Statistica Sinica **31** (2021), 2123-2151
doi:https://doi.org/10.5705/ss.202019.0401

EFFICIENT KERNEL-BASED VARIABLE SELECTION WITH SPARSISTENCY

Xin He, Junhui Wang and Shaogao Lv

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

Abstract: Sparse learning is central to high-dimensional data analysis, and various methods have been developed. Ideally, a sparse learning method should be methodologically flexible, computationally efficient, and provide a theoretical guarantee. However, most existing methods need to compromise some of these properties in order to attain the others. We develop a three-step sparse learning method, involving a kernel-based estimation of the regression function and its gradient functions, as well as a hard thresholding. Its key advantages are that it includes no explicit model assumption, admits general predictor effects, allows efficient computation, and attains desirable asymptotic sparsistency. The proposed method can be adapted to any reproducing kernel Hilbert space (RKHS) with different kernel functions, and its computational cost is only linear in the data dimension. The asymptotic sparsistency of the proposed method is established for general RKHS under mild conditions. The results of numerical experiments show that the proposed method compares favorably with its competitors in both simulated and real examples.

Key words and phrases: Gradient learning, hard thresholding, nonparametric sparse learning, ridge regression, RKHS.

1. Introduction

Sparse learning has attracted much interest from both researchers and practitioners, owing to the availability of large numbers of variables in many real applications. In such scenarios, identifying the truly informative variables for the objective of analysis has become a key part of facilitating statistical modeling and analysis. Ideally, a sparse learning method should be flexible, efficient, and provide a theoretical guarantee. Specifically, the method should not assume restrictive model assumptions, making it applicable to data with complex structures. In addition, its implementation should be computationally efficient, and should take advantage of high performance computing platforms. Furthermore,

Corresponding author: Junhui Wang, School of Data Science, City University of Hong Kong, Hong Kong, China. E-mail: j.h.wang@cityu.edu.hk.

it should provide a theoretical guarantee on its asymptotic consistency in identifying the truly informative variables.

In the literature, many sparse learning methods have been developed in regularization frameworks that assume a certain working model set. Linear models are the most popular working model sets, where the sparse learning task simplifies to identifying nonzero coefficients. Under the linear model assumption, the regularization framework consists of a least squares loss function for the linear model, as well as a sparsity-inducing regularization term. Various regularization terms have been considered, including the least absolute shrinkage and selection operator (Lasso; Tibshirani (1996)), smoothly clipped absolute deviation (SCAD; Fan and Li (2001)), adaptive Lasso (Zou (2006)), minimax concave penalty (MCP; Zhang (2010)), truncated l_1-penalty (TLP; Shen, Pan and Zhu (2012)), and l_0-penalty (Shen et al. (2013)), among others. These methods have also been extended to the nonparametric models to relax the linear model assumption. For example, under the additive model assumption, a number of sparse learning methods have been developed (Shively, Kohn, and Wood (1999); Huang, Horowitz and Wei (2010)), where each component function depends on one variable only. Furthermore, a component selection and smoothing operator method (COSSO; Lin and Zhang (2006)) has been proposed to allow higher-order interaction components in the additive model. However, the higher-order additive models need to enumerate all interaction components, which may be of an exponential order of the number of variables. These nonparametric sparse learning methods, although more flexible than the linear model, still require explicit working model sets.

More recently, attempts have been made to develop nonparametric sparse learning methods to circumvent the dependency on restrictive model assumptions. In particular, sparse learning is formulated in a dimension reduction framework in Li, Zha and Chiaromonte (2005) and Bondell and Li (2009) by searching for the sparse basis of the central dimension reduction space. Fukumizu and Leng (2014) developed a gradient-based dimension-reduction method that can be extended to nonparametric sparse learning. A novel measurement-error-model-based sparse learning method is developed in Stefanski, Wu and White (2014) and Wu and Stefanski (2015) for nonparametric kernel regression models. In addition, gradient learning methods (Rosasco et al. (2013); Yang, Lv and Wang (2016)) have been proposed to conduct sparse learning in a flexible reproducing kernel Hilbert space (RKHS) (Wahba (1998)). Furthermore, a flexible knock-off filter framework (Barber and Candës (2015)) and a recursive feature elimination method using a kernel ridge regression have been proposed (Dasgupta, Goldberg and Kosorok (2019)) that show substantial advantages over most existing methods. However, their lack of selection consistency and computational efficiency remain as obstacles. Interestingly, most existing gradient-based methods (Rosasco et al. (2013); Yang, Lv and Wang (2016)) aim to directly estimate the gradient functions in a regularization framework using some well-designed penalty terms, and thus may not be applicable for analyzing high-dimensional data, owing to their expensive computational cost.

Another popular line of research on high-dimensional data is that on variable screening, which screens out uninformative variables by examining the marginal relationship between the response and each variable. The marginal relationship can be measured by various criteria, including the Pearson's correlation (Fan and Lv (2008)), the empirical functional norm (Fan, Feng and Song (2011)), the distance correlation (Szekely, Rizzo and Bakirov (2007)), and a quantile-adaptive procedure (He, Wang and Hong (2013)). These methods are all computationally very efficient and attain the sure screening property, meaning that all the truly informative variables are retained after screening, with probability tending to one. This is a desirable property, yet slightly weaker than the asymptotic consistency in sparse learning. Another potential weakness of the marginal screening methods is that they may ignore those marginally unimportant, but jointly important variables (He, Wang and Hong (2013)). To remedy this limitation, recent works (Hao and Zhang (2014); Wang and Leng (2016)) have conducted sure screening for variables with interaction effects.

We propose an efficient kernel-based sparse learning method that is methodologically flexible, computationally efficient, and able to achieve asymptotic consistency without requiring any explicit model assumptions. The method consists of three simple steps that include a kernel-based estimation of the regression function and its gradient functions, as well as a hard thresholding. It first fits a kernel ridge regression model in a flexible RKHS to obtain an estimated regression function. Then, it estimates the gradient functions along each variable by taking advantage of the derivative reproducing property (Zhou (2007)). Finally, it hard-thresholds the empirical norm of each gradient function to identify the truly informative variables. This method is flexible in that it can be adapted to any RKHS with different kernel functions to accommodate prior information about the true regression function. The proposed method also enables an efficient estimation of the gradient functions in two steps using the derivative property of the RKHS, which significantly reduces the computational cost and allows for a diverging dimension. The computational cost is only linear in the data dimension. Thus, it is computationally efficient when analyzing data sets with large dimensions. For example, the simulated examples with $p = 100{,}000$ variables can be analyzed efficiently on a standard multi-core PC. More importantly, the asymptotic consistency can be established for the proposed method without requiring any explicit model assumptions. It is clear that the proposed method has advantages over the existing methods, because it achieves methodological flexibility, numerical efficiency, and asymptotic consistency. To the best of our knowledge, this method is the first to achieve these three desirable properties at the same time.

The rest of the paper is organized as follows. In Section 2, we present the proposed general kernel-based sparse learning method and its computational scheme. In Section 3, the asymptotic consistency of the proposed method is established. Two theoretical examples are provided in Section 4. In Section 5, the proposed method is extended to select truly informative interaction terms. Section 6 contains numerical experiments on the simulated and real examples, and Section 7 concludes the paper. All necessary lemmas and technical proofs are provided in the Appendix and in the online Supplementary Materials.

2. Proposed Method

2.1. Regression in an RKHS

Suppose a random sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ comprises independent copies of $\mathcal{Z} = (\mathbf{x}, y)$, drawn from some unknown distribution $\rho_{\mathbf{x},y}$, with $\mathbf{x} = (x^1, \ldots, x^p)^T \in \mathcal{X}$ supported on a compact metric space and $y \in \mathcal{R}$. Consider a general regression setting,

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

where ϵ is a random error, with $\mathrm{E}(\epsilon|\,\mathbf{x}) = 0$ and $\mathrm{Var}(\epsilon|\,\mathbf{x}) = \sigma^2$. Thus $f^*(\mathbf{x}) = \int y d\rho_{y|\,\mathbf{x}}$, with $\rho_{y|\,\mathbf{x}}$ denoting the conditional distribution of y, given $\mathbf{x}$. We further assume that $f^* \in \mathcal{H}_K$, where $\mathcal{H}_K$ is an RKHS induced by some prespecified kernel function $K(\cdot,\cdot)$. For each $\mathbf{x} \in \mathcal{X}$, denote $K_{\mathbf{x}} = K(\mathbf{x}, \cdot) \in \mathcal{H}_K$, and the reproducing property of the RKHS implies that $\langle f, K_{\mathbf{x}}\rangle_K = f(\mathbf{x})$, for any $f \in \mathcal{H}_K$, where $\langle \cdot,\cdot\rangle_K$ is the inner product in $\mathcal{H}_K$.

The RKHS enjoys a number of desirable properties that make it particularly suitable for general nonparametric models, including its approximation ability, functional complexity, and derivative reproducing property. Specifically, many popular kernels, including the Gaussian and Laplace kernels, are universal (Steinwart and Christmann (2008)), meaning that the RKHS each induces is dense in the continuous function space under the infinity norm. This universal approximation property ensures that the kernel-based methods yield nonparametric estimates with a small approximation error when estimating any continuous target function. On the other hand, to characterize the statistical properties of nonparametric models, the notion of functional complexity in an empirical process is widely employed in theoretical analyses. This includes various covering numbers, the VC dimension, and Rademacher complexity (Bartlett and Mendelson (2002)). The RKHS has a very interesting and surprising property that for a unit ball B_1 of the RKHS, its Rademacher complexity (Bartlett and Mendelson (2002)) can be bounded as $R_n(B_1) \leq 2n^{-1/2}(\mathbb{E}(K(X,X)))^{1/2}$, where $R_n(\cdot)$ denotes the global Rademacher complexity. In other words, the functional complexity of the bounded ball in the RKHS is less affected by the dimension of the variables. Thus a small variance estimator can be obtained, without sacrificing the approximation ability of the nonparametric estimation, using kernel-based methods. In addition, in the literature on nonparametric statistics, estimating the gradient function of the target function is, in general, difficult. However, the derivative of any function in a smooth RKHS also has the reproducing property, implying that kernel-based methods have simultaneous convergence behavior in both the function and its gradient function, with the same rate of convergence under the sup norm.

2.2. Gradient-based sparse learning

In sparse modeling, it is generally believed that $f^*(\mathbf{x})$ depends only on a small number of variables, while others are uninformative. Unlike model-based settings, sparse learning for a general regression model is challenging, owing to the lack of explicit regression parameters. Here, we measure the importance of variables in a regression function by examining the corresponding gradient functions. It is crucial to observe that if a variable x^l is deemed uninformative, the corresponding gradient function

$$g_l^*(\mathbf{x}) = \frac{\partial f^*(\mathbf{x})}{\partial x^l}$$

should be exactly zero, almost surely. Thus, the true active set can be defined as

$$\mathcal{A}^* = \{l : \|g_l^*\|_2^2 > 0\},$$

where $\|g_l^*\|_2^2 = \int (g_l^*(\mathbf{x}))^2 \, d\rho_{\mathbf{x}}$ with the marginal distribution $\rho_{\mathbf{x}}$.

The proposed general sparse learning method is presented in Algorithm 1.

We now describe each step in Algorithm 1 on greater detail. To obtain $\widehat{f}$ in

Algorithm 1 General sparse learning method.

```
Step 1: Obtain an estimate $\widehat{f}$ in a smooth RKHS based on the given sample $\mathcal{Z}^n$;
Step 2: Compute $\widehat{g}_l(\mathbf{x}) = \partial \widehat{f}(\mathbf{x})/\partial x^l$ for $l = 1, \ldots, p$;
Step 3: Identify the informative variables by checking the norm of each $\widehat{g}_l$.
```

Step 1, we employ the kernel ridge regression model,

$$\widehat{f}(\mathbf{x}) = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda_n \|f\|_K^2, \tag{2.1}$$

where the first term, denoted as $\mathcal{E}_n(f)$, is an empirical version of $\mathcal{E}(f) = \mathrm{E}(y - f(\mathbf{x}))^2$, and $\|f\|_K = \langle f, f \rangle_K^{1/2}$ is the associated RKHS-norm of $f \in \mathcal{H}_K$. By the representer theorem (Wahba (1998)), the minimizer of (2.1) must have the form

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) = \widehat{\boldsymbol{\alpha}}^T \mathbf{K}_n(\mathbf{x}),$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)^T$ and $\mathbf{K}_n(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x}))^T$. Then, the optimization task in (2.1) can be solved analytically, with

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{K}^2 + n\lambda_n \mathbf{K}\right)^+ \mathbf{K}\,\mathbf{y}, \tag{2.2}$$

where $\mathbf{K} = \left(K(\mathbf{x}_i, \mathbf{x}_j)\right)_{i,j=1}^n$, and $^+$ denotes the Moore–Penrose generalized inverse of a matrix. When $\mathbf{K}$ is invertible, (2.2) simplifies to $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda_n \mathbf{I})^{-1}\mathbf{y}$.

Next, to obtain $\widehat{g}_l$ in Step 2, it follows from Lemma 1 that for any $f \in \mathcal{H}_K$,

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

where $\partial_l K_{\mathbf{x}} = \partial K(\mathbf{x}, \cdot)/\partial x^l$. This implies that the gradient function of any $f \in \mathcal{H}_K$ can be bounded by its K-norm, up to some constant. In other words, if we want to estimate $g_l^*(\mathbf{x})$ within a smooth RKHS, it suffices to estimate f^* itself, without loss of information. Consequently, if $\widehat{f}$ is obtained in Step 1, $g_l^*(\mathbf{x})$ can be estimated as $\widehat{g}_l(\mathbf{x}) = \widehat{\boldsymbol{\alpha}}^T \partial_l \mathbf{K}_n(\mathbf{x})$, for each l, where $\partial_l \mathbf{K}_n(\mathbf{x}) = (\partial_l K_{\mathbf{x}}(\mathbf{x}_1), \ldots, \partial_l K_{\mathbf{x}}(\mathbf{x}_n))^T$.

In Step 3, it is difficult to evaluate $\|\widehat{g}_l\|_2^2$ directly, because $\rho_{\mathbf{x}}$ is usually unknown in practice. We then adopt the empirical norm of $\widehat{g}_l$ as a practical measure,

$$\|\widehat{g}_l\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{g}_l(\mathbf{x}_i)\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\boldsymbol{\alpha}}^T \partial_l \mathbf{K}_n(\mathbf{x}_i)\right)^2.$$

The estimated active set can be set as $\widehat{\mathcal{A}}_{v_n} = \{l : \|\widehat{g}_l\|_n^2 > v_n\}$, for some prespecified v_n. Our method can clearly be regarded as a nonparametric joint screening method that can correctly identify all truly informative variables acting on the response with a general effect, including those that are marginally noninformative, but jointly informative.

The proposed method presented in Algorithm 1 is general in that it can be adapted to any smooth RKHS with different kernel functions, where the choice of the kernel function depends on prior knowledge about f^*. For instance, if f^* is known in advance to be a linear or polynomial function, the RKHS induced by the linear or polynomial kernel can be used. If no prior information about f^* is available, the RKHS induced by the Gaussian kernel can be used, which 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 and Christmann (2008)). In practice, unless some reliable prior information about f^* is known, it is recommended to consider the RKHS induced by the Gaussian kernel owing to its capacity and flexibility.

Remark 1. The proposed method is computationally efficient, with a computational cost of about $O(n^3 + n^2p)$. The complexity $O(n^3)$ comes from inverting an $n \times n$ matrix in (2.2), and the complexity $O(n^2p)$ comes from calculating $\|\widehat{g}_l\|_n^2$, for $l = 1, \ldots, p$. This complexity is particularly attractive in the large-p-small-n scenario, where the computational complexity becomes linear in p, and parallelization can be employed to further speed up the computation. In other scenarios with large n, the $O(n^3)$ complexity can be too demanding. Improvements are available to alleviate the computational burden using some low-rank approximation, such as the random sketch method in Yang, Pilanci and Wainwright (2017). Its computational complexity can be reduced to $O(m^3)$, where $m \ll n$ is the sketch dimension determined in Yang, Pilanci and Wainwright (2017). More importantly, the random sketch method has been proved to be fast and minimax optimal when fitting the kernel ridge regression.

Remark 2. The estimated regression function $\widehat{f}$ is merely an intermediate step for estimating the gradient functions. It provides is a consistent estimate, but converges to the true regression function f^* at some rather slow rate, owing to the inclusion of the noise variables. We also want to emphasize that the data are only used once to estimate the representer coefficients $\widehat{\boldsymbol{\alpha}}$ in (2.2). Then the estimated gradient function $\widehat{g}_l$ can be estimated directly using the derivative reproducing property of the RKHS, by Lemma 1.

2.3. Tuning

The proposed method presented in Algorithm 1 consists of two tuning parameters, the ridge parameter λ_n and the thresholding parameter v_n. Based on our limited numerical experience, the proposed method performs well and is stable when the ridge parameter λ_n is sufficiently small in various scenarios. A similar observation on the choice of the ridge parameter is made in Wang and Leng (2016). Therefore, we set $\lambda_n = 0.001$, and focus on the choice of v_n in the simulated experiments.

To optimize the selection performance of the proposed method, we employ the stability-based criterion (Sun, Wang and Fang (2013)) to select the value of v_n. Its key idea is to measure the stability of sparse learning by randomly splitting the training sample into two parts, and comparing the disagreement between the two estimated active sets. Specifically, given a thresholding value v_n, we randomly split the training sample $\mathcal{Z}^n$ into two parts, $\mathcal{Z}_1^n$ and $\mathcal{Z}_2^n$. Then, the proposed method is applied to $\mathcal{Z}_1^n$ and $\mathcal{Z}_2^n$ to obtain the estimated active sets, $\widehat{\mathcal{A}}_{1,v_n}$ and $\widehat{\mathcal{A}}_{2,v_n}$, respectively. The disagreement between $\widehat{\mathcal{A}}_{1,v_n}$ and $\widehat{\mathcal{A}}_{2,v_n}$ is measured using Cohen's kappa coefficient

$$\kappa(\widehat{\mathcal{A}}_{1,v_n}, \widehat{\mathcal{A}}_{2,v_n}) = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a) = (n_{11} + n_{22})/p$ and $Pr(e) = ((n_{11} + n_{12})(n_{11} + n_{21}))/p^2 + ((n_{12} + n_{22})(n_{21} + n_{22}))/p^2$, with $n_{11} = |\widehat{\mathcal{A}}_{1,v_n} \cap \widehat{\mathcal{A}}_{2,v_n}|, n_{12} = |\widehat{\mathcal{A}}_{1,v_n} \cap \widehat{\mathcal{A}}_{2,v_n}^C|, n_{21} = |\widehat{\mathcal{A}}_{1,v_n}^C \cap \widehat{\mathcal{A}}_{2,v_n}|, n_{22} = |\widehat{\mathcal{A}}_{1,v_n}^C \cap \widehat{\mathcal{A}}_{2,v_n}^C|$, and $|\cdot|$ denoting the set cardinality.

The procedure is repeated B times, and the estimated sparse learning stability is measured as

$$\hat{s}(\Psi_{v_n}) = \frac{1}{B}\sum_{b=1}^{B} \kappa(\widehat{\mathcal{A}}_{1,v_n}^b, \widehat{\mathcal{A}}_{2,v_n}^b).$$

Finally, the thresholding parameter $\widehat{v}_n$ is set as $\widehat{v}_n = \max\{v_n : \hat{s}(\Psi_{v_n})/\max_{v_n}\hat{s}(\Psi_{v_n}) \geq q\}$, where $q \in (0,1)$ is some given percentage. In the simulated experiments, we set $q = 0.95$, as suggested in Sun, Wang and Fang (2013), and the performance of the resultant tuning criterion appears to be satisfactory.

3. Asymptotic Sparsistency

Now, we establish the asymptotic consistency of the proposed method. First, we introduce an integral operator $L_K : \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) \to \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, given by

$$L_K(f)(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{u}) f(\mathbf{u}) d\rho_{\mathbf{x}}(\mathbf{u}),$$

for any $f \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) = \{f : \int f^2(\mathbf{x}) d\rho_{\mathbf{x}} < \infty\}$. Note that if the corresponding RKHS is separable, by the spectral theorem, we have

$$L_K f = \sum_j \mu_j \langle f, e_j \rangle_2 e_j,$$

where $\{e_j\}$ is an orthonormal basis of $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, μ_j is the eigenvalue of the integral operator L_K, and $\langle \cdot, \cdot \rangle_2$ is the inner product in $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$. By Mercer's theorem, under some regularity conditions, the eigen-expansion of the kernel function is $K(\cdot, \cdot) = \sum_{j \geq 1} \mu_j e_j(\cdot) e_j(\cdot)$. Therefore, the RKHS-norm of any $f \in \mathcal{H}_K$ can be written as

$$\|f\|_K^2 = \sum_{j \geq 1} \frac{\langle f, e_j \rangle_2^2}{\mu_j},$$

which implies that the decay rate of μ_j fully characterizes the complexity of the RKHS, and is closely related to various entropy numbers (Steinwart and Christmann (2008)).

We denote the cardinality of the true active set $\mathcal{A}^*$ as $|\mathcal{A}^*| = p_0$, and both p_0 and p are allowed to diverge with n. The following technical assumptions are made.

Assumption 1. *Suppose that f^* is in the range of the rth power of L_K, denoted as L_K^r, for some positive constant $r \in (1/2, 1]$.*

Assumption 2. *There exist some constants κ_1 and κ_2, such that* $\sup_{\mathbf{x} \in \mathcal{X}} \|K_{\mathbf{x}}\|_K \leq \kappa_1$ *and* $\sup_{\mathbf{x} \in \mathcal{X}} \|\partial_l K_{\mathbf{x}}\|_K \leq \kappa_2$*, for any $l = 1, \ldots, p$.*

Assumption 3. *The distribution of ϵ has a q-exponential tail, with some function $q(\cdot)$; that is, there exists some constant $c_1 > 0$, such that $P(|\epsilon| > t) \leq c_1 \exp\{-q(t)\}$, for any $t > 0$.*

In Assumption 1, the operator L_K on $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ is self-adjoint and semi-positive definite, and thus its fractional operator L_K^r is well defined. Furthermore, the range of L_K^r is contained in $\mathcal{H}_K$ if $r \geq 1/2$ (Smale and Zhou (2007)). Thus Assumption 1 implies that there exists some function $h \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ such that $f^* = L_K^r h = \sum_j \mu_j^r \langle h, e_j \rangle_2 e_j \in \mathcal{H}_K$, ensuring strong estimation consistency under the RKHS-norm. Similar assumptions are imposed in Mendelson and Neeman (2010). Assumption 2 assumes the boundedness of the kernel function and its gradient functions, and is satisfied by many popular kernels, including the Gaussian kernel and the Sobolev kernel (Smale and Zhou (2007); Rosasco et al. (2013);

Yang, Lv and Wang (2016)) with the compact support condition. Note that the compact support condition is commonly used in the machine learning literature (Mendelson and Neeman (2010); Rosasco et al. (2013); Dasgupta, Goldberg and Kosorok (2019); Lv et al. (2018)) for mathematical simplicity. However, it may be relaxed by allowing the support to expand with the sample size, which leads to some additional treatment in the asymptotic analysis. Assumption 3 characterizes the tail behaviour of the error distribution, which relaxes the commonly used bound in the machine learning literature (Smale and Zhou (2007); Rosasco et al. (2013); Lv et al. (2018)). It is general and satisfied by a variety of distributions (Wang and Leng (2016); Zhang, Liu and Wu (2016)). For example, if ϵ follows a sub-Gaussian distribution or any bounded distribution, Assumption 3 is satisfied with $q(t) = O(t^2)$; if ϵ follows a sub-exponential distribution, Assumption 3 is satisfied with $q(t) = O(\min\{t/C, t^2/C^2\})$, for some constant C.

Theorem 1. *Suppose Assumptions* 1–3 *are satisfied. Then, with probability at least* $1 - \delta_n/2$*, there holds*

$$\begin{aligned}&\|\widehat{f} - f^*\|_K \\ &\leq 2\log\frac{8}{\delta_n}\left(3\kappa_1\lambda_n^{-1}n^{-1/2}\left(\kappa_1\|f^*\|_K + q^{-1}\left(\log\frac{4c_1n}{\delta_n}\right)\right) + \lambda_n^{r-1/2}\|L_K^{-r}f^*\|_2\right).\end{aligned} \tag{3.1}$$

Additionally, let $\lambda_n = n^{-1/(2r+1)}$*. Then, with probability at least* $1 - \delta_n$*, there holds*

$$\begin{aligned}&\max_{1\leq l\leq p} \left|\|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2\right| \\ &\leq b_{n,1}\max\left\{\kappa_1\|f^*\|_K, q^{-1}\left(\log\frac{4c_1n}{\delta_n}\right)\right\}\log\left(\frac{8p}{\delta_n}\right)n^{-(2r-1)/(2(2r+1))},\end{aligned} \tag{3.2}$$

where $b_{n,1} = 4\max\{\kappa_2^2, \kappa_2^2\|f^*\|_K, \|f^*\|_K^2\}\max\{3\kappa_1, 2\sqrt{2}\kappa_2^2, \|L_K^{-r}f^*\|_2\}$ *and* $q^{-1}(\cdot)$ *denotes the inverse function of* $q(\cdot)$*.*

Theorem 1 establishes the convergence rate of the difference between the estimated regression function and the true regression function in terms of the RKHS-norm. Note that similar results have been established in the learning theory literature (Smale and Zhou (2005, 2007)). However, these results assume that the response is uniformly bounded above, which can be too restrictive in practice. Theorem 1 relaxes the restrictive boundness condition by characterizing the tail behaviour of the error term. Theorem 1 also shows that $\|\widehat{g}_l\|_n^2$ converges to $\|g_l^*\|_2^2$ with high probability, which is crucial to establishing the asymptotic sparsistency. Note that $b_{n,1}$ is spelled out precisely for the subsequent analysis of the asymptotic sparsistency and its dependency on f^*. Note that the convergence result still holds, even when p diverges with n. In addition, the quantities $\|f^*\|_K^2$ and $\|L_K^{-r} f^*\|_2$ in (3.1) and (3.2) may depend on p_0 through f^*, and thus may also diverge with n. However, such dependencies are, in general, difficult to quantify explicitly in a fully general case (Fukumizu and Leng (2014)).

Remark 3. The rate of convergence in Theorem 1 can be strengthened to obtain an optimal strong convergence rate in a minimax sense, as in Fischer and Steinwart (2020). However, it requires that the random error ϵ follows a sub-Gaussian distribution, and that the decay rate of the eigenvalues of L_K has an upper bound of polynomial order; that is, $\mu_j \leq Cj^{-1/\tau}$, for some positive constant C and $\tau \in (0,1)$. Then, the rate of convergence in (3.2) can be further improved.

Assumption 4. *There exists some positive constant* $\xi_1 < (2r-1)/(2(2r+1))$, *such that* $\min_{l \in \mathcal{A}^*} \|g_l^*\|_2^2 > b_{n,1} \max\{\kappa_1 \|f^*\|_K, q^{-1}(\log 4c_1 n/\delta_n)\} n^{-\xi_1} \log p$.

Assumption 4 requires that the true gradient function contains sufficient information about the truly informative variables. Unlike most nonparametric models, we measure the significance of each gradient function to distinguish the informative and uninformative variables without any explicit model specification. Note that the required minimal signal strength in Assumption 4 is much tighter than that in many nonparametric sparse learning methods (Huang, Horowitz and Wei (2010); Yang, Lv and Wang (2016)), which often require the signal to be bounded away from zero.

Now, we establish the asymptotic sparsistency of the proposed sparse learning method.

Theorem 2. *Suppose the assumptions of Theorem* 1 *and Assumption* 4 *are satisfied. Let* $v_n = b_{n,1}/2 \max\{\kappa_1 \|f^*\|_K, q^{-1}(\log 4c_1 n/\delta_n)\} n^{-\xi_1} \log p$. *Then, we have*

$$P\big(\widehat{\mathcal{A}}_{v_n} = \mathcal{A}^*\big) \to 1, \quad as \quad n \to \infty.$$

Theorem 2 shows that the selected active set can exactly recover the true active set with probability tending to one. This result is particularly interesting, given that it is established for any RKHS with different kernel functions. A direct application of the proposed method and Theorem 2 is to conduct nonparametric sparse learning with sparsistency (Szekely, Rizzo and Bakirov (2007); He, Wang and Hong (2013); Yang, Lv and Wang (2016)). If no prior knowledge about the true regression function is available, the proposed method can be applied with an RKHS associated with the Gaussian kernel. Asymptotic sparsistency can be established following Theorem 2, provided that f^* is contained in the RKHS associated with the Gaussian kernel. This 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 and Christmann (2008)). The above theoretical results can be refined further when f^* belongs to a specific RKHS. Some theoretical examples are provided in Section 4.

4. Theoretical Examples

This section provides theoretical examples to illustrate the proposed method with the linear and quadratic kernels. Moreover, we discuss possible treatments to improve the theoretical results, with some additional technical assumptions.

4.1. Linear kernel

Variable selection for linear models is of great interest in the statistical literature, owing to its simplicity and interpretability. In particular, the true regression function is assumed to be a linear function, $f^*(\mathbf{x}) = \beta_0 + \mathbf{x}^T \boldsymbol{\beta}^*$, and the true active set is defined as $\mathcal{A}^* = \{l : \beta_l^* \neq 0\}$. We also centralize the response and each variable, so that β_0 can be discarded from the linear model, for simplicity.

We now apply the general results in Section 3 to establish the sparsistency of the proposed algorithm under the linear model. We first scale the original data as $\widetilde{\mathbf{y}} = p_n^{-1/2}\,\mathbf{y}$ and $\widetilde{\mathbf{x}} = p_n^{-1/2}\,\mathbf{x}$, and let $\mathcal{H}_K$ be the RKHS induced by the scaled linear kernel $K(\widetilde{\mathbf{x}}, \widetilde{\mathbf{u}}) = \widetilde{\mathbf{x}}^T\widetilde{\mathbf{u}} = p_n^{-1}\,\mathbf{x}^T\,\mathbf{u}$. Then, the true regression function can be rewritten as $f^*(\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}}^T\boldsymbol{\beta}^*$. With the scaled data, the ridge regression formula in (2.1) becomes

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - \boldsymbol{\beta}^T\widetilde{\mathbf{x}}_i)^2 + p_n^{-1}\lambda_n\|\boldsymbol{\beta}\|^2. \tag{4.1}$$

By the representer theorem, the solution of (4.1) is

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\big(\mathbf{X}\mathbf{X}^T + n\lambda_n\mathbf{I}_n\big)^{-1}\mathbf{y}, \tag{4.2}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$. This is equivalent to the standard formula for the ridge regression $\widehat{\boldsymbol{\beta}} = \big(\mathbf{X}^T\mathbf{X} + n\lambda_n\,\mathbf{I}_n\,\big)^{-1}\mathbf{X}^T\mathbf{y}$, according to the Sherman–Morrison–Woodbury formula (Wang and Leng (2016)). If we let $\lambda_n = 0$, the estimate in (4.2) is the same as the HOLP estimate in Wang and Leng (2016). In other words, the HOLP method can be regarded as a special case of our proposed algorithm, with the RKHS induced by the linear kernel.

Corollary 1 is a direct application of Theorem 1 under the linear kernel.

Corollary 1. *Suppose that Assumption* S1 *in the Supplementary Material holds. Let* $\lambda_n = O(p_n^{1/3} n^{-(1+\tau_1)/3} (\log n)^{2/3})$*. Then, for any* $\delta_n \geq 4(\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)(\log n)^{-2}$*, there exists some positive constant* c_3 *such that, with probability at least* $1-\delta_n$*, there holds*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \leq c_3 \log\left(\frac{4}{\delta_n}\right) p_n^{1/6} n^{-(1-2\tau_1)/6} (\log n)^{1/3}.$$

Additionally, suppose that Assumption S2 *in the Supplementary Material holds. If we let* $v_n = (s_1/2) p_n^{1/6} n^{-(1-2\tau_1)/6} (\log n)^{\xi_2}$*, then we have*

$$P\left(\widehat{\mathcal{A}}_{v_n} = \mathcal{A}^*\right) \to 1, \quad as \quad n \to \infty,$$

where s_1 *and* ξ_2 *are provided in Assumption* S2.

Note that Corollary 1 holds when p_n diverges at order

$$o(\min\{n^{1-2\tau_1}(\log n)^{-6\xi_2}, n^{1+\tau_1}(\log n)^{-2}\}).$$

In particular, when τ_1 is sufficiently small, p_n can diverge at the polynomial rate $o(n)$. This result is comparable with that of Shao and Deng (2013) under the finite second moment error assumption. The strong convergence rate obtained in Corollary 1 is also comparable with that in Theorem 2 of Shao and Deng (2013), and a similar result holds for the required minimal signal strength.

Remark 4. Note that the proposed algorithm requires $f^* \in \mathcal{H}_K$. Thus $\|\boldsymbol{\beta}^*\|$ needs to be bounded, which implies that p_0 should be fixed in the linear case. Interestingly, if we take $\lambda_n = 0$ and all of the technical assumptions stated in Wang and Leng (2016) are met, including that $\mathbf{x}$ follows a spherically symmetric distribution and that the noise ϵ has a q-exponential tail, we can directly apply the theoretical results of the HOLP method to establish a similar selection consistency in Corollary 1. As a direct consequence, p_n and p_0 are allowed to diverge at some exponential and polynomial rate of n, respectively.

4.2. Quadratic kernel

Variable selection for quadratic models is of great interest in the statistical literature (Hao and Zhang (2014); Kong et al. (2017); She, Wang and Jiang (2018)), where the true regression function is assumed to be $f^*(\mathbf{x}) = \beta_0 + \sum_{l=1}^{p_n} \beta_l^* x^l + \sum_{l \leq k} \gamma_{lk}^* x^l x^k$, where γ_{lk}^* represents the true interaction coefficients, and $\gamma_{lk}^* \neq 0$ implies that x^l and x^k have an interaction effect. The true active set is defined as

$$\mathcal{A}^* = \left\{ l : |\beta_l^*| + \sum_{k=1}^{p_n} |\gamma_{lk}^*| > 0 \right\},$$

which contains variables contributing to f^* through either the main factors or the interaction terms. For simplicity, we denote

$$\overline{\mathbf{x}} = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_{p_n}, x_1^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_{p_n}, x_2^2, \sqrt{2}x_2x_3, \ldots, x_{p_n}^2)^T$$

and $\boldsymbol{\theta}^* = (\beta_0^*, \boldsymbol{\beta}^{*T}, \boldsymbol{\gamma}^{*T})^T$, with $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{p_n}^*)^T/\sqrt{2}$ and

$$\boldsymbol{\gamma}^* = \left(\gamma_{11}^*, \frac{\gamma_{12}^*}{\sqrt{2}}, \ldots, \gamma_{22}^*, \frac{\gamma_{23}^*}{\sqrt{2}}, \ldots, \frac{\gamma_{(p_n-1)p_n}^*}{\sqrt{2}}, \gamma_{p_np_n}^* \right).$$

Then, we scale the original data as $\check{\mathbf{y}} = p_n^{-1}\mathbf{y}$ and $\check{\mathbf{x}} = p_n^{-1}\overline{\mathbf{x}}$, and let $\mathcal{H}_K$ be the RKHS induced by a scaled quadratic kernel $K(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x}^T\mathbf{u})^2/p_n^2 = \check{\mathbf{x}}^T\check{\mathbf{u}}$. The true regression model can be rewritten as $f^*(\check{\mathbf{x}}) = \check{\mathbf{x}}^T\boldsymbol{\theta}^*$. Note that the quadratic model can be transformed into a linear form. Then the established results in Section 4.1 can be applied directly. Specifically, with the scaled data, the ridge regression formula in (2.1) becomes

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (\check{y}_i - \boldsymbol{\theta}^T\check{\mathbf{x}}_i)^2 + p_n^{-2}\lambda_n \|\boldsymbol{\theta}\|^2. \tag{4.3}$$

Then, the estimated active set is defined as $\widehat{\mathcal{A}}_{v_n} = \left\{ l : |\widehat{\beta}_l| + \sum_{k=1}^{p_n} |\widehat{\gamma}_{lk}| > v_n \right\}$, with some prespecified thresholding value v_n.

With a slight modification to the proof of Corollary 1, we obtain the following convergence results for the scaled quadratic kernel.

Corollary 2. *Suppose that Assumption* S3 *in the Supplementary Material holds. Let* $\lambda_n = O(p_n^{2/3} n^{-(1+\tau_2)/3} (\log n)^{2/3})$*. Then, for any* $\delta_n \geq 4(\sigma^2 + \|\boldsymbol{\theta}^*\|_2^2)(\log n)^{-2}$*, there exists some positive constant* c_4 *such that, with probability at least* $1 - \delta_n$*,there holds*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq c_4 \log\left(\frac{4}{\delta_n}\right) p_n^{1/3} n^{-(1-2\tau_2)/6} (\log n)^{1/3}.$$

Additionally, suppose that Assumption S4 *in the Supplementary Material holds. If we let* $v_n = (s_2/2) p_n^{1/3} n^{-(1-2\tau_2)/6} (\log n)^{\xi_3}$*, then we have*

$$P\left(\widehat{\mathcal{A}}_{v_n} = \mathcal{A}^*\right) \to 1, \quad as \ \ n \to \infty,$$

where s_2 and ξ_3 are provided in Assumption S4.

Note that the treatment in this subsection can be extended further to include the polynomial regression model of degree d by using the scaled polynomial kernel $K(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x}^T \mathbf{u})^d / p_n^d$. Similar theoretical results can be established for the proposed algorithm with the scaled polynomial kernel.

5. An extension: Interaction selection

We now extend the proposed method to identify the truly informative interaction effects. In the literature, a number of attempts have been made to identify the true interaction effects in parametric and nonparametric regression models (Lin and Zhang (2006); Choi, Li and Zhu (2010); Radchenko and James (2010); Hao and Zhang (2014); Hao, Feng and Zhang (2018)). However, most existing methods require some prespecified working models, and some are computationally demanding. For example, the COSSO method (Lin and Zhang (2006)) and the SpIn method (Radchenko and James (2010)) assume a second-order additive structure, and need to enumerate $O(p^2)$ two-way interaction terms in the model, making their methods feasible only when p is relatively small. In contrast, our method can be extended directly, and provides an efficient alternative for interaction selection without requiring an explicit model assumption.

Following the idea in Section 2, the true interaction effects can be defined as those with a nonzero second-order gradient function $g_{lk}^*(\mathbf{x}) = \partial^2 f^*(\mathbf{x}) / \partial x^l \partial x^k$. Specifically, given the true active set $\mathcal{A}^*$, we denote

$$\mathcal{A}_2^* = \left\{ l \in \mathcal{A}^* : \|g_{lk}^*\|_2 > 0, \text{ for some } k \in \mathcal{A}^* \right\},$$

which contains the variables that contribute to the interaction effects in f^*. Furthermore, let $\mathcal{A}_1^* = \mathcal{A}^* \setminus \mathcal{A}_2^*$, which contains the variables that contribute to the main effects of f^* only.

Therefore, the main goal of interaction selection is to correctly estimate both $\mathcal{A}_1^*$ and $\mathcal{A}_2^*$. First, let $K(\cdot, \cdot)$ be a fourth-order differentiable kernel function. Then, it follows from Lemma 1 that, for any $f \in \mathcal{H}_K$,

$$g_{lk}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x^l \partial x^k} = \langle f, \partial_{lk} K_{\mathbf{x}} \rangle_K \leq \|\partial_{lk} K_{\mathbf{x}}\|_K \|f\|_K,$$

where $\partial_{lk} K_{\mathbf{x}} = \partial^2 K(\mathbf{x}, \cdot) / (\partial x^l \partial x^k)$. Then, given $\widehat{f}$ from (2.1), its second-order gradient function is

$$\widehat{g}_{lk}(\mathbf{x}) = \frac{\partial^2 \widehat{f}(\mathbf{x})}{\partial x^l \partial x^k} = \widehat{\boldsymbol{\alpha}}^T \partial_{lk} \mathbf{K}_n(\mathbf{x}),$$

where $\partial_{lk}\,\mathbf{K}_n(\mathbf{x}) = \partial^2\mathbf{K}_n(\mathbf{x})/(\partial x^l \partial x^k)$. Its empirical norm is $\|\widehat{g}_{lk}\|_n^2 = (1/n)\sum_{i=1}^n (\widehat{g}_{lk}(\mathbf{x}_i))^2$. With some predefined thresholding value v_n^{int}, the estimated $\mathcal{A}_1^*$ and $\mathcal{A}_2^*$ are set as

$$\widehat{\mathcal{A}}_2 = \left\{l \in \widehat{\mathcal{A}} : \|\widehat{g}_{lk}\|_n^2 > v_n^{int},\ \text{for some } k \in \widehat{\mathcal{A}}\right\} \text{ and } \widehat{\mathcal{A}}_1 = \widehat{\mathcal{A}} \setminus \widehat{\mathcal{A}}_2,$$

respectively. The following technical assumption establishes the interaction selection consistency for the proposed method.

Assumption 5. *There exists some constant* κ_3, *such that* $\sup_{\mathbf{x}\in\mathcal{X}} \|\partial_{lk} K_{\mathbf{x}}\|_K \le \kappa_3$, *for any* l *and* k.

Assumption 5 can be regarded as an extension of Assumption 2 by requiring the boundedness of the second-order gradients of $K_{\mathbf{x}}$.

Theorem 3. *Suppose the assumptions of Theorem* 2 *and Assumption* 5 *hold. Let* $P(\widehat{\mathcal{A}} \neq \mathcal{A}^*) = \Delta_n$. *Then, with probability at least* $1 - \delta_n - \Delta_n$, *there holds*

$$\begin{aligned} &\max_{l,k\in\widehat{\mathcal{A}}} \left|\|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2\right| \\ &\le b_{n,2} \max\left\{\kappa_1\|f^*\|_K, q^{-1}\left(\log\frac{4c_1 n}{\delta_n}\right)\right\} \log\left(\frac{8p_0^2}{\delta_n}\right) n^{-(2r-1)/(2(2r+1))}, \end{aligned}$$

where $b_{n,2} = 4\max\{\kappa_3^2, \|f^*\|_K^2, \kappa_3^2\|f^*\|_K\} \max\{3\kappa_1, 2\sqrt{2}\kappa_3^2, \|L_K^{-r} f^*\|_2\}$.

Theorem 3 shows that $\|\widehat{g}_{lk}\|_n^2$ converges to $\|g_{lk}^*\|_2^2$ with high probability, which is crucial to establishing the interaction selection consistency.

Assumption 6. *There exists some positive constant* $\xi_4 < (2r-1)/(2(2r+1))$, *such that* $\min_{l,k\in\mathcal{A}_2^*} \|g_{lk}^*\|_2^2 > b_{n,2}\max\{\kappa_1\|f^*\|_K, q^{-1}(\log 4c_1 n/\delta_n)\} n^{-\xi_4} \log p_0$.

Assumption 6 can be regarded as an extension of Assumption 3 requiring the true second-order gradient functions to have sufficient information about the interaction effects.

Theorem 4. *Suppose the assumptions of Theorem* 3 *and Assumption* 6 *hold. By taking* $v_n^{int} = (b_{n,2}/2)\max\{\kappa_1\|f^*\|_K, q^{-1}(\log 4c_1 n/\delta_n)\} n^{-\xi_4}\log p_0$, *we have*

$$P\left(\widehat{\mathcal{A}}_2 = \mathcal{A}_2^*, \widehat{\mathcal{A}}_1 = \mathcal{A}_1^*\right) \to 1, \quad \textit{as} \;\; n \to \infty.$$

Theorem 4 shows that the proposed interaction selection method exactly detects the interaction structure with probability tending to one. Note that this result is established without requiring the strong heredity assumption, which is often assumed in existing parametric interaction selection methods

(Choi, Li and Zhu (2010); Hao and Zhang (2014)). It is also clear that the proposed method can be extended to detect higher-order interaction effects, which is of particular interest in real applications (Ritchie et al. (2001)).

6. Numerical Experiments

In this section, we examine the numerical performance of the proposed method and compare it with that of several existing methods, including distance correlation learning (Szekely, Rizzo and Bakirov (2007)) and the quantile-adaptive screening (He, Wang and Hong (2013)). As these two methods are designed for screening only, we truncate them using some thresholding values to conduct sparse learning. For simplicity, we denote these three methods as GM, DC-t, and QaSIS-t, respectively. Note that the computational cost of most existing gradient-based methods (Rosasco et al. (2013); Yang, Lv and Wang (2016)) can be very expensive. Thus they are not included in the numerical comparison with large dimension.

In all simulation examples, no prior knowledge about the true regression function is assumed, and the Gaussian kernel $K(\mathbf{u}, \mathbf{v}) = \exp\left(- \| \mathbf{u} - \mathbf{v} \|^2/2\sigma_n^2\right)$ is used to induce the RKHS, where σ_n is set as the median of all pairwise distances in the training sample. For the proposed method, we set the ridge parameter $\lambda_n = 0.001$ in all simulated examples, and use the stability criterion in Section 2.3 to conduct a grid search for the optimal thresholding parameter v_n, where the grid is set as $\{10^{-3+0.1s} : s = 0, \dots, 60\}$.

6.1. Simulated examples

Two simulated examples are examined under various scenarios.

Example 1. We first generate $x_i = (x_{i1}, \dots, x_{ip})^T$, with $x_{ij} = (W_{ij} + \eta U_i)/(1 + \eta)$, where W_{ij} and U_i are drawn independently from $U(-0.5, 0.5)$. The response y_i is generated as $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $f^*(\mathbf{x}_i) = 6f_1(x_{i1}) + 4f_2(x_{i2})f_3(x_{i3}) + 6f_4(x_{i4}) + 5f_5(x_{i5})$, with $f_1(u) = u, f_2(u) = 2u + 1, f_3(u) = 2u - 1, f_4(u) = 0.1\sin(\pi u) + 0.2\cos(\pi u) + 0.3(\sin(\pi u))^2 + 0.4(\cos(\pi u))^3 + 0.5(\sin(\pi u))^3$, and $f_5(u) = \sin(\pi u)/(2 - \sin(\pi u))$, and ϵ_i is drawn independently drawn from $N(0, 1)$. Clearly, the first five variables are truly informative.

Example 2. The generating scheme is similar to that in Example 1, except that W_{ij} and U_i are drawn independently from $U(0, 1)$ and $f^*(\mathbf{x}) = 20x_1x_2x_3 + 5x_4^2 + 5x_5$. The first five variables are truly informative.

For each example, we consider scenarios with $(n, p) = (400, 500), (400, 1000), (500, 10000), (500, 50000)$, and $(500, 100000)$. For each scenario, $\eta = 0$ and 1

Table 1. The averaged SNR of the simulated examples under different scenarios.

(n, η)	$(400, 0)$	$(400, 1)$	$(500, 0)$	$(500, 1)$
Example 1	5.00	3.87	5.06	3.87
Example 2	3.58	4.23	3.55	4.20

are examined. When $\eta = 0$, the variables are completely independent, whereas when $\eta = 1$, a correlation structure is added to the variables. Each scenario is replicated 50 times. The average signal-to-noise ratios (SNRs) of the simulated examples are summarized in Table 1. The average performance measures are summarized in Tables 2 and 3, where Size is the average number of selected informative variables, TP is the number of truly informative variables selected, FP is the number of truly uninformative variables selected, and C, U, and O are the times of correct fitting, under-fitting, and over-fitting, respectively.

Clearly, the SNRs of the simulated examples are comparable to those in Lin and Zhang (2006); Huang, Horowitz and Wei (2010). GM outperforms the other methods in both examples. In Example 1, GM is able to identify all of the truly informative variables in most replications. However, the other two methods tend to miss some truly informative variables, probably because of the interaction effect between x^2 and x^3. In Example 2, with a three-way interaction term involved in $f^*(\mathbf{x})$, GM is still able to identify all of the truly informative variables with high accuracy, but the other two methods tend to underfit by missing some truly informative variables in the interaction term. Note too that GM tends to overselect the variables in some cases, which is usually less severe than under-selecting truly informative variables.

Note that if we do not threshold DC and QaSIS, they tend to overfit in almost every replication, because both screening methods tend to keep a substantial number of uninformative variables to attain the sure screening property. Furthermore, when the correlation structure with $\eta = 1$ is considered, identifying the truly informative variables becomes more difficult, and both DC-t and QaSIS-t become unstable. However, GM still outperforms these two competitors, and exactly identifies all of the truly informative variables in most replications.

6.2. Supermarket data set

We now apply the proposed method to the supermarket data set of Wang (2009). The data set is collected from a major supermarket located in northern China, and consists of daily sales records of $p = 6{,}398$ products on $n = 464$ days. In this data set, the response is the number of customers on each day, and the

Table 2. The averaged performance measures of various methods in Example 1.

(n, p, η)	Method	Size	TP	FP	C	U	O
(400, 500, 0)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.28	4.28	0.00	22	28	0
	DC-t	4.80	4.80	0.00	40	10	0
(400, 1000, 0)	GM	4.98	4.98	0.00	49	1	0
	QaSIS-t	4.32	4.32	0.00	21	29	0
	DC-t	4.78	4.78	0.00	39	11	0
(500, 10000, 0)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.28	4.28	0.00	24	26	0
	DC-t	4.68	4.68	0.00	36	0	14
(500, 50000, 0)	GM	5.06	4.98	0.08	45	1	4
	QaSIS-t	4.08	4.08	0.00	18	32	0
	DC-t	4.48	4.48	0.00	28	22	0
(500, 100000, 0)	GM	5.18	5.00	0.18	43	0	7
	QaSIS-t	3.98	3.98	0.00	8	42	0
	DC-t	4.52	4.52	0.00	28	22	0
(400, 500, 1)	GM	4.98	4.98	0.00	49	1	0
	QaSIS-t	2.80	2.72	0.08	0	50	0
	DC-t	2.94	2.94	0.00	0	50	0
(400, 1000, 1)	GM	4.96	4.96	0.00	48	2	0
	QaSIS-t	2.34	2.26	0.08	0	50	0
	DC-t	2.96	2.96	0.00	0	50	0
(500, 10000, 1)	GM	4.94	4.94	0.00	47	3	0
	QaSIS-t	2.38	2.28	0.10	0	50	0
	DC-t	3.08	3.08	0.00	0	50	0
(500, 50000, 1)	GM	4.96	4.92	0.04	44	4	2
	QaSIS-t	2.42	2.36	0.08	0	50	0
	DC-t	2.94	2.94	0.00	0	50	0
(500, 100000, 1)	GM	4.94	4.92	0.02	46	3	1
	QaSIS-t	10.26	2.46	7.80	0	50	0
	DC-t	3.12	3.12	0.00	0	50	0

variables are the daily sales volumes of each product. Our primary interest is to identify those products with sale volumes that are related to the number of customers. Then, we design sale strategies based on those products. The data set is pre-processed so that both the response and the predictors have a zero mean and unit variance.

In addition to GM, DC-t, and QaSIS-t, we include comparisons with SCAD (Fan and Li (2001)) and MCP (Zhang (2010)). Because the truly informative variables are unknown for the supermarket data set, we report the prediction performance of each method. Specifically, the data set is split randomly into two

Table 3. The averaged performance measures of various methods in Example 2.

(n, p, η)	Method	Size	TP	FP	C	U	O
(400, 500, 0)	GM	5.00	5.00	0.00	50	0	0
	QaSIS-t	4.26	4.26	0.00	22	28	0
	DC-t	4.92	4.92	0.00	48	2	0
(400, 1000, 0)	GM	5.14	5.00	0.14	44	0	6
	QaSIS-t	4.04	4.04	0.00	20	30	0
	DC-t	4.96	4.96	0.00	48	2	0
(500, 10000, 0)	GM	5.10	5.00	0.10	45	0	5
	QaSIS-t	3.82	3.82	0.00	13	37	0
	DC-t	4.92	4.92	0.00	46	4	0
(500, 50000, 0)	GM	5.40	5.00	0.40	37	0	13
	QaSIS-t	3.04	3.04	0.00	8	42	0
	DC-t	4.66	4.66	0.00	38	12	0
(500, 100000, 0)	GM	5.32	5.00	0.32	41	0	9
	QaSIS-t	3.02	3.02	0.00	5	45	0
	DC-t	4.66	4.66	0.00	34	16	0
(400, 500, 1)	GM	5.00	4.98	0.02	48	1	1
	QaSIS-t	5.78	2.90	2.88	3	38	9
	DC-t	31.30	4.00	27.30	1	0	49
(400, 1000, 1)	GM	5.10	5.00	0.10	45	0	5
	QaSIS-t	7.78	2.22	5.56	1	42	7
	DC-t	38.74	5.00	33.74	2	0	48
(500, 10000, 1)	GM	5.10	4.96	0.14	42	2	6
	QaSIS-t	12.94	2.08	10.86	0	45	5
	DC-t	74.98	5.00	69.98	0	0	50
(500, 50000, 1)	GM	5.16	4.98	0.18	43	1	6
	QaSIS-t	32.52	2.08	30.44	0	42	8
	DC-t	79.62	5.00	74.62	0	1	49
(500, 100000, 1)	GM	5.10	4.96	0.14	44	2	4
	QaSIS-t	42.32	2.54	39.78	0	44	6
	DC-t	79.94	4.88	75.06	0	6	44

parts, with 164 observations for testing, and the remainder used for training. We first apply each method to the full data set to select the informative variables. Then, we refit a kernel ridge regression model for the nonparametric methods and a linear ridge regression for the parametric methods using the variables selected from the training set. The prediction performance of each ridge regression model is measured on the testing set. The procedure is replicated 1,000 times, and the number of selected variables, average prediction errors, and out-of-sample R^2 are summarized in Table 4.

As Table 4 shows, GM selects 10 variables, DC-t and QaSIS-t select seven variables, and the SCAD and MCP select 59 and 28 variables, respectively. The

Table 4. The number of selected variables and the corresponding averaged prediction errors by various methods in the supermarket data set.

Dataset	Method	Size	Testing error (Std)	Out of sample R^2
	GM	10	0.1369 (0.0005)	0.8631
	QaSIS-t	7	0.1674 (0.0006)	0.8326
	DC-t	7	0.1713 (0.0006)	0.8287
	SCAD	59	0.1872 (0.0006)	0.8128
	MCP	28	0.2040 (0.0006)	0.7960

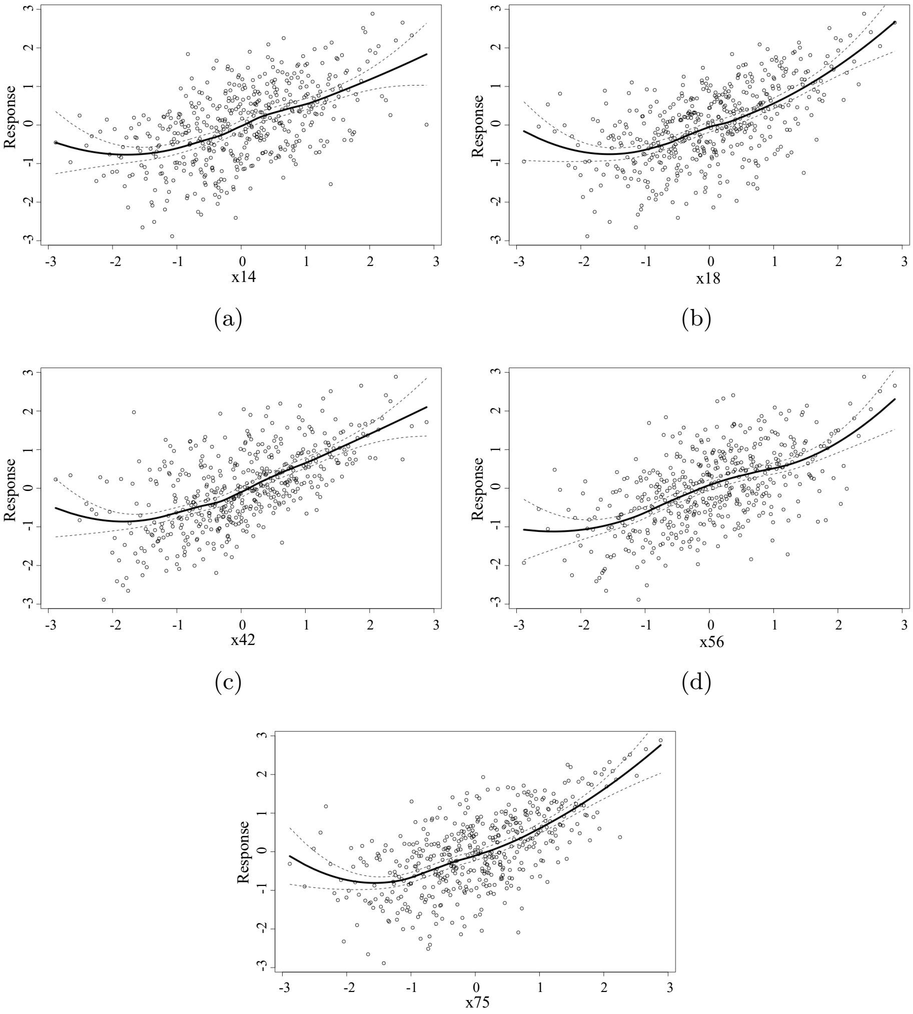
average prediction error of GM is smaller than that of the other four methods. This implies that DC-t and QaSIS-t may miss some truly informative variables, thus reducing their prediction accuracy, and that the SCAD and MCP may include too many noise variables. Specifically, of the 10 variables selected by GM, $X^{14}, X^{18}, X^{42}, X^{56}$, and X^{75} are missed by both DC-t and QaSIS-t. Scatter plots of the response against these five variables are presented in Figure 1.

It is evident that the response and these variables show some clear relationship, which supports the advantage of GM in identifying the truly informative variables.

7. Conclusion

We have proposed a novel gradient-based sparse learning method that simultaneously enjoys methodological flexibility, numerical efficiency, and asymptotic consistency. It provides a novel and promising way to conduct sparse learning for nonparametric models. The proposed method is simple and efficient in that the kernel ridge regression has an analytic solution, and the estimated gradient functions can be computed directly using the derivative reproducing property (Zhou (2007)). It can be scaled easily to analyze data sets with huge dimensions. The theoretical results are established without requiring restrictive model assumptions, which justifies the robustness of the proposed method to the underlying data distribution.

One interesting direction for future work is to consider a more general scenario with f^* out of the specified RKHS $\mathcal{H}_K$, such as a non-differentiable f^*. One possible remedial route is to consider the true active set $\mathcal{A}^* = \{l : D_l(f^*) > 0\}$, where $D_l(f^*) = \max_{\mathbf{x}^{-l}} \left| \max_{x^l} f^*(x^l, \mathbf{x}^{-l}) - \min_{x^l} f^*(x^l, \mathbf{x}^{-l}) \right| > 0$ measures the largest possible change of $f(\mathbf{x})$ along x^l, and $\mathbf{x}^{-l}$ denotes all variables except for x^l. Then, the equivalence between $D_l(f^*)$ and the gradients of some intermediate function $f^0 \in \mathcal{H}_K$ can be examined to bridge the gap between f^* and $\mathcal{H}_K$. We would also like to extend the proposed method to deal with mixed-type



(a)

(b)

(c)

(d)

(e)

Figure 1. Scatter plots of the response against a number of selected variables by the GM in the supermarket data set. The solid lines are the fitted curve by local smoothing, and the dashed lines are the fitted means, plus or minus one standard deviatio

predictors, and $D_l(f^*)$ can be used to measure the significance of each variable.

Supplementary Material

Proofs of Theorems 3 and 4, some necessary lemmas and their proofs, and a verification of the theoretical examples are provided in the online Supplementary Material.

Acknowledgments

Xin He's research was supported in part by NSFC-11901375 and Shanghai Pujiang Program 2019PJC051. Junhui Wang's research was supported in part by HK RGC Grants GRF-11303918, GRF-11300919 and GRF-11304520. Shaogao Lv's research was partially supported by NSFC-11871277. The authors also thank the associate editor and two anonymous referees for their constructive suggestions.

Appendix

Appendix: technical proof

Proof of Theorem 1. For simplicity, we denote two events

$$\begin{aligned}\mathcal{C}_1 = \Bigg\{ &\mathcal{Z}^n : \ \|\widehat{f} - f^*\|_K \\ &\geq 2\log\frac{8}{\delta_n}\left(\frac{3\kappa_1}{n^{1/2}\lambda_n}\left(\kappa_1\|f^*\|_K + q^{-1}\left(\log\frac{2c_1 n}{\delta_n}\right)\right) + \lambda_n^{r-1/2}\|L_K^{-r}f^*\|_2\right)\Bigg\},\\ \mathcal{C}_2 = \Bigg\{ &\mathcal{Z}^n : \ \max_{i=1,\dots,n}|y_i| > \kappa_1\|f^*\|_K + q^{-1}\left(\log\frac{2c_1 n}{\delta_n}\right)\Bigg\}\end{aligned}$$

and $\mathcal{C}_2^c$ denotes the complement of $\mathcal{C}_2$. Then $P(\mathcal{C}_1)$ can be decomposed as

$$P(\mathcal{C}_1) = P\left(\mathcal{C}_1 \cap \mathcal{C}_2\right) + P\left(\mathcal{C}_1 \cap \mathcal{C}_2^c\right) \leq P\left(\mathcal{C}_2\right) + P\left(\mathcal{C}_1 \mid \mathcal{C}_2^c\right) = P_1 + P_2.$$

For P_1, by Assumption 3, we have

$$P\Big(\max_{i=1,\dots,n}|\epsilon_i| \geq t\Big) = P(\cup_{i=1}^n |\epsilon_i| \geq t) \leq nP(|\epsilon_i| \geq t) \leq c_1 n \exp\{-q(t)\}. \quad \text{(A.1)}$$

By Assumption 1 and (A.1), for any $\delta_n \in (0,1)$, with probability at least $1-\delta_n/4$, there holds

$$\max_{i=1,\dots,n}|y_i| \leq \kappa_1\|f^*\|_K + \max_{i=1,\dots,n}|\epsilon_i| \leq \kappa_1\|f^*\|_K + q^{-1}\left(\log\frac{4c_1 n}{\delta_n}\right),$$

implying that $P\left(\mathcal{C}_2\right) \leq \delta_n/4$.

For P_2, note that

$$\|\widehat{f} - f^*\|_K \leq \|\widehat{f} - \widetilde{f}\|_K + \|\widetilde{f} - f^*\|_K.$$

We first bound $\|\widetilde{f} - f^*\|_K$ following the similar treatment as in Smale and Zhou (2005). Suppose $\{\mu_i, e_i\}_{i\geq 1}$ are the normalized eigenpairs of the integral operator

$L_K : \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) \to \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, we have

$$L_K^{1/2} e_i = \sum_{j \geq 1} \mu_j^{1/2} \langle e_i, e_j \rangle_2 e_j = \mu_i^{1/2} e_i \in \mathcal{H}_K,$$

and

$$\|\mu_i^{1/2} e_i\|_K = \left(\sum_{j \geq 1} \frac{\langle \mu_i^{1/2} e_i, e_j \rangle_2^2}{\mu_j} \right)^{1/2} = \langle e_i, e_i \rangle_2 = 1,$$

when $\mu_i > 0$. Thus by Assumption 1, there exists some function $h = \sum_{i \geq 1} \langle h, e_i \rangle_2 e_i \in \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ such that $f^* = L_K^r h = \sum_{i \geq 1} \mu_i^r \langle h, e_i \rangle_2 e_i \in \mathcal{H}_K$. Directly calculation yields to

$$\begin{aligned} \widetilde{f} - f^* &= \big(L_K + \lambda_n I\big)^{-1} L_K f^* - f^* = \big(L_K + \lambda_n I\big)^{-1} \big(- \lambda_n f^*\big) \\ &= - \sum_{i \geq 1} \frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^r \langle h, e_i \rangle_2 e_i. \end{aligned}$$

Therefore, the RKHS-norm of $\widetilde{f} - f^*$ can be bounded as

$$\begin{aligned} \|\widetilde{f} - f^*\|_K^2 &= \sum\nolimits_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^{r-1/2} \langle h, e_i \rangle_2 \right)^2 \|\mu_i^{1/2} e_i\|_K^2 \\ &= \sum\nolimits_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \mu_i^{r-1/2} \langle h, e_i \rangle_2 \right)^2 \\ &= \lambda_n^{2r-1} \sum\nolimits_{i \geq 1} \left(\frac{\lambda_n}{\lambda_n + \mu_i} \right)^{3-2r} \left(\frac{\mu_i}{\lambda_n + \mu_i} \right)^{2r-1} \langle h, e_i \rangle_2^2 \\ &\leq \lambda_n^{2r-1} \sum\nolimits_{i \geq 1} \langle h, e_i \rangle_2^2 = \lambda_n^{2r-1} \|h\|_2^2 = \lambda_n^{2r-1} \|L_K^{-r} f^*\|_2^2. \qquad (A.2) \end{aligned}$$

It then follows from Proposition 1 in the supplemental material that

$$P_2 \leq P\left(\|\widehat{f} - \widetilde{f}\|_K \geq \log \frac{8}{\delta_n} \frac{6\kappa_1}{\lambda_n n^{1/2}} \left(\kappa_1 \|f^*\|_K + q^{-1} \left(\log \frac{4c_1 n}{\delta_n} \right) \right) \mid \mathcal{C}_2^c \right) \leq \frac{\delta_n}{4}.$$

Combining the upper bounds of P_1 and P_2 yields that $P(\mathcal{C}_1) \leq \delta_n/4 + \delta_n/4 \leq \delta_n/2$. Thus, with probability at least $1 - \delta_n/2$, there holds

$$\|\widehat{f} - f^*\|_K \leq 2 \log \frac{8}{\delta_n} \left(\frac{3\kappa_1}{n^{1/2} \lambda_n} \left(\kappa_1 \|f^*\|_K + q^{-1} \left(\log \frac{4c_1 n}{\delta_n} \right) \right) + \lambda_n^{r-1/2} \|L_K^{-r} f^*\|_2 \right).$$

Now we turn to establish the weak convergence rate of $\widehat{g}_l$ in estimating g_l^*. We first introduce some notations. Define the sample operators for gradients $\widehat{D}_l : \mathcal{H}_K \to \mathcal{R}^n$ and their adjoint operators $\widehat{D}_l^* : \mathcal{R}^n \to \mathcal{H}_K$ as

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

respectively. And the integral operators for gradients $D_l : \mathcal{H}_K \to \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X})$ and $D_l^* : \mathcal{L}^2(\rho_{\mathbf{x}}, \mathcal{X}) \to \mathcal{H}_K$ are defined as

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

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

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

Furthermore, we denote $HS(K)$ as a Hilbert space with all the Hilbert-Schmidt operators on $\mathcal{H}_K$, which endows with a norm $\| \cdot \|_{HS}$ such that $\|T\|_K \leq \|T\|_{HS}$ for any $T \in HS(K)$.

With these operators, simple algebra yields that

$$\begin{aligned}
&\big| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2 \big| \\
&= \Big| \frac{1}{n} \sum_{i=1}^n (\widehat{g}_l(\mathbf{x}_i))^2 - \int (g_l^*(\mathbf{x}))^2 \, d\rho_{\mathbf{x}} \Big| \\
&= \Big| \frac{1}{n} \sum_{i=1}^n \widehat{g}_l(\mathbf{x}_i) \big\langle \widehat{f}, \partial_l K_{\mathbf{x}_i} \big\rangle_K - \int g_l^*(\mathbf{x}) \, \langle f^*, \partial_l K_{\mathbf{x}} \rangle_K \, d\rho_{\mathbf{x}} \Big| \\
&= \Big| \big\langle \widehat{f}, \frac{1}{n} \sum_{i=1}^n \widehat{g}_l(\mathbf{x}_i) \partial_l K_{\mathbf{x}_i} \big\rangle_K - \big\langle f^*, \int g_l^*(\mathbf{x}) \partial_l K_{\mathbf{x}} d\rho_{\mathbf{x}} \big\rangle_K \Big| \\
&= \Big| \big\langle \widehat{f} - f^*, \widehat{D}_l^* \widehat{D}_l \widehat{f} \big\rangle_K + \big\langle f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \big\rangle_K + \big\langle f^*, (\widehat{D}_l^* \widehat{D}_l - D_l^* D_l) f^* \big\rangle_K \Big| \\
&= \Big| \big\langle \widehat{f} - f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \big\rangle_K + \big\langle \widehat{D}_l^* \widehat{D}_l f^*, \widehat{f} - f^* \big\rangle_K + \\
&\quad \big\langle f^*, \widehat{D}_l^* \widehat{D}_l (\widehat{f} - f^*) \big\rangle_K + \big\langle f^*, (\widehat{D}_l^* \widehat{D}_l - D_l^* D_l) f^* \big\rangle_K \Big| \\
&\leq \|\widehat{f} - f^*\|_K^2 \|\widehat{D}_l^* \widehat{D}_l\|_{HS} + 2 \|\widehat{f} - f^*\|_K \|f^*\|_K \|\widehat{D}_l^* \widehat{D}_l\|_{HS} + \\
&\quad \|\widehat{D}_l^* \widehat{D}_l - D_l^* D_l\|_{HS} \|f^*\|_K^2,
\end{aligned}$$

where the last inequality follows from the Cauthy-Schwartz inequality. It then suffices to bound the terms in the upper bound of $\big| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_2^2 \big|$ separately. Note that $\|f^*\|_K$ is a bounded quantity, and it follows from Assumption 2 and Rosasco et al. (2013) that $\max_l \big\| \widehat{D}_l^* \widehat{D}_l \big\|_{HS} = \max_l \|\partial_l K_{\mathbf{x}}\|_K^2 \leq \kappa_2^2$. Hence, we have

$$\begin{aligned}&\max_{1\le l\le p}\big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_2^2\big|\\&\le a_1\Big(\|\widehat{f}-f^*\|_K^2+2\|\widehat{f}-f^*\|_K+\max_{1\le l\le p}\|\widehat{D}_l^*\widehat{D}_l-D_l^*D_l\|_{HS}\Big),\end{aligned}$$

where $a_1=\max\{\kappa_2^2,\kappa_2^2\|f^*\|_K,\|f^*\|_K^2\}$. When $\|\widehat{f}-f^*\|_K$ is sufficiently small, the upper bound can be simplified to

$$\max_{1\le l\le p}\big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_2^2\big|\le a_1\Big(3\|\widehat{f}-f^*\|_K+\max_{1\le l\le p}\|\widehat{D}_l^*\widehat{D}_l-D_l^*D_l\|_{HS}\Big),$$

where $\|\widehat{f}-f^*\|_K$ is bounded in the first half of the proof. Furthermore, for any $\epsilon_n\in(0,1)$, by the concentration inequalities for $HS(K)$ (Rosasco et al. (2013)), we have

$$P\Big(\big\|\widehat{D}_l^*\widehat{D}_l-D_l^*D_l\big\|_{HS}\ge\epsilon_n\Big)\le 2p\exp\left(-\frac{n\epsilon_n^2}{8\kappa_2^4}\right),$$

for any $l=1,\ldots,p$. Therefore, with probability at least $1-\delta_n/2$, there holds

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

Combining all the upper bounds above, we have with probability at least $1-\delta_n$, there holds

$$\begin{aligned}\max_{1\le l\le p}\Big|\|\widehat{g}_l\|_n^2-\|g_l^*\|_2^2\Big| \le 2a_1\Bigg(3\log\frac{8}{\delta_n}\bigg(\frac{3\kappa_1}{n^{1/2}\lambda_n}\bigg(\kappa_1\|f^*\|_K+q^{-1}\Big(\log\frac{4c_1n}{\delta_n}\Big)\bigg)+\\ \lambda_n^{r-1/2}\|L_K^{-r}f^*\|_2\bigg)+\left(\frac{2\kappa_2^4}{n}\log\frac{4p}{\delta_n}\right)^{1/2}\Bigg).\end{aligned}$$

This implies the desired results immediately with $\lambda_n=n^{-1/(2r+1)}$.

Proof of Theorem 2. We first show that $\mathcal{A}^*\subset\widehat{\mathcal{A}}$ in probability. If not, suppose there exists some $l'\in\mathcal{A}^*$ but $l'\notin\widehat{\mathcal{A}}$, and thus $\|\widehat{g}_{l'}\|_n^2\le v_n$. By Assumption 4, we have with probability $1-\delta_n$ that

$$\begin{aligned}&\big|\|\widehat{g}_{l'}\|_n^2-\|g_{l'}^*\|_2^2\big|\ge\|g_{l'}^*\|_2^2-\|\widehat{g}_{l'}\|_n^2\\&>b_{n,1}\max\left\{\kappa_1\|f^*\|_K,q^{-1}\bigg(\log\frac{4c_1n}{\delta_n}\bigg)\right\}n^{-\xi_1}\log p-v_n\\&=\frac{b_{n,1}}{2}\max\left\{\kappa_1\|f^*\|_K,q^{-1}\bigg(\log\frac{4c_1n}{\delta_n}\bigg)\right\}n^{-\xi_1}\log p,\end{aligned}$$

which contradicts with Theorem 1. This implies that $\mathcal{A}^* \subset \widehat{\mathcal{A}}$ with probability at least $1-\delta_n$.

Next, we show that $\widehat{\mathcal{A}} \subset \mathcal{A}^*$ in probability. If not, suppose there exists some $l' \in \widehat{\mathcal{A}}$ but $l' \notin \mathcal{A}^*$, which implies $\|\widehat{g}_{l'}\|_n^2 > v_n$ but $\|g_{l'}^*\|_2^2 = 0$, and then with probability at least $1-\delta_n$, there holds

$$\big|\|\widehat{g}_{l'}\|_n^2 - \|g_{l'}^*\|_2^2\big| > v_n = \frac{b_{n,1}}{2}\max\left\{\kappa_1\|f^*\|_K, q^{-1}\left(\log\frac{4c_1 n}{\delta_n}\right)\right\}n^{-\xi_1}\log p.$$

This contradicts with Theorem 1 again, and thus $\widehat{\mathcal{A}} \subset \mathcal{A}^*$ with probability at least $1-\delta_n$. Combining these two results yields the desired sparsistency.

References

Barber, R. and Candës, 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.

Bondell, H. and Li, L. (2009). Shrinkage inverse regression estimation for model free variable selection. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **71**, 287–299.

Choi, N., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364.

Dasgupta, S., Goldberg, Y. and Kosorok, M. (2019). Feature elimination in kernel machines in moderately high dimensions. *The Annals of Statistics* **47**, 497–526.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultrahigh dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **70**, 849–911.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least square algorithm. *Journal of Machine Learning Research* **21**, 1–38.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association* **109**, 359–370.

Hao, N., Feng, Y. and Zhang, H. (2018). Model selection for high dimensional quadratic regression via regularization. *Journal of the American Statistical Association* **113**, 615–625.

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

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

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

Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multiresponse regression via distance correlation. *The Annals of Statistics* **45**, 897–922.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics* **33**, 1580–1616.

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

Lv, S., Lin, H., Lian, H. and Huang, J. (2018). Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *The Annals of Statistics* **2**, 781–813.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics* **38**, 526–565.

Radchenko, P. and James, G. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association* **105**, 1541–1553.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F. et al. (2001). VMultifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* **69**, 138–147.

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

Shao, J. and Deng, X. (2013). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics* **40**, 1821–1831.

She, Y., Wang, Z. and Jiang, H. (2018). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association* **113**, 445–454.

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

Shen, X., Pan, W., Zhu, Y. and Zhou, Z. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics* **65**, 807–832.

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

Smale, S. and Zhou, D. (2005). Shannon sampling II: connections to learning theory. *Applied and Computational Harmonic Analysis* **19**, 285–302.

Smale, S. and Zhou, D. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation* **26**, 153–172.

Stefanski, L., Wu, Y. and White, K. (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association* **109**, 574–589.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machine*. Springer.

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.

Szekely, G., Rizzo, M. and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

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

Wahba, G. (1998). Support vector machines, reproducing kernel hilbert spaces, and randomized GACV. *Advances in Kernel Methods: Support Vector Learning* **58**, 69–88.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of* *the American Statistical Association* **104**, 1512–1524.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78**, 589–611.

Wu, Y. and Stefanski, L. (2015). Automatic structure recovery for additive models. *Biometrika* **102**, 381–395.

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.

Yang, Y., Pilanci, M. and Wainwright, M. (2017). Randomized sketches for kernels: Fast and optimal nonparametic regression. *The Annals of Statistics* **45**, 991–1023.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang, C., Liu, Y. and Wu, Y. (2016). On quantile regression in reproducing kernel Hilbert spaces with data sparsity constraint. *Journal of Machine Learning Research* **17**, 1–45.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **220**, 1418–1429.

Xin He

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

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

Junhui Wang

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

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

Shaogao Lv

School of Statistics and Mathematics, Nanjing Audit University, Nanjing, Jiangsu 211815, China.

E-mail: lvsg716@swufe.edu.cn

(Received October 2019; accepted April 2020)