Electronic Journal of Statistics
Vol. 11 (2017) 2907–2930
ISSN: 1935-7524
DOI: 10.1214/17-EJS1300

Variable selection for partially linear models via learning gradients

Lei Yang

Division of Biostatistics
New York University School of Medicine
e-mail: ly888@nyu.edu

Yixin Fang

Department of Mathematical Sciences
New Jersey Institute of Technology
e-mail: yixin.fang@njit.edu

Junhui Wang

Department of Mathematics
City University of Hong Kong
e-mail: j.h.wang@cityu.edu.hk

and

Yongzhao Shao*

Division of Biostatistics
New York University School of Medicine
e-mail: yongzhao.shao@nyumc.org

Abstract: Partially linear models (PLMs) are important generalizations of linear models and are very useful for analyzing high-dimensional data. Compared to linear models, the PLMs possess desirable flexibility of non-parametric regression models because they have both linear and non-linear components. Variable selection for PLMs plays an important role in practical applications and has been extensively studied with respect to the linear component. However, for the non-linear component, variable selection has been well developed only for PLMs with extra structural assumptions such as additive PLMs and generalized additive PLMs. There is currently an unmet need for variable selection methods applicable to general PLMs without structural assumptions on the non-linear component. In this paper, we propose a new variable selection method based on learning gradients for general PLMs without any assumption on the structure of the non-linear component. The proposed method utilizes the reproducing-kernel-Hilbert-space tool to learn the gradients and the group-lasso penalty to select variables. In addition, a block-coordinate descent algorithm is suggested and some theoretical properties are established including selection consistency and estimation consistency. The performance of the proposed method is further evaluated via simulation studies and illustrated using real data.

*Supported by NIH grants P30 CA016087 and P30 AG008051. The authors thank the editor, the associate editor and the referees for insightful comments and suggestions.

Keywords and phrases: PLM, group Lasso, gradient learning, variable selection, high-dimensional data, reproducing kernel Hilbert space.

Received August 2016.

Contents

1. Introduction

The partially linear model (PLM) is an important generalization of the linear model [8]. During the past decades, it has become a useful tool in statistical analysis for parsimoniously modeling high dimensional data while reflecting nonlinear trend of some continuous covariates [17, 24, 35, 38]. And it has been applied to analyze data in many fields such as econometrics [51], biomedicine [20, 23, 55], and environmetrics [34].

The PLM has the flexibility of a nonparametric regression model, while it contains a linear component whose estimators have desirable asymptotic properties with simple interpretability. These features make it a very useful model for analyzing high-dimensional data where variable selection plays an important role. Variable selection for the linear component of the PLM has been well studied [3, 11, 33, 46]. However, existing variable selection methods for the non-linear component usually rely on some extra structural assumptions, e.g., additivity is a common assumption imposed on the structure of the non-linear component. Variable selection procedures have been developed for both additive partially linear models [27] and generalized additive partially linear models [45]. In addition, many other variable selection procedures proposed for additive models or generalized additive models can be also extended to conduct variable selection in additive partially linear models [18, 19, 26, 37, 47].

In this paper, we will develop a novel variable selection procedure, based on the idea of gradient learning, for partially linear models without imposing any assumption on the structure of the nonparametric component. The method of learning gradients can be traced back to Mukherjee and Zhou [30]. Other de-

velopments include Mukherjee and Wu [29] and De Brabanter et al. [7], which mainly focus on estimating the gradient functions to do regression or classification. Recently, gradient learning technique has been employed to conduct variable selection in non-parametric regression, such as Ying et al. [53], Ye and Xie [52] and Yang et al. [49]. Although gradient learning procedures are model-free, the computational cost for variable selection in the nonlinear component is very high. Therefore, the model-free gradient learning procedures have been only used for low dimensional data. In this paper, using partially linear models as extensions of linear models for high dimensional data with the added flexibility of nonparametric regression models for selected covariates, we can make the model-free gradient learning procedures applicable for high dimensional data.

In the literature, many nonparametric variable selection procedures [2, 6, 22, 28, 36] have been proposed via imposing structural assumptions. Lafferty and Wasserman [22] proposed a greedy method called "rodeo", for simultaneously performing local bandwidth selection and variable selection in nonparametric regression, which starts with a local linear estimator with large bandwidths, incrementally decreases the bandwidth of variables for which the gradient of the estimator with respect to bandwidth is large, and then conducts a sequence of hypothesis tests. Bertin and Lecue[2] proposed an l_1-penalized procedure in the non-parametric Gaussian regression model, but they only considered variable selection at a fixed point. Miller and Hall [28] proposed their "LABAVS" algorithm, along with several variable selection criteria including the local lasso, hard thresholding, and backward stepwise, but its computational cost is very high. Rosasco et al. [36] proposed a general nonparametric variable selection procedure, via modeling the regression function and penalizing its gradients. The difference between [36] and [49] is that the former models the gradients indirectly while the latter models the gradients directly. The current paper improves computational efficiency and convergence on the method proposed in [49] to conduct variable selection in partially linear models without assuming structural constraints on the nonlinear component. Similarly, we should be able to extend the method proposed in [36] to conduct variable selection in partially linear models. Finally, Comminges et al. [6] studied the asymptotic analysis of variable selection in nonparametric regression and revealed two different regimes. The setting considered in this paper belongs to the first regime.

The rest of the paper is organized as follows. In Section 2, we propose a variable selection procedure based on gradient learning for partially linear models, along with an algorithm for implementing the procedure and a method for selecting the tuning parameters. In Section 3, we study some asymptotic properties of the proposed procedure. In Section 4, we evaluate the performance of the proposed procedure via simulation studies and real data applications. We conclude the paper with some summary and discussion in Section 5.

2. Method

Assume data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, are independently generated from the partially linear model,

$$y = \mathbf{z}^T\boldsymbol{\beta}^* + f^*(\mathbf{w}) + \epsilon, \tag{2.1}$$

where y is the response variable, $\mathbf{x} = (\mathbf{z}^T, \mathbf{w}^T)^T$ consists of $d = p + q$ predictors with $\mathbf{z} = (z^{(1)}, \ldots, z^{(p)})^T$ and $\mathbf{w} = (w^{(1)}, \ldots, w^{(q)})^T$, $E(\epsilon) = 0$, and $V(\epsilon) = \sigma^2$. By convention, vectors are denoted by bold letters and their elements by non-bold letters. In this model, we assume that the effects of predictors in $\mathbf{z}$ are linear and the effects of predictors in $\mathbf{w}$ are non-linear, with $\boldsymbol{\beta}^*$ and f^* as their true effects, respectively. There is no assumption about the structure of f^*, which is only assumed to be twice differentiable.

As discussed in [12], statistical accuracy, model interpretability, and computational complexity are three important pillars of any statistical procedures. When the number of predictors d is large, variable selection plays a crucial rule in strengthening these three pillars. For this aim, we assume that the true partially linear model (2.1) is sparse in the sense that some elements of $\mathbf{z}$ and some elements of $\mathbf{w}$ have no effect on the response variable. Specifically, in the linear component, a predictor $z^{(j)}$ is noninformative if the corresponding effect $\beta_j^* = 0$, $j = 1, \cdots, p$, where p can be very large. In the nonlinear component, a predictor $w^{(l)}$ is noninformative if $g_l^*(\mathbf{w}) = \partial f^*(\mathbf{w})/\partial w^{(l)} \equiv 0$, $l = 1, \cdots, q$.

Denote $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and $\mathbf{x}_i = (\mathbf{z}_i, \mathbf{w}_i)$. Let $\mathbf{g} = (g_1, \cdots, g_q)^T$, g_l is the gradient function of f^* corresponding to $w^{(l)}$. The weighted square loss is

$$\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij}\big(y_i - y_j - \boldsymbol{\beta}^T(\mathbf{z}_i - \mathbf{z}_j) - \mathbf{g}^T(\mathbf{w}_i)(\mathbf{w}_i - \mathbf{w}_j)\big)^2$$

where the kernel weight $w_{ij} = e^{-\| \mathbf{x}_i - \mathbf{x}_j \|_2^2/\tau_n^2}$ depends on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ and some pre-specified parameter τ_n. In order to conduct variable selection in both the linear and nonlinear components, we propose to consider a penalized procedure, minimizing the following objective function over the vector of parameters $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and functions $\mathbf{g} = (g_1, \cdots, g_q)^T$,

$$\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) + \lambda_1 \sum_{l=1}^{p} \pi_{\mathbf{z},l}|\beta_l| + \lambda_2 \sum_{l=1}^{q} \pi_{\mathbf{w},l}\|g_l\|_K, \tag{2.2}$$

where $\|\cdot\|_K$ is the norm of some reproducing kernel Hilbert space (RKHS) with kernel $K(\cdot,\cdot)$. In addition, $\pi_{\mathbf{z},l}$ and $\pi_{\mathbf{w},l}$ are some weights to be discussed in Subsection 2.1, and λ_1 and λ_2 are tuning parameters that control the compromise between goodness-of-fit and parsimony of the selected model to be discussed in Subsection 2.2. The RKHS method is a very popular tool for modeling nonparametric function. By the representer theorem [43], the minimizer of (2.2) over g_l satisfies $\widehat{g}_l(\mathbf{w}) = \sum_{i=1}^n \alpha_i^{(l)} K(\mathbf{w}, \mathbf{w}_i)$. Denoting $\mathbf{K} = (K(\mathbf{w}_i, \mathbf{w}_j))_{n\times n} = (\mathbf{K}_1, \ldots, \mathbf{K}_n)$, $\boldsymbol{\alpha}_l = (\alpha_1^{(l)}, \ldots, \alpha_n^{(l)})^T$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q)$, the minimization in (2.2) is equivalent to

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ L(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_1 \sum_{l=1}^{p} \pi_{\mathbf{z},l}|\beta_l| + \lambda_2 \sum_{l=1}^{q} \pi_{\mathbf{w},l}\|\mathbf{K}^{1/2}\boldsymbol{\alpha}_l\|_2 \right\}, \tag{2.3}$$

where $L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i,j=1}^n w_{ij}\big(y_i - y_j - \boldsymbol{\beta}^T(\mathbf{z}_i - \mathbf{z}_j) - \mathbf{K}_i^T\boldsymbol{\alpha}(\mathbf{w}_i - \mathbf{w}_j)\big)^2/n(n-1)$.

Here are some remarks on the objective functions (2.2 and 2.3). First, the proposed partially linear model is a special case of the nonparametric model considered in Yang et al. [49], if we write the RKHS considered in Yang et al. [49] as the direct sum of an RKHS of linear functions with respect to $\mathbf{z}$ and an arbitrary one with respect to $\mathbf{w}$. However, this special case is still worth investigating, because (1) the number of parameters is reduced from $(p+q)n$ in a general RKHS in to $p+qn$ in this special RKHS. Therefore, compared with Yang et al. [49], our proposed method can be implemented much faster and can be applied to high dimensional data, (2) the resulting penalty term can be written as the summation of the well-known adaptive lasso penalty [56] and the adaptive group-lasso penalty [44] and (3) the tuning can be respectively considered for the linear component and the non-linear component.

Second, that the above penalized objective function can be used to conduct variable selection is due to the two sparsity-inducing penalties that are incorporated. The first penalty is the lasso type of penalty, which is the lasso penalty without weights $\pi_{\mathbf{z},l}$ [41] and is the adaptive lasso penalty with weights [56]. The second penalty is the group-lasso type of penalty, which is the group lasso penalty without weights $\pi_{\mathbf{w},l}$ [54] and is the adaptive group-lasso penalty with weights [44]. In particular, the group-lasso type of penalty on the nonparametric gradient functions has the so-called "all-in-all-out" property. That is, when λ_2 is large, some individual terms of the group-lasso penalty will become zero, that is $\|g_l\|_K = 0$ for some l's, implying that $g_l(\mathbf{w}) \equiv 0$, for any $\mathbf{w}$.

Third, the kernel weights w_{ij} are introduced, because $f^*(\mathbf{w}_j)$ can be locally approximated by $f^*(\mathbf{w}_i)+g^*(\mathbf{w}_i)^T(\mathbf{w}_j-\mathbf{w}_i)$. This idea of approximation comes from kernel weighted regression [9]. For simplicity, the parameter τ_n in the kernel weights is not considered as a tuning parameter, and can be set as the median over the pairwise distances among all the sample points [30].

2.1. Implementation

First we consider weights $\pi_{\mathbf{z},l}$ and $\pi_{\mathbf{w},l}$. As in [56], we select them as $\pi_{\mathbf{z},l} = 1/|\widetilde{\beta}_l|^{\gamma_1}$ and $\pi_{\mathbf{w},l} = 1/\|\widetilde{g}_l\|_2^{\gamma_2}$, where $\widetilde{\beta}_l$ and $\widetilde{g}_l$ are the initial estimates of β_l and g_l via the following,

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ L(\boldsymbol{\alpha},\boldsymbol{\beta}) + \widetilde{\lambda}_1\|\boldsymbol{\beta}\|_2^2 + \widetilde{\lambda}_2 \sum_{l=1}^{q} \|g_l\|_2^2 \right\}, \tag{2.4}$$

where tuning parameters $\widetilde{\lambda}_1$ and $\widetilde{\lambda}_2$ can be determined using cross validation or generalized cross-validation [14]. The computation of the above minimization problem with ridge type of penalties is fast, because the solutions have explicit formulae.

Now we are ready to describe an algorithm to implement (2.3). Although the proximal algorithm [1, 32] can be used, here we use the coordinate descent algorithm [13], which was also used in [49]. For this aim, we alternatively update

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When $\boldsymbol{\alpha}$ is given, we update $\boldsymbol{\beta}$ via the following minimization,

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ L_{\boldsymbol{\alpha}}(\boldsymbol{\beta}) + \lambda_1 \sum_{l=1}^{p} \pi_{\mathbf{z},l} |\beta_l| \right\}, \tag{2.5}$$

where $L_{\boldsymbol{\alpha}}(\boldsymbol{\beta}) = L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the loss function of $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$. This minimization can be solved by using R package "glmnet". Next we consider updating $\boldsymbol{\alpha}$ given $\boldsymbol{\beta}$. Denote $\bar{\boldsymbol{\alpha}} = \mathbf{K}^{1/2}\boldsymbol{\alpha} = (\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_q)$ and $\mathbf{M}_{ij} = (\mathbf{I}_p \otimes \mathbf{K}^{-1/2})\big((\mathbf{w}_i - \mathbf{w}_j) \otimes \mathbf{K}_i\big)$. Given $\boldsymbol{\beta}$, we can update $\bar{\boldsymbol{\alpha}}$ via the following minimization,

$$\underset{\bar{\boldsymbol{\alpha}}}{\operatorname{argmin}}\, \bar{L}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\alpha}}) + \lambda_2 \sum_{l=1}^{q} \pi_{\mathbf{w},l} \|\bar{\boldsymbol{\alpha}}_l\|_2, \tag{2.6}$$

where $\bar{L}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\alpha}}) = \sum_{i \neq j} w_{ij}[y_i - y_j - \boldsymbol{\beta}^T(\mathbf{z}_i - \mathbf{z}_j) - \bar{\boldsymbol{\alpha}}^T \mathbf{M}_{ij}]^2 / n(n-1)$. This is because, by the representer theorem, $\mathbf{g}(\mathbf{w}_i)^T(\mathbf{w}_i - \mathbf{w}_j) = \boldsymbol{\alpha}^T\big((\mathbf{w}_i - \mathbf{w}_j) \otimes \mathbf{K}_i\big)$. This minimization can also be solved by using the R package "gglasso".

2.2. Tuning

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and $\lambda = (\lambda_1, \lambda_2)^T$. Assume that some appropriately tuning parameters, $\widehat{\lambda} = (\widehat{\lambda}_1, \widehat{\lambda}_2)$, are selected, say, by the procedure discussed in Subsection 2.2. Let $\widehat{\boldsymbol{\theta}}_{\widehat{\lambda}} = (\widehat{\boldsymbol{\alpha}}_{\widehat{\lambda}}^T, \widehat{\boldsymbol{\beta}}_{\widehat{\lambda}}^T)^T$ be the minimizer of (2.2 and 2.3). Then $\mathcal{S}_{\widehat{\lambda}} = \{l : \widehat{\theta}_{\widehat{\lambda}l} \neq 0, l = 1, \ldots, d\}$ is the set of those selected variables for the partially linear model under consideration, where $\mathcal{A}_{\widehat{\lambda}} = \{l : \widehat{\boldsymbol{\alpha}}_{\widehat{\lambda}l} \neq 0, l = 1, \ldots, q\}$ is for the non-linear component and $\mathcal{B}_{\widehat{\lambda}} = \{l : \widehat{\beta}_{\widehat{\lambda}l} \neq 0, l = 1, \ldots, p\}$ is for the linear component. We propose to select tuning parameters λ_1 and λ_2 using the selection stability procedure proposed by [39]. Here we briefly describe this tuning procedure, and the reader is referred to [39] for more details. Given $\lambda = (\lambda_1, \lambda_2)$, let $\mathcal{S}_\lambda(\mathcal{D})$ denote the subset of variables selected for the partially linear model under consideration based on training dataset $\mathcal{D}$. Randomly partition the original dataset $\mathcal{D}$ into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$, we have $\mathcal{S}_\lambda(\mathcal{D}_1)$ and $\mathcal{S}_\lambda(\mathcal{D}_2)$. Then we use Cohen's kappa [5] to measure the agreement of these two subsets and denote it as $\kappa(\mathcal{S}_\lambda(\mathcal{D}_1), \mathcal{S}_\lambda(\mathcal{D}_2))$. When B random partitions are repeated, we obtain B copies of kappa measures, and denote their average as $\text{stab}(\lambda)$, which measures the selection stability given λ. Finally, we consider its maximizer, $\widehat{\lambda} = \text{argmax}_\lambda\{\text{stab}(\lambda)\}$, as an estimate for the tuning parameters.

3. Asymptotic theory

In this section, we derive the estimation consistency and variable selection consistency of the proposed method under the following assumptions.

Assumption A1. The support $\mathcal{Z}$ of $\mathbf{Z}$ and support $\mathcal{W}$ of $\mathbf{W}$ are non-degenerate compact subsets of $\mathcal{R}^p$ and $\mathcal{R}^q$, respectively. Also, $\sup_{\mathbf{w}} \|\mathbf{H}^*(\mathbf{w})\|_2 \leq c_1$ for some constant c_1, where $\mathbf{H}^*(\mathbf{w}) = \nabla^2 f^*(\mathbf{w})$ and $\|\cdot\|_2$ is the l_2 norm, denoted as the largest eigenvalue of the matrix.

Assumption A2. For some constant c_2, the probability density $p(\mathbf{x})$ of $\mathbf{x}$ exists and satisfies $|p(\mathbf{x}) - p(\mathbf{x}')| \leq c_2 d_x(\mathbf{x}, \mathbf{x}')$, for any $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{X}$, where $d_x(\cdot, \cdot)$ is the Euclidean distance on $\mathcal{X}$.

Assumption A3. There exist constants c_3 and c_4 such that $c_3 \leq \lim_{n\to\infty} \min_{1\leq l\leq p_0} \pi_{\mathbf{z},l} \leq \lim_{n\to\infty} \max_{1\leq l\leq p_0} \pi_{\mathbf{z},l} \leq c_4$, $c_3 \leq \lim_{n\to\infty} \min_{1\leq l\leq q_0} \pi_{\mathbf{w},l} \leq \lim_{n\to\infty} \max_{1\leq l\leq q_0} \pi_{\mathbf{w},l} \leq c_4$, $\lambda_1 \pi_{\mathbf{z},l} \to \infty$ for $l > p_0$ and $n^{-3/2}\lambda_2 \pi_{\mathbf{w},l} \to \infty$ for $l > q_0$.

About the above three assumptions, Assumption A1 is often used in the literature of partially linear models [16], which can simplify the technical proof significantly. Assumption A2 specifies the smoothness of the density function of $\mathbf{x}$ by the regular Lipschitz condition. Under Assumptions A1 and A3, the density $p(\mathbf{x})$ is continuous and bounded on $\mathcal{X}$, and therefore there exists some constant c_5 such that $\sup_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \leq c_5$. Moreover, if we denote $\mathcal{X}_t = \{\mathbf{x} \in \mathcal{X} : d_X(\mathbf{x}, \partial\mathcal{X}) < t\}$, where $\partial\mathcal{X}$ is the boundary of $\mathcal{X}$ and $d_x(\mathbf{x}, \partial\mathcal{X}) = \inf_{\mathbf{u}\in\partial\mathcal{X}} d_X(\mathbf{x}, \mathbf{u})$, then there exists some constant c_6 such that $\text{Prob}(\mathcal{X}_t) \leq c_6 t$ for any t. Assumption A3 provides the convergence rate of adaptive Lasso weight, which will guarantee the estimation and selection consistency.

Theorem 3.1. *Under Assumptions A1–A3, if* $\lambda_1 = n^{-1/8}$, $\lambda_2 = n^{-1/8}$ *and* $\tau_n = n^{-\frac{1}{16(p+q+3)}}$, *then* $\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|_2 = O_p\big(n^{-\frac{1}{16(p+q+3)}}\big)$ *and* $\|\widehat{\mathbf{g}}_\lambda - \mathbf{g}^*\|_2 = O_p\big(n^{-\frac{1}{16(p+q+3)}}\big)$ *in probability, as* $n \to \infty$.

Compared with weak estimation consistency in Yang et al. [49], the strong estimation convergence rate is established under Kernel norm regularization in this paper. However, we only consider the estimation consistency in Theorem 3.1, which is the common practice in literature [10, 49, 56] and the sparsity level will be discussed in Theorem 3.2.

Next, let $\mathcal{S}^* = \{j : \beta_j^* \neq 0, j = 1, \cdots, p\} \cup \{l : g_l^*(\mathbf{w}) \neq 0 \text{ for some } \mathbf{w}, l = 1, \cdots, q\}$ be the true subset consisting of all the truly informative variables in the linear and nonlinear components, and let $\widehat{\mathcal{S}}_\lambda = \{j : \widehat{\beta}_{\lambda j} \neq 0, j = 1, \cdots, p\} \cup \{l : \widehat{g}_{\lambda l}(\mathbf{w}) \neq 0 \text{ for some } \mathbf{w}, l = 1, \cdots, q\}$ be the estimated subset for given λ.

Theorem 3.2. *Under Assumptions A1–A3, if* $\lambda_1 = n^{-1/8}$, $\lambda_2 = n^{-1/8}$ *and* $\tau_n = n^{-\frac{1}{16(p+q+3)}}$, *then Prob*$(\widehat{\mathcal{S}}_\lambda = \mathcal{S}^*) \to 1$, *as* $n \to \infty$.

Theorem 3.2 assures that, with probability tending to 1, the selected variables is exactly the same as the truly informative variables.

4. Numerical results

4.1. Simulation studies

We examine the performance of the proposed variable selection method for partially linear models (referred to as PL), comparing against some other popular variable selection methods in literature, including the variable selection method for additive models proposed by [47], Cosso by [26], model free gradient learning method by [49] and regular gradient learning method [30], referred to as

Add, Cosso, MF-GL and GL respectively. Note that most of the partially linear model based variable selection procedure, such as Cheng et al. [4], Liang and Li [25], Liu et al. [27], Ni et al. [33] and Wang et al. [45], do not select variables in the nonlinear part. Therefore, we will not include these methods because their performance of nonlinear part selection is inferior than our proposed method.

In all the numerical studies, Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/2\sigma_n^2}$ is used, where scalar parameter σ_n^2 is set as the median over the pairwise distances among all the sample points [30]. For all the four methods considered, the tuning parameters are selected using the same criteria, the variable selection stability criteria proposed by [39], as discussed in Subsection 2.2. For simplicity, we set $\lambda_1 = \lambda_2$ in all the simulation examples. We consider the following two simulation examples. The data generating model is additive partially linear in Example 1, while is non-additive partially linear in Example 2.

Example 1. In this example, first generate predictors $\mathbf{z}_i = (z_i^{(1)}, \cdots, z_i^{(p)})^T$ and $\mathbf{w}_i = (w_i^{(1)}, \cdots, w_i^{(q)})^T$, where $z_i^{(j)}$ and $w_i^{(l)}$ are independently generated from $U(-0.5, 0.5)$, $j = 1, \cdots, p$ and $l = 1, \cdots, q$, for each $i = 1, \cdots, n$. Then set $f^*(\mathbf{w}_i) = 2\sin(\pi w_i^{(1)}) + 2\exp(-2w_i^{(2)})$ and generate response $y_i = 4\sum_{j=1}^5 z_i^{(j)} + f^*(\mathbf{w}_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Example 2. In this example, first generate predictors $\mathbf{z}_i = (z_i^{(1)}, \cdots, z_i^{(p)})^T$ and $\mathbf{w}_i = (w_i^{(1)}, \cdots, w_i^{(q)})^T$, where $z_i^{(j)}$ and $w_i^{(l)}$ are independently generated from $N(0, 1)$, $j = 1, \cdots, p$ and $l = 1, \cdots, q$, for each $i = 1, \cdots, n$. Then set $f^*(\mathbf{w}_i) = (3w_i^{(1)} - 1)(3w_i^{(2)} - 1)$ and generate response $y_i = 4\sum_{j=1}^5 z_i^{(j)} + f^*(\mathbf{w}_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

For each example, 9 different scenarios are considered, where $(n, p, d) = (150, 6, 10)$, $(225, 16, 20)$ or $(300, 46, 50)$ and $\sigma^2 = 0.1$, 0.25 or 1. Here n is the sample size, d is the total number of predictors, and p is the number of predictors in the linear component. In both examples, the true submodel for the linear component is $\mathcal{A}^* = \{1, \cdots, 5\}$, the true submodel for the non-linear component is $\mathcal{B}^* = \{1, 2\}$, and therefore the correct number of informative predictors is 7.

Each scenario is replicated 50 times, and the results are reported in Table 1 for Example 1 and Table 2 for Example 2. Specifically, column "Size" reports the averaged number of selected variables, "TLP" and "TNP" reports the number of selected truly informative variables in the linear and nonlinear components respectively, and "FLP" and "FNP" reports the number of selected truly non-informative variables in the linear and nonlinear components respectively. Columns "C", "U", and "O" report the times, out of 50 times, of correct-fitting, under-fitting, and over-fitting, respectively.

From Table 1, we see that our newly proposed method (PL) is comparable with the others when the data generating model is additive partially linear model. From Table 2, we see that our method outperforms the others when the data generating model is nonadditive partially linear model. First, the average size of selected subsets by PL is closer to 7 than the other methods. Second, most of TLP and TNP from our method are close to 5 and 2, while both FLP

TABLE 1
The summarized results from 50 repetitions of Example 1

(n, p, d, σ^2)	Methods	Size	TLP	FLP	TNP	FNP	C	U	O
(150,6,10,0.1)	PL	7.260	**5.000**	0.060	**2.000**	0.200	38	0	12
	Cosso	7.100	**5.000**	**0.000**	**2.000**	**0.100**	45	0	5
	Add	7.100	**5.000**	**0.000**	**2.000**	**0.100**	45	0	5
	MF-GL	7.120	**5.000**	**0.000**	**2.000**	0.120	44	0	6
	GL	6.800	4.600	0.240	1.840	0.320	12	18	20
(150,6,10,0.25)	PL	7.160	**5.000**	0.040	**2.000**	0.120	42	0	8
	Cosso	6.960	4.960	**0.000**	**2.000**	**0.000**	48	2	0
	Add	7.040	**5.000**	0.040	**2.000**	**0.000**	48	0	2
	MF-GL	7.040	**5.000**	0.040	**2.000**	**0.000**	48	0	2
	GL	6.880	4.680	0.160	1.840	0.200	14	18	18
(150,6,10,1)	PL	7.200	4.960	0.080	**2.000**	0.160	36	2	12
	Cosso	7.100	4.960	0.040	**2.000**	0.100	41	2	7
	Add	7.080	**5.000**	**0.000**	**2.000**	**0.080**	46	0	4
	MF-GL	7.100	**5.000**	0.020	**2.000**	**0.080**	45	0	5
	GL	6.780	4.600	0.200	1.800	0.180	17	15	18
(225,16,20,0.1)	PL	7.120	**5.000**	0.100	**2.000**	0.020	45	0	5
	Cosso	6.840	**5.000**	**0.000**	1.840	**0.000**	42	8	0
	Add	7.040	**5.000**	0.020	**2.000**	0.020	48	0	2
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.840	4.700	0.220	1.840	0.080	20	18	12
(225,16,20,0.25)	PL	7.240	4.980	0.200	**2.000**	0.060	36	1	13
	Cosso	6.820	**5.000**	**0.000**	1.820	**0.000**	41	9	0
	Add	7.100	**5.000**	0.100	**2.000**	**0.000**	45	0	5
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.960	4.700	0.320	1.840	0.100	17	18	15
(225,16,20,1)	PL	7.100	4.980	0.160	1.940	0.020	38	3	9
	Cosso	7.020	**5.000**	0.200	1.820	**0.000**	35	9	6
	Add	7.020	**5.000**	**0.040**	**1.980**	**0.000**	47	1	2
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.960	4.600	0.400	1.800	0.160	10	25	15
(300,46,50,0.1)	PL	7.100	**5.000**	0.100	**2.000**	**0.000**	45	0	5
	Cosso	7.200	**5.000**	0.100	**2.000**	0.100	40	0	10
	Add	7.040	**5.000**	**0.040**	**2.000**	**0.000**	48	0	2
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.780	4.880	0.100	1.800	**0.000**	31	14	5
(300,46,50,0.25)	PL	7.100	**5.000**	0.100	**2.000**	**0.000**	45	0	5
	Cosso	7.300	**5.000**	0.300	**2.000**	**0.000**	35	0	15
	Add	7.000	**5.000**	**0.000**	**2.000**	**0.000**	50	0	0
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.640	4.700	0.100	1.840	**0.000**	25	20	5
(300,46,50,1)	PL	7.060	**5.000**	0.060	**2.000**	**0.000**	47	0	3
	Cosso	7.220	**5.000**	0.180	**2.000**	0.040	39	0	11
	Add	7.040	**5.000**	**0.040**	**2.000**	**0.000**	48	0	2
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.640	4.700	0.140	1.800	**0.000**	24	20	6

and FNP are close to zero, which means that our method is good in selecting the true subsets in both linear component and non-linear component. Third, from column "C', we see that our method has the largest number of times when the true subset is selected exactly. Fourth, from columns "U" and "O", we see Cosso and Add are often under-fitting when $n = 150$, are often over-fitting

TABLE 2
The summarized results from 50 repetitions of Example 2

(n, p, d, σ^2)	Methods	Size	TLP	FLP	TNP	FNP	C	U	O
(150,6,10,0.1)	PL	7.040	**4.940**	0.120	1.900	**0.080**	35	6	9
	Cosso	6.700	4.600	**0.000**	1.700	0.400	5	30	15
	Add	7.200	4.900	0.100	1.960	0.240	30	7	13
	MF-GL	7.340	**4.940**	0.100	**2.000**	0.300	34	3	13
	GL	7.680	4.760	0.560	**2.000**	0.360	6	8	36
(150,6,10,0.25)	PL	7.160	**4.960**	0.100	**1.960**	**0.140**	37	3	10
	Cosso	6.380	4.500	**0.000**	1.700	0.180	20	21	9
	Add	7.100	4.900	0.100	1.840	0.260	26	12	12
	MF-GL	7.200	**4.960**	0.120	**1.960**	0.160	35	3	12
	GL	7.600	4.840	0.600	**1.960**	0.280	8	8	34
(150,6,10,1)	PL	7.260	**4.980**	0.080	1.900	0.300	32	4	14
	Cosso	6.180	4.500	0.180	1.400	**0.100**	5	40	5
	Add	6.400	4.700	**0.000**	1.600	**0.100**	27	18	5
	MF-GL	7.180	**4.980**	**0.000**	1.900	0.200	35	5	10
	GL	7.760	4.920	0.600	**2.000**	0.240	8	4	38
(225,16,20,0.1)	PL	7.080	**4.980**	**0.120**	1.940	0.040	39	4	7
	Cosso	6.780	4.800	0.380	1.600	**0.000**	14	25	11
	Add	7.160	4.920	0.440	1.760	0.040	24	10	16
	MF-GL	–	–	–	–	–	–	–	–
	GL	8.280	**4.980**	0.900	**2.000**	0.400	10	1	39
(225,16,20,0.25)	PL	7.200	**5.000**	**0.240**	**1.960**	**0.000**	36	2	12
	Cosso	6.620	4.700	0.300	1.620	**0.000**	5	30	15
	Add	7.200	4.860	0.500	1.700	0.140	24	10	16
	MF-GL	–	–	–	–	–	–	–	–
	GL	7.900	4.920	0.800	**1.960**	0.220	9	6	35
(225,16,20,1)	PL	7.080	4.900	**0.300**	**1.880**	**0.000**	31	8	11
	Cosso	6.600	4.800	**0.300**	1.500	0.100	5	35	10
	Add	7.580	**4.940**	0.640	1.800	0.200	16	12	22
	MF-GL	–	–	–	–	–	–	–	–
	GL	8.400	4.960	1.000	**2.000**	0.440	8	2	40
(300,46,50,0.1)	PL	7.100	**5.000**	**0.100**	**2.000**	**0.000**	45	0	5
	Cosso	7.600	**5.000**	0.600	**2.000**	**0.000**	25	0	25
	Add	6.400	**5.000**	0.600	1.800	**0.000**	15	10	25
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.720	4.880	**0.080**	1.760	**0.000**	30	16	4
(300,46,50,0.25)	PL	6.920	**5.000**	0.060	1.860	**0.000**	41	6	3
	Cosso	7.400	**5.000**	0.400	**1.900**	0.100	20	5	25
	Add	6.960	4.760	0.400	1.800	**0.000**	10	20	20
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.900	**5.000**	**0.040**	1.720	**0.000**	34	14	2
(300,46,50,1)	PL	6.940	**5.000**	**0.000**	**1.940**	**0.000**	47	3	0
	Cosso	7.300	5.000	0.500	1.800	**0.000**	30	10	10
	Add	6.400	4.600	0.200	1.600	**0.000**	17	25	8
	MF-GL	–	–	–	–	–	–	–	–
	GL	6.600	4.880	0.080	1.640	**0.000**	24	22	4

when $n = 300$, and are often over-fitting or under-fitting when $n = 225$. Fifth, we find that GL always tends to over-fitting or under-fitting. This may be due to the fact that GL selection procedure depends on the prespecified truncated value. From both tables, we see that although the performance for MF-GL is promising, the computational cost is relatively high, especially as the sample

size and dimension increase. In addition, we see that as variance σ^2 becomes larger, it is more challenging to select the true subset.

4.2. Real data applications

We further examine the performance of the proposed variable selection method using two real data applications, the digit recognition data [40] and Japanese industrial chemical firms data [48], both of which are publicly available.

4.2.1. Digit recognition data

In the digit recognition data, each digit is described by an 8×8 gray-scale image with each entry ranging from 0 to 16. Due to their similarity, it is challenging to distinguish digits 3 and 5. Therefore, in this real data application, we only consider digits 3 and 5, and the resultant dataset consists of 365 observations and 64 attributes. Consider outcome variable $y = 3$ for digit 3 and $y = 5$ for digit 5. For the purpose of demonstration, we consider the partially linear model as the predictive model, although other classification models may be more appropriate. By some descriptive analysis, we find that these two digits mainly differ on dimensions 19, 21, 22, 27 and 54. Therefore, in the partially linear model, we put these 5 dimensions in the nonlinear component and the others in the linear component.

In this analysis, all the variables are standardized, all the missing value observations are ignored, and all the four aforementioned variable selection procedures (PL, Cosso, Add and MF-GL) are applied. The performance of the variable selection procedures are compared based on the averaged prediction errors using only the selected variables. The averaged prediction errors are estimated as what follows. First, the dataset is randomly split into two parts, $m = 35$ observations for testing and the remaining 330 observations for training. Second, partially linear model is fitted based on the training dataset, one submodel is selected, and the average prediction error estimate is obtained. The results are summarized in Table 3.

From Table 3, we see that the proposed variable selection procedure, PL, selects less variables and has smaller prediction error than the other variable selection procedures. Figure 1 shows two randomly selected digits of 3 and 5 in the right and middle panels respectively, and the two finally selected contributes are displayed in the left panel. We can see the the two digits are clear different at the two selected contributes. Although MF-GL provides competitive performance, its computational cost is about 10 times higher than PL.

4.2.2. Japanese industrial chemical firms data

The Japanese industrial chemical firms dataset consists of 186 Japanese industrial chemical firms listed on the Tokyo stock exchange, and the goal is to check whether concentrated shareholding is associated with lower expenditure on ac-

TABLE 3
Digit Recognition Data – the number of selected variables and the prediction errors by four variable selection procedures

	PL	Cosso	Add	MF-GL	GL
No. of variables	2	8	48	2	2
Pred.Err	**1.832**	1.878	1.871	1.857	1.857
Pred.Std	**0.033**	0.031	0.031	0.032	0.032

FIG 1. *Digit Recognition Data – Left panel displays an example of digit 3, middle panel displays an example of digit 5, and right panel display the two selected contributes using the propose method*

tivities with scope for managerial private benefits. The response variable is MH5 (the general sales and administrative expenses deflated by sales), and 12 predictors are ASSETS (logarithm of assets), AGE (the age of the firm), LEVERAGE (ratio of debt to total assets), VARS (variance of operating profits to sales), OPERS (operating profits to sales), TOP10 (the percentage of ownership held by the 10 largest shareholders), TOP5 (the percentage of ownership held by the 5 largest shareholders), OWNIND (ownership Herfindahl index), AOLC (amount owed to largest creditor), SHARE (share of debt held by largest creditor), BDHIND (bank debt Herfindahl index), and BDA (bank debt to assets).

For the purpose of demonstration, we consider the partially linear model. From some descriptive analysis, we find that variables LEVERAGE, VARS, OPER2 and BDA are highly correlated with the MH5, and that the marginal relationships of LEVERAGE, OPER2 and BDA with the response variable are strongly linear. For other less correlated variables, the marginal scatter plot shows that associations between MH5 with ASSETS, AGE, TOP5 and TOP10 are seemly linear. Therefore, we put LEVERAGE, OPER2, BDA, ASSETS, AGE, TOP5 and TOP10 in linear part, while others in nonlinear part.

We apply the same strategy to compare those four variable selection procedures. The only difference is that now we consider 100 times random split of the dataset into a test dataset of $m = 24$ observations and a traning dataset of the remaining observations. The results are summarized in Table 4. From Table 4, we can see that PL has the smallest prediction error. Although Cosso, Add and MF select less variables, their prediction errors are larger. Especially, the PL method selects variable OWNIND, which are not selected by the other methods. Figure 2 displays the scatterplot of MH5 against OWNIND. It seems that the mean of MH5 does not change much with OWNIND, while its variance appears to shrink as OWNIND increases.

TABLE 4
The number of selected variables and the prediction errors by various selection methods in Japanese industrial chemical firms dataset.

	PL	Cosso	Add	MF-GL	GL
ASSETS					
AGE					√
LEVERAGE	√	√	√	√	√
VARS	√	√	√	√	√
OPER2	√		√	√	√
TOP10					√
TOP5					√
OWNIND	√				
AOLC					
SHARE			√		
BDHIND			√		
BDA	√			√	√
Pred.Err	**0.513**	0.555	0.561	0.554	0.574
Pred.Std	**0.0179**	0.0194	0.0193	0.0191	0.0280

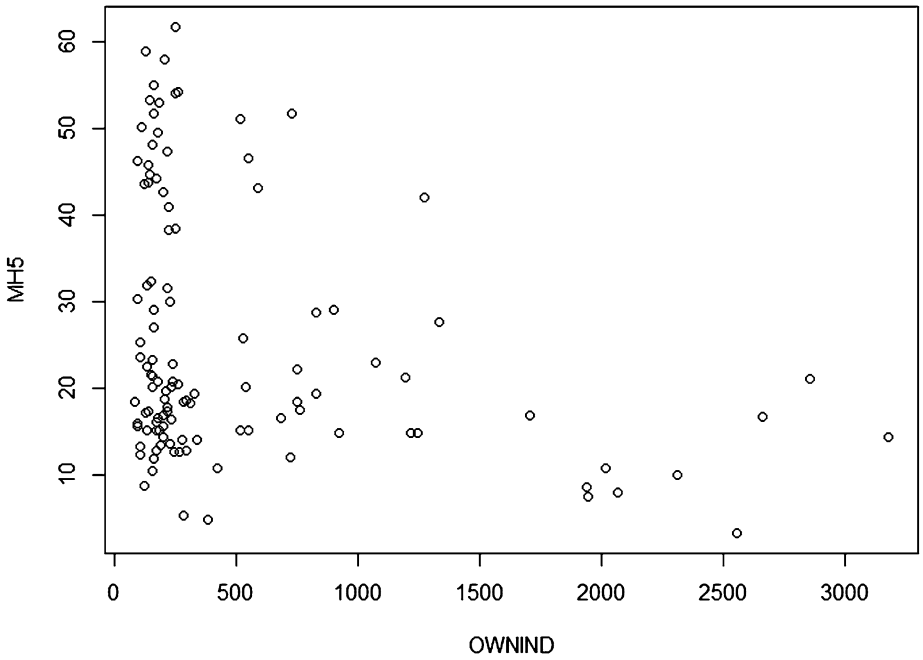

FIG 2. *Japanese Industrial Chemical Firms Data – The scatterplot of MH5 versus OWNIND*

5. Summary

There has been an unmet need for developing a novel variable selection procedure for partially linear models without imposing structural constraints on the non-parametric component. Both computational efficiency and model flexibility are important for analyzing high dimensional data. Recently, the idea of gradient

learning has become popular for variable selection, because it is model free [49]. However, model-free gradient learning is computationally expensive for high-dimensional data in pure non-parametric setting. Therefore, gradient learning is particularly suitable for partially linear models. In this paper, we propose a variable selection procedure based on gradient learning for partially linear models. The proposed procedure incorporates two sparsity-inducing penalties, one for variable selection in the linear component and the other for variable selection in the nonlinear component. Since the computational cost of variable selection in the linear component is cheap, the proposed procedure is computationally feasible and applicable for the analysis of high-dimensional data. At the same time, the variable selection in the nonlinear component is model-free, thus without the need to impose any assumption on the structure of the nonlinear component.

The proposed procedure is formulated in terms of reproducing kernel Hilbert space, which provides a general framework for developing efficient implementation algorithm and deriving desirable theoretical properties. Specifically, a block-wise coordinate decent algorithm is developed and estimation consistency and selection consistency are both established. Two issues are crucial in applying the proposed methods. One issue is selecting the tuning parameters. We propose to use stability selection for the tuning parameters, and other tuning methods such as cross-validation can also be used. The other issue is the specification of the partially linear model, that is deciding which variables are in the linear component and which variables are in the nonlinear component. In this manuscript, we focus on variable selection and assume that the partially linear model has been pre-specified. The reader is referred to [16] for the discussion of partially linear model specification. Under similar assumptions, some existing general variable selection procedure, such as Rosasco et al. [36], can also be extended to partially linear models in a similar fashion.

Appendix: Technical proofs

Denote the first term in objective function (2.2) as $\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta},\mathbf{g})$, and denote its expectation as $\mathcal{E}(\boldsymbol{\beta},\mathbf{g})$, which is equal to

$$2\sigma_s^2 + E\left\{w(\mathbf{x},\mathbf{x}')\left[(\boldsymbol{\beta}^*-\boldsymbol{\beta})^T(\mathbf{z}-\mathbf{z}') + f^*(\mathbf{w}) - f^*(\mathbf{w}') - \mathbf{g}(\mathbf{w})^T(\mathbf{w}-\mathbf{w}')\right]^2\right\},$$

where $w(\mathbf{x},\mathbf{x}') = \exp\{-\|\mathbf{x}-\mathbf{x}'\|_2^2/\tau^2\}$, $\sigma_s^2 = \sigma^2 E\{w(\mathbf{x},\mathbf{x}')\}$, and the expectation is over random predictors.

Lemma 1. *Let* $\varphi_0(\mathbf{z},\mathbf{w}) = \mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}^*,\mathbf{g}^*) - \mathcal{E}(\boldsymbol{\beta}^*,\mathbf{g}^*)$, $\varphi_1(\mathbf{z},\mathbf{w}) = \mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - \mathcal{E}_{\mathbf{z},\mathbf{w}}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})$, *and* $J(\boldsymbol{\beta},\mathbf{g}) = \lambda_1\sum_{l=1}^{p}\pi_{\mathbf{z},l}|\beta_l| + \lambda_2\sum_{l=1}^{q}\pi_{\mathbf{w},l}\|g_l\|_K$. *Then the following inequality holds,*

$$\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) + J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 \le \varphi_1(\mathbf{z},\mathbf{w}) + \varphi_0(\mathbf{z},\mathbf{w}) + \Lambda,$$

where $\Lambda = \mathcal{E}(\boldsymbol{\beta}^*,\mathbf{g}^*) - 2\sigma_s^2 + \lambda_1\sum_{l=1}^{p_0}\pi_{\mathbf{z},l}|\beta_l^*| + \lambda_2\sum_{l=1}^{q_0}\pi_{\mathbf{w},l}\|g_l{}^*\|_K$.

Proof of Lemma 1.

$$\begin{aligned}
& \mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})+J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})-2\sigma_s^2 \\
= \;& \mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})-\mathcal{E}_{\mathbf{z},\mathbf{w}}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})+\mathcal{E}_{\mathbf{z},\mathbf{w}}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})+J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})-2\sigma_s^2 \\
\le \;& \mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})-\mathcal{E}_{\mathbf{z},\mathbf{w}}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})+\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}^*,\mathbf{g}^*)+J(\boldsymbol{\beta}^*,\mathbf{g}^*)-2\sigma_s^2 \\
= \;& \varphi_1(\mathbf{z},\mathbf{w})+\varphi_0(\mathbf{z},\mathbf{w})+\mathcal{E}(\boldsymbol{\beta}^*,\mathbf{g}^*)+J(\boldsymbol{\beta}^*,\mathbf{g}^*)-2\sigma_s^2 \\
= \;& \varphi_1(\mathbf{z},\mathbf{w})+\varphi_0(\mathbf{z},\mathbf{w})+\Lambda.
\end{aligned}$$

The first inequality holds because $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{g}}$ are the minimizer of $\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta},\mathbf{g})+J(\boldsymbol{\beta},\mathbf{g})$, the last equality is due to the fact that $\beta_l^*=0$ for any $l>p_0$ and $g_l^*=0$ for any $l>q_0$. □

Lemma 2. *(McDiarmid's Inequality) Let $V_1,...,V_n$ be independent random variables taking values in a set $\mathcal{V}$, and assume that $\mathbf{f}:\mathcal{V}^n\to\mathbb{R}$ satisfies*

$$\sup_{v_1,...,v_n,v_i'\in\mathcal{V}}|\mathbf{f}(v_1,...,v_n)-\mathbf{f}(v_1,...,v_i',...,v_n)|\le C_i,$$

for every $i\in\{1,2,...,n\}$. Then, for every $t>0$,

$$\mathbb{P}\{|\mathbf{f}(V_1,...,V_n)-\mathbb{E}\left(\mathbf{f}(V_1,...,V_n)\right)|\ge t\}\le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n C_i^2}\right).$$

Lemma 3. *Let $\mathcal{S}(R,\lambda_1,\lambda_2)=\sup\{|\mathcal{E}(\boldsymbol{\beta},\mathbf{g})-\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta},\mathbf{g})|:(\boldsymbol{\beta},\mathbf{g})\in\mathcal{H}_R^d\}$, where $\mathcal{H}_R^d=\{(\boldsymbol{\beta},\mathbf{g}):\boldsymbol{\beta}\in\mathcal{R}^p,\mathbf{g}\in\mathcal{H}_K,J(\boldsymbol{\beta},\mathbf{g})\le R\}$. If $|y|\le M_n$ and Assumptions A1–A3 hold, then we have, for any constant $R>0$ and $t>0$,*

$$P(|\mathcal{S}(R,\lambda_1,\lambda_2)-\mathbb{E}(\mathcal{S}(R,\lambda_1,\lambda_2))|\ge t)\le 2\exp\left(\frac{-nt^2}{8\big(M_n+\frac{c_zR}{c_3\lambda_1}+\frac{c_wR}{c_3\lambda_2}\big)^4}\right),$$

$$P(|\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}^*,\mathbf{g}^*)-\mathcal{E}(\boldsymbol{\beta}^*,\mathbf{g}^*)|\ge t)\le 2\exp\left(\frac{-nt^2}{8\big(M_n+\sum_{l=1}^p|\beta_l^*|+\sum_{l=1}^q\|g_l^*\|_K\big)^4}\right).$$

Proof of Lemma 3. It suffices to verify the conditions required by the McDiarmid's inequality. Denote $(\mathbf{z}',\mathbf{w}',y')$ as a sample point drawn from the distribution $\rho(\mathbf{z},\mathbf{w},y)$ and independent of $(\mathbf{z}_i,\mathbf{w}_i,y_i)$. For any fixed $(\boldsymbol{\beta},\mathbf{g})\in\mathcal{H}_R^d$, let $h(\mathbf{z}_i,\mathbf{w}_i,\mathbf{z}_j,\mathbf{w}_j)=w_{ij}(y_i-y_j-(\mathbf{z}_i-\mathbf{z}_j)^T\boldsymbol{\beta}-\mathbf{g}(\mathbf{x}_i)^T(\mathbf{w}_i-\mathbf{w}_j))^2$. Decompose $n(n-1)\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta},\mathbf{g})$ as follows,

$$\sum_{k\ne i,j\ne i}^n h(\mathbf{z}_k,\mathbf{z}_j,\mathbf{w}_k,\mathbf{w}_j)+\sum_{j=1}^n h(\mathbf{z}_i,\mathbf{z}_j,\mathbf{w}_i,\mathbf{w}_j)+\sum_{k=1}^n h(\mathbf{z}_k,\mathbf{z}_i,\mathbf{w}_k,\mathbf{w}_i).$$

Note that $\lambda_1c_3\sum_{l=1}^p|\beta_l|\le\lambda_1\sum_{l=1}^p\pi_{\mathbf{z},l}|\beta_l|\le R$ and by assumption A3,

$$\lambda_2c_3\sum_{l=1}^q\|g_l\|_K\le\lambda_2\sum_{l=1}^q\pi_{\mathbf{w},l}\|g_l\|_K\le R$$

Then if $(\mathbf{z}, \mathbf{w})$ is replaced by $(\mathbf{z}', \mathbf{w}')$,

$$\begin{aligned}\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) - \mathcal{E}_{\mathbf{z}',\mathbf{w}'}(\boldsymbol{\beta}, \mathbf{g}) &\leq \frac{4(M_n + c_z \sum_{l=1}^{p} |\beta_l| + c_w \sum_{l=1}^{q} \|g_l\|_K)^2}{n} \\ &\leq \frac{4\big(M_n + \frac{c_z R}{c_3 \lambda_1} + \frac{c_w R}{c_3 \lambda_2}\big)^2}{n},\end{aligned}$$

where the first inequality holds because supports $\mathcal{Z}$ and $\mathcal{W}$ are compact. Interchanging the roles of $(\mathbf{z}, \mathbf{w})$ and $(\mathbf{z}', \mathbf{w}')$ yields that, for $(\boldsymbol{\beta}, \mathbf{g}) \in \mathcal{H}_R^d$,

$$|\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) - \mathcal{E}_{\mathbf{z}',\mathbf{w}'}(\boldsymbol{\beta}, \mathbf{g})| \leq \frac{4\big(M_n + \frac{c_z R}{c_3 \lambda_1} + \frac{c_w R}{c_3 \lambda_2}\big)^2}{n}.$$

Then applying the McDiarmid's inequality, we obtain the first result of Lemma 3. For the second result of Lemma 3, we can easily verify that

$$C_i = 4(M_n + c_z \sum_{l=1}^{p} |\beta_l^*| + c_w \sum_{l=1}^{q} \|g_l^*\|_K)^2 / n.$$

Thus plugging C_i into the McDiarmid's inequality, we obtain the second result of the lemma. □

Lemma 4. *If $|y| \leq M_n$, then there exists a constant c_7 such that*

$$|\mathbb{E}(\mathcal{S}(R, \lambda_1, \lambda_2))| \leq c_7 \frac{(M_n + \frac{c_{\mathbf{z}} R}{c_3 \lambda_1} + \frac{c_{\mathbf{w}} R}{c_3 \lambda_2})^2}{\sqrt{n}}.$$

Proof of Lemma 4. Denote $\xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}', \mathbf{w}', y') = w(\mathbf{x}, \mathbf{x}')(y - y' - \boldsymbol{\beta}(\mathbf{z} - \mathbf{z}') - \mathbf{g}(\mathbf{w})(\mathbf{w} - \mathbf{w}'))^2$, $\mathcal{E}(\boldsymbol{\beta}, \mathbf{g}) = \mathbb{E}_{(\mathbf{z},\mathbf{w},y)} \mathbb{E}_{(\mathbf{z}',\mathbf{w}',y')} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}', \mathbf{w}', y')$,

$$\mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \xi(\mathbf{z}_i, \mathbf{w}_i, y_i, \mathbf{z}_j, \mathbf{w}_j, y_j),$$

and then we get

$$\begin{aligned}&|\mathcal{S}(R, \lambda_1, \lambda_2)| \\ \leq\ & \sup_{\boldsymbol{\beta}, \mathbf{g} \in \mathcal{H}_R^d} \Big| \mathcal{E}(\boldsymbol{\beta}, \mathbf{g}) - \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(\mathbf{z},\mathbf{w},y)} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}_j, \mathbf{w}_j, y_j) \Big| \\ & + \sup_{\boldsymbol{\beta}, \mathbf{g} \in \mathcal{H}_R^d} \Big| \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(\mathbf{z},\mathbf{w},y)} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}_j, \mathbf{w}_j, y_j) - \mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) \Big| \\ \leq\ & \sup_{\boldsymbol{\beta}, \mathbf{g} \in \mathcal{H}_R^d} \mathbb{E}_{(\mathbf{z},\mathbf{w},y)} \Big| \mathbb{E}_{(\mathbf{z}',\mathbf{w}',y')} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}', \mathbf{w}', y') - \frac{1}{n} \sum_{j=1}^{n} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}_j, \mathbf{w}_j, y_j) \Big| \\ & + \frac{1}{n} \sum_{j=1}^{n} \sup_{\boldsymbol{\beta}, \mathbf{g} \in \mathcal{H}_R^d} \sup_{(\mathbf{z}',\mathbf{w}',y')} \Big| \mathbb{E}_{(\mathbf{z},\mathbf{w},y)} \xi(\mathbf{z}, \mathbf{w}, y, \mathbf{z}', \mathbf{w}', y')\end{aligned}$$

$$- \frac{1}{n-1}\sum_{i\neq j}^{n}\xi(\mathbf{z}_i,\mathbf{w}_i,y_i,\mathbf{z}',\mathbf{w}',y')\Big|$$

$$= S_1 + S_2,$$

where the first inequality dues to the triangle inequality and the second inequality dues to the relationship between expectation and absolute value. Let σ_i, $i=1,\ldots,n$ to be independent Rademacher variables. For the first term, by using the properties of Rademacher complexities [42], we have

$$\begin{aligned}
\mathbb{E}(S_1) &= \mathbb{E}_{(\mathbf{z},\mathbf{w},y)}\sup_{\boldsymbol{\beta},\mathbf{g}\in\mathcal{H}_R^d}\Big|\mathbb{E}_{(\mathbf{z}',\mathbf{w}',y')}\xi(\mathbf{z},\mathbf{w},y,\mathbf{z}',\mathbf{w}',y') \\
&\qquad - \frac{1}{n}\sum_{j=1}^{n}\xi(\mathbf{z},\mathbf{w},y,\mathbf{z}_j,\mathbf{w}_j,y_j)\Big| \\
&\leq 2\sup_{(\mathbf{z},\mathbf{w},y)}\mathbb{E}\sup_{\boldsymbol{\beta},\mathbf{g}\in H_R^d}\Big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j w(\mathbf{x},\mathbf{x}_j)(y_j - y - \boldsymbol{\beta}(\mathbf{z}_j-\mathbf{z}) \\
&\qquad - \mathbf{g}(\mathbf{w}_j)(\mathbf{w}_j-\mathbf{w}))^2\Big| \\
&\leq 4\left(M_n + \frac{c_{\mathbf{z}}R}{c_3\lambda_1} + \frac{c_{\mathbf{w}}R}{c_3\lambda_2}\right)\Big(\sup_{(\mathbf{z},\mathbf{w},y)}\mathbb{E}\sup_{\boldsymbol{\beta},\mathbf{g}\in H_R^d}\frac{1}{n}\sum_{j=1}^{n}\sigma_j(y_j - y \\
&\qquad - \boldsymbol{\beta}(\mathbf{z}_j-\mathbf{z}) - \mathbf{g}(\mathbf{w}_j)(\mathbf{w}_j-\mathbf{w})) + \frac{M_n}{\sqrt{n}}\Big) \\
&\leq c_7\frac{(M_n + \frac{c_{\mathbf{z}}R}{c_3\lambda_1} + \frac{c_{\mathbf{w}}R}{c_3\lambda_2})^2}{\sqrt{n}}.
\end{aligned}$$

Similary we can verify the second term S_2 and therefore, $|\mathbb{E}(\mathcal{S}(R,\lambda_1,\lambda_2))| \leq c_7\frac{(M_n+\frac{c_{\mathbf{z}}R}{c_3\lambda_1}+\frac{c_{\mathbf{w}}R}{c_3\lambda_2})^2}{\sqrt{n}}$, where c_7 is a fixed constant. □

Lemma 5. *Assume the Assumptions A1–A3 are satisfied. If* $\mathcal{E}_{\mathbf{z}}(0) = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}(y_i-y_j)^2$ *is bounded by* M_0*, then there exists a constant* c_8 *such that for any* $\delta\in(0,1)$*, with probability at least* $1-\delta$*,*

$$\begin{aligned}
J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) &\leq c_8\sqrt{\log(2/\delta)}\Big(M_n^2 n^{-1/2} + n^{-1/2}\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} + \tau_n^{p+q+4} \\
&\qquad + \lambda_1 + \lambda_2\Big), \\
\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 &\leq c_8\sqrt{\log(2/\delta)}\Big(M_n^2 n^{-1/2} + n^{-1/2}\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} + \tau_n^{p+q+4} \\
&\qquad + \lambda_1 + \lambda_2\Big).
\end{aligned}$$

Proof of Lemma 5. By Lemma 4, have

$$\mathbb{E}(\mathcal{S}(R,\lambda_1,\lambda_2)) \leq c_7\left(M_n + \frac{c_z + c_w}{c_3}\left(\frac{R}{\lambda_1}+\frac{R}{\lambda_2}\right)\right)^2/\sqrt{n},$$

which, together with Lemma 3, implies that with probability at least $1-\delta$,

$$\varphi_1(\mathbf{z},\mathbf{w}) \le |\mathcal{S}(R,\lambda_1,\lambda_2)| \le c_7\sqrt{\frac{\log(2/\delta)}{n}}\left(M_n + \frac{c_z+c_w}{c_3}\left(\frac{R}{\lambda_1}+\frac{R}{\lambda_2}\right)\right)^2.$$

According to the second result in Lemma 3, we know that with probability at least $1-\delta$,

$$\begin{aligned}\varphi_0(\mathbf{z},\mathbf{w}) &\le |\mathcal{E}(\boldsymbol{\beta}^*,\mathbf{g}^*) - \mathcal{E}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}^*,\mathbf{g}^*)| \le 3\sqrt{\frac{\log(2/\delta)}{n}}\Big(M_n + \sum_{l=1}^{p}|\beta_l^*| \\ &\quad + \sum_{l=1}^{q}\|g_l^*\|_K\Big)^2.\end{aligned}$$

For Λ in Lemma 1, we can easily check that

$$\begin{aligned}\Lambda &= E(\mathbf{x},\mathbf{x}')w(\mathbf{x},\mathbf{x}')\left(f^*(\mathbf{w}) - f^*(\mathbf{w}') - \mathbf{g}^*(\mathbf{w})^T(\mathbf{w}-\mathbf{w}')\right)^2 + J(\boldsymbol{\beta}^*,\mathbf{g}^*) \\ &\le c_1 E(\mathbf{x},\mathbf{x}')w(\mathbf{x},\mathbf{x}')\|\mathbf{x}-\mathbf{x}'\|_2^2 + J(\boldsymbol{\beta}^*,\mathbf{g}^*) \\ &\le c_1\tau_n^{p+q+4} + p_0c_6\lambda_1 + q_0c_6\lambda_2 = O(\tau_n^{p+q+4}+\lambda_1+\lambda_2).\end{aligned}$$

By Lemma 1, we can see that there exists a constant c_9 such that

$$\begin{aligned}&\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) + J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 \le |\varphi_1(\mathbf{z},\mathbf{w})| + |\varphi_0(\mathbf{z},\mathbf{w})| + \Lambda \\ = \;& c_9\sqrt{\log(2/\delta)}\left(M_n^2 n^{-1/2} + n^{-1/2}R^2\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} + \tau_n^{p+q+4} + \lambda_1 + \lambda_2\right).\end{aligned}$$

Noting that $\mathcal{E}_{\mathbf{z},\mathbf{w}}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) + J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) \le \mathcal{E}_{\mathbf{z},\mathbf{w}}(0,0) + J(0,0) \le M_0$, we have $\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}} \in \mathcal{H}_R^d$, where $R = M_0$. Therefore, there exists a constant c_8 such that

$$\begin{aligned}\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) + J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 &\le c_8\sqrt{\log(2/\delta)}\Big(M_n^2 n^{-1/2} + n^{-1/2}\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} \\ &\quad + \tau_n^{p+q+4} + \lambda_1 + \lambda_2\Big).\end{aligned}$$

Combining the fact that $\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 \ge 0$ and $J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) \ge 0$, we obtain the bounds for $J(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}})$ and $\mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2$ as stated in Lemma (5). □

Proof of Theorem 3.1. For given constant $c_8 > 0$, denote event $\mathcal{C}$ as

$$\begin{aligned}\mathcal{C} &= \Big\{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}} : \mathcal{E}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{g}}) - 2\sigma_s^2 > c_8\sqrt{\log(4/\delta)}\Big(n^{-1/4} + n^{-1/2}\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} \\ &\quad + \tau_n^{p+q+4} + \lambda_1 + \lambda_2\Big)\Big\}.\end{aligned}$$

Denote $U = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}(y_i - y_j)^2$ and $M_0 = 4B^2 + 2\sigma^2 + 1$ with B an upper bound of $\mathbf{z}^T\boldsymbol{\beta}^* + f^*(\mathbf{w})$. The constant B exists because f^* is continuous and set $\mathcal{Z}$ and $\mathcal{W}$ are compact. Then we split $\mathcal{C}$ into three different events as follows,

$$P(\mathcal{C}) = P\Big(\mathcal{C} \cap \{|y| \le n^{1/8}, U \le M_0\}^c\Big) + P\Big(\mathcal{C} \cap \{|y| \le n^{1/8}, U \le M_0\}\Big)$$

$$\leq \quad P(|y| > n^{1/8}) + P(|y| \leq n^{1/8}, U > M_0) + P\Big(\mathcal{C} \cap \big\{|y| \leq n^{1/8}, U \leq M_0\big\}\Big),$$

Same as the proof in Yang et al. [49], we get $P(|y| > n^{\frac{1}{8}}) = E(P(\mathbf{z}^T\boldsymbol{\beta}^* + f^*(\mathbf{x}) + \epsilon > n^{\frac{1}{8}}|\,\mathbf{x})) \leq O(n^{-\frac{1}{4}})$ and $P(|y| \leq n^{1/8}, U > M_0) \leq P(U > E(U) + 1\big||y| \leq n^{1/8}) \leq \exp\big\{-\frac{1}{16}n^{3/4}\big\}$. For the third term, by Lemma 5, with probability at least $1-\delta$,

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2 \quad \leq \quad c_8\sqrt{\log(2/\delta)}\Big(n^{-1/4} + n^{-1/2}\lambda_1^{-2} + n^{-1/2}\lambda_2^{-2} + \tau_n^{p+q+4} + \lambda_1 + \lambda_2\Big).$$

Therefore, $P(\mathcal{C}) \leq O\Big(n^{-1/4} + \exp\{-\frac{1}{16}n^{3/4}\} + \delta/2\Big)$. By Theorem 5 in Ye and Xie [52] and assumption A2, we can verify that in set $\mathcal{C}^c$,

$$\int_{\mathbf{z},\mathbf{w}} \Big\|\Big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \widehat{\mathbf{g}}(\mathbf{w}) - \mathbf{g}^*(\mathbf{w})\Big)\Big\|_2 d\rho_{\mathbf{z},\mathbf{w}} \leq \left(M_0^2\tau_n + \tau_n + \frac{\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2}{\tau_n^{p+q+3}}\right),$$

which indicates that

$$\begin{aligned}
\int_{\mathbf{z},\mathbf{w}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 d\rho_{\mathbf{z},\mathbf{w}} \quad &\leq \quad \left(M_0^2\tau_n + \tau_n + \frac{\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2}{\tau_n^{p+q+3}}\right), \\
\int_{\mathbf{z},\mathbf{w}} \|\widehat{\mathbf{g}}(\mathbf{w}) - \mathbf{g}^*(\mathbf{w})\|_2 d\rho_{\mathbf{z},\mathbf{w}} \quad &\leq \quad \left(M_0^2\tau_n + \tau_n + \frac{\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2}{\tau_n^{p+q+3}}\right),
\end{aligned}$$

where $\rho_{\mathbf{z},\mathbf{w}}$ is the joint CDF of $\mathbf{z}$ and $\mathbf{w}$. Therefore, we conclude that

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \quad &\leq \quad \left(M_0^2\tau_n + \tau_n + \frac{\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2}{\tau_n^{p+q+3}}\right), \\
\|\widehat{\mathbf{g}} - \mathbf{g}^*\|_2 \quad &\leq \quad \left(M_0^2\tau_n + \tau_n + \frac{\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{g}}) - 2\sigma_s^2}{\tau_n^{p+q+3}}\right).
\end{aligned}$$

As $\lambda_1 = \lambda_2 = n^{-1/8}$ and $\tau_n = n^{-\frac{1}{16(p+q+3)}}$, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq O\big(n^{-\frac{1}{16(p+q+3)}}\big)$ and $\|\widehat{\mathbf{g}} - \mathbf{g}^*\|_2 \leq O\big(n^{-\frac{1}{16(p+q+3)}}\big)$. Then as $n \to \infty$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \to 0$, $\|\widehat{\mathbf{g}} - \mathbf{g}^*\|_2 \to 0$ and $P(\mathcal{C}^c) \to 1$. Thus Theorem 3.1 is proved. □

Proof of Theorem 3.2. By Theorem 3.1, $|\hat{\beta}_l| > 0$ for any $l \leq p_0$. Now we show that $|\beta_l| = 0$ for any $l > p_0$ by contradiction. Assume that $|\beta_l| > 0$ for some $l > p_0$ in the linear part. Taking the derivative of the objective function with respect to β_l, we get

$$\frac{2}{n(n-1)}\sum_{i,j=1}^{n} w_{ij}\big(y_i - y_j - (\mathbf{z}_i - \mathbf{z}_j)^T\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{g}}(\mathbf{w}_i)^T(\mathbf{w}_i - \mathbf{w}_j)\big)(z_{il} - z_{jl})$$

$$= -\frac{\lambda_1 \pi_{\mathbf{z},l} \hat{\beta}_l}{|\hat{\beta}_l|}. \quad (5.1)$$

Note that the norm of the right-hand side of (5.1) is $\lambda_1 \pi_{\mathbf{z},l}$, which diverges to ∞ by assumption A3. Denote $\mathcal{B}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g}) = \frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \Big(y_i - y_j - \boldsymbol{\beta}^T (\mathbf{z}_i - \mathbf{z}_j) - \mathbf{g}(\mathbf{w}_i)^T (\mathbf{w}_i - \mathbf{w}_j) \Big)$. By Theorem 3.1, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \to 0$ and $J(\hat{\boldsymbol{\beta}}, \hat{\mathbf{g}}) \le M_0$ in probability, which implies that $c_3 \sum_{l=1}^{q} \|\hat{g}_l\|_K \le M_0$, $\|\hat{g}_l\|_K$ and $|\hat{\beta}_l|$ is bounded by a constant c_9 in probability. Therefore, as $w_{ij} \le 1$ and the support $\mathcal{Z}$ and $\mathcal{W}$ are compact, $|\mathcal{B}_{\mathbf{z},\mathbf{w}}(\boldsymbol{\beta}, \mathbf{g})| \le \frac{4}{n} \sum_{i=1}^{n} |y_i| + (p_0 + q_0) c_9$. Then by the central limit theorem, the left hand side of (5.1) is bounded in probability. Thus, the contradiction appears, and we conclude that $|\hat{\beta}_l| = 0$ for any $l > p_0$ in the linear component.

Similarly, we can show the selection consistency for the nonlinear component. For this aim, taking the first derivative with respect to $\boldsymbol{\alpha}^{(l)}$, we have

$$\frac{2}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \big(y_i - y_j - \boldsymbol{\beta}^T (\mathbf{z}_i - \mathbf{z}_j) - \widehat{\mathbf{g}}(\mathbf{x}_i)^T (\mathbf{x}_i - \mathbf{x}_j) \big) (w_{il} - w_{jl}) \mathbf{K}_l$$

$$= -\frac{\lambda_2 \pi_l \, \mathbf{K} \, \hat{\boldsymbol{\alpha}}^{(l)}}{\| \mathbf{K} \, \hat{\boldsymbol{\alpha}}^{(l)} \|_2}. \quad (5.2)$$

Same as the argument above, each element in left-hand-side of (5.2) is bounded in probability and therefore, the norm of left-hand-side of (5.2) divided by $n^{1/2}$ is bounded in probability. Suppose the Kernel matrix $\mathbf{K}$ has smallest eigenvalue n^{-1}, then the norm of the right-hand-side is larger than $\lambda_2 \pi_l n^{-1}$. Because $n^{-3/2} \lambda_2 \pi_l \to \infty$ by Assumption A3, the norm of right-hand-side divided by $n^{1/2}$ goes to infinity, the contradiction appears. Thus, Theorem 3.2 is proved. □

A remark on Assumption A3. We will show that Assumption A3 can be satisfied based on initial estimates from (2.4). If initial estimates $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\mathbf{g}}$ are via (2.4), following the same argument as in the proof of Lemma 5, we have, given $|y| \le n^{1/8}$ and $U < M_0$,

$$\mathcal{E}(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{g}}) - 2\sigma_s^2 \quad \le \quad c_{10} \sqrt{\log(2/\delta)} \Big(M_n^2 n^{-1/2} + n^{-1/2} \tilde{\lambda}_1^{-1} + n^{-1/2} \tilde{\lambda}_2^{-1} + \tau_n^{p+q+4} + \tilde{\lambda}_1 + \tilde{\lambda}_2 \Big).$$

Following the same argument as in the proof of Theorem 3.1, we can show that the probability of the following event, denoted by as $\widetilde{\mathcal{C}}$, goes to one as as $n \to \infty$,

$$\Big\{ \widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{g}} : \mathcal{E}(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{g}}) - 2\sigma_s^2 \quad > \quad c_{10} \sqrt{\log(4/\delta)} \Big(n^{-1/4} + n^{-1/2} \tilde{\lambda}_1^{-1} + n^{-1/2} \tilde{\lambda}_2^{-1} + \tau_n^{p+q+4} + \tilde{\lambda}_1 + \tilde{\lambda}_2 \Big) \Big\}.$$

Same as the argument in theorem 3.1, we can verify that in set $\widetilde{\mathcal{C}}^c$,

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 &\leq \left(M_0^2\tau_n+\tau_n+\frac{\mathcal{E}(\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{g}})-2\sigma_s^2}{\tau_n^{p+q+3}}\right),\\ \|\widetilde{\mathbf{g}}-\mathbf{g}^*\|_2 &\leq \left(M_0^2\tau_n+\tau_n+\frac{\mathcal{E}(\widetilde{\boldsymbol{\beta}},\widetilde{\mathbf{g}})-2\sigma_s^2}{\tau_n^{p+q+3}}\right).\end{aligned}$$

As $\tilde{\lambda}_1=\tilde{\lambda}_2=n^{-1/4}$ and $\tau_n=n^{-\frac{1}{4(p+q+4)}}$, $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\leq O\big(n^{-\frac{1}{4(p+q+4)}}\big)$ and $\|\widetilde{\mathbf{g}}-\mathbf{g}^*\|_2\leq O\big(n^{-\frac{1}{4(p+q+4)}}\big)$. Therefore, letting $\gamma_1=4(p+q+4)$ and $\gamma_2=8(p+q+4)$, Assumption A3 can be satisfied. □

References

[1] Bach, Francis and Jenatton, Rodolphe and Mairal, Julien and Obozinski, Guillaume and others. (2004). Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, **5**, 19–53.

[2] Bertin, Karine and Lecué, Guillaume. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, **5**, 19–53. MR2461900

[3] Bunea, Florentina and Wegkamp, Marten H. (2004). Two-stage model selection procedures in partially linear regression. *The Canadian Journal of Statistics*, **32**, 105–118. MR2064395

[4] Cheng, Guang and Zhang, Hao Helen and Shang, Zuofeng. (2015). Sparse and efficient estimation for partial spline models with increasing dimension. *Annals of the Institute of Statistical Mathematics*. MR3297860

[5] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. **20**, 37–46.

[6] Comminges, Laëtitia and Dalalyan, Arnak S and others. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistic*. **40**, 2667–2696.

[7] De Brabanter, Kris and De Brabanter, Jos and De Moor, Bart and Gijbels, Irène. (2013). Derivative estimation with local polynomial fitting. *The Journal of Machine Learning Research*. **14**, 281–301. MR3033332

[8] Engle, R. F. and Granger, C. W. J. and Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*. **81**, 310–320.

[9] Fan, Jianqing and Gijbels, I. (2003). Local polynomial modelling and its applications. *CRC Press*, Boca Raton. MR1383587

[10] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348–1360. MR1946581

[11] Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of the American Statistical Association*, **99**, 710–723. MR2090905

[12] Fan, Jianqing and Lv, Jinchi. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101–148.

[13] Friedman, Jerome and Hastie, Trevor and Tibshirani, Rob. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**.

[14] Golub, Gene H and Heath, Michael and Wahba, Grace. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

[15] Härdle, Wolfgang and Gasser, Theo. (1985). On robust kernel estimation of derivatives of regression functions. *Scandinavian journal of statistics*, 233–240. MR0817941

[16] Härdle, Wolfgang and Liang, Hua and Gao, Jiti. (2000). Partially Linear Models. *Physica-Verlag*, Heidelberg.

[17] Härdle, W. and Müller, M. and Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models. *Springer-Verlag*, New York. MR2061786

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

[19] Huang, Jianhua Z and Yang, Lijian. (2004). Identification of nonlinear additive autoregressive models. *Journal of the Royal Statistical Society: Series B*, **66**, 463–477.

[20] Hunsberger, Sally and Albert, Paul S. and Follmann, Dean A. and Suh, Edward. (2002). Parametric and semiparametric approaches to testing for seasonal trend in serial count data. *Biostatistic*, **3**, 289–298.

[21] Jarrow, Robert and Ruppert, David and Yu, Yan. (2004). Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association*, **99**, 57–66.

[22] Lafferty, John and Wasserman, Larry. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 28–63. MR2387963

[23] Liang, Feng and Paulo, Rui and Molina, German and Clyde, Merlise A. and Berger, Jim O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410–423.

[24] Liang, H. and Härdle, W. and and Carroll, R.J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics*, **27**, 1519–1535.

[25] Liang, Hua and Li, Runze. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, **104**, 234–248.

[26] Y. Lin and H. H. Zhang. (2006). Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models. *Applied Statistics*, **34**, 2272–2297.

[27] Liu, X. and Wang, Li and Liang, H. (2011). Estimation and Variable Selection for Semiparametric Additive Partial Linear Models. *Statistica Sinica*, **21**, 1225–1248.

[28] Miller, Hugh and Hall, Peter. (2006). Local polynomial regression and variable selection. *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, 216–233. MR2798521

[29] Mukherjee, Sayan and Wu, Qiang. (2006). Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, **7**, 2481–2514.

[30] Mukherjee, S. and Zhou, D. (2006). Learning coordinate covariates via gradient. *Journal of Machine Learning Research*, **7**, 419–549. MR2274377

[31] Müller, Hans-Georg and Stadtmüller, U and Schmitt, Thoma. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, **74**, 743–749.

[32] Nesterov, Yu. (2005). Smooth minimization of non-smooth functions. *Mathematical programming*, **103**, 127–152. MR2166537

[33] Ni, Xiao and Zhang, Hao Helen and Zhang, Daowen. (2009). Automatic model selection for partially linear models. *Journal of Multivariate Analysis*, **100**, 2100–2111. MR2543089

[34] Prada-Sánchez, J.M. and Febrero-Bande, M. and Cotos-Yáñez, T. and González-Manteiga, W. and Bermúdez-Cela, J.L. and Lucas-Dominguez, T. (2000). Prediction of SO_2 pollution incidents near a power station using partially linear models and an historical matrix of predictor-response vectors. *Environmetrics*, **11**, 209–225.

[35] Robinson, P. M. (1998). Root n-Consistent Semiparametric Regression. *Econometrica*, **56**, 931–954.

[36] Rosasco, Lorenzo and Villa, Silvia and Mosci, Sofia and Santoro, Matteo and Verri, Alessandro. (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, **14**, 1665–1714. MR3104492

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

[38] Speckman, P. E. (1998). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B*, **50**, 413–436.

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

[40] Tang, E Ke and Suganthan, Ponnuthurai N and Yao, Xin and Qin, A Kai. (2005). Linear dimensionality reduction using relevance weighted LDA. *Pattern recognition*, **38**, 485–493.

[41] Tibshirani, Robert. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–288. MR1379242

[42] Van Der Vaart, Aad W and Wellner, Jon A. (1996). Weak convergence. *Weak Convergence and Empirical Processes*, 16–28.

[43] G. Wahba. (1990). Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia. MR1045442

[44] Wang, Hansheng and Leng, Chenlei. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis* , **52**, 5277–5286.

[45] Wang, L. and Liu, X. and Liang, H. and Carroll, R. (2011). Estimation and Variable Selection for Generalized Additive Partial Linear Models. *The Annals of Statistics*, **39**, 1827–1851.

[46] Xie, Huiliang and Huang, Jian. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, **37**, 673–696.

[47] Xue, Lan. (2009). Consistent variable selection in additive models. *Statistica Sinica*, **19**, 1281–1296. MR2536156

[48] Yafeh, Yishay and Yosha, Oved. (2003). Large Shareholders and Banks: Who monitors and How? *The Economic Journal*, **113**, 128–146.

[49] Yang, Lei and Lv, Shaogao and Wang, Junhui. (2016). Model free variable selection in reproducing Kernel Hilbert space. *Journal of Machine Learning Research*, **17**, 1–24. MR3517105

[50] Yang, Yi and Zou, Hui. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, **25**, 1129–1141.

[51] A. Yatchew and J. A. No. (2001). Household Gasoline Demand in Canada. *Econometrica*, **69**, 1697–1709.

[52] Ye, Guibo and Xie, Xiaohui. (2012). Learning sparse gradients for variable selection and dimension reduction. *Machine Learning Journal*, **87**, 303–355.

[53] Ying, Yiming and Wu, Qiang and Campbell, Colin. (2012). Learning the coordinate gradients. *Advances in Computational Mathematics*, **37**, 355–378.

[54] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B*, **68**, 49–67. MR2212574

[55] Zeger, S.L. and Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.

[56] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.