boldsymbol{\mu}}(\mathbf{c})=\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})+\lambda_0R(\mathbf{c})+\lambda_1\widehat{\Omega}^*_{[p],\mu_2}(\mathbf{c})$, the derivative of which can be directly obtained, as follows:

$$\nabla F_{\boldsymbol{\mu}}=\nabla\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})+\lambda_0\nabla R(\mathbf{c})+\lambda_1\nabla\widehat{\Omega}^*_{[p],\mu_2}(\mathbf{c}),$$

where $\nabla\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}(\mathbf{c})=-(1/n)\sum_{i=1}^n y_i\mathbf{M}_n(\mathbf{x}_i)u_i^*$, $\nabla R(\mathbf{c})=2\widetilde{\mathbf{M}}\,\mathbf{c}$, and $\nabla\widehat{\Omega}^*_{[p],\mu_2}(\mathbf{c})=n^{-1/2}\sum_{l=1}^p\mathbf{U}^l\mathbf{v}_l^*$. Moreover, an upper bound of the gradient Lipschitz constant of $\nabla F_{\boldsymbol{\mu}}$ is given by

$$\mathcal{C}_{\boldsymbol{\mu}} = \mathcal{C}_{\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}} + \lambda_0 \mathcal{C}_R + \lambda_1 \mathcal{C}_{\widehat{\Omega}^*_{\mu_2}},$$

where $\mathcal{C}_{\mathcal{E}^*_{\mathcal{Z}^n,\mu_1}} = \max_{1\le i\le n}(\left\|\mathbf{M}_n(\mathbf{x}_i)\mathbf{M}_n(\mathbf{x}_i)^T\right\|_F / n\mu_1), \mathcal{C}_R = \|\widetilde{\mathbf{M}}\|_F, \mathcal{C}_{\widehat{\Omega}^*_{\mathcal{Z}^n,\mu_2}} = (1/n\mu_2)\| \sum_{l=1}^p \pi_l^2 \mathbf{U}^l(\mathbf{U}^l)^T\|_F$, and $\|\cdot\|_F$ denotes the Frobenius norm. Note that the gradient Lipschitz constant is the spectral norm of the matrix, which is upper bounded by its Frobenius norm. In the computing algorithm, we adopt the Frobenius norm to compute the step size of the proposed gradient descent algorithm, for computational simplicity.

We are now ready to use the gradient descent algorithm to optimize $F_{\boldsymbol{\mu}}(\mathbf{c})$. In the kth iteration, we consider the following update sequence:

$$\mathbf{a}^k = \mathbf{c}^k - \frac{\nabla F_{\boldsymbol{\mu}}(\mathbf{c}^k)}{\mathcal{C}_{\boldsymbol{\mu}}}, \quad \mathbf{b}^k = \mathbf{c}^0 - \sum_{t=1}^{k} \frac{t+1}{2\mathcal{C}_{\boldsymbol{\mu}}} \nabla F_{\boldsymbol{\mu}}(\mathbf{c}^t), \quad \mathbf{c}^{k+1} = \frac{2\mathbf{b}^k + (k+1)\mathbf{a}^k}{k+3},$$

where $\mathbf{a}^k$ and $\mathbf{b}^k$ are the solutions to the following auxiliary optimization problems:

$$\min_{\mathbf{a}\in\mathcal{R}^{n(p+1)}} \langle \mathbf{a} - \mathbf{c}^k, \nabla F_{\boldsymbol{\mu}}(\mathbf{c}^k)\rangle + \frac{\mathcal{C}_{\boldsymbol{\mu}}}{2} \|\, \mathbf{a} - \mathbf{c}^k\|_2^2,$$

$$\min_{\mathbf{b}\in\mathcal{R}^{n(p+1)}} \frac{\mathcal{C}_{\boldsymbol{\mu}}}{2} \|\, \mathbf{b} - \mathbf{c}^0\|_2^2 + \sum_{t=1}^{k} \frac{t+1}{2} \left(F_{\boldsymbol{\mu}}(\mathbf{c}^t) + (\mathbf{b} - \mathbf{c}^t)^T \nabla F_{\boldsymbol{\mu}}(\mathbf{c}^k) \right).$$

The proposed algorithm is summarized as follows. As a computational re-

Algorithm 1: Computing algorithm.

```
given: parameters $\lambda_0$, $\lambda_1$, $\mu_1$, $\mu_2$ and $\pi_l$; $l = 1, 2, \ldots, p$.
initialize: $\mathbf{c}^1 = \mathbf{0}$, $\mathbf{d}^1 = \mathbf{0}$, $k = 1$.
for $k \ge 1$, repeat
    $\mathbf{a}^{k+1} = \mathbf{c}^k - \frac{\nabla F_{\boldsymbol{\mu}}(\mathbf{c}^k)}{\mathcal{C}_{\boldsymbol{\mu}}}$,
    $\mathbf{d}^{k+1} = \mathbf{d}^k + (k+1)\nabla F_{\boldsymbol{\mu}}(\mathbf{c}^k)$,
    $\mathbf{b}^{k+1} = \mathbf{c}^0 - \frac{\mathbf{d}^{k+1}}{2\mathcal{C}_{\boldsymbol{\mu}}}$,
    $\mathbf{c}^{k+1} = \frac{2\mathbf{b}^{k+1} + (k+1)\mathbf{a}^{k+1}}{k+3}$.
until $\mathbf{c}^k$ converges.
```

mark, at the kth iteration, the combination of $\mathbf{a}^k$ and $\mathbf{b}^k$ is used to adjust the descent. Thus, the algorithm has a convergence rate $O(1/k^2)$, which can be derived based on Theorem 2 in Nesterov (2005).

3. Statistical Properties

In this section, the asymptotic results of the estimation and subset selection of our proposed method are established under some regularity assumptions. To ensure the uniqueness of f^*, we further define $f^* = \operatorname{argmin}_{f\in\mathcal{B}} \|f\|_K^2$, with $\mathcal{B} = \{f \in \mathcal{H}_K : f = \operatorname{argmin}_{f\in\mathcal{H}_K} \mathcal{E}(f)\}$. Note that $\mathcal{A}^*$ can be rewritten as $\mathcal{A}^* = \{l, \|g_l^*\|_{\rho_{\mathbf{x}}}^2 > 0\}$. Without loss of generality, we assume the first p_0 variables of $\mathbf{x}$ are truly informative, and thus the cardinality of $\mathcal{A}^*$ is $|\mathcal{A}^*| = p_0 < p$. The following technical assumption is made.

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

Assumption 1 assumes the boundedness of the kernel functions, which is commonly used in the machine learning literature (Rosasco et al. (2013); Yang, Lv and Wang (2016)). Such a condition can be verified for many kernel functions, including the Gaussian kernel, if they satisfy certain smoothness conditions.

Denote the function space $\mathcal{F}_M = \big\{f \in \mathcal{H}_K : \|f - f^*\|_K \le M\big\}$, with some $M > 0$. Now, we establish the estimation consistency of the proposed method.

Theorem 1. *Suppose Assumption* 1 *is met and* $\widehat{f} \in \mathcal{F}_M$. *For some* $\delta_n \in (0,1)$, *with probability at least* $1 - \delta_n$, *we have*

$$\begin{aligned}\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \le & 2\sqrt{2}\big((M + \|f^*\|_K)\sqrt{\kappa_1} + \kappa_1 M\big)\sqrt{\frac{\log(1/\delta_n)}{n}} + 69\kappa_1 \frac{M\log(1/\delta_n)}{2n} \\ & + 2\lambda_0 M\|f^*\|_K + 2\lambda_1\kappa_2 M \sum_{l\in\mathcal{A}^*} \pi_l.\end{aligned}$$

Theorem 1 illustrates how the excess risk of the proposed method explicitly depends on the tuning parameters $(\lambda_0, \lambda_1, \pi_l)$ and the sample size n. The assumption $\widehat{f} \in \mathcal{F}_M$ is fairly weak. Thus, we consider two special cases in the corollaries below, where M is either a constant, or is obtained from the regularized approach in (2.6).

Corollary 1. *Suppose the assumptions of Theorem* 1 *and case (i) are satisfied. Let* $\lambda_0 = n^{-1/2}$ *and* $\lambda_1\pi_l = n^{-1/2}$, *for any* $\delta_n > 0$, *with probability at least* $1 - \delta_n$;

then, there holds

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \le C_0 n^{-1/2}\Big(\sqrt{\log\frac{1}{\delta_n}} + |\mathcal{A}^*|\Big),$$

where $C_0 = 2\sqrt{2\kappa_1}\|f^*\|_K + 2M\big(\sqrt{2\kappa_1} + \kappa_1\sqrt{2} + 35/2\kappa_1 + \|f^*\|_K + \kappa_2\big)$.

We omit the proof of Corollary 1 here because it can be obtained directly from Theorem 1 under case (i). The boundedness assumption on hypothesis spaces is standard in the statistical literature; see Tarigan and Van De Geer (2006) for high-dimensional SVMs and Horowitz and Mammen (2007) for generalized additive models. Compared with the variable selection approaches based on a local derivative estimation (Yang, Lv and Wang (2016)), our convergence rate in Corollary 1 does not depend on the dimension of the covariates p. This indicates that our proposed method is more robust to the ambient dimension.

We now consider a relaxation of Corollary 1, with essentially weaker conditions. When we do not specify M, we take $M = {\lambda_0}^{-1/2} + \|f^*\|_K$ from our proposed method in (2.6), by the fact that $\mathcal{E}_{\mathcal{Z}_n}(\hat{f}) + \lambda_0\|\hat{f}\|_K^2 + \lambda_1\widehat{\Omega}_{[p]}(\hat{f}) \le \mathcal{E}_{\mathcal{Z}_n}(0) + \lambda_0\|0\|_K^2 + \lambda_1\widehat{\Omega}_{[p]}(0) \le 1$. In this case, the following corollary can be obtained easily from Theorem 1.

Corollary 2. *Suppose the assumptions of Theorem* 1 *and case (ii) are satisfied. Let* $\lambda_0 = n^{-1/2}$, $\lambda_1\pi_l = n^{-1/2}$*; then, for any* $\delta_n > 0$*, with probability at least* $1 - \delta_n$*, there holds*

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) = O_p\left(n^{-1/4}\left(\left(\log\frac{1}{\delta_n}\right)^{1/2} + |\mathcal{A}^*|\right)\right).$$

Although the convergence rate in Corollary 2 is not as tight as some existing results for the classical SVM, it is established under much weaker conditions than those of most existing methods. For example, the method of Tarigan and Van De Geer (2006) requires an extra margin condition to achieve a tighter rate. In fact, a tighter convergence rate can be obtained in Corollary 1 by assuming M is a constant.

Furthermore, note that Theorem 1 only provides a type of weak convergence in estimating f^* and, thus, is not sufficient to imply selection consistency, which requires an error bound of $\|\hat{f} - f^*\|_K$. For this purpose, two additional assumptions are made for our analysis. Let $\widetilde{f}$ be the minimizer of $\mathcal{E}(f) + \lambda_0\|f\|_K^2 + \lambda_1\Omega_{[p]}(f)$ over $\mathcal{H}_K$ as an intermediate function.

Assumption 2. *There exist some positive constants* $\theta_1, \theta_2, C_1, C_2, C_3,$ *and* C_4, *such that* $\|\widetilde{f}-f^*\|_K = C_1 \max\{\lambda_0, \lambda_1\}^{\theta_1}$, $C_2 \leq \max_{l \in \mathcal{A}^*} \pi_l \leq C_3$, *and* $\min_{l \notin \mathcal{A}^*} \pi_l = C_4 n^{\theta_2}$.

Assumption 3. *There exist some positive constants* C_5 *and* $\xi > 1/2$, *such that* $\min_{l \in \mathcal{A}^*} \|g_l^*\|_{\rho_{\mathbf{x}}} > C_5 \max\{p^{1/2} n^{-3/32+\theta_2/2}, n^{-\theta_1/8}\}\big(\log(6p/\delta_n)\big)^{\xi}$.

Assumption 2 controls the approximation error of the proposed method, which has been well studied in Cuker and Zhou (2007). Note that a similar assumption on the approximation error is imposed in Rosasco et al. (2013). Assumption 2 also characterizes the property of the adaptive tuning weight π_l, which is designed to converge to some constant for informative variables, and to diverge to infinity for noninformative variables. As suggested by Zou (2006), an initial estimator $\widetilde{g}_l$ can be obtained by setting $\lambda_0 = \lambda_1 = 0$ in (5). Then we choose $\pi_l = \|\widetilde{g}_l\|_n^{-\gamma}$, with γ as some positive constant satisfying Assumption 2. Assumption 2 can be verified in a similar manner to the proof of Lemma 1. Assumption 3 requires that the true derivative function contain sufficient information about the truly informative variables. Now, we establish the asymptotic sparsistency of the proposed method.

Lemma 1. *Suppose Assumptions* 1–2 *are met. There exists some constant* C_6, *such that, with probability at least* $1 - \delta_n$, *there holds*

$$\begin{aligned}
&\max_{l=1,\ldots,p} \big| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_{\rho_{\mathbf{x}}}^2 \big| \\
&\leq C_6 \left(M^{1/2} \left(\frac{1}{n^{1/2}\lambda_1} + \frac{\sum_{l=1}^{p} \pi_l}{n^{1/4}} \right)^{1/2} + \max\{\lambda_0, \lambda_1\}^{\theta_1} \right) \left(\log \frac{6p}{\delta_n} \right)^{1/2}.
\end{aligned}$$

Additionally, if case (ii) is satisfied by taking $M = \lambda_0^{-1/2} + \|f^*\|_K$ *and letting* $\lambda_0 = n^{-1/8}$ *and* $\lambda_1 = n^{-1/4}$, *then there exists some constant* C_7, *such that, with probability at least* $1 - \delta_n$, *there holds*

$$\max_{l=1,\ldots,p} \big| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_{\rho_{\mathbf{x}}}^2 \big| \leq C_7 \max\{p^{1/2} n^{-3/32+\theta_2/2}, n^{-\theta_1/8}\} \left(\log \frac{6p}{\delta_n} \right)^{1/2}.$$

Lemma 1 establishes the convergence relationship between $\|\widehat{g}_l\|_n^2$ and $\|g_l^*\|_{\rho_{\mathbf{x}}}^2$, which is crucial to establishing the asymptotic subset selection in Theorem 2.

Theorem 2. *Suppose the assumptions of Lemma* 1 *and Assumption* 3 *are met. Then, we conclude that*

$$P\left(\mathcal{A}^* \subset \widehat{\mathcal{A}}\right) \to 1.$$

Theorem 2 shows that the selected variables contain all of the truly informative variables asymptotically, which can be regarded as a safe filter. The guarantee in Theorem 2 is the same as the SURE screening property, which has been studied extensively in the literature (Fan and Song (2010); Li, Zhong and Zhu (2012); Wang et al. (2015)). However, it remains open to the bound $P\big(\mathcal{A}^* \supset \widehat{\mathcal{A}}\big)$ in order to guarantee variable selection consistency. Recall that a standard technique for proving variable selection consistency is the so-called construction approach. This approach first constructs a local estimator $\hat{\theta}_{\mathcal{A}^*}$ based on the proposed estimation restricted to $\mathcal{A}^*$, and then shows that $(\hat{\theta}_{\mathcal{A}^*}, \mathbf{0})$ is a feasible solution to the proposed estimation. However, in our infinite-dimensional case, $\hat{\theta}_{\mathcal{A}^*}$ may not be within the specific RKHS, unlike any finite-dimensional vectors. In other words, our proposed method can be viewed as an approximation strategy for variable selection beyond those offered by additive models. Note that $(\hat{\theta}_{\mathcal{A}^*}, \mathbf{0})$ is not well defined, which may cause difficulty. In addition, existing penalty terms associated with sparsity, such as the group lasso and multiple kernel learning (Bach (2008)), are based on the fact that the penalties are induced by orthogonal projection operators. Unfortunately, our regularizer is not associated with such operators. Addressing these challenge is beyond the scope of this study and, thus, is left to future research.

4. Numerical Examples

In this section, we compare the numerical performance of the proposed method against that of several existing methods, including the sparse additive machine (Zhao and Liu (2012)), the SURE independence screening (Fan and Song (2010)), and conditional distance correlation (Wang et al. (2015)), denoted as MF, SAM, SIS, and CDC, respectively. The corresponding R packages "SAM", "SIS" and "cdcsis" are used for the three existing methods. For MF, we set $\mu_1 = \mu_2 = 0.2$, and for CDC, we set 0.05 as the threshold value to identify the informative variables. Note that Zhao and Liu (2012) conduct a numerical comparison, showing that SAM outperforms several existing methods, including the Cosso-SVM (Zhang (2006)). In all the simulated examples presented here, no prior knowledge about the true target function is assumed; thus the Gaussian kernel $K(\mathbf{s}, \mathbf{t}) = e^{-\|\mathbf{s} - \mathbf{t}\|^2/(2\sigma_n^2)}$ is used to induce the RKHS, where σ_n is set as the median of all pairwise distances within the training sample (Jaakkola, Diekhans and Haussler (1999)). Because the choice of tuning parameters highly affects the performance of the methods, they are selected using the variable selection stability criterion (Sun, Wang and Fang (2013)). This criterion measures the stability of variable selection by randomly splitting the training sample into two parts, and then comparing the disagreement between the two selected variable sets. The maximization of the stability criterion is conducted via a grid search, where the grid is set as $\{10^{-2+0.1s} : s = 0, \dots, 40\}$.

4.1. Simulated examples

Three simulated examples are examined. We first generate $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, with $x_{ij} = (W_{ij} + \eta U_i)/(1+\eta)$, where W_{ij} and U_i are drawn independently from $U(0,1)$. The following cases are considered:

- *Example 1:* $f^*(\mathbf{x}) = 2/\cos(2\pi x_1 x_2) - 1$;
- *Example 2:* $f^*(\mathbf{x}) = 6x_1 - \cos(\pi x_1) + 2x_2 + 8x_2^2 + 6\sin(\pi(x_3 - x_4)) - 8$;
- *Example 3:* $f^*(\mathbf{x}) = 20x_1x_2x_3 + 4x_4^2 + 4x_5 - 5$.

Then, we generate $y \sim Bernoulli(1/(1+e^{-f^*(\mathbf{x})}))$. Clearly, the first two variables are truly informative in the first example, with an interaction effect. In the second example, an additive model structure with an interaction term is examined, and the first four variables are truly informative. In the third example, a three-way interaction term is considered, and the first five variables are truly informative.

For each example, we consider scenarios with $(n, p) = (200, 10), (200, 20)$, and $(300, 40)$. In each scenario, $\eta = 0$ and $\eta = 0.1$. When $\eta = 0$, the data are generated to be completely independent, whereas when $\eta = 0.1$, a correlation structure is added between the variables. Each scenario is replicated 50 times, and the averaged performance measures are summarized in Tables 1–3. Specifically, Size represents the averaged number of selected informative variables, TP represents the number of truly informative variables selected, FP represents the number of truly noninformative variables selected, and C, U, and O are the numbers of times correct-fitting, under-fitting, and over-fitting occurred, respectively. The AUCs of all methods are also reported for comparison. For each replication, we refit a standard SVM with the selected variables identified by each method, and then compute the AUC of the fitted model using a testing set with $10,000$ records, which are generated independently and identically as the training set.

Clearly, MF outperforms the other methods in most scenarios. In Example 1, MF is able to identify the two truly informative variables in almost every replication, which have an interaction effect in the denominator of $f^*(\mathbf{x})$. In Example 2, where $f^*(\mathbf{x})$ is additive with one interaction term involved, MF is

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

(n,p,η)	Method	Size	TP	FP	C	U	O	AUC
(200,10,0)	MF	1.92	1.90	0.02	44	5	1	0.8416
	SAM	3.02	2.00	1.02	25	0	25	0.8353
	SIS	2.18	1.98	0.30	41	1	8	0.8510
	CDC	2.04	1.98	0.06	46	0	4	0.8543
(200,20,0)	MF	2.10	1.94	0.16	42	3	5	0.8447
	SAM	3.72	2.00	1.72	13	0	37	0.8272
	SIS	2.28	1.98	0.30	36	1	13	0.8477
	CDC	4.42	2.00	2.22	2	0	48	0.8080
(300,40,0)	MF	2.24	2.00	0.24	42	0	8	0.8619
	SAM	4.20	2.00	2.20	8	0	42	0.8227
	SIS	2.36	2.00	0.36	38	0	12	0.8596
	CDC	20.82	2.00	18.82	0	0	50	0.7531
(200,10,0.1)	MF	2.02	2.00	0.02	49	0	1	0.8807
	SAM	3.00	2.00	1.00	16	0	34	0.8590
	SIS	2.16	2.00	0.16	43	0	7	0.8764
	CDC	2.06	2.00	0.06	47	0	3	0.8784
(200,20,0.1)	MF	2.06	1.96	0.10	44	2	4	0.8702
	SAM	3.60	2.00	1.60	14	0	36	0.8495
	SIS	2.24	2.00	0.24	43	0	7	0.8757
	CDC	4.44	2.00	2.44	3	0	47	0.8387
(300,40,0.1)	MF	2.04	2.00	0.04	48	0	2	0.8842
	SAM	4.06	2.00	2.06	6	0	44	0.8469
	SIS	2.16	2.00	0.16	42	0	8	0.8826
	CDC	22.28	2.00	20.28	0	0	50	0.7862

able to identify all truly informative variables with high accuracy. In Example 3, MF correctly identifies the truly informative variables in a three-way interaction effect, and identifies square effect and linear effect variables included in most replications. However, SAM tends to overfit, and thus wrongly includes some noninformative variables in most scenarios. SIS selects fewer redundant variables than SAM does, but still tends to include some noninformative variables. CDC performs well when (n,p) is small, but the performance worsens as (n,p) increases. These conclusions are supported by the AUC comparison. Furthermore, when the correlation structure with $\eta = 0.1$ is considered, identifying the truly informative variables becomes more difficult; nevertheless, MF still outperforms its competitors in most scenarios.

As computational remarks, to obtain the adaptive weight π_l, we estimate $\widetilde{g}_l$ by setting $\lambda_0 = \lambda_1 = 0$ in Algorithm 1, and take $\gamma = 1$ for computational sim-

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

(n, p, η)	Method	Size	TP	FP	C	U	O	AUC
(200,10,0)	MF	4.02	4.00	0.02	49	0	1	0.9604
	SAM	5.06	4.00	1.06	19	0	31	0.9527
	SIS	4.22	4.00	0.22	42	0	8	0.9583
	CDC	3.96	3.92	0.04	44	4	2	0.9558
(200,20,0)	MF	4.00	3.96	0.04	46	2	2	0.9578
	SAM	5.92	4.00	1.92	10	0	40	0.9468
	SIS	4.66	4.00	0.66	37	0	13	0.9556
	CDC	5.84	3.98	1.86	6	0	44	0.9474
(300,40,0)	MF	4.02	4.00	0.02	49	0	1	0.9667
	SAM	6.72	4.00	2.72	7	0	43	0.9484
	SIS	4.54	4.00	0.54	40	0	10	0.9628
	CDC	21.54	4.00	17.54	0	0	50	0.9220
(200,10,0.1)	MF	3.90	3.90	0.00	47	3	0	0.9561
	SAM	5.04	4.00	1.04	20	0	30	0.9568
	SIS	4.20	4.00	0.20	42	0	8	0.9612
	CDC	3.88	3.84	0.04	40	8	2	0.9540
(200,20,0.1)	MF	3.98	3.94	0.04	46	3	1	0.9612
	SAM	5.56	3.98	1.58	11	1	38	0.9535
	SIS	4.22	4.00	0.22	44	0	6	0.9625
	CDC	6.36	3.98	2.38	3	0	47	0.9519
(300,40,0.1)	MF	4.00	3.96	0.04	46	2	2	0.9674
	SAM	5.78	4.00	1.78	11	0	39	0.9587
	SIS	4.82	4.00	0.82	36	0	14	0.9652
	CDC	23.52	4.00	19.52	0	0	50	0.9302

plicity in all numerical examples. Furthermore, owing to Nesterovs's smoothing approximation on the regularizer term, the estimated $\|\widehat{g}_l\|_n$ for those noninformative variables is not exactly zero. Therefore, we truncate the estimated $\|\widehat{g}_l\|_n$ as $(\|\widehat{g}_l\|_n - \lambda_1)_+$ in all simulations.

4.2. Real–data analysis

The proposed method is applied to three real-data examples: the BUPA Liver Disorder (BUPA) data set , the Indian Liver Patient (ILP) data set and the Wisconsin breast cancer (WBC) data set. The BUPA data set contains 345 observations, with 145 liver-disorder and 200 liver-normal. The ILP data set collects 416 liver patient records and 167 non-liver patient records with 10 features. The WBC data set consists of 569 cases with two diagnoses, 212 malignant and 357 benign. The variables in each data set are standardized and scaled to $[0, 1]$.

Table 3. The averaged performance measures of various variable selection methods in Example 3.

(n, p, η)	Method	Size	TP	FP	C	U	O	AUC
(200,10,0)	MF	5.06	5.00	0.06	47	0	3	0.9327
	SAM	6.58	4.96	1.62	8	1	41	0.9175
	SIS	5.34	5.00	0.34	41	0	9	0.9302
	CDC	4.66	4.62	0.04	32	17	1	0.9085
(200,20,0)	MF	5.16	4.94	0.22	37	3	10	0.9264
	SAM	8.12	4.96	3.16	6	1	43	0.9054
	SIS	5.44	5.00	0.44	36	0	14	0.9287
	CDC	6.94	4.96	1.98	8	2	40	0.9143
(300,40,0)	MF	5.14	4.94	0.20	39	3	8	0.9347
	SAM	8.96	5.00	3.96	6	0	44	0.9152
	SIS	5.20	5.00	0.20	43	0	7	0.9385
	CDC	22.54	5.00	17.00	0	0	50	0.8794
(200,10,0.1)	MF	5.16	4.94	0.22	37	3	10	0.9237
	SAM	6.62	5.00	1.62	8	0	42	0.9176
	SIS	5.18	4.98	0.20	41	0	9	0.9258
	CDC	4.62	4.58	0.04	30	20	0	0.9030
(200,20,0.1)	MF	5.04	4.84	0.20	36	5	9	0.9161
	SAM	8.32	5.00	3.32	5	0	45	0.9039
	SIS	5.46	4.98	0.48	34	1	15	0.9216
	CDC	7.44	4.94	2.50	4	3	43	0.9058
(300,40,0.1)	MF	5.06	4.94	0.12	41	3	6	0.9329
	SAM	9.18	5.00	4.18	2	0	48	0.9096
	SIS	5.32	5.00	0.32	45	0	5	0.9349
	CDC	24.20	5.00	19.20	0	0	50	0.8754

Then, we apply the MF, SAM, and SIS methods to select the informative variables. Note that owing to the computational burden of CDC, we do not report its performance for these data sets. In addition, we add a standard kernel SVM, denoted as KSVM, to report the prediction performance if all variables are included. When the informative variables have been selected, we randomly split each data set, with 200 observations for testing and the remainder for training, and refit a standard kernel SVM. The splitting is replicated 1,000 times. The averaged prediction results are summarized in Table 4.

As Table 4 shows, MF identifies three informative variables in the ILP data set, three informative variables in the BUPA data set, and 16 informative variables in the WBC data set. The number of informative variables selected by SAM and SIS for the data sets are 8, 6, 7 and 6, 6, 10, respectively. However, the averaged testing errors based on the selected sets of MF are smaller than those of

Table 4. The selected variables as well as the corresponding averaged prediction errors by various selection methods in the dataset.

Dataset	Method	Number selected	Testing error (Std)
BUPA	MF	3	0.3196 (0.0001)
	SAM	6	0.3391 (0.0001)
	SIS	6	0.3391 (0.0001)
	KSVM	6	0.3391 (0.0001)
ILP	MF	3	0.2840 (0.0008)
	SAM	8	0.2863 (0.0008)
	SIS	6	0.2891 (0.0008)
	KSVM	9	0.2875 (0.0008)
WBC	MF	16	0.0257 (0.0003)
	SAM	7	0.0307 (0.0003)
	SIS	10	0.0287 (0.0003)
	KSVM	30	0.0278 (0.0003)

SAM, SIS, and KSVM in all three examples, suggesting that SAM and SIS may wrongly identify some variables, and that KSVM may include some noninformative variables, both of which are detrimental to their prediction performance.

5. Conclusion

In this study, we focus on the variable selection problem for nonparametric classification beyond additive models, based on a general criterion that defines the truly informative variables for classification using the optimal Bayes rule. Furthermore, we propose an infinite-dimensional regularized framework within a smooth RKHS, where the sparsity is naturally induced by our derivative-based penalty. In contrast to many existing variable selection methods for nonparametric classification, our proposed method is a global estimation within an RKHS, and can be solved using a finite-dimensional convex optimization. Thus, it is much more robust to the ambient dimension. More importantly, the proposed estimation enjoys nice statistical properties, such as estimation consistency and the subset selection property, without imposing an explicit model assumption. As a result, the proposed method takes full advantage of the mathematical properties of the RKHS, at the cost of increased computational complexity. In future research, we will improve the computational efficiency of the proposed method.

Acknowledgements

Xin He's research was supported in part by NSFC-11901375. Junhui Wang's research was supported in part by HK RGC Grants GRF-11331016, GRF-11303918, and GRF-11300919. Shaogao Lv is the corresponding author, and his research was partially supported by NSFC-11871277. The authors also thank the associate editor and two anonymous referees for their constructive suggestions.

A. Appendix

A.1. Computational details

The RKHS $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_{\mathbf{x}} : K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ with the inner product $\langle K_{\mathbf{x}}, K_{\mathbf{u}} \rangle_K = K(\mathbf{x}, \mathbf{u})$. Then the reproducing property takes the form $\langle f, K_{\mathbf{x}} \rangle_K = f(\mathbf{x})$, for any $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}_K$. From Theorem 1 of Zhou (2007), the partial derivative reproducing property holds true that

$$g_l(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x^l} = \langle f, \partial_l K_{\mathbf{x}} \rangle_K \,,$$

where $\partial_l K_{\mathbf{x}}(\cdot) = (\partial K(\mathbf{s}, \cdot)/\partial s^l)|_{\mathbf{s}=\mathbf{x}}$. It implies that

$$\begin{aligned} \langle \partial_l K_{\mathbf{x}}, \partial_{l'} K_{\mathbf{x}'} \rangle_K &= \partial_l \left(\partial_{l'} K_{\mathbf{x}'} \right) (\mathbf{x}) = \frac{\partial K(\mathbf{s}, \mathbf{t})}{\partial s^l \partial t^{l'}} \Big|_{\mathbf{s}=\mathbf{x}, \mathbf{t}=\mathbf{x}'}, \\ \langle \partial_l K_{\mathbf{s}}, K_{\mathbf{x}} \rangle_K &= \frac{\partial K(\mathbf{t}, \mathbf{x})}{\partial \mathbf{t}^l} \Big|_{\mathbf{t}=\mathbf{s}} = \partial_l K_{\mathbf{s}}(\mathbf{x}). \end{aligned}$$

Following these properties, directly calculation yields that

$$\begin{aligned} \|f\|_K^2 &= \mathbf{c}^T \left\langle \mathbf{M}_n(\cdot), \mathbf{M}_n(\cdot) \right\rangle_K \mathbf{c} = \mathbf{c}^T \begin{bmatrix} \mathbf{K}_{n\times n} & \mathbf{DK}_{n\times np} \\ \mathbf{DK}^T_{np\times n} & \mathbf{D^2K}_{np\times np} \end{bmatrix} \mathbf{c} = \mathbf{c}^T \widetilde{\mathbf{M}} \, \mathbf{c}, \\ \|g_l\|_n &= \left(\frac{1}{n} \sum_{j=1}^{n} \left(\mathbf{c}^T \mathbf{U}^l_{\mathbf{x}_j} \right)^2 \right)^{1/2}, \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix, $\mathbf{DK} = (\partial_1 \mathbf{K}, \ldots, \partial_p \mathbf{K})$, $\mathbf{U}^l_{\mathbf{x}_j} = (\partial_l \mathbf{K}_{\mathbf{x}_j}, \partial_{1l} \, \mathbf{K}_n(\mathbf{x}_j), \ldots, \partial_{pl} \, \mathbf{K}_n(\mathbf{x}_j))$ and $\mathbf{D^2K}= (\partial_{ll'} \, \mathbf{K})^p_{l,l'=1}$ with $\partial_l \mathbf{K} = ((\partial K(\mathbf{s}, \mathbf{x}_j)/\partial s^l)|_{\mathbf{s}=\mathbf{x}_i})^n_{i,j=1}$, $\partial_{ll'} \mathbf{K} = ((\partial K(\mathbf{s}, \mathbf{t})/\partial s^l \partial t^{l'})|_{\mathbf{s}=\mathbf{x}_i, \mathbf{t}=\mathbf{x}_j})^n_{i,j=1}$ and $\partial_l \mathbf{K}_{\mathbf{x}_j} = (\partial_l K_{\mathbf{x}_j}(\mathbf{x}_1), \ldots, \partial_l K_{\mathbf{x}_j} (\mathbf{x}_n))^T$.

If the kernel is set to be the linear kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, then we have

$$\begin{aligned}
&\partial_l K_{\mathbf{x}_i}(\mathbf{x}_j) = x_j^l, \\
&\frac{\partial K(\mathbf{s}, \mathbf{t})}{\partial s^l \partial t^k}\Big|_{\mathbf{s}=\mathbf{x}_i, \mathbf{t}=\mathbf{x}_j} = \begin{cases} 0, \text{ if } l \neq k; \\ 1, \text{ if } l = k. \end{cases}
\end{aligned}$$

If the kernel is set to be the polynomial kernel with degree d, $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^d$, then we have

$$\begin{aligned}
&\partial_l K_{\mathbf{x}_i}(\mathbf{x}_j) = d(1 + \mathbf{x}_i^T \mathbf{x}_j)^{d-1} x_j^l, \\
&\frac{\partial K(\mathbf{s}, \mathbf{t})}{\partial s^l \partial t^k}\Big|_{\mathbf{s}=\mathbf{x}_i, \mathbf{t}=\mathbf{x}_j} = \begin{cases} d(d-1)(1 + \mathbf{x}_i^T \mathbf{x}_j)^{d-2} x_i^k x_j^l, & \text{if } l \neq k; \\ d(1 + \mathbf{x}_i^T \mathbf{x}_j)^{d-2}\Big((d-1) x_i^k x_j^l + 1 + \mathbf{x}_i^T \mathbf{x}_j\Big), & \text{if } l = k. \end{cases}
\end{aligned}$$

If the kernel is set to be the Gaussian kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma_n^2)}$, then we have

$$\begin{aligned}
&\partial_l K_{\mathbf{x}_i}(\mathbf{x}_j) = -e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma_n^2)} \left(\frac{\mathbf{x}_i^l - \mathbf{x}_j^l}{\sigma_n^2}\right), \\
&\frac{\partial K(\mathbf{s}, \mathbf{t})}{\partial s^l \partial t^k}\Big|_{\mathbf{s}=\mathbf{x}_i, \mathbf{t}=\mathbf{x}_j} = \begin{cases} -e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma_n^2)} \left(\frac{\mathbf{x}_i^l - \mathbf{x}_j^l}{\sigma_n^2}\right)\left(\frac{\mathbf{x}_i^k - \mathbf{x}_j^k}{\sigma_n^2}\right), \text{ if } l \neq k; \\ -e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma_n^2)} \left(\left(\frac{\mathbf{x}_i^l - \mathbf{x}_j^l}{\sigma_n^2}\right)^2 - \frac{1}{\sigma_n^2}\right), \text{ if } l = k. \end{cases}
\end{aligned}$$

A.2. Technical proofs

Proof of Theorem 1. Using the short-hand notation ,$\lambda_1 \widehat{\Omega}_{\mathcal{S}}(f) = \lambda_1 \sum_{l \in \mathcal{S}} \pi_l \|g_l\|_n$ for any $\mathcal{S} \subseteq [p]$, we denote our proposed scheme (2.6) by $\mathcal{E}_{\mathcal{Z}_n}(f) + \lambda_0 \|f\|_K^2 + \lambda_1 \widehat{\Omega}_{[p]}(f)$, whose minimizer exists uniquely based on the strict convex optimization and is denoted by $\hat{f}$. The definition of minimization problem (2.6) directly implies that

$$\mathcal{E}_{\mathcal{Z}_n}(\widehat{f}) + \lambda_0 \|\widehat{f}\|_K^2 + \lambda_1 \widehat{\Omega}_{[p]}(\widehat{f}) \leq \mathcal{E}_{\mathcal{Z}_n}(f^*) + \lambda_0 \|f^*\|_K^2 + \lambda_1 \widehat{\Omega}_{[p]}(f^*),$$

which is equivalent to the following one by sparse assumption on $\partial_l f^*$ that

$$\mathcal{E}_{\mathcal{Z}_n}(\widehat{f}) + \lambda_0 \|\widehat{f} - f^*\|_K^2 + \lambda_1 \widehat{\Omega}_{[p]}(\widehat{f}) \leq \mathcal{E}_{\mathcal{Z}_n}(f^*) - 2\lambda_0 \langle \widehat{f} - f^*, f^* \rangle_K + \lambda_1 \widehat{\Omega}_{\mathcal{A}^*}(f^*).$$

Adding $\mathcal{E}(\hat{f}) - \mathcal{E}(f^*)$ to both sides of the last inequality, we obtain

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) + \lambda_0 \|\widehat{f} - f^*\|_K^2 + \lambda_1 \widehat{\Omega}_{[p]}(\widehat{f}) \leq \big[\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) - (\mathcal{E}_{\mathcal{Z}_n}(\widehat{f}) - \mathcal{E}_{\mathcal{Z}_n}(f^*))\big]$$

$$- 2\lambda_0\langle \widehat{f} - f^*, f^*\rangle_K + \lambda_1 \widehat{\Omega}_{\mathcal{A}^*}(f^*). \tag{A.1}$$

On the other hand, we have $\lambda_1\widehat{\Omega}_{(\mathcal{A}^*)^c}(\hat{f}) = \lambda_1\widehat{\Omega}_{(\mathcal{A}^*)^c}(\hat{f} - f^*)$ by the fact that $g_l^*(\mathbf{x}) = 0$, for any $l \in (\mathcal{A}^*)^c$. By the triangle inequality, we obtain from (A.1) that

$$\begin{aligned} &\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) + \lambda_0\|\hat{f} - f^*\|_K^2 + \lambda_1\widehat{\Omega}_{[p]}(\widehat{f} - f^*) \\ &\leq \left[\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) - (\mathcal{E}_{\mathcal{Z}_n}(\widehat{f}) - \mathcal{E}_{\mathcal{Z}_n}(f^*))\right] - 2\lambda_0\langle \widehat{f} - f^*, f^*\rangle_K + 2\lambda_1\widehat{\Omega}_{\mathcal{A}^*}(\widehat{f} - f^*). \end{aligned} \tag{A.2}$$

To bound the first term in (A.2), we need to use empirical process theory and particularly concentration inequalities, since $\hat{f}$ is a random function within $\mathcal{H}_K$ without an explicit expression. The proof of the following proposition is delegated to Appendix A.3.

Proposition 1. *Suppose Assumption* 1 *is satisfied, and the minimizer* $\widehat{f}$ *of* (2.6) *is computed over the bounded ball* $\mathcal{F}_M$. *Then with probability at least* $1 - \delta_n$, *we have*

$$\begin{aligned} &\left|\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) - (\mathcal{E}_{\mathcal{Z}_n}(\widehat{f}) - \mathcal{E}_{\mathcal{Z}_n}(f^*))\right| \\ &\leq 2(M + \|f^*\|_K)\sqrt{\frac{2\kappa_1}{n}} + \kappa_1 M\sqrt{\frac{8\log(1/\delta_n)}{n}} + 69\kappa_1\frac{M\log(1/\delta_n)}{2n}. \end{aligned}$$

Note that, the result of Proposition 1 does not rely on the smoothness of RKHS, which appears to be a rough quantity but it suffices to analyze our excess risk for classification problems.

Now we plug the result of Proposition 1 into (A.2), and with probability at least $1 - \delta_n$, there holds

$$\begin{aligned} &\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) + \lambda_0\|\widehat{f} - f^*\|_K^2 + \lambda_1\widehat{\Omega}_{[p]}(\widehat{f} - f^*) \\ &\leq 2(M + \|f^*\|_K)\sqrt{\frac{2\kappa_1}{n}} + \kappa_1 M\sqrt{\frac{8\log(1/\delta_n)}{n}} + 69\kappa_1\frac{M\log(1/\delta_n)}{2n} \\ &\quad - 2\lambda_0\langle \widehat{f} - f^*, f^*\rangle_K + 2\lambda_1\widehat{\Omega}_{\mathcal{A}^*}(\widehat{f} - f^*). \end{aligned} \tag{A.3}$$

Additionally, note that $\widehat{\Omega}_{\mathcal{A}^*}(\widehat{f} - f^*) = \sum_{l\in\mathcal{A}^*}\pi_l\|\widehat{g}_l - g_l^*\|_n$. An application of reproducing property of partial derivative functions from RKHS implies that

$$\|\widehat{g}_l - g_l^*\|_n \leq \|\widehat{g}_l - g_l^*\|_\infty \leq \sup_{\mathbf{x}} \|\partial_l K_{\mathbf{x}}\|_K \|\hat{f} - f^*\|_K \leq \kappa_2 \|\hat{f} - f^*\|_K,$$

by $\|\partial_l K_{\mathbf{x}}\|_K \leq \kappa_2$ in Assumption 1. This immediately implies that

$$\lambda_1 \widehat{\Omega}_{\mathcal{A}^*}(\hat{f} - f^*) \leq \lambda_1 \kappa_2 \|\hat{f} - f^*\|_K \leq \lambda_1 \kappa_2 M \sum_{l \in \mathcal{A}^*} \pi_l. \tag{A.4}$$

Finally, plugging (A.4) into (A.3) yields that

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 2(M + \|f^*\|_K)\sqrt{\frac{2\kappa_1}{n}} + \kappa_1 M \sqrt{\frac{8\log(1/\delta_n)}{n}} + \\ 69\kappa_1 \frac{M\log(1/\delta_n)}{2n} + 2\lambda_0 M \|f^*\|_K + 2\lambda_1 \kappa_2 M \sum_{l \in \mathcal{A}^*} \pi_l,$$

with probability at least $1 - \delta_n$. Thus the proof of Theorem 1 ends.

To establish the asymptotic selection results, the following operators are introduced. Define the sample operator for derivatives, $\widehat{D}_l : \mathcal{H}_K \to \mathcal{R}^n$ and its adjoint operator $\widehat{D}^* : \mathcal{R}^n \to \mathcal{H}_K$ as

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

respectively. Accordingly, the integral operators for derivatives, $D_l : \mathcal{H}_K \to \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X})$, where $\mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X}) = \{f : \int_{\mathcal{X}} f(\mathbf{x})^2 d\rho_{\mathbf{x}} < \infty\}$ and $D_l^* : \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X}) \to \mathcal{H}_K$ are defined as

$$D_l f = \langle f, \partial_l K_{\mathbf{x}} \rangle_K \text{ and } D_l^* f = \int \partial_l K_{\mathbf{x}} f(\mathbf{x}) d\rho_{\mathbf{x}}.$$

Note that D_l and $\widehat{D}_l$ are the Hilbert-Schimdt operators by Propositions 12 and 13 of Rosasco et al. (2013) and we have

$$D_l^* D_l f = \int \partial_l K_{\mathbf{x}} g_l(\mathbf{x}) d\rho_{\mathbf{x}} \text{ and } \widehat{D}_l^* \widehat{D}_l f = \frac{1}{n}\sum_{i=1}^n \partial_l K_{\mathbf{x}_i} g_l(\mathbf{x}_i).$$

By Assumption 1 and the direct calculation as in Theorem 7 of Rosasco, Belkin and De vito (2010), there holds

$$\left\|\widehat{D}_l^* \widehat{D}_l\right\|_{HS}^2 = \|\partial_l K\|_K^4 \leq \kappa_2^4. \tag{A.5}$$

Moreover, by the concentration inequalities in the Hilbert-Schmidt space $HS(K)$ on $\mathcal{H}_K$, where $HS(K)$ is a Hilbert space with all the Hilbert-Schmidt operators on $\mathcal{H}_K$, endowed with norm $\|\cdot\|_{HS}$ (Rosasco, Belkin and De vito (2010)), for any $\epsilon_n \in (0,1)$, we have

$$P\Big(\big\|\widehat{D}_l^*\widehat{D}_l - D_l^*D_l\big\|_{HS} \geq \epsilon_n\Big) \leq 2\exp\left(-\frac{n\epsilon_n^2}{8\kappa_2^4}\right). \tag{A.6}$$

Note that the following property holds: $\|T\|_K \leq \|T\|_{HS}$ for any $T \in HS(K)$.

Lemma 2. *Suppose Assumption* 1 *is satisfied. For some* $\delta_n \in (0,1)$, *with probability at least* $1-\delta_n$, *there holds*

$$\sup_{f\in\mathcal{F}_M} |\widehat{\Omega}_{[p]}(f) - \Omega_{[p]}(f)| \leq \sum_{l=1}^p \pi_l (M + \|f^*\|_K)\frac{2\kappa_2}{n^{1/4}}\left(\log\frac{2p}{\delta_n}\right)^{1/4}. \tag{A.7}$$

Proof of Lemma 2: A direct calculation yields that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_M} |\widehat{\Omega}_{[p]}(f) - \Omega_{[p]}(f)| \\
&\leq \sum_{l=1}^p \pi_l\big(|\|g_l\|_n^2 - \|g_l\|_{\rho_{\mathbf{x}}}^2|\big)^{1/2} \\
&= \sum_{l=1}^p \pi_l\left(\left|\frac{1}{n}\sum_{i=1}^n g_l(\mathbf{x}_i)\langle f, \partial_l K_{\mathbf{x}_i}\rangle_K - \int g_l(\mathbf{x})\langle f, \partial_l K_{\mathbf{x}}\rangle_K d\rho_{\mathbf{x}}\right|\right)^{1/2} \\
&= \sum_{l=1}^p \pi_l\left(\left|\left\langle f, \frac{1}{n}\sum_{i=1}^n g_l(\mathbf{x}_i)\partial_l K_{\mathbf{x}_i} - \int g_l(\mathbf{x})\partial_l K_{\mathbf{x}} d\rho_{\mathbf{x}}\right\rangle_K\right|\right)^{1/2} \\
&= \sum_{l=1}^p \pi_l\left(\left|\langle f, (\widehat{D}_l^*\widehat{D}_l - D_l^*D_l)f\rangle_K\right|\right)^{1/2} \leq \sum_{l=1}^p \pi_l\|f\|_K\|\widehat{D}_l^*\widehat{D}_l - D_l^*D_l\|_K^{1/2},
\end{aligned}$$

where the inequalities are trivial. Note that $f \in \mathcal{F}_M$ implies that $\|f\|_K \leq M + \|f^*\|_K$ by the triangle inequality and by (A.6), for some $\delta_n \in (0,1)$, with probability at least $1-\delta_n$, there holds

$$\max_{l=1,\ldots,p}\big\|\widehat{D}_l^*\widehat{D}_l - D_l^*D_l\big\|_{HS} \leq \left(\frac{8\kappa_2^4}{n}\log\frac{2p}{\delta_n}\right)^{1/2}. \tag{A.8}$$

The desired result follows immediately.

Lemma 3. *Suppose that Assumptions* 1 *and* 2 *are satisfied. For some* $\delta_n \in (0,1)$, *with probability at least* $1-\delta_n$*, there holds*

$$\|\widehat{f}-f^*\|_K \leq C_8\left(M^{1/2}\left(\frac{1}{n^{1/2}\lambda_1}+\frac{\sum_{l=1}^p \pi_l}{n^{1/4}}\right)^{1/2}\left(\log\frac{4p}{\delta_n}\right)^{1/2}+\max\{\lambda_0,\lambda_1\}^{\theta_1}\right),$$

where $C_8=\max\{C_1, 4\big(\max\{1,2\sqrt{2}\kappa_2^2\}\max\{\sqrt{2\kappa_1}, 69\kappa_1/4, 2\kappa^2\}\big)^{1/2}\}$.

Proof of Lemma 3: By the triangle inequality, We have $\|\widehat{f}-f^*\|_K \leq \|\widehat{f}-\widetilde{f}\|_K + \|\widetilde{f}-f^*\|_K$. By Assumption 2, we have $\|\widetilde{f}-f^*\|_K = C_1\max\{\lambda_0,\lambda_1\}^{\theta_1}$. To bound $\|\widehat{f}-\widetilde{f}\|_K$, we first denote $\mathcal{E}^{\lambda_1}(f)=\mathcal{E}(f)+\lambda_0\|f\|_K^2+\lambda_1\Omega_{[p]}(f)$ and $\mathcal{E}^{\lambda_1}_{\mathcal{Z}_n}(f)=\mathcal{E}(f)+\lambda_0\|f\|_K^2+\lambda_1\widehat{\Omega}_{[p]}(f)$ for simplicity. Directly followed by Proposition 2 and Theorems 2.6 and 2.7 in Villa et al. (2012), there holds

$$\begin{aligned}\Psi^{\diamond}_{\lambda_1}\left(\|\widehat{f}-\widetilde{f}\|_K\right) &\leq 4\sup_{f\in\mathcal{F}_M}\left|t_{\mathcal{E}^{\lambda_1}}\mathcal{E}^{\lambda_1}(f)-t_{\mathcal{E}^{\lambda_1}}\mathcal{E}^{\lambda_1}_{\mathcal{Z}^n}(f)\right| \\ &\leq 4\sup_{f\in\mathcal{F}_M}|\mathcal{E}(f)-\widehat{\mathcal{E}}(f)|+4\lambda_1\sup_{f\in\mathcal{F}_M}|\widehat{\Omega}_{[p]}(f)-\Omega_{[p]}(f)|, \quad \text{(A.9)}\end{aligned}$$

where $\Psi^{\diamond}_{\lambda_1}(t)=\inf\{(\lambda/2)s^2+|t-s|: s\in[0,\infty)\}$ and $t_{\mathcal{E}^{\lambda_1}}$ is the translation map defined as $t_{\mathcal{E}^{\lambda_1}}G(f)=G(f+\widetilde{f})-\mathcal{E}^{\lambda_n}(f)$ for all $G:\mathcal{H}_K\to\mathcal{R}$.

Now we bound (A.9) separately. For the first part, note that by the proof of Proposition 1, with probability at least $1-\delta_n/2$, we have

$$\sup_{f\in\mathcal{F}_M}|\mathcal{E}(f)-\widehat{\mathcal{E}}(f)| \leq (M+\|f^*\|_K)\left(\frac{\sqrt{2\kappa_1}}{n^{1/2}}+\frac{\sqrt{2}\kappa_1}{n^{1/2}}\left(\log\frac{2}{\delta_n}\right)^{1/2}+\frac{69\kappa_1}{4n}\log\frac{2}{\delta_n}\right).$$

For the second part, by Lemma 2, we have with probability at least $1-\delta_n/2$ that

$$\sup_{f\in\mathcal{F}_M}|\widehat{\Omega}_{[p]}(f)-\Omega_{[p]}(f)| \leq \sum_{l=1}^p \pi_l(M+\|f^*\|_K)\frac{2\kappa_2}{n^{1/4}}\left(\log\frac{4p}{\delta_n}\right)^{1/4}. \quad \text{(A.10)}$$

Moreover, since $\Psi^{\diamond}_{\lambda_1}$ is invertible and increasing, we can write its inverse explicitly as $\left(\Psi^{\diamond}_{\lambda_1}\right)^{-1}$ as

$$\left(\Psi^{\diamond}_{\lambda_1}\right)^{-1}(t)=\begin{cases}\sqrt{\frac{2t}{\lambda_1}}, & \text{if } t<\frac{1}{2\lambda_1};\\ t+\frac{1}{2\lambda_1}, & \text{otherwise.}\end{cases}$$

And thus, when the upper bound of (A.9) is sufficiently small, with probability at least $1-\delta_n$, there holds

$$
\begin{aligned}
&\|\widehat{f}-\widetilde{f}\|_K\\
&\leq 2\sqrt{2}\lambda_1^{-1/2}\left(\sup_{f\in\mathcal{F}_M}|\mathcal{E}(f)-\widehat{\mathcal{E}}(f)|+\lambda_1\sup_{f\in\mathcal{F}_M}|\widehat{\Omega}_{[p]}(f)-\Omega_{[p]}(f)|\right)^{1/2}\\
&\leq \frac{2\sqrt{2}(M+\|f^*\|_K)^{1/2}}{\lambda_1^{1/2}}\left(\frac{\sqrt{2\kappa_1}}{n^{1/2}}+\frac{\sqrt{2}\kappa_1}{n^{1/2}}+\frac{69\kappa_1}{4n}+\lambda_1\sum_{l=1}^p\pi_l\frac{2\kappa_2}{n^{1/4}}\right)^{1/2}\left(\log\frac{4p}{\delta_n}\right)^{1/2}.\\
&\leq C_9M^{1/2}\left(\frac{1}{n^{1/2}\lambda_1}+\frac{\sum_{l=1}^p\pi_l}{n^{1/4}}\right)^{1/2}\left(\log\frac{4p}{\delta_n}\right)^{1/2}.
\end{aligned}
$$

where $C_9=4\big(\max\{1,2\sqrt{2}\kappa_2^2\}\max\{\sqrt{2\kappa_1},69\kappa_1/4,2\kappa^2\}\big)^{1/2}$. This completes the proof.

Proof of Lemma 1: By representer theorem of derivative functions in RKHS in (4), a direct calculation yields that

$$
\begin{aligned}
&\max_{l=1,\dots,p}\big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_{\rho_{\mathbf{x}}}^2\big|\\
&=\max_{l=1,\dots,p}\left|\left\langle\widehat{f},\frac{1}{n}\sum_{i=1}^n\widehat{g}_l(\mathbf{x}_i)\partial_lK_{\mathbf{x}_i}\right\rangle_K-\left\langle f^*,\int g_l^*(\mathbf{x})\partial_lK_{\mathbf{x}}d\rho_{\mathbf{x}}\right\rangle_K\right|\\
&=\max_{l=1,\dots,p}\Big|\big\langle\widehat{f}-f^*,\widehat{D}_l^*\widehat{D}_l(\widehat{f}-f^*)\big\rangle_K+2\big\langle f^*,\widehat{D}_l^*\widehat{D}_l(\widehat{f}-f^*)\big\rangle_K+\\
&\qquad\qquad\big\langle f^*,(\widehat{D}_l^*\widehat{D}_l-D_l^*I_l)f^*\big\rangle_K\Big|\\
&\leq\kappa_2^2\|\widehat{f}-f^*\|_K^2+2\kappa_2\|f^*\|_K\|\widehat{f}-f^*\|_K+\|f^*\|_K^2\max_{l=1,\dots,p}\|\widehat{D}_l^*\widehat{D}_l-D_l^*D_l\|_{HS},
\end{aligned}
$$

where the inequality follows from Cauthy-Schwartz inequality. By $(A.5)$ and $(A.8)$, when $\|\widehat{f}-f^*\|_K$ is sufficient small, with probability at least $1-\delta_n/3$, we have

$$
\max_{l=1,\dots,p}\big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_2^2\big|\leq\frac{2\sqrt{2}\kappa_2^2\|f^*\|_K^2}{n^{1/2}}\left(\log\frac{6p}{\delta_n}\right)^{1/2}+\kappa_2(2\|f^*\|_K+\kappa_2)\|\hat{f}-f^*\|_K.
$$

Hence, associated with Lemma 3 and under case (ii), for some positive constant C_6, with probability at least $1-\delta_n$, there holds

$$
\big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_2^2\big|\leq C_6\left(M^{1/2}\left(\frac{1}{n^{1/2}\lambda_1}+\frac{\sum_{l=1}^p\pi_l}{n^{1/4}}\right)^{1/2}+\max\{\lambda_0,\lambda_1\}^{\theta_1}\right)\left(\log\frac{6p}{\delta_n}\right)^{1/2},\tag{A.11}
$$

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

Specially, under case (ii) by taking $M = \lambda_0^{-1/2} + \|f^*\|_K$, by Assumption 2 and let $\lambda_0 = n^{-1/8}$ and $\lambda_1 = n^{-1/4}$, the desired result follows immediately.

Proof of Theorem 2: Now we show that $\mathcal{A}^* \subset \widehat{\mathcal{A}}$ in probability. If not, suppose that there exists some $l' \in \mathcal{A}^*$ but $l' \notin \widehat{\mathcal{A}}$, which implies $\|\widehat{g}_{l'}\|_n^2 = 0$. By Assumption 3, we have with probability at least $1 - \delta_n$ that

$$\big|\|\widehat{g}_{l'}\|_n^2 - \|g_{l'}^*\|_2^2\big| = \|g_{l'}^*\|_2^2 > C_5 \max\{p^{1/2} n^{-3/32+\theta_2/2}, n^{-\theta_1/8}\} \left(\log \frac{6p}{\delta_n}\right)^{\xi},$$

which contradicts with Lemma 1 . This implies that $\mathcal{A}^* \subset \mathcal{A}$ with probability at least $1 - \delta_n$, $\mathcal{A}^* \subset \widehat{\mathcal{A}}$. This completes the proof.

A.3. Concentration inequalities

Our main tool is a result about concentration inequality from Bousquet (2002), involving the expectation of the supremum of the empirical process.

Lemma 4. *Let $Z_1, \dots, Z_n$ be independent and identically distributed copies of a random variable $Z \in \mathcal{Z}$. Let Γ be a class of real-valued functions on $\mathcal{Z}$ satisfying* $\sup_z |\gamma(z)| \le D$ *for all $\gamma \in \Gamma$. Define*

$$\mathbf{Z} := \sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \{\gamma(Z_i) - \mathbb{E}[\gamma(Z_i)]\} \right|$$

and

$$\tau^2 := \sup_{\gamma \in \Gamma} var(\gamma(Z)).$$

Then, for any positive t,

$$P\left(\mathbf{Z} \ge 2\mathbb{E}[\mathbf{Z}] + \tau\sqrt{\frac{8t}{n}} + \frac{69Dt}{2n}\right) \le \exp(-t).$$

The key step using Lemma 4 is to deal with the expectation, which appeals to be complicated. However, one may derive bounds for it based on symmetrization technique and the Lipschitz property of Rademacher complexity.

Proof of proposition 1. In order to apply Lemma 4, we take

$$\Gamma = \big\{\gamma_f : f \in \mathcal{H}_K \,\text{with}\, \|f - f^*\|_K \le M\big\}, \ \text{where}\, \gamma_f(x, y) = \phi(yf(x)) - \phi(yf^*(x))$$

and where $\phi(z) = (1 - z)_+$ is the hinge loss function. Since ϕ is Lipschitz and

$\|g\|_\infty \leq \kappa_1 \|g\|_K$ for any $g \in \mathcal{H}_K$ by reproducing property of RKHS, we have $\sup_z |\gamma_f(z)| \leq \kappa_1 M$ for all $f \in \mathcal{H}_K$, which means that $D = \kappa_1 M$ in Lemma 4. This also implies that $\tau = \kappa_1 M$ can be set. Then, the conclusion of Lemma 4 tells us that, with probability at least $1 - \delta_n$

$$\mathbf{Z} \leq 2\mathbb{E}[\mathbf{Z}] + \kappa_1 M \sqrt{\frac{8 \log(1/\delta_n)}{n}} + 69\kappa_1 \frac{M \log(1/\delta_n)}{2n}, \tag{A.12}$$

where we take $t = \log(1/\delta_n)$. To complete the proof, it remains to give an upper bound of $\mathbb{E}[\mathbf{Z}]$. By symmetrization theorem in Van de Geer (2000), we have

$$\mathbb{E}[\mathbf{Z}] \leq 2\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \gamma(Z_i)\right|\right] \leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}_M} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right|\right],$$

where the second inequality follows from the Lipschitz property of Rademacher complexity. Recall the following result was proved in Bartlett and Mendelson (2002). Suppose that the kernel K is bounded uniformly by κ_1, then there holds

$$\mathbb{E}\left[\sup_{f:\|f\|_K \leq 1} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right|\right] \leq \sqrt{\frac{2\kappa_1}{n}}.$$

Then, we obtain the desired result immediately in Proposition 1 by plugging the upper bound of Rademacher complexity into (A.12).

References

Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research* **9**, 1179–1225.

Bach, F. (2009). High-dimensional non-linear variable selection through hierarchical kernel learning. *Technical Report HAL 00413473*, INRIA.

Barber, R. and Candès, E. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43**, 2055–2085.

Bartlett, P. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482.

Bi, J., Bennett, K., Embrechts, M., Breneman, C. and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research* **3**, 1229–1243.

Bousquet, O. (2002). A Bennet concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique* **334**, 495–550.

Cuker, F. and Zhou D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint.* Cambridge University Press.

Ekeland, I. and Temam, R. (1976). *Convex Analysis and Variational Problems.* North-Holland Publishing Company, Amsterdam.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics* **36**, 2605–2637.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Horowitz, J. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics* **35**, 2589–2619.

Jaakkola, T., Diekhans, M. and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of Seventh International Conference on Intelligent Systems for Molecular Biology*, 149–158.

Lee, K., Li, B. and Zhao, H. (2016). Variable selection via additive conditional independence. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **78**, 1037–1055.

Li, Y. and Liu, J. (2018). Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, in press.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association* **107**, 1129–1139.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34**, 22–72.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming* **103**, 127–152.

Rosasco, L., Belkin, M. and De vito, E. (2010). On learning with integral operators. *Journal of Machine Learning Research* **11**, 905–934.

Rosasco, L., Villa, S., Mosci, S., Santoro, M. and Verri, A. (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research* **14**, 1665–1714.

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

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

Tarigan, B. and Van De Geer, S. (2006). Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli* **12**, 1045–1076.

Van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

Villa, S., Rosasco, L., Mosci, S., and Verri. S. (2012). Consistency of learning algorithms using Attouch-Wets convergence. *Optimization* **61**, 287–305.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. *Advances in Kernel Methods: Support Vector Learning*, 69–88. MIT Press.

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association* **110**, 1726–1734.

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

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

Zhang, H. (2006). Variable selection for support vector machines via smoothing spline anova. *Statistica Sinica* **16**, 659–674.

Zhang, X., Wu, Y., Wang, L. and Li, R. (2016). Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78**, 53–76.

Zhao, T. and Liu, H. (2012). Sparse additive machine. In *International Conference on Artificial Intelligence and Statistics.*

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

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004). 1-norm support vector machines. *Proceedings of the Neural Information Processing Systems.*

Zou, H. (2007). An improved 1-norm SVM for simultaneous classification and variable selection. *Journal of Machine Learning Research* **2**, 675–681.

Zou, H. and Trevor, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **67**, 301–320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1419–1429.

Zou, H. and Yuan, M. (2008). The f-infinity norm support vector machine. *Statistica Sinica* **18**, 379–398.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: he.xin17@mail.shufe.edu.cn

School of Statistics and Mathematics, Nanjing Audit University, China.

E-mail: lvsg716@swufe.edu.cn

School of Data Science, City University of Hong Kong, Hong Kong, China.

E-mail: j.h.wang@cityu.edu.hk

(Received March 2018; accepted December 2018)