Electronic Journal of Statistics
Vol. 17 (2023) 2386–2415
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2153

Structure learning via unstructured kernel-based M-estimation*

Xin He, Yeheng Ge and Xingdong Feng†

School of Statistics and Management, Shanghai University of Finance and Economics
e-mail: he.xin17@mail.shufe.edu.cn; geyh96@163.sufe.edu.cn; feng.xingdong@mail.shufe.edu.cn

Abstract: In statistical learning, identifying underlying structures of true target functions based on observed data plays a crucial role to facilitate subsequent modeling and analysis. Unlike most of those existing methods that focus on some specific settings under certain model assumptions, a general and novel framework is proposed for recovering the true structures of target functions by using unstructured M-estimation in a reproducing kernel Hilbert space (RKHS) in this paper. This framework is inspired by the fact that gradient functions can be employed as a valid tool to learn underlying structures, including sparse learning, interaction selection and model identification, and it is easy to implement by taking advantage of some nice properties of the RKHS. More importantly, it admits a wide range of loss functions, and thus includes many commonly used methods as special cases, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification, which is also computationally efficient by solving convex optimization tasks. The asymptotic results of the proposed framework are established within a rich family of loss functions without any explicit model specifications. The superior performance of the proposed framework is also demonstrated by a variety of simulated examples and a real case study.

MSC2020 subject classifications: Primary 62G05.
Keywords and phrases: Convex optimization, consistency, gradient learning, reproducing kernel Hilbert space, structure learning.

Received June 2022.

Contents

arXiv: 1901.00615v2

*Xin He's research is supported in part by NSFC-11901375, Shanghai Pujiang Program 2019PJC051, the Fundamental Research Funds for the Central Universities, and Program for Innovative Research Team of Shanghai University of Finance and Economics. Xingdong Feng's research is supported in part by NSFC-11971292. This research is also supported by Shanghai Research Center for Data Science and Decision Technology.

†Xingdong Feng is the corresponding author of this paper.

1. Introduction

In statistical learning, true target functions are often assumed to have some specific structures to facilitate the following statistical modeling and analysis. Thus, tremendous interests have been paid to recover underlying structures from observed data, including learning sparse structures (He, Wang and Hong, 2013; Pan et al., 2019; Tang, Xue and Qu, 2021; Lemhadri et al., 2021; Deng et al., 2022; Li and Xu, 2023), interaction effects (Radchenko and James, 2010; Hao and Zhang, 2014; Kong et al., 2017; Hao, Feng and Zhang, 2018) or identifying linear and nonlinear effects (Zhang, Cheng and Liu, 2011; Lian, Liang and Ruppert, 2015; He and Wang, 2020).

However, most existing methods are designed for learning some specific structures, and their successes either rely on restrictive model assumptions or require intensive computational efforts. For example, various attempts have been made to learn the sparsity of the conditional mean function by regularization (Fan and Lv, 2010), screening (Fan and Lv, 2008), or checking variable robustness against added noises (Barber and Candës, 2015). The counterparts of these methods have also been proposed in the context of quantile regression (Wu and Liu, 2009), margin-based classification (Steinwart and Christmann, 2008a; Zhang et al., 2016), generalized linear models (Li and Liu, 2019) and so on. Furthermore, the additive model assumption is often imposed to relax the linear assumption in pursuing sparsity (Huang, Horowitz and Wei, 2010; Lv et al., 2018). However, all these methods are only designed for some specific learning tasks and lack of universality. Most recently, tremendous attentions have been paid to tackle the universality issue. Loh (2017) focuses on the theoretical aspect of regularized linear M-estimators within a family of robust loss functions. Dasgupta, Goldberg and Kosorok (2019) propose a recursive feature elimination method via repeatedly fitting a kernel ridge regression for a general loss

function, but it is less efficient in computation. Han (2019) proposes a novel non-parametric screening method under a strictly convex loss function family, but it requires the loss function to be differentiable almost everywhere, which excludes some popular loss functions, such as the hinge loss. Other popularly assumed structures of the true target function include the interaction structure (Hao, Feng and Zhang, 2018; Hao and Zhang, 2014; Kong et al., 2017; Dong and Wu, 2022) and the model identification by identifying linear or nonlinear effects (Zhang, Cheng and Liu, 2011; Lian, Liang and Ruppert, 2015; He and Wang, 2020), and these methods are developed similar to the sparse learning. However, these methods are also designed for some specific scenarios, and the lack of theoretical consistency or computational efficiency become their main obstacles.

Recently, many kernel-based sparse learning methods have been motivated by the fact that gradient functions provide an appropriate tool to identify informative variables in a model-free fashion, and thus various strategies have been adopted to learn the gradient functions under some specific scenarios. For example, Rosasco et al. (2013) add an empirical functional penalty on the gradients in a standard kernel ridge regression, and He, Lv and Wang (2020) further extend it to learn the sparse structure in support vector machines. Yang, Lv and Wang (2016) employ a pair-wise learning task to estimate gradient functions and use a functional group lasso penalty to induce sparsity, and He and Wang (2020) extend it to learn interaction structures. Most recently, He, Wang and Lv (2021) propose an efficient two-step sparse learning method in the least square regression and Chen, He and Wang (2021) adopt the similar strategy to learn the sparse structure of the conditional distribution. It is worthy pointing out that both of the methods use the derivative reproducing property for the computation of the gradient function to mainly conduct sparse learning for the mean regression and multiple quantile regression, respectively. Moreover, their theoretical consistencies are established by using some specific technical treatments, which only work for their pre-specified loss function. It is non-trivial to extend their theoretical results to accommodate a general loss family for tackling the universality issue, and the theoretical derivation should be much more involved since the general loss family is considered. Clearly, all these sparse learning methods are methodologically flexible in the sense that they make no model specifications, and thus are applicable to datasets with complicated dependence patterns. However, these methods are developed under specific scenarios in regression or classification, and their high computational costs, lack of consistency and universality are still unsolved issues.

This paper proposes a novel structure learning framework via the regularized M-estimation for a general family of loss functions in a flexible RKHS. The proposed framework is inspired by the fact that gradient functions characterize structures of their corresponding true target functions without explicit model assumptions, and the derivative reproducing properties of the RKHS (Zhou, 2007) facilitate the computation of gradient functions. The proposed framework is methodologically simple, and computationally easy to implement, which can also be scaled up by a parallelization procedure. Specifically, it consists of es-

timation of the regularized M-regression in a RKHS, and computation of the gradient functions with one-step matrix multiplication. It is computationally efficient by only fitting a standard kernel ridge regression via solving a convex optimization problem, and thus scalable to analyze large-scale datasets. More importantly, the asymptotic properties of the proposal in sparse learning, interaction selection and model identification are established based on a general family of loss functions without imposing any explicit model assumptions.

The major contributions of the proposed framework are four-fold.

(i) It works for a general loss function family including most commonly used ones in literature, such as the squared loss, check loss, hinge loss, Huber loss, logistic loss, ϵ-insensitive loss, exponential loss and so on.

(ii) It establishes a unified framework for learning the underlying structures of true target functions and admits general dependence structures. The proposed framework employs gradient functions to recover the structures in a model-free fashion and can be regarded as a joint screening method and thus is able to identify all the informative variables acting on the response with a general dependence structure, including those marginally non-informative but jointly informative ones.

(iii) It is methodologically simple and computationally easy to implement. Specifically, it avoids directly estimating gradient functions, but solving a kernel-based convex optimization problem. Then, the estimated gradient functions can be efficiently obtained by using the derivative reproducing properties of the RKHS, which significantly reduces the computational costs. For instance, in Examples 1 and 2 of our simulation study, the proposed framework is efficiently implemented to sparse learning with dimensionality up to 10^5.

(iv) It provides theoretical guarantees for structure learning under mild conditions. Specifically, the proposed framework requires the existence of the gradients of true target functions, and thus those gradient functions are adopted as a valid tool to define the corresponding true structures. Then, the proposed framework is designed to learn the underlying structures that can be well characterized by their gradient functions. With the help of empirical process and functional operators in learning theory, the estimation consistency of gradient functions is established for a general loss function family. More importantly, as a direct consequence, the asymptotic consistencies of sparse learning, interaction selection and model identification are established without imposing any explicit model specifications.

The rest of this paper is organized as follows. Section 2 introduces the rich family of loss functions and illustrates the connections between gradient functions and the corresponding functional structures. Section 3 introduces the motivations and the proposed structure learning framework. All the computational details are summarized in Section 4. In Section 5, the asymptotic theoretical results of the proposed method are given. The simulated examples and a real case study are provided in Section 6. A brief discussion is given in Section 7, and extra numerical results and all the technical proofs of Theorems 1–5 are

deferred to the Supplementary Material.

2. Preambles and methodology

2.1. A rich family of loss functions

Suppose a random pair $\mathcal{Z} = (\mathbf{x}, y)$ is drawn from some unknown joint distribution $\rho_{\mathbf{x},y}$, with $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathcal{X}$ supported on a compact subset of $\mathcal{R}^p$ and $y \in \mathcal{Y} \subset \mathcal{R}$. In statistical learning, the true target function f^* is often defined as the minimizer of the following expected error

$$f^* = \operatorname{argmin} \mathcal{E}^L(f) = \operatorname{argmin} EL(y, f(\mathbf{x})), \tag{2.1}$$

where $L(\cdot,\cdot) : \mathcal{Y} \times \mathcal{R} \to \mathcal{R}^+$ is the loss function of our interests. We first impose the following conditions on the loss L.

Assumption 1. The loss function L satisfies the following two conditions.

(1) There exist some positive constants c_1 and $q \geq 1$ such that $L(y, \omega) \leq c_1(|y|^q + |\omega|^q)$, for any $y \in \mathcal{Y}$ and $\omega \in \mathcal{R}$.
(2) $L(y, \cdot)$ is convex, and locally Lipschitz continuous; that is, for any $V \geq 0$, there exists a constant $c_2 > 0$ such that $|L(y, \omega) - L(y, \omega')| \leq c_2|\omega - \omega'|$, for any $\omega, \omega' \in [-V, V]$ and $y \in \mathcal{Y}$.

Note that the above conditions are mild and commonly used in literature to characterize loss functions (Hang and Steinwart, 2018; Dasgupta, Goldberg and Kosorok, 2019). The loss space satisfying these two conditions include many popular losses:

(i) **Squared loss**: $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ with $c_2 = 2(M_y + V)$ and $q = 2$, for any $|y| \leq M_y$ with a positive constant M_y;
(ii) **Check loss**: $L_\tau(y, f(\mathbf{x})) = (y - f(\mathbf{x}))(\tau - I_{\{y<f(\mathbf{x})\}})$ with $c_2 = 1$ and $q = 1$;
(iii) **Huber loss**: $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$, if $|y - f(\mathbf{x})| \leq \delta$; $\delta|y - f(\mathbf{x})| - \frac{1}{2}\delta^2$, otherwise, with $c_2 = \delta$ and $q = 1$;
(iv) **ϵ-insensitive loss**: $L(y, f(\mathbf{x})) = \max\{0, |y - f(\mathbf{x})| - \epsilon\}$ with $c_2 = 1$ and $q = 1$;
(v) **Logistic loss**: $L(y, f(\mathbf{x})) = (\log 2)^{-1} \log\left(1 + \exp\left(-yf(\mathbf{x})\right)\right)$ with $c_2 = (\log 2)^{-1} e^V/(1 + e^V)$ and $q = 1$;
(vi) **Hinge loss**: $L(y, f(\mathbf{x})) = (1 - yf(\mathbf{x}))_+$ with $c_2 = 1$ and $q = 1$;
(vii) **Exponential loss**: $L(y, f(\mathbf{x})) = \exp\left(-yf(\mathbf{x})\right)$ with $c_2 = e^V$ and $q = 1$.

The explicit form of f^* varies from one loss function to another. For example, when the squared loss is used, $f^*(\mathbf{x}) = E(y|\,\mathbf{x})$; when the check loss is used, $f^*(\mathbf{x}) = Q_\tau(y|\,\mathbf{x})$ with $Q_\tau(y|\,\mathbf{x}) = \inf\{y : P(Y \leq y|\,\mathbf{x}) \geq \tau\}$; and when the hinge loss is used, $f^*(\mathbf{x}) = \operatorname{sign}(P(y = 1|\,\mathbf{x}) - 1/2)$, where $\operatorname{sign}(\cdot)$ is the sign function. In this paper, we further assume that $f^* \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the RKHS induced by a pre-specified kernel function $K(\cdot,\cdot)$ satisfying $K(\cdot,\cdot) \in C^4(\mathcal{X}, \mathcal{X})$

with C^4 referring to the class of functions whose fourth derivative is continuous. Note that the requirement that $f^* \in \mathcal{H}_K$ is commonly used in statistical learning literature (Rosasco et al., 2013; Yang, Lv and Wang, 2016; Dasgupta, Goldberg and Kosorok, 2019), and the requirement $K(\cdot,\cdot) \in C^4(\mathcal{X},\mathcal{X})$ ensures that for any function belonging to the induced RKHS, the corresponding first- and second-order gradients exist (Zhou, 2007), which is naturally satisfied by many kernels including the Gaussian kernel. It is well-known that the RKHS induced by some universal kernels, such as the Gaussian kernel, is a fairly large functional space in that any continuous function can be arbitrarily well approximated by an intermediate function in its induced RKHS under the infinity norm (Steinwart, 2005).

2.2. Structure learning via gradient functions

In statistical analysis, the true target function f^* is often assumed to have a specific structure and tremendous interests have been paid to recover the structure of f^* from the observed data, including learning the sparse/interaction structure of f^* or identifying the linear and nonlinear effects in f^*. Unlike most of existing methods that only work under specific settings and model assumptions, we observe that the gradient functions can be employed as an efficient and flexible tool to meet these statistical interests. Precisely, for the true target function f^* defined in (2.1) with a loss function satisfying Assumption 1, we focus on the first- and second- order gradient functions of f^* that

$$g_l^*(\mathbf{x}) = \frac{\partial f^*(\mathbf{x})}{\partial x^l} \text{ and } g_{lk}^*(\mathbf{x}) = \frac{\partial^2 f^*(\mathbf{x})}{\partial x^l \partial x^k}, \tag{2.2}$$

for $l,k = 1,\ldots,p$. In the following, we illustrate how to use $g_l^*(\mathbf{x})$ and $g_{lk}^*(\mathbf{x})$ to conduct sparse learning, interaction selection and model identification in the sequential.

Example 2.1. *In sparse learning, it is generally believed that only a few variables have effect on f^*, while others are noises (Shen, Pan and Zhu, 2012; Shen et al., 2013; He, Wang and Hong, 2013). By using the first-order gradient function in (2.2), we observe that a variable x^l makes no contributions to the true target function f^* if and only if*

$$\|g_l^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} = \int_{\mathcal{X}} (g_l^*(\mathbf{x}))^2 d\rho_{\mathbf{x}} = 0, \tag{2.3}$$

where $\|\cdot\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}$ denotes the $\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})$-norm and $\rho_{\mathbf{x}}$ denotes the marginal distribution of the covariate $\mathbf{x}$. Thus, evaluating the importance of a variable turns to measure the importance of the corresponding gradient function, and thus $\|g_l^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}$ can be adopted as a valid measure to distinguish the informative and non-informative variables in f^*. Then the true active set can be defined as $\mathcal{A}^* = \{l : \|g_l^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} > 0\}$.

Example 2.2. *For interaction selection, many attempts have been made to identify the true interaction effects in underlying models (Radchenko and James, 2010; Hao and Zhang, 2014; Kong et al., 2017; Hao, Feng and Zhang, 2018). We observe that the true interaction effects on f^* can be evaluated by the second-order gradient functions. Specifically, given the true active set $\mathcal{A}^*$, if a variable x^l has no interaction effect on f^*, the corresponding second-order gradient functions among all the other variables should be zero almost surely in the sense that*

$$\|g_{lk}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = \int_{\mathcal{X}} (g_{lk}^*(\mathbf{x}))^2 d\rho_{\mathbf{x}} = 0, \tag{2.4}$$

for any $k \in \mathcal{A}^*$. Thus, the active set containing all the variables that contribute to the two-way interaction effects in f^* can be defined as

$$\mathcal{A}_2^* = \big\{ l \in \mathcal{A}^* : \|g_{lk}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 > 0, \text{ for some } k \in \mathcal{A}^* \big\},$$

Moreover, we further denote the set containing the variables that only contribute to the main effects of f^* as $\mathcal{A}_1^* = \mathcal{A}^* \setminus \mathcal{A}_2^*$. It is interesting to point out that the definitions of $\mathcal{A}_1^*$ and $\mathcal{A}_2^*$ are general and reduce to those in Kong et al. (2017) when the true structure of f^* is a quadratic function.

Example 2.3. *Identifying the linear and nonlinear effects in f^* has also attracted many attentions in the literature of partially linear models (PLMs) (Zhang, Cheng and Liu, 2011; Lian, Liang and Ruppert, 2015; He and Wang, 2020). Generally, a PLM considers*

$$f^*(\mathbf{x}) = \mathbf{x}_{\mathcal{L}^*}^\top \boldsymbol{\beta}^* + h^*(\mathbf{x}_{\mathcal{N}^*}),$$

where $\mathbf{x} = (\mathbf{x}_{\mathcal{L}^*}^\top, \mathbf{x}_{\mathcal{N}^*}^\top)^\top \in \mathcal{R}^p$, $\mathcal{L}^*$ and $\mathcal{N}^*$ denote the sets of linear and nonlinear effects, $\mathbf{x}_{\mathcal{L}}^\top \boldsymbol{\beta}^*$ is the linear part and $h^*(\mathbf{x}_{\mathcal{N}}^*)$ is the nonlinear part. One of the primal interests is to correctly identify the linear and nonlinear effects in a PLM. We notice that the true linear and nonlinear effects can be distinguished by evaluating the corresponding second-order gradient functions. Specifically, given the true active set $\mathcal{A}^*$, we observe that if a variable x_l has a linear effect on f^*, the corresponding second-order gradient functions among all the other variables should be zero almost surely that is $\|g_{lk}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = 0$ for any $k \in \mathcal{A}^*$. Thus, the sets of true linear effects and true nonlinear effects can be defined as $\mathcal{L}^* = \{l : \|g_{lk}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = 0, \text{ for any } k \in \mathcal{A}^*\}$ and $\mathcal{N}^* = \mathcal{A}^* \setminus \mathcal{L}^*$.

As demonstrated in the above examples, the gradient functions can be employed as an efficient and flexible tool to learn the interested structure of f^* and more importantly, it provides appropriate definitions of the interested structure of f^* in a "model-free" sense, which avoids the risk of potential model misspecifications. Thus, it suffices to learn the corresponding gradient functions consistently and efficiently for identifying the underlying structure of f^*. It is worthy pointing out that in the aforementioned examples, each corresponding gradient function is evaluated given all the other variables in the sense that $\|g_l^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = \int_{\mathcal{X}} (\frac{\partial f^*(\mathbf{x})}{\partial x^l})^2 d\rho_{\mathbf{x}}$ and $\|g_{lk}^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = \int_{\mathcal{X}} (\frac{\partial^2 f^*(\mathbf{x})}{\partial x^l \partial x^k})^2 d\rho_{\mathbf{x}}$, and thus

the defined underlying structures are consisting of those variables, whose corresponding true gradient function should contain sufficient information given all the other variables.

Note that for structure learning, f^* needs be defined in advance since the underlying structure affects the response through f^*. In this paper, f^* is given by (2.1) with some pre-specified loss function, which may depend on the problem of interest and the knowledge of the obtained data, and thus corresponds to different learning tasks. In the following sections, a general framework is proposed to learn the underlying structure of f^* defined by a loss function belonging to a rich loss family introduced in Section 2.1 and its theoretical consistencies are also established. In practice, based on the problem of interest and prior knowledge, the user can firstly specify some loss belonging to the considered family, and then the proposed framework as well as the established theoretical guarantees directly apply.

3. The proposed framework

Most existing learning gradient methods formulate the task into a regularized framework (Rosasco et al., 2013; Yang, Lv and Wang, 2016; He and Wang, 2020; He, Lv and Wang, 2020) with some carefully designed functional penalties on the gradient functions. However, these methods usually suffer computational burdens due to the employed local pair-wise learning tasks or the added complicated empirical functional penalties. On the contrary, the proposed framework provides an efficient alternative to learning the structure of f^*. It is motivated by the key observations that the derivative reproducing properties in RKHS (Zhou, 2007) assure that if $K(\cdot,\cdot) \in C^2(\mathcal{X},\mathcal{X})$, then for any $f \in \mathcal{H}_K$, there holds

$$g_l(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x^l} = \langle f, \partial_l K_{\mathbf{x}} \rangle_K, \tag{3.1}$$

where C^2 denotes the class of functions whose second derivative is continuous and $\partial_l K_{\mathbf{x}}(\cdot) = \frac{\partial K(\mathbf{x},\cdot)}{\partial x^l} \in \mathcal{H}_K$. Moreover, if $K(\cdot,\cdot) \in C^4(\mathcal{X},\mathcal{X})$, there also holds that

$$g_{lk}(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x^l \partial x^k} = \langle f, \partial_{lk} K_{\mathbf{x}} \rangle_K, \tag{3.2}$$

where $\partial_{lk} K_{\mathbf{x}} = \frac{\partial K(\mathbf{x},\cdot)}{\partial x^l \partial x^k} \in \mathcal{H}_K$. Note that the facts (3.1) and (3.2) assure that to estimate the interested gradient functions within the induced RKHS, it suffices to estimate the target function f^* itself, and then the gradient functions can be directly obtained. In the rest of this paper, we focus on the applications of the first- and second-order gradient functions to learn the structure of f^* and thus require $K(\cdot,\cdot) \in C^4(\mathcal{X},\mathcal{X})$, which is naturally satisfied by many kernels, including the Gaussian kernel. Note that it is trivial to extend the proposed framework to estimate arbitrary higher-order gradient functions, which may be useful in some real applications (Ritchie et al., 2001).

Motivated by these key facts, we propose an efficient framework to learn the underlying structure of the true target function f^*, which involves a regularized M-estimation in the induced RKHS and the fast computation of corresponding gradient functions. Suppose that the random sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independent copies of the random pair $(\mathbf{x}, y)$. Firstly, we consider the regularized M-estimation in a RKHS to estimate f^* by solving the following optimization problem that

$$\widehat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \lambda \|f\|_K^2, \tag{3.3}$$

where the first term is denoted as $\mathcal{E}^L_{\mathcal{Z}^n}(f)$ and $\|\cdot\|_K$ denotes the induced RKHS-norm. Although the regularized M-estimation procedure in a RKHS (3.3) uses all the variables to learn f^*, only the reproducing kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ is needed, which represents the inner product of some (possibly unknown) feature map function mapping the the original input into some high-dimensional feature space. This is known as the kernel trick and is ensured by Mercer's Theorem (Steinwart and Christmann, 2008b). More importantly, by the representer theorem (Wahba, MIT Press, 1998), the solution of (3.3) must have a finite form that

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) = \widehat{\boldsymbol{\alpha}}^\top \mathbf{K}_n(\mathbf{x}), \tag{3.4}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)^\top$ denotes the representer coefficients and $\mathbf{K}_n(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x}))^\top$ is the n-dimensional kernel vector.

The representer theorem converts the original optimization task (3.3) over an infinite functional space $\mathcal{H}_K$ into an optimization task over a finite n-dimensional vector space of $\boldsymbol{\alpha}$. Thus, by plugging (3.4) into (3.3), solving the optimization task (3.3) is equivalent to solving the following task that

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathcal{R}^n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \boldsymbol{\alpha}^\top \mathbf{K}_n(\mathbf{x}_i)\right) + \lambda \boldsymbol{\alpha}^\top \mathbf{K}\, \boldsymbol{\alpha}, \tag{3.5}$$

where $\mathbf{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathcal{R}^{n \times n}$. The employed computing algorithm for (3.5) varies from one loss to another, and the details are given in Section 4.1.

Then, we apply the derivative reproducing properties (3.1) and (3.2) to facilitate the computation of gradient functions of our interests. Specifically, once $\widehat{\boldsymbol{\alpha}}$ is obtained, we can efficiently compute the estimated first- and second-order gradient functions that

$$\widehat{g}_l(\mathbf{x}) = \frac{\partial \widehat{f}(\mathbf{x})}{\partial x^l} = \widehat{\boldsymbol{\alpha}}^\top \partial_l \mathbf{K}_n(\mathbf{x}) \quad \text{and} \quad \widehat{g}_{lk}(\mathbf{x}) = \frac{\partial^2 \widehat{f}(\mathbf{x})}{\partial x^l \partial x^k} = \widehat{\boldsymbol{\alpha}}^\top \partial_{lk} \mathbf{K}_n(\mathbf{x}), \tag{3.6}$$

where $\partial_l \mathbf{K}_n(\mathbf{x}) = \frac{\partial \mathbf{K}_n(\mathbf{x})}{\partial x^l} \in \mathcal{R}^n$ and $\partial_{lk}\, \mathbf{K}_n(\mathbf{x}) = \frac{\partial^2 \mathbf{K}_n(\mathbf{x})}{\partial x^l \partial x^k} \in \mathcal{R}^n$. Note that once the kernel function $K(\cdot, \cdot)$ is pre-specified, the corresponding gradients $\partial_l\, \mathbf{K}_n(\mathbf{x})$ and $\partial_{lk}\, \mathbf{K}_n(\mathbf{x})$ are also analytically determined. Clearly, the gradients

can be directly computed by using the estimated coefficient $\widehat{\boldsymbol{\alpha}}$ to evaluate the importance of corresponding variable without any extra estimation procedure.

Now we illustrate how to apply the estimated gradient functions to recover the underlying structure of the true target function f^* in Examples 2.1–2.3. Precisely, for sparse learning, we adopt the empirical norm of $\widehat{g}_l$ as a practical measure by computing $\|\widehat{g}_l\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(\widehat{g}_l(\mathbf{x}_i)\big)^2$, and thus the estimated active set is defined as $\widehat{\mathcal{A}} = \Big\{l : \|\widehat{g}_l\|_n^2 > v_n\Big\}$, where v_n denotes some pre-specified thresholding value; for interaction selection, we adopt the empirical norm of the estimated second-order gradient function by computing $\|\widehat{g}_{lk}\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(\widehat{g}_{lk}(\mathbf{x}_i)\big)^2$, and thus the sets of active interaction and main effects in f^* can be estimated as

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

respectively, where v_n^{int} denotes some pre-specified thresholding value. Moreover, we can also apply the estimated second-order gradient functions to conduct model identification, and thus the estimated nonlinear and linear effect sets $\mathcal{N}^*$ and $\mathcal{L}^*$ are identified as $\widehat{\mathcal{N}} = \big\{l \in \widehat{\mathcal{A}} : \|\widehat{g}_{lk}\|_n^2 > v_n^{iden}, \text{ for some } k \in \widehat{\mathcal{A}}\big\}$ and $\widehat{\mathcal{L}} = \widehat{\mathcal{A}} \setminus \widehat{\mathcal{N}}$, respectively, where v_n^{iden} is the pre-defined thresholding value. Note that the structure learning performance of the proposed method highly relies on the choice of pre-specified thresholding values, which can be appropriately determined through a stability-based selection criterion (Sun, Wang and Fang, 2013) and more details are provided in Section 4.2.

It is worthy noting that the proposed framework is as computationally efficient as the screening-type methods but the strong marginal correlation condition is not necessary any longer. Differently, the proposed framework can be treated as a non-parametric joint screening method in the sense that each gradient function is computed given all the other variables.

Remark 1. It is interesting to point out that the proposed framework is motivated from the definition of true structures that is defined based on derivatives of the true regression function f^*. And thus, the existence of the corresponding gradients of f^* ensures that they can be adopted as a valid tool to define the true structures. Then, the task of learning the underlying structure is converted to learning the corresponding gradient functions. Clearly, the proposed framework is designed to learn the underlying structures that can be well defined by the corresponding gradient functions. Note that if f^* has a finite number of non-differentiable points, it can be well approximated by some continuous function, which can be further approximated by an intermediate function $f^0 \in \mathcal{H}_K$ arbitrarily well when K is a universal kernel (Steinwart and Christmann, 2008b). It directly implies that all the informative variables act on f^* can be correctly identified by examining the gradients of f^0. Some additional numerical experiments are provided in the Supplementary Material, where the f^* is the step function, and the proposed framework can still correctly identify the underlying structure in all the replications.

4. Computational issues

In this section, we provide all the computational details as well as the tuning procedures of the proposed framework.

4.1. Computing algorithms

It is clear that the proposed framework is computationally efficient in that we only need to solve a convex optimization problem (3.5), and then the estimated gradient functions can be directly obtained with the derivative reproducing property of the RHKS. Note that the employed computing algorithm for (3.5) varies from one loss function to another. For example, for the squared loss function, the solution to (3.5) has an explicit form that $\widehat{\boldsymbol{\alpha}} = (\mathbf{K}^2 + n\lambda \mathbf{I}_n)^{-1} \mathbf{K}\mathbf{y}$; for the check and hinge loss functions, the dual optimization can be considered (Takeuchi et al., 2006), which converts (3.5) to a quadratic programming problem with certain linear constraints; for the logistic loss, the kernel-based weighted least-square iterations (Zhu and Hastie, 2005) can be employed. Note that the optimization task (3.5) can be efficiently implemented by using some disciplined convex optimization algorithms, and the R package *CVXR* (Fu, Narasimhan and Boyd, 2020) is used to carry out the optimization of the proposed framework in all the numerical examples of this paper. More importantly, we provide an R package implementing the proposed framework under various losses belonging to the family introduced in Section 2.1, which is available at `https://github.com/geyh96/GSLM/`.

4.2. Tuning procedure

It is interesting to notice that the proposed structure learning framework involves two tuning parameters that the parameter λ in (3.5) and the pre-defined thresholding value used to define active sets of variables. Due to our limited numerical experience, the performance of the proposed framework is satisfactory when λ is sufficiently small in various scenarios. Similar observation has also been made in Wang and Leng (2016). Thus, we use $\lambda = 10^{-5}$ in Section 6, which yields satisfying performance. Based on our limited numerical experience, the performance of the proposed framework is less sensitive to λ when it is sufficiently small. It is also worthy pointing out that λ can also be chosen via cross-validation, which may slightly improve the numerical performance, but at the cost of much increased computational time.

Moreover, we employ the stability-based criterion (Sun, Wang and Fang, 2013) to select the optimal value of thresholding parameter. Its key idea is to measure the stability of sparse learning by randomly splitting the training sample into two independent parts and comparing the disagreement between these 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 $\mathcal{Z}_2^n$ and obtain two 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 by Cohen's kappa coefficient and the procedure is repeated for multiple times, and then the optimal thresholding value can be determined correspondingly. We refer to Sun, Wang and Fang (2013) for more details, which is originally designed to select the optimal tuning parameter for the purpose of variable selection with theoretical guarantee.

5. Statistical properties

In this section, we provide the theoretical guarantees of the proposed structure learning framework. Precisely, we establish the estimation consistency of gradient functions and provide the asymptotic consistencies of sparse learning, interaction selection and model identification under mild conditions, respectively.

We start with a brief introduction about some basic knowledge in learning theory. Specifically, we have $K(\mathbf{x},\cdot) \in \mathcal{H}_K$ for any $\mathbf{x} \in \mathcal{X}$, and $\langle f, K_{\mathbf{x}}\rangle_K = f(\mathbf{x})$ for any $f \in \mathcal{H}_K$. By Mercer's theorem (Steinwart and Christmann, 2008b), under some regularity conditions, the eigen-expansion of the kernel function is given by

$$K(\mathbf{x},\mathbf{x}) = \sum_{k=1}^{\infty} \mu_k \phi_k(\mathbf{x})\phi_k(\mathbf{x}), \tag{5.1}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are non-negative eigenvalues, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}}) = \big\{f : \|f\|_2^2 < \infty\big\}$. The RKHS-norm of any $f \in \mathcal{H}_K$ then can be written as

$$\|f\|_K^2 = \sum_{k\geq 1} \frac{1}{\mu_k} \langle f, \phi_k\rangle_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2,$$

which implies that the decay rate of μ_k fully characterizes the complexity of the RKHS induced by K, and has a close relationship with various entropy numbers (Steinwart and Christmann, 2008b). Therefore, for any $f \in \mathcal{H}_K$, we have $f(\mathbf{x}) = \sum_{k=1}^{\infty} a_k\phi_k(\mathbf{x})$, where $a_k = \langle f, \phi_k\rangle_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} = \int_{\mathcal{X}} f(\mathbf{x})\phi_k(\mathbf{x})d\rho_{\mathbf{x}}$ are Fourier coefficients. Note that these results require that $\mathcal{H}_K \subset \mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})$, which is automatically satisfied if $\sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x},\mathbf{x})$ is bounded. Moreover, the solution of (2.1) may not be unique, and thus we further define $f^* = \operatorname{argmin}_{f\in\mathcal{B}} \|f\|_K^2$ with $\mathcal{B} = \{f : f = \operatorname{argmin}_{h\in\mathcal{H}_K} \mathcal{E}^L(h)\}$ to ensure the uniqueness of f^* in the sequel. Furthermore, we denote $\widetilde{f} = \operatorname{argmin}_{f\in\mathcal{H}_K} \mathcal{E}^L(f) + \lambda\|f\|_K^2$. We now rewrite λ as λ_n in the rest of this paper, to emphasize its dependency on n.

5.1. Estimation consistency of gradient functions

The following technical assumptions are made to establish the estimation consistencies of gradient functions, which is crucial to ensure the asymptotic consistency of the proposed structure learning framework. We further introduce following assumptions.

Assumption 2. 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 3. There exist some positive constants c_3 and θ such that the approximation error $\|\widetilde{f} - f^*\|_K = c_3\lambda_n^{\theta}$.

Assumption 2 imposes the boundedness condition on the kernel function as well as the corresponding gradient functions. This assumption is commonly used in machine learning literature (Rosasco et al., 2013; Yang, Lv and Wang, 2016) and satisfied by many kernels with the compact support condition, including the Gaussian kernel, Sobolev kernel, scaled linear kernel, scaled quadratic kernel and so on. Note that the requirement of compact support is usually assumed in machine learning literature for mathematical simplicity, and many efforts have been made to extend it to the non-compact setting (Simon-Gabriel and Schölkopf, 2018). Assumption 3 quantifies the approximation error as a function of the tuning parameter λ_n, which is sensible as $\lim_{\lambda_n\to 0} \|\widetilde{f} - f^*\|_K^2 = 0$ in general. Similar assumptions are also used in literature to control the approximation error rate (Rosasco et al., 2013; Zhang, Liu and Wu, 2016; Dasgupta, Goldberg and Kosorok, 2019), and it is analogous to the approximation error under the squared loss function. Particularly, Mendelson and Neeman (2010) prove that the approximation error under the squared loss function can be explicitly quantified as $O(\lambda_n^{r-1/2})$ with $r \in (1/2, 1]$. Further investigations about the approximation error rate are provided in Eberts and Steinwart (2013) by imposing some additional technical assumptions.

Theorem 1. *Suppose that Assumptions 1–3 are satisfied. Let $\lambda_n = n^{-1/(4q)}$, then for any $\delta_n \ge 2(\log n)^{-1/q}E|y|$, there exists some positive constant c_4 such that with probability at least $1 - \delta_n$, such that the following inequality holds*

$$\max_{l=1,\ldots,p} \left| \|\widehat{g}_l\|_n^2 - \|g_l^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 \right| \le c_4 \left(\log \frac{4p}{\delta_n} \right)^{1/2} (\log n)^{q/2} n^{-\Theta}, \tag{5.2}$$

where $\Theta = \min\{\frac{3}{16}, \frac{\theta}{4q}\}$, q and θ are given in Assumptions 1 and 3, respectively.

Theorem 1 establishes the estimation consistency of the estimated first-order gradients $\widehat{g}_l$, $l = 1, \ldots, p$, in the sense that $\|\widehat{g}_l\|_n^2$ converges to $\|g_l^*\|_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2$ with high probability, and is crucial to recovery the underlying structure of f^*. Note that this convergence result is established without any model assumption on f^* and holds true for a general loss L satisfying Assumption 1, which includes many scenarios as its special cases, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification. Specially, for the binary classification, the upper bound in Theorem 1 reduces to $c_4\big(\log \frac{4p}{\delta}\big)^{1/2} n^{-\Theta}$ for any $\delta \in (0, 1)$. It is also worthy pointing out that once the squared loss is used, the convergence rate in Theorem 1 can be further strengthened to obtain a faster strong convergence rate (Fischer and Steinwart, 2020) if some additional technical assumptions, such as the decay rate of μ_k in (5.1), are met.

Remark 2. Some theoretical results of gradient estimation are also established in the literature based on various estimation approaches in different scenarios. Mukherjee and Zhou (2006) and Yang, Lv and Wang (2016) adopt a pair-wise learning task to estimate gradient functions with the squared loss, and the estimation consistency is established by imposing the upper bound of $\sum_{l=1}^{p} \|\widehat{g}_l - g_l^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}$ as $O(n^{-\Theta_p})$ with Θ_p denoting a linear or polynomial function of p, which is further improved by assuming that $\mathcal{X}$ is a d-dimensional connected compacted C^∞ submanifold of $\mathcal{R}^p$ (Mukherjee, Wu and Zhou, 2010; Ye and Xie, 2012). The estimation consistency in Deng et al. (2022) is established under the Bayesian posterior framework with Gaussian prior, which requires that the representer coefficients and noise term are normally distributed. Rosasco et al. (2013) and He, Lv and Wang (2020) use a regularization framework with an empirical functional penalty on the gradients to conduct the estimation, and their theoretical results achieve a much slower convergence rate because of their complicated estimation framework. It is also interesting to point out that the selection consistency for sparse learning can not be established in Rosasco et al. (2013) and He, Lv and Wang (2020) due to the theoretical difficulty of dealing with the empirical regularizer. He, Wang and Lv (2021) and Chen, He and Wang (2021) use the derivative reproducing property to compute the gradient function, yet their theoretical consistencies are given under some specific technical treatments, which only work for their pre-specified loss. Specifically, He, Wang and Lv (2021) use the integral operator in learning theory to obtain the theoretical consistency, which can only be used for the squared loss, and Chen, He and Wang (2021) use the property of the pre-specified check loss to establish the theoretical consistency. We want to emphasize that it is nontrivial to extend the theoretical results in He, Wang and Lv (2021) and Chen, He and Wang (2021) to accommodate a general loss family as considered in Theorem 1, where the theoretical derivation is much more challenging.

The following technical assumption is made to establish the estimation consistency of the second-order gradient functions.

Assumption 4. There exists some constant κ_3 such that $\sup_{\mathbf{x}\in\mathcal{X}} \|\partial_{lk} K_{\mathbf{x}}\|_K \leq \kappa_3$, for any $l, k = 1, \ldots, p$.

Assumption 4 can be regarded as the extension of Assumptions 2 by requiring the boundedness of the second-order gradients of $K_{\mathbf{x}}$, and is also naturally satisfied by all the kernels discussed after Assumption 2.

Theorem 2. *Suppose all the assumptions of Theorem 1 as well as Assumption 4 are met. Then, there exists some positive constant c_5 such that with probability at least $1 - \delta_n$, there holds*

$$\max_{l,k=1,\ldots,p} \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} \right| \leq c_5 \left(\log \frac{4p^2}{\delta_n} \right)^{\frac{1}{2}} (\log n)^{q/2} n^{-\Theta},$$

where $\Theta = \min\{\frac{3}{16}, \frac{\theta}{4q}\}$, δ_n, q and θ are given in Theorem 1.

Theorem 2 shows that the estimated second-order gradient function $\|\widehat{g}_{lk}\|_n^2$

converges to $\|g_{lk}^*\|_2^2$ with high probability, which is crucial to establish the consistency for the application to interaction selection and model identification. In literature, only a few attentions have been paid to investigate the estimation results of the second-order gradient function. Specifically, He and Wang (2020) adopt the pair-wise learning task by using the second-order Taylor expansion to estimate the second-order gradient functions with the squared loss, and the consistency result is given under the $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$ norm, which is similar to those in Mukherjee and Zhou (2006) and Yang, Lv and Wang (2016). He, Wang and Lv (2021) establish a theoretical estimation result similar to Theorem 2, but they use the integral operator to obtain the convergence result which only holds for the squared loss. It is worthy pointing out that the estimation consistency of arbitrary higher-order gradient functions can also be established by requiring the boundedness of corresponding higher-order gradients of $K_{\mathbf{x}}$, which is naturally satisfied by many popularly used kernels, such as Gaussian kernel. Note that the asymptotic consistency of gradient estimation can be extended to arbitrary order with the similar argument of the proof of Theorem 2 and Theorem 1 in Zhou (2007) with slight modifications, but with the additional requirement on the higher order differentiability of f^*.

5.2. Theoretical property of sparse learning

In this section, we use the obtained theoretical results in Section 5.1 to establish the selection consistency of the proposal in the sparse learning given in Example 2.1 of Section 2.1 by using the first-order gradient functions. The following technical assumption is needed to establish the theoretical result.

Assumption 5. There exist some positive constants c_6 and $\xi_1 > \frac{q}{2}$ such that $\min_{l \in \mathcal{A}^*} \|g_l^*\|^2_{\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})} > c_6 \big(\log \frac{4p}{\delta_n} \big)^{1/2} (\log n)^{\xi_1} n^{-\Theta}$, where Θ is given in Theorem 1.

Assumption 5 requires that the true gradient function g_l^* should contain sufficient information about the truly informative variables, and it can be regarded as a condition on the required minimal signal strength, which may go to zero with the increase of sample size. This assumption automatically rules out those cases where variables are highly correlated or some variables share exactly the same information with others, since each gradient function is evaluated given all the other variables in the sense that $\|g_l^*\|^2_{\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})} = \int_{\mathcal{X}} (\frac{\partial f^*(\mathbf{x})}{\partial x^l})^2 d\rho_{\mathbf{x}}$. Note that this assumption is crucial to establish the selection consistency and is much weaker than that of many non-parametric sparse learning methods (Huang, Horowitz and Wei, 2010; Yang, Lv and Wang, 2016), which often require the signal is bounded away from zero by some positive constants.

Theorem 3 (Sparse learning)**.** *Suppose all the assumptions of Theorem 1 as well as Assumption 5 are satisfied. Let* $v_n = \frac{c_6}{2} \big(\log \frac{4p}{\delta_n} \big)^{1/2} (\log n)^{\xi_1} n^{-\Theta}$*, then we have*

$$P(\widehat{\mathcal{A}} = \mathcal{A}^*) \to 1. \tag{5.3}$$

Theorem 3 shows that the estimated informative set in sparse learning can exactly recover the true active set with high probability. This result is fascinating and attractive given the fact that it holds true for a general loss function satisfying Assumption 1, and thus includes many scenarios as its special cases, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification. Moreover, this result is established without requiring any pre-specified model assumption and allows general dependence structures among variables and response in a model-free fashion. Note that the above theoretical results, including the required technical conditions, are established for a general loss function family, and some tighter results may be obtained by using specific technical tools for the specific loss function. Moreover, it is interesting to notice that the required Assumptions 1 and 2 in Sun, Wang and Fang (2013) can also be verified under our framework, and thus the selection guarantee of the employed tuning method established in Sun, Wang and Fang (2013) still holds.

5.3. Theoretical guarantees for interaction selection/model identification

In this section, we use the obtained theoretical results in Section 5.1 to establish the consistency of the proposal in interaction selection and model identification by using the second-order gradient functions in Theorems 4 and 5, respectively.

Firstly, we consider the interaction selection as given in Example 2.2 of Section 2.1 and the following technical assumption is required.

Assumption 6. There exist some positive constants c_7 and $\xi_2 > \frac{q}{2}$ such that $\min_{l,k\in\mathcal{A}_2^*} \|g_{lk}^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} > c_7\Big(\log \frac{4p_0^2}{\delta_n}\Big)^{1/2}(\log n)^{\xi_2} n^{-\Theta}$, where Θ is given in Theorem 1.

Assumption 6 can be regarded as the extension of Assumption 5 by requiring the true second-order gradient functions have sufficient information about the interaction effects.

Theorem 4 (Interaction selection)**.** *Suppose that the assumptions of Theorem 3 as well as Assumption 6 are met. By taking* $v_n^{int} = \frac{c_7}{2}\Big(\log \frac{4p_0^2}{\delta_n}\Big)^{1/2}(\log n)^{\xi_2} n^{-\Theta}$, *we have*

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

Theorem 4 shows that the proposal used in the interaction selection can exactly detect all the interaction effects with high probability. Note that this result is established without imposing the strong heredity assumption, which is often assumed by the existing parametric interaction selection methods (Hao and Zhang, 2014). Clearly, the proposed method can be extended to detect higher-order interaction effects, which is of particular interests in some real applications (Ritchie et al., 2001). It is also worthy pointing out that the interaction selection

consistency is established for a rich loss function family with a general kernel, which allows detecting general interaction structures among variables for various scenarios.

Finally, we turn to establish the consistency of model identification as illustrated in Example 2.3 of Section 2.1 and the following technical assumption is introduced.

Assumption 7. There exist some positive constants c_8 and $\xi_3 > \frac{q}{2}$ such that $\min_{l,k\in\mathcal{N}^*} \|g_{lk}^*\|^2_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})} > c_8\Big(\log\frac{4p_0^2}{\delta_n}\Big)^{1/2}(\log n)^{\xi_2}n^{-\Theta}$, where Θ is given in Theorem 1.

Assumption 7 requires that the gap of signal strengths between linear and nonlinear effects in the sense that the corresponding second-order gradient functions of the linear effects are exactly zero, and those of the nonlinear effects are lower bounded.

Theorem 5 (Linear and nonlinear effects)**.** *Suppose that all the assumptions in Theorem 3 as well as Assumption 7 are met. By taking* $v_n^{iden} = \frac{c_8}{2}\Big(\log\frac{4p_0^2}{\delta_n}\Big)^{1/2}(\log n)^{\xi_2}n^{-\Theta}$*, we have*

$$P\Big(\widehat{\mathcal{L}} = \mathcal{L}^*, \widehat{\mathcal{N}} = \mathcal{N}^*\Big) \to 1, \quad as \quad n \to \infty.$$

Theorem 5 shows that the underlying model structure can be exactly identified with probability tending to 1. This theoretical result is also established for a general loss function satisfying Assumption 1 without any explicit model specifications. It provides strong theoretical support for automatically discovering the model structure for the PLMs, which is particularly attractive in the field of partially linear models. It is worthy pointing out that Theorems 4 and 5 are established under the case that noise variables are also included in the collected variable set that $|\mathcal{A}^*| \ll p$, and thus the proposal in sparse learning is used to recover all the informative variables at first, and then either interaction selection or model identification with the proposed method are conducted based on the variables identified at the first step. In some other scenarios, where all the collected variables are believed to be related with the response that $\mathcal{A}^* = \{1, \ldots, p\}$, the proposed method can be directly applied without applying the sparse learning and the similar theoretical results can be obtained.

6. Numerical studies

In this section, the proposed framework is applied to sparse learning and interaction selection, and its numerical performance are compared with various state-of-the-art competitors under several settings. For the proposed framework, the RKHS induced by the Gaussian kernel $K(\mathbf{u},\mathbf{v}) = \exp\Big(-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma_n^2}\Big)$ is adopted in all the examples, where σ_n is set as the median of all the pairwise distances among the training sample as suggested by Jaakkola, Diekhans and Haussler

(1999), Yang, Lv and Wang (2016), and Mukherjee and Zhou (2006). Note that σ_n can also be chosen by cross-validation, which may slightly improve the numerical performance, but at the cost of much increased computational burden. Other tuning parameters such as the thresholding value are selected by the stability-based criterion (Sun, Wang and Fang, 2013) as introduced in Section 4.2 via a grid search, where the grid is set as $\{10^{-3+0.1s} : s = 0, \ldots, 60\}$.

6.1. Application to sparse learning

In this subsection, the application of the proposed framework to sparse learning is considered. Specifically, we consider regression with the squared loss, the check loss with $\tau = 0.5$ and the Huber loss, and classification with the hinge loss and the logistic loss, due to their popularity and importance in statistical machine learning, and denoted as GSLM-SQ, GSLM-QA, GSLM-HB, GSLM-SVM and GSLM-LOG for simplicity. Under regression setting, we consider five competitors, including distance correlation learning (DC, Li, Zhong and Zhu, 2012), the quantile-adaptive screening (QaSIS, He, Wang and Hong, 2013), the sure independence rank screening (SIRS, Zhu et al., 2011), the modified Blum-Kiefer-Rossenblatt correlation (MBKR, Zhou and Zhu, 2018) and the generic sure independence screening (Ball, Pan et al., 2019). Under classification setting, we also consider five competitors, including DC, SIRS, MBKR, the screening procedure based on empirical conditional distributions (MV-SIS, Cui, Li and Zhong, 2015), and Kolmogolov Filter (Kol-Filter, Mai and Zou, 2013). Note that the screening-based competitors are suggested to keep the first $\lceil n/\log n \rceil$ variables to assure the sure screening property, and for fair comparison and saving the space, we here report the results for those competitors truncated by the thresholding values based on the stability-based criterion as introduced at the beginning of this section to conduct sparse learning, and we denote the truncated versions with the suffix "-t", such as DC-t and QaSIS-t. More numerical results of those competing methods implemented as suggested by their authors are reported in the Supplementary Material.

The following two simulated examples are examined under various scenarios.

Example 1 (Regression)**.** We first generate $x_i = (x_{i1}, \ldots, x_{ip})^\top$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently drawn from $U(-0.5, 0.5)$. The response y_i is generated as $y_i = 8f_1(x_{i1}) + 4f_2(x_{i2})f_3(x_{i3}) + 6f_4(x_{i4}) + 5f_5(x_{i5}) + \epsilon_i$, where $f_1(u) = u, f_2(u) = 2u + 1, f_3(u) = 2u - 1, f_4(u) = 0.1\sin(\pi u) + 0.2\cos(\pi u) + 0.3(\sin(\pi u))^2 + 0.4(\cos(\pi u))^3 + 0.5(\sin(\pi u))^3$, $f_5(u) = \sin(\pi u)/(2 - \sin(\pi u))$, and ϵ_i's are independently drawn from $N(0, 1)$. Clearly, the first five variables are truly informative.

Example 2 (Classification)**.** We generate $x_i = (x_{i1}, \ldots, x_{ip})^\top$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently drawn from $U(0, 1)$. Then we generate $y \sim$ Bernoulli $\left(\frac{1}{1+e^{-f^*(\mathbf{x})}}\right)$ with the true conditional logit function $f^*(\mathbf{x}) = 8x_1 + 4x_1^2 - 2\cos(\pi x_1/2) + 6\sin(\pi(x_2 - x_3)) - 4$. Clearly, the first three variables are truly informative.

For both examples, we consider different combinations $(n, p) = (500, 5000)$, $(500, 10000)$, $(500, 50000)$ and $(500, 100000)$, and for each case, $\eta = 0$ and $\eta = 0.5$ are examined. When $\eta = 0$, the variables are completely independent, whereas when $\eta = 0.5$, correlation structures are added among the variables. Under each setting, the experiment is replicated 100 times and the averaged performance measures are summarized in Tables 1–4, where "MeanSize" and "MaxSize" denotes the averaged and largest numbers of selected informative variables, "X_i" refers to the frequency of selecting the corresponding i-th covariate variable, and "C", "U", "O" are the frequency of correct-fitting, under-fitting, and over-fitting, respectively.

TABLE 1
The averaged performance measures in Example 1 with $n = 500$ and $\eta = 0$.

p	Method	X_1	X_2	X_3	X_4	X_5	MaxSize	U	O	C	MeanSize
5000	GSLM-SQ	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.03	0.97	5.03
	GSLM-QA	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.04	0.96	5.04
	GSLM-HB	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.02	0.98	5.02
	SIRS-t	1.00	0.99	1.00	0.96	1.00	5.00	0.04	0.00	0.96	4.95
	MBKR-t	1.00	1.00	1.00	0.96	1.00	5.00	0.04	0.00	0.96	4.96
	DC-t	1.00	0.99	1.00	0.92	1.00	5.00	0.08	0.00	0.92	4.91
	Ball-t	0.96	0.99	1.00	0.87	0.99	5.00	0.15	0.00	0.85	4.81
	QaSIS-t	0.93	0.97	0.96	0.85	0.97	5.00	0.26	0.00	0.74	4.68
10000	GSLM-SQ	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.02	0.98	5.02
	GSLM-QA	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.02	0.98	5.02
	GSLM-HB	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.02	0.98	5.02
	SIRS-t	1.00	1.00	1.00	0.97	0.97	5.00	0.05	0.00	0.95	4.94
	MBKR-t	1.00	1.00	1.00	0.98	1.00	5.00	0.02	0.00	0.98	4.98
	DC-t	1.00	1.00	1.00	0.97	0.97	5.00	0.05	0.00	0.95	4.94
	Ball-t	0.96	1.00	0.99	0.81	0.91	5.00	0.25	0.00	0.75	4.67
	QaSIS-t	0.91	0.92	0.90	0.83	0.91	5.00	0.40	0.00	0.60	4.47
50000	GSLM-SQ	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.13	0.87	5.13
	GSLM-QA	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.16	0.84	5.16
	GSLM-HB	1.00	1.00	1.00	1.00	1.00	6.00	0.00	0.16	0.84	5.16
	SIRS-t	1.00	0.97	0.99	0.91	0.97	5.00	0.11	0.00	0.89	4.84
	MBKR-t	1.00	0.99	0.99	0.93	0.99	5.00	0.07	0.00	0.93	4.90
	DC-t	1.00	0.99	1.00	0.92	0.99	5.00	0.09	0.00	0.91	4.90
	Ball-t	0.96	0.90	0.91	0.72	0.81	5.00	0.48	0.00	0.52	4.30
	QaSIS-t	0.88	0.84	0.82	0.65	0.70	5.00	0.69	0.00	0.31	3.89
100000	GSLM_SQ	1.00	1.00	1.00	1.00	1.00	9.00	0.00	0.22	0.78	5.27
	GSLM_QA	1.00	1.00	1.00	1.00	1.00	9.00	0.00	0.25	0.75	5.33
	GSLM_HB	1.00	1.00	1.00	1.00	1.00	9.00	0.00	0.22	0.78	5.29
	SIRS-t	1.00	0.98	0.98	0.90	0.97	5.00	0.12	0.00	0.88	4.83
	MBKR-t	1.00	0.99	0.99	0.95	1.00	5.00	0.07	0.00	0.93	4.93
	DC-t	1.00	1.00	1.00	0.94	1.00	5.00	0.06	0.00	0.94	4.94
	Ball-t	0.92	0.94	0.93	0.66	0.82	5.00	0.52	0.00	0.48	4.27
	QaSIS-t	0.79	0.74	0.82	0.51	0.66	5.00	0.82	0.00	0.18	3.52

It is evident that the proposed method outperforms other competing methods in both examples. In Example 1, GSLM-SQ, GSLM-QA and GSLM-HB are able to exactly identify all those truly informative variables in most replications. Yet, all the other competitors tend to miss some truly informative variables. It is also interesting to notice that from Table 2, all the other competing methods are consistently missing X_2 but selecting X_3. This is largely due to the fact that X_2 and X_3 contribute to the response differently under the correlated scenario. For instance, under the case $(n, p, \eta) = (500, 5000, 0.5)$, the averaged estimated correlations over 100 replications are $\widehat{Corr}(y, X_2) = -0.08$ and $\widehat{Corr}(y, X_3) = 0.552$. This may lead to the consistently missing since most of

TABLE 2
The averaged performance measures in Example 1 with $n = 500$ and $\eta = 0.5$.

p	Method	X_1	X_2	X_3	X_4	X_5	MaxSize	U	O	C	MeanSize
5000	GSLM-SQ	1.00	1.00	1.00	0.99	1.00	7.00	0.01	0.03	0.96	5.03
	GSLM-QA	1.00	1.00	1.00	1.00	1.00	7.00	0.00	0.03	0.97	5.04
	GSLM-HB	1.00	1.00	1.00	1.00	1.00	7.00	0.00	0.03	0.97	5.04
	SIRS-t	0.99	0.00	0.96	0.58	0.98	4.00	1.00	0.00	0.00	3.51
	MBKR-t	1.00	0.00	0.98	0.72	0.99	4.00	1.00	0.00	0.00	3.69
	DC-t	0.98	0.00	0.99	0.62	0.98	4.00	1.00	0.00	0.00	3.57
	Ball-t	0.92	0.00	0.94	0.43	0.95	4.00	1.00	0.00	0.00	3.24
	QaSIS-t	0.79	0.00	0.94	0.48	0.96	4.00	1.00	0.00	0.00	3.17
10000	GSLM-SQ	1.00	1.00	1.00	0.98	1.00	7.00	0.02	0.04	0.94	5.03
	GSLM-QA	1.00	1.00	1.00	0.99	1.00	7.00	0.01	0.04	0.95	5.04
	GSLM-HB	1.00	1.00	1.00	0.98	1.00	7.00	0.02	0.04	0.94	5.03
	SIRS-t	1.00	0.00	0.99	0.66	1.00	4.00	1.00	0.00	0.00	3.65
	MBKR-t	1.00	0.00	1.00	0.72	1.00	4.00	1.00	0.00	0.00	3.72
	DC-t	1.00	0.00	1.00	0.64	1.00	4.00	1.00	0.00	0.00	3.64
	Ball-t	0.97	0.00	0.94	0.47	0.96	4.00	1.00	0.00	0.00	3.34
	QaSIS-t	0.83	0.00	0.86	0.52	0.84	4.00	1.00	0.00	0.00	3.05
50000	GSLM-SQ	1.00	1.00	1.00	0.99	1.00	6.00	0.01	0.05	0.94	5.04
	GSLM-QA	1.00	1.00	1.00	0.99	1.00	7.00	0.01	0.11	0.88	5.11
	GSLM-HB	1.00	1.00	1.00	0.99	1.00	6.00	0.01	0.04	0.95	5.03
	SIRS-t	0.92	0.00	0.96	0.39	0.99	4.00	1.00	0.00	0.00	3.26
	MBKR-t	0.94	0.00	0.97	0.51	1.00	4.00	1.00	0.00	0.00	3.41
	DC-t	0.93	0.00	0.98	0.40	0.99	4.00	1.00	0.00	0.00	3.30
	Ball-t	0.85	0.00	0.90	0.20	0.89	4.00	1.00	0.00	0.00	2.84
	QaSIS-t	0.57	0.00	0.70	0.28	0.73	4.00	1.00	0.00	0.00	2.27
100000	GSLM-SQ	1.00	1.00	1.00	0.97	1.00	7.00	0.03	0.08	0.89	5.06
	GSLM-QA	1.00	1.00	1.00	0.96	1.00	7.00	0.04	0.17	0.79	5.16
	GSLM-HB	1.00	1.00	1.00	0.99	1.00	7.00	0.01	0.09	0.90	5.10
	SIRS-t	0.96	0.00	0.96	0.39	0.95	4.00	1.00	0.00	0.00	3.26
	MBKR-t	0.94	0.00	0.95	0.45	0.98	4.00	1.00	0.00	0.00	3.32
	DC-t	0.95	0.00	0.97	0.41	0.95	4.00	1.00	0.00	0.00	3.28
	Ball-t	0.85	0.00	0.84	0.16	0.87	4.00	1.00	0.00	0.00	2.72
	QaSIS-t	0.53	0.00	0.63	0.25	0.76	4.00	1.00	0.00	0.00	2.17

these compteting methods are designed to evaluated the marginal relationships between the response and covariates. In Example 2, GSLM-SVM and GSLM-LOG are also able to identify all those truly informative variables acting on the true conditional logit function with high probability, but all the other competitors tend to be underfiting since they miss some important covariates. Furthermore, when the correlation structure with $\eta = 0.5$ is considered, identifying those truly informative variables becomes more challenging, yet the proposed method still outperforms the other methods in most scenarios. As indicated in Table 4, the proposed method tends to choose more variables when the correlation among covariates presents. This is probably due to the fact that the binary responses are less informative than the continuous one (Fan, Samworth and Wu, 2009; Fan and Song, 2010), and thus in general, it becomes more challenging to conduct estimation and sparse learning for the task of classification. In practice, choosing some extra variables is usually much less severe than missing some important variables, and the performance of the proposed method may be further improved with finer tuning or a larger sample size.

More specifically, the existing methods have included almost all the informative variables as $\eta = 0$, but they tend to miss some truly important variables as $\eta = 0.5$ even by keeping the first $\lceil n/\log n \rceil$ variables. It is worthy pointing out that the proposed method is computationally efficient, which is clearly

TABLE 3
The averaged performance measures in Example 2 with $n = 500$ and $\eta = 0$.

p	Method	X_1	X_2	X_3	MaxSize	U	O	C	MeanSize
5000	GSLM-SVM	1.00	0.95	0.97	5.00	0.07	0.20	0.73	3.18
	GSLM-LOG	1.00	0.93	0.95	5.00	0.11	0.05	0.84	2.94
	SIRS-t	0.99	0.60	0.67	3.00	0.49	0.00	0.51	2.26
	MBKR-t	0.99	0.62	0.68	3.00	0.47	0.00	0.53	2.29
	DC-t	0.98	0.61	0.68	3.00	0.48	0.00	0.52	2.27
	MVxy-t	0.98	0.60	0.67	3.00	0.49	0.00	0.51	2.25
	Kol. Filter-t	0.94	0.51	0.59	3.00	0.66	0.00	0.34	2.04
10000	GSLM-SVM	1.00	0.96	0.96	6.00	0.07	0.35	0.58	3.38
	GSLM-LOG	1.00	0.96	0.94	4.00	0.09	0.07	0.84	2.97
	SIRS-t	1.00	0.65	0.59	3.00	0.54	0.00	0.46	2.24
	MBKR-t	1.00	0.63	0.58	3.00	0.53	0.00	0.47	2.21
	DC-t	1.00	0.65	0.65	3.00	0.51	0.00	0.49	2.30
	MVxy-t	1.00	0.66	0.63	3.00	0.52	0.00	0.48	2.29
	Kol. Filter-t	0.98	0.57	0.45	3.00	0.71	0.00	0.29	2.00
50000	GSLM-SVM	1.00	0.93	0.98	10.00	0.08	0.15	0.77	3.50
	GSLM-LOG	1.00	0.90	0.92	6.00	0.13	0.24	0.63	3.14
	SIRS-t	0.98	0.42	0.45	3.00	0.82	0.00	0.18	1.85
	MBKR-t	0.98	0.44	0.44	3.00	0.81	0.00	0.19	1.86
	DC-t	0.98	0.45	0.47	3.00	0.78	0.00	0.22	1.90
	MVxy-t	0.98	0.47	0.48	3.00	0.78	0.00	0.22	1.93
	Kol. Filter-t	0.90	0.31	0.35	3.00	0.92	0.00	0.08	1.56
100000	GSLM-SVM	1.00	0.89	0.95	15.00	0.13	0.14	0.73	3.47
	GSLM-LOG	1.00	0.84	0.97	6.00	0.18	0.32	0.50	3.28
	SIRS-t	0.99	0.36	0.42	3.00	0.82	0.00	0.18	1.77
	MBKR-t	0.99	0.32	0.38	3.00	0.87	0.00	0.13	1.69
	DC-t	0.99	0.36	0.44	3.00	0.82	0.00	0.18	1.79
	MVxy-t	0.99	0.37	0.44	3.00	0.82	0.00	0.18	1.80
	Kol. Filter-t	0.85	0.27	0.29	3.00	0.97	0.00	0.03	1.41

demonstrated by the computing times given in Table 5 based on a computation machine with 8 cores Intel Xeon E5-2695 CPU and 16 GB memory.

6.2. Application to interaction selection

In this subsection, we study the interaction selection. Specifically, we consider the regression with squared loss, check loss with $\tau = 0.5$ and Huber loss, and assess the performance of four methods: the regularized interaction selection method (RAMP, Hao, Feng and Zhang, 2018), the interaction pursuit with distance correlation (IPDC, Kong et al., 2017), the forward selection methods (iFort and iForm, Hao and Zhang, 2014), and the proposed method. For IPDC, we also report the truncated results and denote it as IPDC-t. Note that the computational cost of the existing nonparametric interaction selection methods (Radchenko and James, 2010; Dong and Wu, 2022) is very expensive, and thus they are not included in this numerical study where large dimensions are considered.

The following simulated example is examined under various scenarios.

Example 3. The data generating scheme is the same as Example 1 except that the response y_i is generated as $y_i = 2\,(f(x_{i1}) - f(x_{i2}) + f(x_{i3}) - f(x_{i4})) + 5\pi\,(g(x_{i1}, x_{i2}) - g(x_{i2}, x_{i3}) + g(x_{i3}, x_{i4})) + \epsilon_i$, where $f(u) = \exp(u)$, $g(u, v) = \cos^2(\pi uv)$, and ϵ_i's are independently drawn from $N(0, 1)$. Clearly, the first four variables are truly informative and $(X_1X_2,\ X_2X_3,\ X_3X_4)$ are important

TABLE 4
The averaged performance measures of the proposed framework and its competitors in Example 2 with $n = 500$ and $\eta = 0.5$.

p	Method	X_1	X_2	X_3	MaxSize	U	O	C	MeanSize
5000	GSLM-SVM	0.96	1.00	1.00	8.00	0.04	0.33	0.63	3.58
	GSLM-LOG	0.95	1.00	1.00	9.00	0.05	0.17	0.78	3.20
	SIRS-t	0.55	0.95	0.16	3.00	0.90	0.00	0.10	1.66
	MBKR-t	0.53	0.95	0.18	3.00	0.89	0.00	0.11	1.66
	DC-t	0.54	0.94	0.18	3.00	0.89	0.00	0.11	1.66
	MVxy-t	0.55	0.94	0.19	3.00	0.88	0.00	0.12	1.68
	Kol. Filter-t	0.36	0.83	0.17	3.00	0.97	0.00	0.03	1.36
10000	GSLM-SVM	0.95	0.99	1.00	10.00	0.06	0.28	0.66	3.69
	GSLM-LOG	0.94	0.99	1.00	6.00	0.07	0.27	0.66	3.28
	SIRS-t	0.63	0.94	0.08	3.00	0.94	0.00	0.06	1.65
	MBKR-t	0.63	0.95	0.07	3.00	0.95	0.00	0.05	1.65
	DC-t	0.63	0.93	0.09	3.00	0.93	0.00	0.07	1.65
	MVxy-t	0.61	0.93	0.10	3.00	0.92	0.00	0.08	1.64
	Kol. Filter-t	0.37	0.73	0.10	3.00	0.98	0.00	0.02	1.20
50000	GSLM-SVM	0.92	1.00	1.00	27.00	0.08	0.37	0.55	4.68
	GSLM-LOG	0.88	0.99	0.99	43.00	0.14	0.41	0.45	4.40
	SIRS-t	0.40	0.95	0.04	3.00	0.99	0.00	0.01	1.39
	MBKR-t	0.40	0.93	0.07	3.00	0.97	0.00	0.03	1.40
	DC-t	0.40	0.90	0.07	3.00	0.98	0.00	0.02	1.37
	MVxy-t	0.39	0.90	0.07	3.00	0.98	0.00	0.02	1.36
	Kol. Filter-t	0.28	0.61	0.06	2.00	1.00	0.00	0.00	0.96
100000	GSLM-SVM	0.86	0.99	1.00	56.00	0.15	0.45	0.40	4.14
	GSLM-LOG	0.84	0.99	0.97	14.00	0.19	0.31	0.50	4.33
	SIRS-t	0.44	0.92	0.02	3.00	0.98	0.00	0.02	1.38
	MBKR-t	0.43	0.91	0.02	3.00	0.98	0.00	0.02	1.36
	DC-t	0.43	0.89	0.04	3.00	0.97	0.00	0.03	1.36
	MVxy-t	0.41	0.89	0.04	3.00	0.97	0.00	0.03	1.34
	Kol. Filter-t	0.23	0.62	0.05	3.00	1.00	0.00	0.00	0.91

TABLE 5
Comparison of all the methods in terms of averaged run-time (in seconds) in Examples 1 and 2.

	p	GSLM-SQ	GSLM-QA	GSLM-HB	MBKR	SIRS	DC	Ball	QaSIS
Ex. 1	5000	6.2	7.2	6.9	655.4	21.2	114.4	9.4	11.9
	10000	9.4	9.8	9.0	1106.9	41.8	226.4	19.0	24.7
	50000	43.0	41.4	36.4	5656.6	200.8	1059.0	91.5	114.0
	100000	108.8	97.9	93.9	11494.1	468.4	2612.0	213.7	274.8
		GSLM-SVM	GSLM-LOG		MBKR	SIRS	DC	MV-SIS	Kol. Filter
Ex. 2	5000	5.9	9.2		266.7	18.6	116.2	71.6	54.6
	10000	8.8	11.9		523.0	36.0	231.3	139.5	105.1
	50000	35.7	38.1		2783.0	179.9	1106.3	691.5	515.2
	100000	98.7	101.3		6443.3	405.2	2607.9	1542.8	1146.2

interaction factors. In this example, we consider the same scenarios as those of Section 6.1, and the averaged performance measures for the case $\eta = 0.5$ are summarized in Tables 6 and the results for the case with $\eta = 0$ is given in Tables S1 of the Supplementary Materials. In these tables, "S_M" denotes the frequency of covering all the four main effects, "NumMain" denotes the average number of selected main effects, "X_iX_j" refers to the frequency of selecting the corresponding interaction effects between the i-th and j-th covariates, "NumInter" and "MaxInter" are the averaged and maximum numbers of selected interaction effects, and "C_I", "U_I", "O_I" are the frequency of correct-fitting, under-fitting, and over-fitting in terms of interaction effects, respectively.

TABLE 6

The averaged performance measures in Example 3 with $n = 500$ and $\eta = 0.5$ (IPDC selects $\lceil n/\log n \rceil$ main effects and $\lceil n/\log n \rceil$ interaction factors).

p	Method	S_M	NumMain	X_1X_2	X_2X_3	X_3X_4	U_I	O_I	C_I	NumInter	MaxInter
5000	GSLM-SQ	1.00	4.07	1.00	1.00	1.00	0.00	0.17	0.83	3.17	4.00
	GSLM-QA	1.00	4.07	1.00	0.98	1.00	0.02	0.15	0.83	3.14	5.00
	GSLM-HB	1.00	4.04	1.00	1.00	1.00	0.00	0.22	0.78	3.22	4.00
	RAMP	1.00	4.02	0.48	0.09	0.06	1.00	0.00	0.00	0.63	2.00
	iFort	1.00	4.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	iForm	1.00	4.00	0.14	0.00	0.00	1.00	0.00	0.00	0.14	1.00
	IPDC	1.00	81.00	0.87	0.35	0.91	0.79	0.21	0.00	81.00	81.00
	IPDC-t	0.91	3.88	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
10000	GSLM-SQ	1.00	4.11	1.00	1.00	1.00	0.00	0.14	0.86	3.14	4.00
	GSLM-QA	1.00	4.10	0.98	0.95	1.00	0.06	0.14	0.80	3.08	5.00
	GSLM-HB	1.00	4.09	1.00	1.00	0.99	0.01	0.14	0.85	3.14	5.00
	RAMP	1.00	4.03	0.47	0.13	0.09	0.97	0.00	0.03	0.69	3.00
	iFort	1.00	4.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	iForm	1.00	4.00	0.12	0.01	0.00	1.00	0.00	0.00	0.13	1.00
	IPDC	1.00	81.00	0.94	0.39	0.87	0.73	0.27	0.00	81.00	81.00
	IPDC-t	0.83	3.77	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
50000	GSLM-SQ	0.99	4.12	0.99	0.99	1.00	0.01	0.13	0.86	3.14	5.00
	GSLM-QA	1.00	4.21	1.00	0.94	0.97	0.08	0.14	0.78	3.06	4.00
	GSLM-HB	0.99	4.12	0.99	0.99	1.00	0.01	0.15	0.84	3.16	5.00
	RAMP	1.00	4.03	0.33	0.10	0.05	0.99	0.00	0.01	0.48	3.00
	iFort	1.00	4.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	iForm	1.00	4.00	0.11	0.00	0.00	1.00	0.00	0.00	0.11	1.00
	IPDC	1.00	81.00	0.91	0.40	0.90	0.69	0.31	0.00	81.00	81.00
	IPDC-t	0.60	3.45	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
100000	GSLM-SQ	1.00	4.19	1.00	0.99	1.00	0.01	0.12	0.87	3.11	4.00
	GSLM-QA	1.00	4.21	0.97	0.96	0.99	0.06	0.13	0.81	3.08	5.00
	GSLM-HB	0.99	4.22	0.99	0.98	0.99	0.02	0.12	0.86	3.10	4.00
	RAMP	1.00	4.01	0.32	0.06	0.02	1.00	0.00	0.00	0.40	2.00
	iFort	1.00	4.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
	iForm	1.00	4.00	0.06	0.00	0.00	1.00	0.00	0.00	0.06	1.00
	IPDC	1.00	81.00	0.86	0.50	0.89	0.66	0.34	0.00	81.00	81.00
	IPDC-t	0.68	3.48	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00

It is clear that the proposed framework outperforms all its competitors in that it can identify all the non-linear interaction effects with high probability, while the other methods tend to be under-fitting. As indicated in Table 6, the proposed method still achieves the best performance in all the scenarios. The IPDC method, which keeps $\lceil n/\log n \rceil = 81$ interaction effects, still tends to be underfiting. However, IPDC-t fails in all the scenarios, which implies that the ranking given by IPDC may not be accurate. The poor performance of the methods RAMP, iFort and iForm, in identifying interaction effects, is probably due to the fact that they are designed for the parametric cases and the marginal linear correlations among those interaction effects and the response in Example 3 are quite weak. We refer to Example S1 of the Supplementary Material for some additional results under a parametric setting, where the similar conclusion can be drawn.

6.3. Real application to the human breast cancer study

In this section, we use the proposed method to analyze a real dataset on the human breast cancer study (Zhang et al., 2016), which can be downloaded at https://www.ncbi.nlm.nih.gov/geo/ with accessing number GSE20194. It

consists of 278 patients, whereas 164 of them have positive oestrogen receptor status and the other 114 have negative oestrogen receptor status, and each patient is characterized by 22283 probs. A patient has positive oestrogen receptor status if the receptors for estrogen are detected, which suggests that estrogen may send signals to the cancer cells among normal breast cells to promote their growth. It has been shown that roughly 80 percent of the patients diagnosed with breast cancers, have the positive estrogen receptor status. Consequently, the main interest of the study is to identify those genes related with the oestrogen receptor status.

For interpretability, we map the prob IDs to the gene symbol and delete those IDs that cannot be mapped. The map relationship is also provided by `https://www.ncbi.nlm.nih.gov/geo/`. Finally, 19820 genes are considered in this study. Clearly, the response variable in this dataset is binary, and thus we apply all the methods used in Example 2 to identify the informative genes. The genes chosen by all considered methods are reported in Table 7.

Clearly, GSLM-SVM chooses 10 genes and GSLM-LOG selects 26 important genes while all the other screening-based methods keep 36 genes as suggested and their truncated versions select at most 7 genes. It is interesting to point out that four genes, including PRKD3, TNNT1, HOXA1 and IRX4, are identified by GSLM-SVM and GLSM-LOG, but missed by all the other competing methods. More importantly, the existing literature suggests that these four genes have important biological implications. Specifically, PRKD3 functions as an important oncogenic driver in the invasive breast cancer (Liu et al., 2017); TNNT1 facilitates proliferation of breast cancer cells by promoting the G1/S phase transition (Shi et al., 2018); HOXA1 upregulation is associated with poor prognosis and tumor progression in the breast cancer (Liu, Liu and Lu, 2019); Corrêa et al. (2017) discovers the high levels expression of IRX4 in the breast cancer plasma samples.

To assess the performance in prediction, we also report some accuracy measures of the proposed framework and all the screening-based methods given their chosen genes. Specifically, we randomly split the dataset with 84 (30%) patients for testing and the rest for training, and refit a standard kernel SVM by using the R package *kernlab*. The splitting process is replicated 100 times, and the boxplots of the prediction errors are given in the left panel of Figure 1. Since the oestrogen receptor status plays an important role in assisting diagnosis for the breast cancer, it is more severe to miss-classify the patients with positive oestrogen receptor status to be negative. Therefore, we also summarize the false negative rates in the right panel of Figure 1.

It is clear from Figure 1 that GSLM-LOG and GSLM-SVM outperforms all their competing methods in predictions. Note that GSLM-LOG selects 26 genes and GSLM-SVM selects 10 genes while the other screening based methods choose 36 genes. This implies that the proposed method has probably identified some important genes which are missed by its competing methods.

TABLE 7

The genes selected by different methods in the application to the human breast cancer study.

Method	Number	Selected Genes					
GSLM-SVM	10	CDH3	ESR1	GREB1	AGR2	PRKD3	TNNT1
		NAT1	HOXA1	VGLL1	IRX4		
GSLM-LOG	26	SCCPDH	SPTLC2	CDH3	CA12	PTPRG	ESR1
		REPS2	GREB1	ZIC1	SCGB1D2	SLC15A1	SEPT9
		AGR2	ABCC3	BLZF1	PRKD3	ANXA9	TNNT1
		NAT1	HOXA1	VGLL1	VAV3	SLC37A1	MBNL3
		IRX4	NPAS2				
SIRS-t	7	SLC39A6	CA12	ESR1	AGR2	GATA3	TBC1D9
		NAT1					
MBKR-t	4	ESR1	GATA3	TBC1D9	CA12		
MV-SIS-t	4	ESR1	GATA3	TBC1D9	CA12		
DC-t	1	ESR1					
Kol.Filter-t	1	ESR1					
SIRS	36	ESR1	GATA3	TBC1D9	CA12	NAT1	SLC39A6
		AGR2	FOXA1	GREB1	MLPH	DNAJC12	VAV3
		C6orf211	XBP1	VGLL1	KDM4B	ANXA9	CDH3
		DNALI1	IL6ST	UGCG	TFF1	MKL2	SCCPDH
		EVL	IGF1R	TTC39A	METRN	GFRA1	MYB
		PBX1	CERS6	WWP1	MCCC2	IGFBP4	ABAT
MBKR	36	ESR1	GATA3	TBC1D9	CA12	NAT1	C6orf211
		SLC39A6	FOXA1	DNAJC12	GREB1	KDM4B	IGF1R
		UGCG	VAV3	MKL2	EVL	IL6ST	ANXA9
		AGR2	ABAT	GFRA1	TTC39A	MAGED2	MLPH
		MCCC2	WWP1	XBP1	SCCPDH	RABEP1	CDH3
		EGFR	TFF1	VGLL1	DNALI1	DACH1	MYB
MV-SIS	36	ESR1	GATA3	TBC1D9	CA12	NAT1	C6orf211
		SLC39A6	DNAJC12	FOXA1	GREB1	IGF1R	KDM4B
		UGCG	VAV3	MKL2	IL6ST	EVL	ANXA9
		GFRA1	ABAT	MCCC2	AGR2	MAGED2	WWP1
		TFF1	EGFR	DNALI1	XBP1	TTC39A	RABEP1
		MLPH	SCCPDH	CDH3	VGLL1	DACH1	COX6C
DC	36	ESR1	GATA3	TBC1D9	CA12	NAT1	C6orf211
		SLC39A6	FOXA1	AGR2	DNAJC12	GREB1	MLPH
		KDM4B	VAV3	MKL2	IL6ST	EVL	IGF1R
		ANXA9	VGLL1	UGCG	XBP1	GFRA1	DNALI1
		TFF1	TTC39A	ABAT	WWP1	CDH3	MCCC2
		SCCPDH	MAGED2	RABEP1	MYB	METRN	PBX1
Kol.Filter	36	ESR1	GATA3	TBC1D9	NAT1	CA12	C6orf211
		IL6ST	SLC39A6	UGCG	ANXA9	DNAJC12	EVL
		GREB1	TFF1	ABAT	FOXA1	MKL2	VAV3
		IGF1R	KDM4B	MYB	MLPH	GFRA1	MCCC2
		VGLL1	DNALI1	COX6C	RARA	BTG3	SLC44A4
		WWP1	CLSTN2	XBP1	EGFR	AGR2	SCCPDH

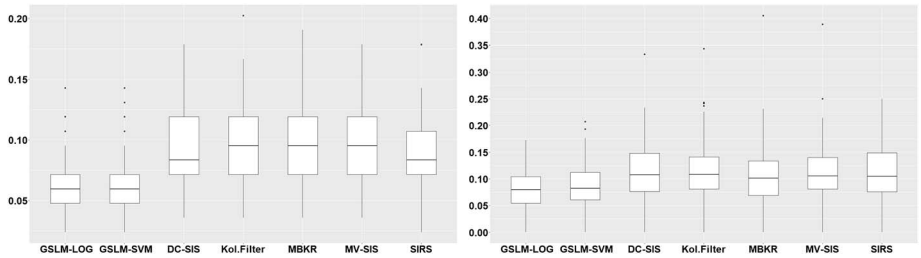

FIG 1. *The boxplots of the testing errors (left panel) and the false negative rates (right panel) of all the methods considered in Section 6.3.*

7. Discussion

It is known that continuous functions can be well approximated by those functions of the RKHS induced by some universal kernels under the infinity norm. We thus propose a general structure learning framework within the induced RKHS, which can be used to solve many interesting statistical problems, such as sparse learning, interaction selection, model identification and so on. The proposed framework is inspired by the fact that gradient functions can be employed to define the underlying structures of true target functions without model specifications, and the nice properties of the RKHS facilitate the whole computation of the proposed framework. It is methodologically and computationally simple, and thus can efficiently process large-scale datasets. More importantly, it attains many advantages that it works for a general family of loss functions, and admits general dependence structures with theoretical guarantees under weaker conditions than existing methods. In our future work, we may consider more general derivatives to learn much more complicated structures in some challenging areas such as manifold learning and graph estimation.

Acknowledgments

The authors thank the associate editor and two anonymous referees for their constructive suggestions, which significantly improve this paper. The authors also thank Professor Junhui Wang for helpful and valuable comments on the initial draft of this work.

Supplementary Material

Supplement to "Structure learning via unstructured kernel-based M-estimation"
(doi: 10.1214/23-EJS2153SUPP; .pdf).

References

Aubin, J. (1993). *An Introduction to Non-linear Analysis (translated from the French by Stephen Wilson*. Springer-Verlag, Berlin.

Barber, R. and Candës, E. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics* **43** 2055–2085. MR3375876

Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research* **3** 463–482. MR1984026

Chen, F., He, X. and Wang, J. (2021). Learning sparse conditional distribution: An efficient kernel-based approach. *Electronic Journal of Statistics* **15** 1610–1635. MR4255310

Corrêa, S., Panis, C., Binato, R., Herrera, A., Pizzatti, L. and Abdelhay, E. (2017). Identifying potential markers in breast cancer subtypes using plasma label-free proteomics. *Journal of Proteomics* **151** 33–42.

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110** 630–641. MR3367253

Dasgupta, S., Goldberg, Y. and Kosorok, M. (2019). Feature elimination in kernel machines in moderately high dimensions. *Annals of Statistics* **47** 497-526. MR3909940

Deng, W., Cocker, B., Mukherjee, R., Liu, J. and Coull, B. (2022). Towards a unified framework for uncertainty-aware nonlinear variable selection with theoretical guarantees. *Advances in Neural Information Processing Systems* **35** 27636–27651.

Dong, Y. and Wu, Y. (2022). Nonparametric interaction selection. *Statistica Sinica* **32** 1563–1582. MR4449906

Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics* **7** 1–42. MR3020412

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B* **70** 849–911. MR2530322

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica* **20** 101–148. MR2640659

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10** 2013-2038. MR2550099

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38** 3567–3604. MR2766861

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research* **21** 1–38. MR4209491

Fu, A., Narasimhan, B. and Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software* **94** 1–34.

Han, X. (2019). Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data. *Annals of Statistics* **47** 1995–2022. MR3953442

Hang, H. and Steinwart, I. (2018). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Annals of Statistics* **45** 708–743. MR3650398

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. MR3832213

Hao, N. and Zhang, H. (2014). Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association* **109** 1285–1301. MR3265697

He, X., Lv, S. and Wang, J. (2020). Variable selection for classification with derivative-induced regularization. *Statistica Sinica* **30** 2075–2103. MR4260756

He, X., Wang, L. and Hong, H. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics* **41** 342–369. MR3059421

He, X. and Wang, J. (2020). Discovering model structure for partially linear models. *Annals of the Institute of Statistical Mathematics* **72** 45–63. MR4052650

He, X., Wang, J. and Lv, S. (2021). Efficient kernel-based variable selection with sparsistency. *Statistica Sinica* **31** 2123–2151. MR4328855

Huang, J., Horowitz, J. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38** 2282–2313. MR2676890

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.

Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Annals of Statistics* **45** 897–922. MR3650404

Lemhadri, I., Ruan, F., Abraham, L. and Tibshirani, R. (2021). LassoNet: a neural network with feature sparsity. *Journal of Machine Learning Research* **22** 1–29. MR4279778

Li, Y. and Liu, J. (2019). Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association* **114** 271–286. MR3941254

Li, X. and Xu, C. (2023). Feature screening with conditional rank utility for big-data classification. *Journal of the American Statistical Association* In Press 1–35.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900

Lian, H., Liang, H. and Ruppert, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica* **25** 591–607. MR3379090

Liu, J., Liu, J. and Lu, X. (2019). HOXA1 upregulation is associated with poor prognosis and tumor progression in breast cancer. *Experimental and Therapeutic Medicine* **17** 1896–1902.

Liu, Y., Li, J., Zhang, J., Yu, Z., Yu, S., Wu, L., Wang, Y., Gong, X., Wu, C., Cai, X., Mo, L., Wang, M., Gu, J. and Chen, L. (2017). Oncogenic protein kinase D3 regulating networks in invasive breast cancer. *International Journal of Biological Sciences* **13** 748–758.

Loh, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics* **45** 866–896. MR3650403

Lv, S., Lin, H., Lian, H. and Huang, J. (2018). Oracle inequalities for sparse

additive quantile regression in reproducing kernel Hilbert space. *Annals of Statistics* **46** 781–813. MR3782384

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *Annal of Statistics* **38** 526–565. MR2590050

Mukherjee, S., Wu, Q. and Zhou, D. (2010). Learning gradients on manifolds. *Bernoulli* **16** 181–207. MR2648754

Mukherjee, S. and Zhou, D. (2006). Learning coordinate covariances via gradients. *Journal of Machine Learning Research* **7** 519–549. MR2274377

Pan, W., Wang, X., Xiao, W. and Zhu, H. (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association* **114** 928-937. MR3963192

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. MR2796570

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F. and Moore, J. (2001). Multifactor-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. MR3104492

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232. MR2949354

Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics* **65** 807–832. MR3105798

Shi, Y., Zhao, Y., Zhang, Y., AiErken, N., Shao, N., Ye, R., Lin, Y. and Wang, S. (2018). TNNT1 facilitates proliferation of breast cancer cells by promoting G1/S phase transition. *Life Sciences* **208** 161–166.

Simon-Gabriel, C. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research* **19** 1–29. MR3874152

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory* **51** 128–142. MR2234577

Steinwart, I. and Christmann, A. (2008a). Sparsity of SVMs that use the epsilon-insensitive loss. *Advances in Neural Information Processing Systems* **21** 1569–1576.

Steinwart, I. and Christmann, A. (2008b). *Support Vector Machine*. Springer. MR2796580

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. MR3144467

Takeuchi, I., Le, Q., Sears, T. and Smola, A. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research* **7** 1231–1264. MR2274404

Tang, X., Xue, F. and Qu, A. (2021). Individualized multidirectional variable selection. *Journal of the American Statistical Association* **116** 1280–1296. MR4309272

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

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B* **78** 589–611. MR3506794

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica* **19** 801–817. MR2514189

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. MR3517105

Ye, G. and Xie, X. (2012). Learning sparse gradients for variable selection and dimension reduction. *Machine Learning* **87** 303–355. MR2917060

Zhang, H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106** 1099–1112. MR2894767

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. MR3491134

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* **78** 53–76. MR3453646

Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics* **220** 456–463. MR2444183

Zhou, Y. and Zhu, L. (2018). Model-free feature screening for ultrahigh dimensional data through a modified Blum-Kiefer-Rosenblatt correlation. *Statistica Sinica* **28** 1351–1370. MR3821008

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics* **14** 185–205. MR2137897

Zhu, L., Li, L., Li, R. and Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106** 1464–1475. MR2896849