Annals of the Institute of Statistical Mathematics (2020) 72:45–63
https://doi.org/10.1007/s10463-018-0682-9

Discovering model structure for partially linear models

Xin He[1] · **Junhui Wang**[2]

Received: 6 December 2017 / Revised: 4 June 2018 / Published online: 30 July 2018

Abstract
Partially linear models (PLMs) have been widely used in statistical modeling, where prior knowledge is often required on which variables have linear or nonlinear effects in the PLMs. In this paper, we propose a model-free structure selection method for the PLMs, which aims to discover the model structure in the PLMs through automatically identifying variables that have linear or nonlinear effects on the response. The proposed method is formulated in a framework of gradient learning, equipped with a flexible reproducing kernel Hilbert space. The resultant optimization task is solved by an efficient proximal gradient descent algorithm. More importantly, the asymptotic estimation and selection consistencies of the proposed method are established without specifying any explicit model assumption, which assure that the true model structure in the PLMs can be correctly identified with high probability. The effectiveness of the proposed method is also supported by a variety of simulated and real-life examples.

Keywords Lasso · Gradient learning · Partially linear models · Proximal gradient descent · Reproducing kernel Hilbert space (RKHS)

1 Introduction

Linear and nonlinear models are two main structures for statistical modeling. The linear models are simple and interpretable, but their efficiency highly relies on the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0682-9) contains supplementary material, which is available to authorized users.

✉ Xin He
xinhe6-c@my.cityu.edu.hk

Junhui Wang
j.h.wang@cityu.edu.hk

[1] School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China

[2] Department of Mathematics, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong 999077, Hong Kong, China

validity of the linearity assumption. The nonlinear models are more flexible and less dependent on the model assumptions, but they are often computationally challenging and less interpretable. In order to enjoy both interpretability and flexibility, partially linear models (PLMs; Engle et al. 1986) were proposed, where some variables are linearly related with the response, while the others are nonlinearly related. PLMs have a wide range of applications in practice, ranging from economics (Schmalensee and Stoker 1999), biomedical studies (Lin and Ying 1994) and environmental science (Prada-Sánchez et al. 2000).

Generally, a PLM considers

$$y = \mu + \mathbf{x}_{\mathcal{L}}^{\mathrm{T}}\boldsymbol{\beta} + h(\mathbf{x}_{\mathcal{N}}) + \epsilon,$$

where y is the response, μ is the intercept, $\mathbf{x} = (\mathbf{x}_{\mathcal{L}}^{\mathrm{T}}, \mathbf{x}_{\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{R}^p$, $\mathcal{L}$ and $\mathcal{N}$ denote the sets of linear and nonlinear variables, $\mathbf{x}_{\mathcal{L}}^{\mathrm{T}}\boldsymbol{\beta}$ is the linear part, $h(\mathbf{x}_{\mathcal{N}})$ is the nonlinear part, $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. In the literature, it is often assumed that $\mathcal{L}$ and $\mathcal{N}$ are pre-specified in the PLMs. For example, in modeling the household gasoline demands, Schmalensee and Stoker (1999) treated age and household income as two nonlinear variables. However, in practice the true model structure is usually unknown, and the choice of nonlinear variables is often justified intuitively. Therefore, a key challenge to PLMs is to correctly determine the linear and nonlinear variables, which has motivated a number of structure selection methods in recent years. Hypothesis testing (Rotnitzky and Jewell 1990) and screening (Fan et al. 2011) are two commonly used methods. The hypothesis testing method constructs multiple hypotheses on different variables for linear against nonlinear fits, and the screening method uses the marginal fits of individual variables to visually determine the model structure. Yet as pointed out in Zhang et al. (2011), it is hard to construct powerful tests when the number of variables is large, whereas the screening method largely depends on the marginal relationship. Zhang et al. (2011) proposed the LAND method which adopts a new regularization framework in the context of smoothing spline ANOVA models to select the model structure for the PLMs. Huang et al. (2012) proposed a model pursuit method by using a group minimax concave penalty to distinguish linear and nonlinear variables. Wu and Stefanski (2015) proposed a kernel-based stepwise structure selection method through local polynomial smoothing, which can identify polynomial nonlinear effect. Lian et al. (2015) proposed a double penalization term for the PLMs to conduct variable selection and structure selection simultaneously. It is interesting to note that all of the aforementioned methods are all based on the additive model assumptions.

In this paper, we propose a new model-free structure selection method for the PLMs, which is capable of discovering the model structure without specifying any model assumptions. Its key idea is to check whether the second-order gradient of the regression function along any variable is exactly 0, implying the variable has linear or nonlinear effects in the PLMs. More importantly, interaction effects can be detected if the second-order gradients along the corresponding variables are nonzero. The proposed method is formulated in a framework of gradient learning, equipped with a flexible reproducing kernel Hilbert space (RKHS; Wahba 1998). It also has a com-

bined penalty term, which couples functional norm and group lasso penalties on the first- and second-order gradient functions, respectively. The resultant optimization task is solved by the accelerated proximal gradient descent algorithm (Rockafellar, 1970). A variety of simulated examples and real applications, together with the asymptotic estimation and selection consistencies, support the advantage of the proposed method over other existing competitors. Particularly, the theoretical results assure that the proposed method is able to discover the true model structure of the PLMs with probability tending to one.

One salient feature of the proposed method is that it examines the second-order gradient functions to distinguish the linear and nonlinear variables and achieves the structure selection purpose. This can be viewed as an extension of the existing gradient learning methods (Ye and Xie 2012; Yang et al. 2016), which mainly focus on the first-order gradient functions to select informative variables. The asymptotic structure selection consistency is established, which is in contrast to most existing methods whose theoretical results are built upon various model assumptions. A useful by-product of the proposed method is that it is able to identify the interaction terms and thus help with the post-structure selection modeling.

The rest of the article is organized as follows. In Sect. 2, we provide a general framework of the proposed method along with the accelerated proximal gradient descent algorithm. Section 3 presents the asymptotic estimation and structure selection consistencies for the proposed method. The numerical experiments on the simulated examples and real applications are contained in Sect. 4. A brief discussion is provided in Sect. 5, and technical proofs are given in "Appendix" and a supplementary file.

2 Methodology

2.1 Preambles

Consider $\mathcal{Z} = (\mathbf{x}, y)$, where $\mathbf{x} = (x^1, \ldots, x^p)^{\mathrm{T}} \in \mathcal{X} \subset \mathcal{R}^p$ and $y \in \mathcal{R}$ are drawn from some unknown distribution. A PLM is formulated as

$$y = f^*(\mathbf{x}) + \epsilon = \mu^* + \mathbf{x}_{\mathcal{L}^*}^{\mathrm{T}} \boldsymbol{\beta}^* + h^*(\mathbf{x}_{\mathcal{N}^*}) + \epsilon,$$

where $f^* : \mathcal{R}^p \to \mathcal{R}$ is the true regression function consisting of an intercept μ^*, a linear coefficient $\boldsymbol{\beta}^*$ and a nonlinear regression function h^*, $\mathcal{L}^*$ is the true linear variable set, and $\mathcal{N}^*$ is the true nonlinear variable set. Here we assume $\mathcal{L}^* \cup \mathcal{N}^* = \{1, \ldots, p\} = \mathcal{S}$, therefore all the variables in $\mathcal{S}$ are truly informative and the goal is to determine whether they belong to $\mathcal{L}^*$ or $\mathcal{N}^*$.

The proposed model-free structure selection method is developed based on the relationship between variables and their corresponding gradient functions. Note that if $l \in \mathcal{L}^*$, all the corresponding second-order gradient functions $\nabla^2_{ll'} f^*(\mathbf{x}) = \partial^2 f^*(\mathbf{x}) / \partial x^l \partial x^{l'} = 0$ for any $\mathbf{x}$ and $l' \in \mathcal{S}$. Therefore, $\mathcal{L}^* = \{l : \nabla^2_{ll'} f^*(\mathbf{x}) = 0, \text{ for any } \mathbf{x} \text{ and } l' \in \mathcal{S}\}$, and $\mathcal{N}^* = \mathcal{S} \backslash \mathcal{L}^*$. This is the key fact that motivates the proposed structure selection method in a gradient learning framework.

Denote $\mathbf{g}^*(\mathbf{x}) = \nabla f^*(\mathbf{x}) = \left(\nabla f_1^*(\mathbf{x}), \ldots, \nabla f_p^*(\mathbf{x})\right)^{\mathrm{T}}$ and $\mathbf{H}^*(\mathbf{x}) = \nabla^2 f^*(\mathbf{x}) = \left(\nabla^2_{ll'} f^*(\mathbf{x})\right)_{l,l'=1}^{p}$ as the true gradient function and the true Hessian matrix of $f^*(\mathbf{x})$, respectively. Note that for any $\mathbf{u}$, $f^*(\mathbf{u})$ can be approximated by $f^*(\mathbf{x}) + \mathbf{g}^*(\mathbf{x})^{\mathrm{T}}(\mathbf{u} - \mathbf{x}) + \frac{1}{2}(\mathbf{u} - \mathbf{x})^{\mathrm{T}}\mathbf{H}^*(\mathbf{x})(\mathbf{u} - \mathbf{x})$ by Taylor's expansion. Then the estimation error of $(\mathbf{g}, \mathbf{H})$ can be denoted as

$$\begin{aligned}\mathcal{E}(\mathbf{g}, \mathbf{H}) &= E_{(\mathbf{x},y),(\mathbf{u},v)} w(\mathbf{x}, \mathbf{u})\big(y - v + \mathbf{g}(\mathbf{x})^{\mathrm{T}}(\mathbf{u} - \mathbf{x}) \\ &\quad + \frac{1}{2}(\mathbf{x} - \mathbf{u})^{\mathrm{T}}\mathbf{H}(\mathbf{x})(\mathbf{x} - \mathbf{u})\big)^2 \\ &= 2\sigma_s^2 + E_{\mathbf{x},\mathbf{u}} w(\mathbf{x}, \mathbf{u})\big(f^*(\mathbf{x}) - f^*(\mathbf{u}) + \mathbf{g}(\mathbf{x})^{\mathrm{T}}(\mathbf{u} - \mathbf{x}) \\ &\quad + \frac{1}{2}(\mathbf{u} - \mathbf{x})^{\mathrm{T}}\mathbf{H}(\mathbf{x})(\mathbf{u} - \mathbf{x})\big)^2, \end{aligned} \tag{1}$$

where $(\mathbf{u}, v)$ is an independent copy of $(\mathbf{x}, y)$, $\sigma_s^2 = E\big(w(\mathbf{x}, \mathbf{u})(y - f^*(\mathbf{x}))^2\big)$, and $w(\mathbf{x}, \mathbf{u}) = e^{-\|\mathbf{x}-\mathbf{u}\|^2/s^2}$ with a pre-specified parameter s, assuring the neighborhood of $\mathbf{x}$ contributes more to estimating $\mathbf{g}^*$ and $\mathbf{H}^*$. Clearly, $\mathcal{E}(\mathbf{g}, \mathbf{H})$ provides an appropriate evaluation metric on the closeness between $(\mathbf{g}, \mathbf{H})$ and $(\mathbf{g}^*, \mathbf{H}^*)$. Besides, We restrict the true regression function f^* to be contained in a RKHS $\mathcal{H}_K$ induced by a kernel function $K(\cdot, \cdot)$. Following the derivative reproducing properties (Zhou 2007), $\mathbf{g}^*$ and $\mathbf{H}^*$ should also be contained in $\mathcal{H}_K^p$ and $\mathcal{H}_K^{p\times p}$, respectively.

2.2 Proposed formulation

Given a training sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the proposed method is formulated as

$$\operatorname{argmin}_{\mathbf{g},\mathbf{H}} \frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij}\Big(y_i - y_j + \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\mathbf{x}_j - \mathbf{x}_i) \tag{2}$$

$$+ \frac{1}{2}(\mathbf{x}_j - \mathbf{x}_i)^{\mathrm{T}}\mathbf{H}(\mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)\Big)^2 + J(\mathbf{g}, \mathbf{H}), \tag{3}$$

which mimics the minimization of $\mathcal{E}(\mathbf{g}, \mathbf{H})$. Here the first term of (3), denoted as $\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})$, can be regarded as an empirical version of $\mathcal{E}(\mathbf{g}, \mathbf{H})$, $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$, and $J(\mathbf{g}, \mathbf{H})$ is a penalty term on both $\mathbf{g}$ and $\mathbf{H}$. Clearly, if $H_{ll'}(\mathbf{x}) = 0$ for all $\mathbf{x}$ and $l' \in \mathcal{S}$, x^l only has linear effect in the PLMs, and otherwise x^l has nonlinear effect in the PLMs.

By the representer theorem of RKHS (Wahba 1998), the minimizer of (3) must have the following forms,

$$\widehat{g}_l(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i^l K(\mathbf{x}, \mathbf{x}_i), \quad \widehat{H}_{ll'}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{c}_i^{ll'} K(\mathbf{x}, \mathbf{x}_i), \quad l, l' = 1, \ldots, p, \tag{4}$$

where $\widehat{\boldsymbol{\alpha}}^l = (\widehat{\alpha}_1^l, \ldots, \widehat{\alpha}_n^l)^{\mathrm{T}} \in \mathcal{R}^n$ and $\widehat{\mathbf{c}}^{ll'} = (\widehat{c}_1^{ll'}, \ldots, \widehat{c}_n^{ll'})^{\mathrm{T}} \in \mathcal{R}^n$ denote the representer coefficients. To achieve the structure selection purpose, we consider a composite regularization term which penalizes the complexity of both $\mathbf{g}$ and $\mathbf{H}$. Then the proposed formulation becomes

$$\operatorname{argmin}_{\mathbf{g},\mathbf{H}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}) + \lambda_0 \sum_{l=1}^{p} \|g_l\|_{\mathcal{H}_K}^2 + \lambda_1 \sum_{l,l'=1}^{p} \pi_{ll'} \|H_{ll'}\|_{\mathcal{H}_K}. \tag{5}$$

Here $\mathbf{K} = \left(K(\mathbf{x}_i, \mathbf{x}_j)\right)_{i,j=1}^n$ is a kernel matrix, $\|g_l\|_{\mathcal{H}_K}^2 = (\boldsymbol{\alpha}^l)^{\mathrm{T}} \mathbf{K} \boldsymbol{\alpha}^l$ is a RKHS norm of g_l , and $\|H_{ll'}\|_{\mathcal{H}_K} = \left((\mathbf{c}^{ll'})^{\mathrm{T}} \mathbf{K} \mathbf{c}^{ll'}\right)^{1/2}$ can be regarded as a group Lasso penalty (Yuan and Lin 2006) on the representer coefficients $\mathbf{c}^{ll'}$, which pushes all or none element of $\mathbf{c}^{ll'}$ to be exactly 0 and thus attains the structure selection purpose. More importantly, $\mathbf{H}$ provides useful information about the model structure of the PLMs. For instances, if $\|H_{ll'}\|_{\mathcal{H}_K} = 0$ for any l', a linear term on x^l shall be included in $f^*(\mathbf{x})$; a nonzero $\|H_{ll'}\|_{\mathcal{H}_K}$ suggests an interaction effect between x^l and $x^{l'}$; and if $\|H_{ll'}\|_{\mathcal{H}_K} = 0$ for all $l' \neq l$ but $\|H_{ll}\|_{\mathcal{H}_K} \neq 0$, a nonlinear term involving only x^l shall be included in $f^*(\mathbf{x})$. Furthermore, λ_0 and λ_1 are two tuning parameters, and $\pi_{ll'}$ is adaptively chosen to assign different weights to $H_{ll'}$ to seek consistent structure selection performance.

2.3 Computing algorithm

Denote $\Omega(\mathbf{H}) = \lambda_1 \sum_{l,l'=1}^{p} \pi_{ll'} \|H_{ll'}\|_{\mathcal{H}_K}$, then the objective function in (5) simplifies to

$$\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}) + \lambda_0 \sum_{l=1}^{p} \|g_l\|_{\mathcal{H}_K}^2 + \Omega(\mathbf{H}).$$

Its minimization can be done using the forward–backward splitting algorithm (Combettes and Wajs 2005). Specifically, at the tth iteration, the update procedure is

$$\mathbf{g}^{t+1} = \operatorname{argmin} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}^t) + \lambda_0 \sum_{l=1}^{p} \|g_l\|_{\mathcal{H}_K}^2, \tag{6}$$

$$\mathbf{H}^{t+1} = \operatorname{prox}_{\frac{1}{D}\Omega}\left(\widetilde{\mathbf{H}}^t - \frac{1}{D}\nabla_{\mathbf{H}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^{t+1}, \widetilde{\mathbf{H}}^t)\right),$$

$$\widetilde{\mathbf{H}}^{t+1} = \mathbf{H}^{t+1} + \frac{t}{t+3}(\mathbf{H}^{t+1} - \mathbf{H}^t), \tag{7}$$

where $(\mathbf{g}^t, \mathbf{H}^t)$ are the current solutions, prox is a proximal operator (Moreau 1962), $\frac{t}{t+3}$ is an accelerated parameter, and D is a Lipschitz upper bound of $\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})$. Following the representer theorem and the notation in (4), the optimization task in (6) can be solved through updating the coefficients $\boldsymbol{\alpha}^l$ as

$$\boldsymbol{\alpha}^{t+1}$$
$$= \left(\frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \Big((\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j))^{\mathrm{T}} \Big) + \lambda_0 \mathbf{I}_p \otimes \mathbf{K} \right)^{-1}$$
$$\left(\frac{1}{n(n-1)} \sum_{i,j=1}^{n} w_{ij} \Big(y_i - y_j + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{H}^t(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j) \Big) \mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j) \right),$$

where $\mathbf{I}_p$ is a p-dimensional identity matrix, and $\mathbf{K}_{\mathbf{x}} = (K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x}))^{\mathrm{T}}$. Moreover, let $\bar{\mathbf{H}}^t = \widetilde{\mathbf{H}}^t - \frac{1}{D} \nabla_{\mathbf{H}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^{t+1}, \widetilde{\mathbf{H}}^t)$, then the solution of (7) is

$$[\mathbf{H}^{t+1}]_{ll'} = \frac{[\bar{\mathbf{H}}^t]_{ll'}}{\|[\bar{\mathbf{H}}^t]_{ll'}\|_{\mathcal{H}_K}} \left(\|[\bar{\mathbf{H}}^t]_{ll'}\|_{\mathcal{H}_K} - \frac{\pi_{ll'}\lambda_1}{D} \right)_+ .$$

The details of calculating $\nabla_{\mathbf{H}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^{t+1}, \widetilde{\mathbf{H}}^t)$ and the related computational issues are provided in the supplementary file. The proposed algorithm then iteratively updates $\mathbf{g}$ and $\mathbf{H}$ as in (6) and (7) until convergence.

Note that the update of $\mathbf{H}^t$ requires a constant D, whose exact value is often difficult to determine in large-scale problems. We can adopt a backtracking scheme (Boyd and Vandenberghe 2004) as a more efficient alternative to approximate the value of D. Furthermore, to optimally determine the tuning parameters in (5), we adopt the stability selection criterion (Sun et al. 2013) to select the tuning parameter achieving the largest variable selection stability.

3 Asymptotic theory

This section establishes the asymptotic estimation and selection consistency for the proposed method in the fixed dimension scenario. Denote the minimizer of (5) as $(\widehat{\mathbf{g}}, \widehat{\mathbf{H}})$, the selected linear variable set as $\widehat{\mathcal{L}} = \{l : \|\widehat{H}_{ll'}\|_{L^2_{\rho_{\mathbf{x}}}} = 0, \text{ for any } l' \in \mathcal{S}\}$, and the selected nonlinear variable set as $\widehat{\mathcal{N}} = \mathcal{S} \backslash \widehat{\mathcal{L}}$, where $\|\widehat{H}_{ll'}\|^2_{L^2_{\rho_{\mathbf{x}}}} = \int (\widehat{H}_{ll'}(\mathbf{x}))^2 \mathrm{d}\rho_{\mathbf{x}}$. The following technical assumptions are made.

Assumption 1 The support $\mathcal{X}$ is a non-degenerate compact subset of $\mathcal{R}^p$. Further, there exists a constant c_0 such that $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{H}^*(\mathbf{x})\|_2 \leq c_0$ with $\|\cdot\|_2$ being the matrix-2 norm, and for any $\mathbf{x}, \mathbf{u} \in \mathcal{X}$,

$$\left| f^*(\mathbf{x}) - f^*(\mathbf{u}) + \mathbf{g}^*(\mathbf{x})^{\mathrm{T}}(\mathbf{u} - \mathbf{x}) + \frac{1}{2}(\mathbf{u} - \mathbf{x})^{\mathrm{T}} \mathbf{H}^*(\mathbf{x})(\mathbf{u} - \mathbf{x}) \right| \leq c_0 \|\mathbf{x} - \mathbf{u}\|_2^3.$$

Assumption 2 For some constants c_1 and θ, the marginal density $p(\mathbf{x})$ satisfies

$$|p(\mathbf{x}) - p(\mathbf{u})| \leq c_1 \Big(d_{\mathbf{x}}(\mathbf{x}, \mathbf{u}) \Big)^{\theta}, \quad \text{for any } \mathbf{x}, \quad \mathbf{u} \in \mathcal{X},$$

where $d_{\mathbf{x}}(\cdot, \cdot)$ is the Euclidean distance.

Assumption 3 There exist constants c_2 and c_3 such that

$$c_2 \leq \lim_{n\to\infty} \min_{l,l'\in\mathcal{N}^*} \pi_{ll'} \leq \lim_{n\to\infty} \max_{l,l'\in\mathcal{N}^*} \pi_{ll'} \leq c_3.$$

In Assumption 1, the boundness condition prevents the loss function from diverging too fast, which extends a similar assumption in Yang et al. (2016). Assumption 2 regularizes the smoothness of the underlying distribution of $\mathbf{x}$ by imposing a Lipschitz condition. By Assumptions 1 and 2, there exists a constant c_4 such that $\sup_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \leq c_4$, and thus the marginal density is bounded everywhere. Assumption 3 restricts the behavior of the adaptive weights when n diverges to infinity.

Theorem 1 *Suppose Assumptions 1–3 are met. For any* $\delta_n \in (0,1)$, $\lambda_0 = O(n^{-\frac{1}{4}})$, $\lambda_1 = O(n^{-\frac{1}{4(p+2)}})$ *and* $s = O(n^{-\frac{1}{4(p+6)(p+2+2\theta)}})$, *there must exist constants* c_5 *and* c_6 *such that with probability at least* $1-\delta_n$,

$$0 \leq \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2 \leq c_5 \left(\log\frac{4}{\delta_n}\right)^{1/2} n^{-\frac{1}{4(p+2+2\theta)}}.$$

And with probability at least $1-\delta_n$,

$$|\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*)| \leq c_6 \left(\log\frac{4}{\delta_n}\right)^{1/2} n^{-\frac{1}{4(p+2+2\theta)}}.$$

Theorem 1 establishes the estimation consistency of the proposed method by showing that with properly chosen λ_0, λ_n and s, the distance between the estimation errors of the estimated and true functions converges to zero at some rate theoretically. It is worthy to point out that the convergence rate in Theorem 1 can be further improved by assuming that the support $\mathcal{X}$ is a d-dimensional connected compacted C^∞ submanifold of $\mathcal{R}^p$ which is isometrically embedded. This assumption maps the high-dimensional data into a low-dimensional manifold, and then the convergence rate in Theorem 1 can be improved to $O_P(n^{-\frac{1}{4(d+2+2\theta)}})$ by Ye and Xie (2012).

Assumption 4 For any $l \in \mathcal{L}^*$, $H^*_{ll'}(\mathbf{x}) \equiv 0$ for any $\mathbf{x}$ and $l' \in \mathcal{S}$; and for any $l \in \mathcal{N}^*$, there exists at least one $l' \in \mathcal{S}$ and some constant $t > 0$ such that $\int_{\mathcal{X}\setminus\mathcal{X}_t} \left(H^*_{ll'}(\mathbf{x})\right)^2 \mathrm{d}\rho_{\mathbf{X}} > 0$, where $\mathcal{X}_t = \{\mathbf{x} \in \mathcal{X} : d_{\mathbf{x}}(\mathbf{x}, \partial\mathcal{X}) < t\}$ and $\partial\mathcal{X}$ is the boundary of $\mathcal{X}$.

Assumption 5 For any $l \in \mathcal{L}^*$ and $l' \in \mathcal{S}$, $n^{-1/2}\lambda_1\psi_{\min}\psi_{\max}^{-1/2}\min\pi_{ll'} \to \infty$, where $\psi_{\max}$ and $\psi_{\min}$ denote the largest and smallest eigenvalues of $\mathbf{K} = \left(K(\mathbf{x}_i, \mathbf{x}_j)\right)_{i,j=1}^n$.

Assumption 4 requires that the corresponding second-order gradient functions along the linear variables are exactly 0, and those along the nonlinear variables are bounded away from zero. Assumption 5 characterizes the behavior of the adaptive weight. For instance, when the second-order sobolev kernel is used, $\psi_{\max}$ and $\psi_{\min}$ are of order $O_p(n)$ and $O_p(n^{-1})$ (Braun 2006; Raskutti et al. 2012), then Assumption 5 is satisfied with $\pi_{ll'} = \|\widetilde{H}_{ll'}\|_2^{-\gamma}$, where γ can be determined by $\psi_{\max}$ and $\psi_{\min}$, and $\widetilde{H}_{ll'}$ is the

solution of (5) with $\lambda_0 = 0$ and $\lambda_1 = 0$. The verification can be done similarly as the proof of Lemma 4 in the supplementary file.

Theorem 2 *Suppose Assumptions 1–5 are met. As* $n \to \infty$, $P(\widehat{\mathcal{L}} = \mathcal{L}^*, \widehat{\mathcal{N}} = \mathcal{N}^*) \to 1$.

Theorem 2 shows that the proposed method can exactly recover the true model structure with probability tending to 1. This theoretical result is established without any explicit model assumption, and provides strong theoretical support for the proposed method in automatically discovering the model structure for the PLMs.

4 Numerical experiments

In this section, we examine the numerical performance of the proposed method in both simulated and real examples, comparing against the double-penalized PLM (Lian et al. 2015), denoted as MF and DPLM, respectively. Note that Lian et al. (2015) showed that DPLM delivers superior performance over a number of existing methods in their numerical experiments. For both MF and DPLM, the kernel function is set as the Gaussian kernel, $K(\mathbf{s}, \mathbf{t}) = e^{-\frac{\|\mathbf{s}-\mathbf{t}\|^2}{\tau^2}}$, where τ is set as the median of all the pairwise distances among the training sample (Jaakkola et al. 1999). As the performance of both methods relies on the choice of tuning parameters, they are determined by the variable selection stability (Sun et al. 2013). This stability criterion conducts a cross-validation scheme and measures the stability by randomly splitting the training sample into two parts. The maximization of stability is conducted via a grid search, where the grid is set as $\{10^{-2+0.1s} : s = 0, \ldots, 40\}$.

4.1 Simulated examples

Two simulated examples are examined. In the first example, the true regression function is additive, whereas in the second example the true regression function contains a three-way interaction term.

Example 1 First, generate $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \in \mathcal{R}^p$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$, where W_{ij} and U_i are independently from $U(0, 1)$. Next, set $f^*(\mathbf{x}_i) = \cos(2\pi x_{i1}) + 10x_{i2}(1 - x_{i2}) + 3x_{i3} + 2x_{i4} - 2x_{i5}$, and then generate $y_i = f^*(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$, where $\sigma = 1$ and 1.37. Clearly, the true regression model is additive, where the first two variables are nonlinear, and the remaining three are linear.

Example 2 The data are generated similarly as in Example 1, except that the true regression function is $f^*(\mathbf{x}_i) = 10x_{i1}x_{i2}x_{i3} + 4x_{i4}^2 + x_{i5} - x_{i6}$, and $\sigma = 0.5$ and 1. Clearly, the true regression model is no longer additive, where the first three variables have interaction effect on the response, the fourth variable is nonlinear, and the last two are linear.

Table 1 The averaged performance measures of MF and DPLM in Example 1

(n, ρ, σ)	Method	NumL	CorrL	NumN	CorrN	CorrI (%)
(200, 0, 1)	MF	3.00	3.00	2.00	2.00	100
	DPLM	3.00	3.00	2.00	2.00	100
(200, 0.1, 1)	MF	3.00	3.00	2.00	2.00	100
	DPLM	3.02	3.00	1.98	1.98	98
(300, 0, 1)	MF	3.00	3.00	2.00	2.00	100
	DPLM	3.00	3.00	2.00	2.00	100
(300, 0.1, 1)	MF	3.00	3.00	2.00	2.00	100
	DPLM	3.00	3.00	2.00	2.00	100
(200, 0, 1.37)	MF	2.94	2.86	2.06	1.92	84
	DPLM	3.24	2.94	1.76	1.70	70
(200, 0.1, 1.37)	MF	2.66	2.60	2.34	1.94	76
	DPLM	3.16	2.82	1.84	1.66	54
(300, 0, 1.37)	MF	2.94	2.94	2.06	2.00	96
	DPLM	3.08	2.94	1.92	1.86	82
(300, 0.1, 1.37)	MF	3.02	2.98	1.98	1.96	94
	DPLM	3.18	2.92	1.82	1.74	68

In each example, scenarios with $n = 200$ and 300 are examined, and η is set as 0 or 0.1, where positive η imposes correlation structure among the variables. Each scenario is repeated 50 times, and the averaged performance measures are summarized in Tables 1 and 2. Specially, NumL and NumN represent averaged number of linear and nonlinear variables selected, CorrL and CorrN represent averaged number of correct linear and nonlinear variables selected, and CorrI represents the percentage that the correct nonlinear variables are exactly identified over the 50 replications.

It is evident that MF has delivered superior performance and outperforms DPLM in most scenarios. In Example 1 where the regression function is additive, MF yields similar structure selection performance as DPLM when $\sigma = 1$, and both of them almost identify the correct model structure perfectly. With $\sigma = 1.37$, the performance of MF is still satisfactory, whereas DPLM tends to overselect linear variables. In Example 2 where the regression function contains a three-way interaction term, MF becomes more advantageous in structure selection against DPLM. DPLM discovers the variable which has square effect and the linear variables with high probability, but it identifies the first three interaction variables as linear in most replications. On the contrary, MF is able to recover the interaction terms among x^1, x^2 and x^3 through the corresponding entries of the estimated Hessian matrix in most replications and distinguishes the variables which have linear or nonlinear effects almost perfectly. Furthermore, in both examples with $\eta = 0.1$, the correlation structure increases the difficulty of identifying the model structure, yet MF still delivers superior performance and outperforms DPLM in most scenarios.

Table 2 The averaged performance measures of MF and DPLM in Example 2

(n, ρ, σ)	Method	NumL	CorrL	NumN	CorrN	CorrI (%)
(200, 0, 0.5)	MF	2.00	2.00	4.00	4.00	100
	DPLM	4.72	1.94	1.28	1.22	2
(200, 0.1, 0.5)	MF	2.00	2.00	4.00	4.00	100
	DPLM	4.76	1.94	1.24	1.18	0
(300, 0, 0.5)	MF	2.00	2.00	4.00	4.00	100
	DPLM	4.86	1.98	1.14	1.12	0
(300, 0.1, 0.5)	MF	2.00	2.00	4.00	4.00	100
	DPLM	4.88	2.00	1.12	1.12	0
(200, 0, 1)	MF	1.62	1.48	4.38	3.86	50
	DPLM	4.08	1.66	1.92	1.58	0
(200, 0.1, 1)	MF	1.60	1.50	4.40	3.90	48
	DPLM	4.22	1.58	1.78	1.36	0
(300, 0, 1)	MF	1.88	1.88	4.12	4.00	92
	DPLM	4.74	1.88	1.26	1.14	0
(300, 0.1, 1)	MF	1.80	1.70	4.20	3.90	64
	DPLM	4.22	1.70	1.78	1.48	0

4.2 Real data analysis

In this section, we analyze the Japanese industrial chemical firm dataset (JICF; Yafeh and Yosha 2003) and the Boston Housing dataset (BH) by using the proposed method. The JICF dataset includes 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 activities with scope for managerial private benefits. The dataset consists of a response variable MH5 (the general sales and administrative expenses deflated by sales) and 12 variables: ASSETS (log(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), OWN-IND (ownership Herfindahl index), AOLC (amount owed to largest creditor), SHARE (share of debt held by largest creditor), BDHIND (bankdebt Herfindahl index) and BDA (bank debt to assets). The dataset is available online through the Economic Journal at http://www.res.org.uk. The BH dataset concerns the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970. It consists of a response variables MEDV (Median value of owner-occupied homes in $1000's) and 13 covariates, including CRIM (per capita crime rate by town), ZN (proportion of residential land zoned for lots over 25,000 square feet), INDUS (proportion of non-retail business acres per town), CHAS (Charles River dummy variable), NOX (nitric oxides concentration), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston

Table 3 The selected structure as well as the corresponding averaged prediction errors by MF and DPLM in the JICF dataset

Variables	MF	DPLM
LEVERAGE	$\mathcal{N}$	$\mathcal{L}$
VARS	$\mathcal{N}$	$\mathcal{N}$
OPERS	$\mathcal{L}$	$\mathcal{N}$
SHARE	$\mathcal{N}$	$\mathcal{N}$
BDHIND	$\mathcal{N}$	$\mathcal{N}$
Pred. Err.	8.0343 (0.1698)	9.2547 (0.1754)

Table 4 The selected structure as well as the corresponding averaged prediction errors by MF and DPLM in the BH dataset

Variables	MF	DPLM
CRIM	$\mathcal{L}$	$\mathcal{L}$
NOX	$\mathcal{N}$	$\mathcal{L}$
RM	$\mathcal{N}$	$\mathcal{N}$
DIS	$\mathcal{L}$	$\mathcal{L}$
TAX	$\mathcal{N}$	$\mathcal{N}$
PTRATIO	$\mathcal{N}$	$\mathcal{L}$
LSTAT	$\mathcal{N}$	$\mathcal{N}$
Pred. Err.	2.9330 (0.0381)	6.0679 (0.7872)

employment centers), index of accessibility to radial highways (RAD), TAX (full-value property-tax rate per $10,000), PTRATIO (pupil–teacher ratio by town), B (the proportion of blacks by town), and LSTAT (lower status of the population).

To proceed, we first apply the variable selection method proposed by Xue (2009), and the variables LEVERAGE, VARS, OPERS, SHARE and BDHIND are identified for JICH dataset, and the variables CRIM, NOX, RM, DIS, TAX, PTRATIO and LSTAT are identified for BH dataset. We then apply the structure selection methods to the identified variables for both datasets. For evaluating the prediction accuracy, we refit the PLM with the selected structure, and randomly split the datasets into two parts with 34 observations of the JICH dataset and 206 observations of the BH dataset for testing and the remaining are for training. The splitting is replicated 1000 times, and the structure selection performance and the averaged prediction errors are summarized in Tables 3 and 4.

As Tables 3 and 4 show, for the JICH dataset, MF identifies OPERS as linear variable, and the other four variables as nonlinear, whereas DPLM identifies LEVERAGE as linear variable and the remaining three as nonlinear. For the BH dataset, MF identifies CRIM and DIS as linear variables and the other five as nonlinear, whereas DPLM detects CRIM, NOX, DIS and PTRATIO as linear variables and the other three as nonlinear. The averaged prediction errors of MF for both datasets are smaller than that of DPLM, suggesting that DPLM may select a wrong model structure that deteriorates

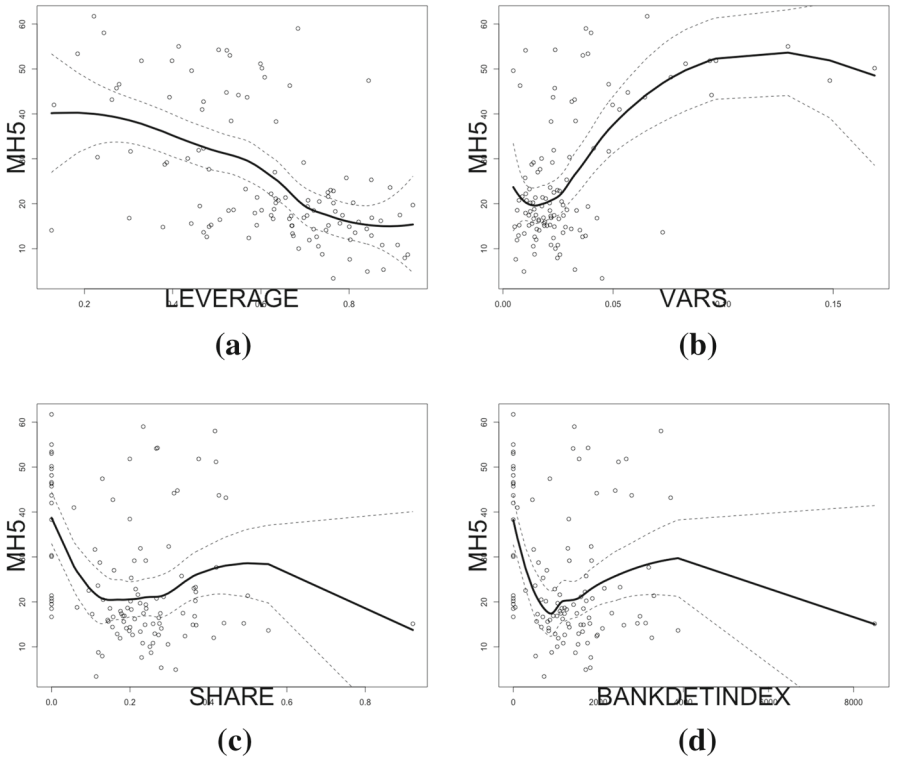

Fig. 1 The scatter plots of MH5 against the selected nonlinear variables in the JICF dataset. The solid line is a fitted curve by local smoothing, and the dashed lines display the fitted mean plus or minus one standard deviation

its prediction performance. Figures 1 and 2 display scatter plots of the response against the identified nonlinear variables by MF, where nonlinear pattern is evident.

It is also interesting to note that MF detects that LEVERAGE and VARS, LEVERAGE and SHARE, LEVERAGE and BDHIND, and SHARE and BDHIND have interaction effects on the response in the JICH dataset, and RM and NOX, RM and TAX, RM and PTRATIO, RM and LSTAT, TAX and PTRATIO, and TAX and LSTAT have interact effects on the response in the BH dataset. The detected interaction terms are summarized in Tables 5 and 6. This is another advantage of MF over DPLM, as DPLM can only detect nonlinear variables without giving detailed model structure.

To scrutinize the detected interaction effects by MF, Fig. 3 displays panel plots of MEDV against NOX or PTRATIO given different ranges of RM in the BH dataset. It is clear that MEDV only slightly decreases with NOX or PTRATIO when RM is small, but sharply decreases with NOX or PTRATIO when RM is large. Evidently, this supports the findings of MF that MEDV is affected by the interactions between RM and NOX, and RM and PTRATIO. Similar patterns can be observed in the panel plots for other interaction terms.

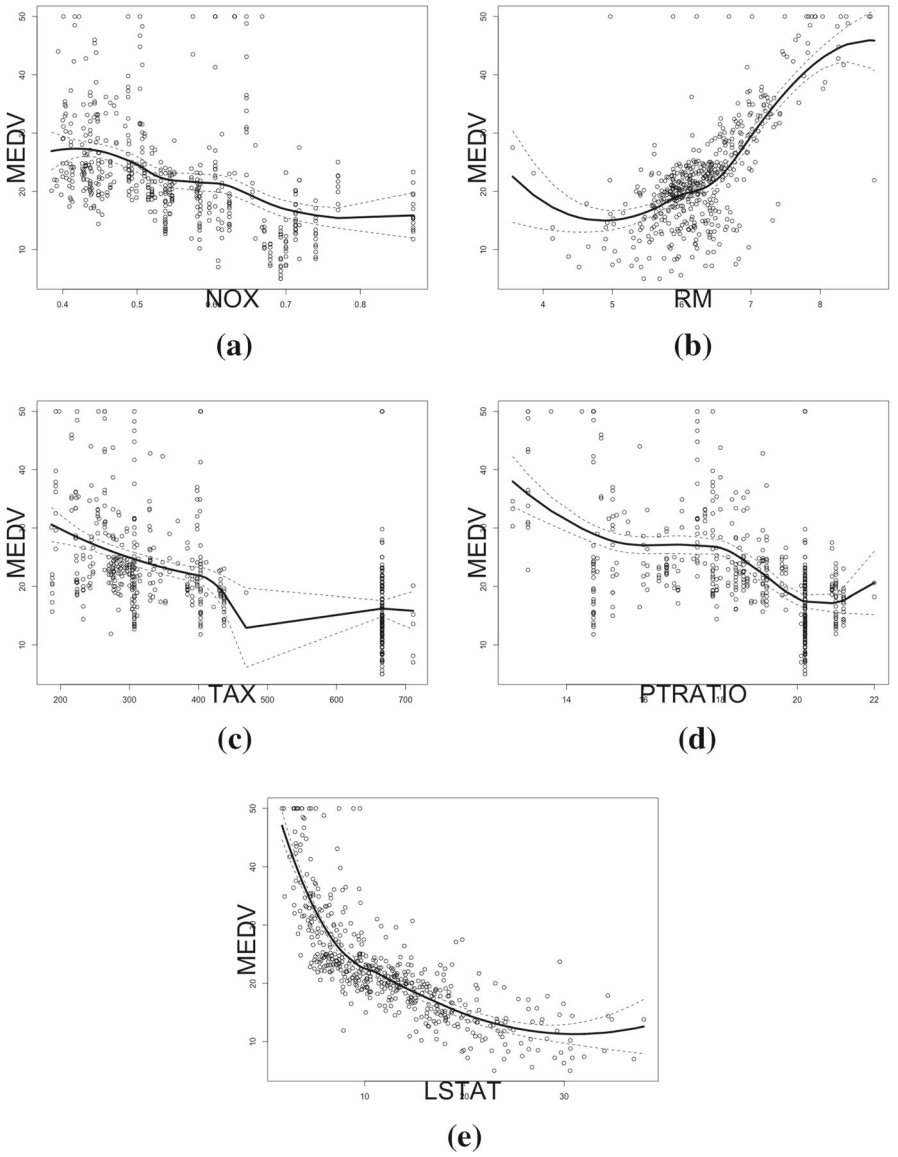

Fig. 2 The scatter plots of MEDV against the selected nonlinear variables in the BH dataset. The solid line is a fitted curve by local smoothing, and the dashed lines display the fitted mean plus or minus one standard deviation

5 Summary

This paper proposes a model-free structure selection method to discover model structure for the PLMs. The proposed method is formulated in a gradient learning framework with a flexible RKHS, and an efficient proximal gradient descent algo-

Table 5 The selected structure by MF in the JICF dataset

Variables	LEVERAGE	VARS	OPERS	SHARE	BDHIND
LEVERAGE	√	√	–	√	√
VARS	√	√	–	–	–
OPERS	–	–	–	–	–
SHARE	√	–	–	√	√
BDHIND	√	–	–	√	√

The √ denotes the detected interaction effect between variables

Table 6 The identified structure by MF in the BH dataset

Variables	CRIM	NOX	RM	DIS	TAX	PTRATIO	LSTAT
CRIM	–	–	–	–	–	–	–
NOX	–	–	√	–	–	–	–
RM	–	√	√	–	√	√	√
DIS	–	–	–	–	–	–	–
TAX	–	–	√	–	√	√	√
PTRATIO	–	–	√	–	√	–	–
LSTAT	–	–	√	–	√	–	√

The √ denotes the detected interaction effect between variables

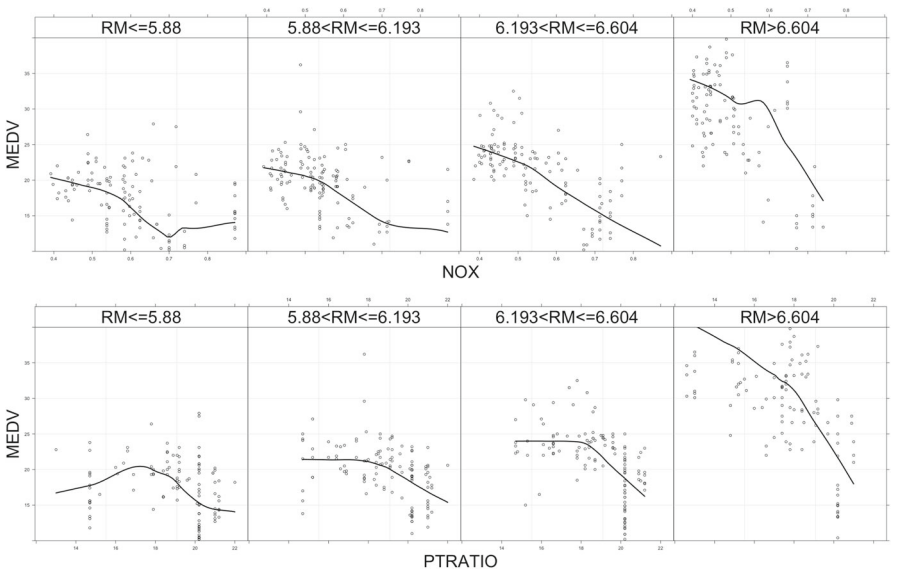

Fig. 3 The scatter plots of MEDV against various interaction terms in the BH dataset. The solid line is the fitted curve by local smoothing

rithm is implemented to tackle the large-scale optimization problem. Its asymptotic estimation and selection consistencies are established without specifying any restrictive model assumptions. Numerical experiments on both simulated and real examples are also supportive of the effectiveness of the proposed method. Note that the asymptotic results for the proposed method are established in the scenario of fixed dimension. To extend it to the scenario with diverging dimension, some screening procedures (Fan and Lv 2008; He et al. 2018) can be employed to first eliminate the noise variables, and then the proposed method can be applied within the selected variable set to identify the true model structure.

Acknowledgements This research is supported in part by HK GRF-11302615, HK GRF-11331016, and City SRG-7004865. The authors would like to thank the associate editor and two anonymous referees for their constructive suggestions. The authors would also like to thank Dr. Heng Lian (City University of Hong Kong) for sharing his code on the DPLM method.

Appendix: technical proofs

Proof of Theorem 1 For some constant a_1, denote

$$\mathcal{C} = \Big\{\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2 \geq a_1 \left(\log \frac{4}{\delta_n}\right)^{1/2} \left(n^{-1/4} + n^{-1/2}\lambda_0^{-1} + n^{-1/2}\lambda_1^{-2} + s^{p+6} + \lambda_0 + \lambda_1\right)\Big\}.$$

Then it suffices to bound $P(\mathcal{C})$. First,

$$\begin{aligned} P(\mathcal{C}) &= P\left(\mathcal{C} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}\right) \\ &\quad + P\left(\mathcal{C} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}^C\right) \\ &\leq P\left(|y| > n^{1/8}\right) + P\left(|y| \leq n^{1/8} \text{ and } U_n > M_0\right) \\ &\quad + P\Big(\mathcal{C} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}\Big) = P_1 + P_2 + P_3, \end{aligned}$$

where $U_n = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} (y_i - y_j)^2$, and $M_0 = 4A^2 + 2\sigma^2 + 1$ with A being the upper bound of $f^*(\mathbf{x})$ on $\mathcal{X}$ and $\sigma^2 = \text{Var}(\epsilon)$. Next, we bound P_1, P_2, and P_3 separately. To bound P_1, we have $P_1 \leq E(|y|)n^{-1/8}$ by the Markov's inequality, where $E(|y|)$ is a bounded quantity. To bound P_2, note that $E(U_n) = E(E(U_n|\mathbf{x}_i, \mathbf{x}_j)) = E((f^*(\mathbf{x}_i) - f^*(\mathbf{x}_j))^2) + E((\epsilon_i - \epsilon_j)^2) \leq 4A^2 + 2\sigma^2$. And thus by Bernstein's inequality for U-statistics (Hoeffding 1963), we have that

$$\begin{aligned} P_2 &\leq P\left(U_n > M_0 \big| |y| \leq n^{1/8}\right) \\ &\leq P\left(U_n - E(U_n) > 1 \big| |y| \leq n^{1/8}\right) \leq \exp\left(-\frac{n^{1/2}}{16}\right). \end{aligned}$$

To bound P_3, within the set $\{|y| \le n^{\frac{1}{8}}$ and $U_n \le M_0\}$, equlaity (1) and by Lemma 3 in the supplementary file, we have with probability at least $1-\delta_n$ that

$$0 \le \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2 \le a_1 \left(\log \frac{4}{\delta_n}\right)^{1/2} \left(n^{-1/4} + M_0^2 n^{-1/2}\lambda_0^{-1} + M_0^2 n^{-1/2}\lambda_1^{-2} + s^{p+6} + \lambda_0 + \lambda_1\right),$$

which implies $P_3 \le \delta_n$, and thus $P(\mathcal{C}) \le \delta_n + O(n^{-1/8})$ for some constant a_1. Specially, when $\lambda_0 = n^{-1/4}$, $\lambda_1 = n^{-1/4(p+2)}$ and $s = n^{-1/4(p+6)(p+2+2\theta)}$, there exists a constant c_5 such that with probability at least $1-\delta_n$

$$0 \le \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2 \le c_5 \left(\log \frac{4}{\delta_n}\right)^{1/2} n^{-\frac{1}{4(p+2+2\theta)}}.$$

Next, we establish the estimation consistency. By Assumptions 1 and 2 and equlaity (1), for some constant a_2 there holds

$$\begin{aligned} 0 \le \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - 2\sigma_s^2 &\le \iint w(\mathbf{x}, \mathbf{u}) c_0^2 \|\mathbf{x}-\mathbf{u}\|_2^6 \rho_{\mathbf{x}} \rho_{\mathbf{u}} \\ &\le c_0^2 c_4 s^{p+6} \int e^{-\mathbf{t}^{\mathrm{T}}\mathbf{t}} \mathbf{t}^{\mathrm{T}}\mathbf{t} \mathrm{d}\mathbf{t} \le a_2 s^{p+6}, \end{aligned}$$

where $a_2 = c_0^2 c_4 \int e^{-\mathbf{t}^{\mathrm{T}}\mathbf{t}} \mathbf{t}^{\mathrm{T}}\mathbf{t} \mathrm{d}\mathbf{t}$, $\mathbf{t} = (\mathbf{u}-\mathbf{x})/s$ and $\int e^{-\mathbf{t}^{\mathrm{T}}\mathbf{t}} \mathbf{t}^{\mathrm{T}}\mathbf{t} \mathrm{d}\mathbf{t}$ is a bounded quantity. Specially, with $s = n^{-1/4(p+6)(p+2+2\theta)}$, we have $\mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - 2\sigma_s^2 \le a_2 n^{-1/4(p+2+2\theta)}$. Therefore, for some constant c_6, triangle inequality implies that

$$\begin{aligned} |\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*)| &\le |\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2| + |\mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - 2\sigma_s^2| \\ &\le c_6 \left(\log \frac{4}{\delta_n}\right)^{1/2} n^{-\frac{1}{4(p+2+2\theta)}}. \end{aligned}$$

This completes the Proof of Theorem 1. □

Proof of Theorem 2: First we show that for any $l \in \mathcal{L}^*$, $\|\widehat{H}_{ll'}\|_{L^2_{\rho_{\mathbf{x}}}} = 0$ for any $l' \in \mathcal{S}$. Note that $\|\widehat{c}_{ll'}\|_2 = 0$ implies that $\|\widehat{H}_{ll'}\|_{L^2_{\rho_{\mathbf{x}}}} = 0$ based on the representer theorem for the RKHS, and thus it suffices to show $\|\widehat{\mathbf{c}}_{ll'}\|_2 = 0$ for any $l \in \mathcal{L}^*$ and $l' \in \mathcal{S}$.

Suppose $\|\widehat{\mathbf{c}}_{ll'}\|_2 > 0$ for some $l \in \mathcal{L}^*$ and $l' \in \mathcal{S}$. The derivative of (5) with respect to $\mathbf{c}_{ll'}$ yields that

$$\sum_{i,j=1}^{n} x_{il} x_{jl'} A_1(\mathbf{x}_i, y_i, \mathbf{x}_j, y_j) \mathbf{K}_{\mathbf{x}_i} = \lambda_1 A_2(\widehat{\mathbf{c}}_{ll'}), \tag{8}$$

where

$$A_1(\mathbf{x}_i, y_i, \mathbf{x}_j, y_j) = \frac{1}{n(n-1)} w_{ij}(y_i - y_j - \widehat{\mathbf{g}}(\mathbf{x}_i)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\widehat{\mathbf{H}}(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j)),$$

$A_2(\widehat{\mathbf{c}}_{ll'}) = \frac{\pi_{ll'}\mathbf{K}\widehat{\mathbf{c}}_{ll'}}{\left(\widehat{\mathbf{c}}_{ll'}^{\mathrm{T}}\mathbf{K}\widehat{\mathbf{c}}_{ll'}\right)^{1/2}}$, and $x_{ijl} = x_{il} - x_{jl}$. For the right-hand side of (8), its norm divided by $n^{1/2}$ is $n^{-1/2}\lambda_1\|A_2(\widehat{\mathbf{c}}_{ll'})\|_2 \geq n^{-1/2}\lambda_1\pi_{ll'}\psi_{min}\psi_{max}^{-1/2}$, which diverges to infinity by Assumption 5. For the left-hand side of (8), by Assumption 1, x_{ijl}, $x_{jil'}$, and every elements of $\mathbf{K}_{\mathbf{x}}$ are bounded. Denote $A_{\mathcal{Z}^n}(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{c}}) = \sum_{i,j=1}^{n} A_1(\mathbf{x}_i, y_i, \mathbf{x}_j, y_j)$, we will show that $|A_{\mathcal{Z}^n}(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{c}})|$ is bounded as well.

For some constant a_3 and $\delta_n \in (0, 1)$, denote

$$\mathcal{D} = \left\{|A_{\mathcal{Z}^n}(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{c}})| > a_3\left(\log\frac{2}{\delta_n}\right)^{1/2} \left(n^{-1/8(p+2+2\theta)} + n^{-3/8} + n^{-1/2}\lambda_0^{-1/2} + n^{-1/2}\lambda_1^{-1}\right)\right\},$$

and thus it suffices to bound $P(\mathcal{D})$. First, we have

$$\begin{aligned} P(\mathcal{D}) &= P\left(\mathcal{D} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}\right) \\ &\quad + P\left(\mathcal{D} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}^C\right) \\ &\leq P\left(|y| > n^{1/8}\right) + P\left(|y| \leq n^{1/8} \text{ and } U_n > M_0\right) \\ &\quad + P\left(\mathcal{D} \cap \{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}\right) \leq P_1 + P_2 + P_4, \end{aligned}$$

where U_n and M_0 are defined as in Theorem 1. Note that $P_1 + P_2 = O(n^{-1/8})$ as in the Proof of Theorem 1. To bound P_4, by Cauchy–Schwarz inequality, we conclude that

$$E(A_{\mathcal{Z}^n}(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{c}})) \leq \left(\iint w(\mathbf{x}, \mathbf{u})\Big(f^*(\mathbf{x}) - f^*(\mathbf{u}) - \widehat{\mathbf{g}}(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{u}) + \frac{1}{2}(\mathbf{x} - \mathbf{u})^{\mathrm{T}}\widehat{\mathbf{H}}(\mathbf{x})(\mathbf{x} - \mathbf{u})\Big)^2 \mathrm{d}\rho_{\mathbf{x}}\mathrm{d}\rho_{\mathbf{u}}\right)^{1/2} = \left(\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2\right)^{1/2}.$$

Within the set $\{|y| \leq n^{1/8} \text{ and } U_n \leq M_0\}$, following similar proofs of Lemma 1 and Proposition 2, we have for some constant a_3, with probability at least $1 - \delta_n$ there holds

$$|A_{\mathcal{Z}^n}(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{c}})| \leq a_3\left(\log\frac{2}{\delta_n}\right)^{1/2}\left(n^{-1/8(p+2+2\theta)} + n^{-3/8} + n^{-1/2}\lambda_0^{-1} + n^{-1/2}\lambda_1^{-1}\right),$$

which implies $P_4 \leq \delta_n$, and thus we have $P(\mathcal{D}) \leq \delta_n + O(n^{-1/8})$. Combining the above results, the norm of the left-hand side of (8) divided by $n^{1/2}$ converges to zero

in probability, which contradicts with the fact that the right-hand side of (8) diverges to infinity when $n \to \infty$. Therefore, for any $l \in \mathcal{L}^*$ and $l' \in \mathcal{S}$, $\|\widehat{\mathbf{c}}_{ll'}\|_2 \equiv 0$, implying $\|\widehat{H}_{ll'}\|_{L^2_{\rho_\mathbf{x}}} = 0$ for any $l \in \mathcal{L}^*$ and $l' \in \mathcal{S}$, and thus there holds $\widehat{\mathcal{L}} \subset \mathcal{L}^*$.

Next, we show $\|\widehat{H}_{ll'}\|^2_{L^2_{\rho_\mathbf{x}}} \neq 0$ for any $l \in \mathcal{N}^*$ and some $l' \in \mathcal{S}$. By Lemma 4, for the set $\mathcal{X}_s = \{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \partial\mathcal{X}) > s,\ p(\mathbf{x}) > s + c_1 s^\theta\}$ and some constant as in the supplementary file, there holds

$$\int_{\mathcal{X}_s} \|\widehat{\mathbf{H}}(\mathbf{x}) - \mathbf{H}^*(\mathbf{x})\|^2_F \mathrm{d}\rho_\mathbf{x} \leq \frac{b_6}{s^{p+5}}(s^{6+p} + \mathcal{E}\left(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}\right) - 2\sigma_s^2),$$

which converges to zero in probability by Theorem 1. Suppose that there exist some $l \in \mathcal{N}^*$ such that $\|\widehat{H}_{ll'}\|^2_{L^2_{\rho_X}} = 0$ for any $l' \in \mathcal{S}$, then

$$\int_{\mathcal{X}_s} (H^*_{ll'}(\mathbf{x}))^2 \mathrm{d}\rho_\mathbf{x} \leq \int_{\mathcal{X}_s} \|\widehat{\mathbf{H}}(\mathbf{x}) - \mathbf{H}^*(\mathbf{x})\|^2_F \mathrm{d}\rho_\mathbf{x}.$$

However, Assumption 4 implies that for some $l' \in \mathcal{S}$, $\int_{\mathcal{X}_s} (H^*_{ll'}(\mathbf{x}))^2 \mathrm{d}\rho_\mathbf{x} > \int_{\mathcal{X}\setminus\mathcal{X}_t} (H^*_{ll'}(\mathbf{x}))^2 \mathrm{d}\rho_\mathbf{x}$ when s is sufficiently small, which is a positive constant, and thus leads to contradiction. Therefore, $\|\widehat{H}_{ll'}\|^2_{L^2_{\rho_\mathbf{x}}} \neq 0$ for any $l \in \mathcal{N}^*$ and some $l' \in \mathcal{S}$, and thus there holds $\widehat{\mathcal{N}} \subset \mathcal{N}^*$.

Finally, since $\mathcal{S} = \mathcal{L}^* \cup \mathcal{N}^* = \widehat{L} \cup \widehat{\mathcal{N}}$ and $\mathcal{L}^* \cap \mathcal{N}^* = \widehat{L} \cap \widehat{\mathcal{N}} = \varnothing$, combining with the above results we have $P(\widehat{\mathcal{L}} = \mathcal{L}^*) \to 1$ and $P(\widehat{\mathcal{N}} = \mathcal{N}^*) \to 1$ when n diverges. Moreover, we have $P(\widehat{\mathcal{L}} = \mathcal{L}^*, \widehat{\mathcal{N}} = \mathcal{N}^*) \geq 1 - P(\widehat{\mathcal{L}} \neq \mathcal{L}^*) - P(\widehat{\mathcal{N}} \neq \mathcal{N}^*) \to 1$ as $n \to \infty$. This completes the Proof of Theorem 2. □

References

Boyd, S., Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Braun, M. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, *7*, 2303–2328.

Combettes, P., Wajs, V. (2005). Signal recovery by proximal forward–backward splitting. *Multiscale Modeling and Simulation*, *4*, 1168–1200.

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

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

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*, *58*, 13–30.

He, X., Wang, J., Lv, S. (2018). Scalable kernel-based variable selection with sparsistency. arXiv:1802.09246.

Huang, J., Wei, F., Ma, S. (2012). Semiparametric regression pursuit. *Statistica Sinica*, *22*, 1403–1426.

Jaakkola, T., Diekhans, M., 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* (pp. 149–158).

Lian, H., Liang, H., Ruppert, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica*, *25*, 591–607.
Lin, D., Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, *81*, 61–71.
Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *Reports of the Paris Academy of Sciences, Series A*, *255*, 2897–2899.
Prada-Sánchez, J., Febrero-Bande, M., Cotos-Yáñez, T., Gonzlez-Manteiga, W., Bermúdez-Cela, J., Lucas-Dominguez, T. (2000). Prediction of SO2 pollution Incidents near a power station using partially linear models and an historical matrix of predictor-response Vectors. *Environmetrics*, *11*, 209–225.
Raskutti, G., Wainwright, M., Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, *13*, 389–427.
Rotnitzky, A., Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized Linear models for cluster correlated Data. *Biometrika*, *77*, 485–497.
Schmalensee, R., Stoker, M. (1999). Household gasoline demand in the united states. *Econometrica*, *67*, 645–662.
Sun, W., Wang, J., Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, *14*, 3419–3440.
Wahba, G. (1998). Support vector machines, reproducing kernel hilbert spaces, and randomized GACV. In B. Scholkopf, C. Burges, A. Smola (Eds.), *Advances in kernel methods: support vector learning* (pp. 69–88). Cambridge: MIT Press.
Wu, Y., Stefanski, L. (2015). Automatic structure recovery for additive models. *Biometrika*, *102*, 381–395.
Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica*, *19*, 1281–1296.
Yafeh, Y., Yosha, O. (2003). Large shareholders and banks: Who monitors and how? *The Economic Journal*, *113*, 128–146.
Yang, L., Lv, S., Wang, J. (2016). Model-free variable selection in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, *17*, 1–24.
Ye, G., Xie, X. (2012). Learning sparse gradients for variable selection and dimension reduction. *Machine Learning*, *87*, 303–355.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with group variables. *Journal of the Royal Statistical Society Series B*, *68*, 49–67.
Zhang, H., Cheng, G., Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of American Statistical Association*, *106*, 1099–1112.
Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, *220*, 456–463.