Annals of the Institute of Statistical Mathematics (2020) 72:827–854
https://doi.org/10.1007/s10463-019-00711-9

Nonparametric variable selection and its application to additive models

Zhenghui Feng[1] · Lu Lin[2,3] · Ruoqing Zhu[4] · Lixing Zhu[5,6]

Received: 25 February 2016 / Revised: 24 January 2019 / Published online: 29 March 2019

Abstract
Variable selection for multivariate nonparametric regression models usually involves parameterized approximation for nonparametric functions in the objective function. However, this parameterized approximation often increases the number of parameters significantly, leading to the "curse of dimensionality" and inaccurate estimation. In this paper, we propose a novel and easily implemented approach to do variable selection in nonparametric models without parameterized approximation, enabling selection consistency to be achieved. The proposed method is applied to do variable selection for additive models. A two-stage procedure with selection and adaptive estimation is proposed, and the properties of this method are investigated. This two-stage algorithm is adaptive to the smoothness of the underlying components, and the estimation consistency can reach a parametric rate if the underlying model is really parametric. Simulation studies are conducted to examine the performance of the proposed method. Furthermore, a real data example is analyzed for illustration.

Keywords Nonparametric regression · Variable selection · Nonparametric additive model · Adaptive estimation

1 Introduction

For multivariate nonparametric regression models with many predictors, even with moderate number of predictors, estimation can be very inefficient, see Härdle (1990). Therefore, when the model is sparse, it is necessary to select "active" predictors and remove "inactive" ones from a parsimonious working model such that further statistical analysis can be performed efficiently. "Active predictors" refer to those components X_i of $X = (X_1, \ldots, X_p)^\top$ that have an effect on the response variable Y. "Inactive predictors" are the other components X_i that are not in the model and do not have

✉ Lixing Zhu
lzhu@hkbu.edu.hk

Extended author information available on the last page of the article

effect on Y. For parametric models, the most promising methodologies use various penalized objective functions for simultaneous selection and estimation. Among them, the proven effective methods are LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), and Dantzig selector (Candés and Tao 2007).

There have been several efforts attempts to apply or extend these methods to handle multivariate nonparametric models. However, when the number (or dimension) of predictors is large, these methods can become inefficient for simultaneous variable selection and estimation. This is because of the following reasons. To obtain the residuals needed to construct an objective function, there must be some approximation of the underlying nonparametric regression function. Any approximation is a parameterization of a nonparametric regression function, and thus its approximation accuracy merely depends on the extent of data denseness in the space and the smoothness of the regression function. As the dimension increases, the number of parameters will increase dramatically and the model fitting may not be sufficiently accurate, which would seriously affect the accuracy of further variable selection and estimation. That is, a meaningful nonparametric approximation is often not possible. Thus, both Lin and Zhang (2006) and Storlie et al. (2011) effectively focused on the additive model and the two-way interaction model, rather than purely multivariate nonparametric regression models. Another strategy is to use ranking and screening to reduce the dimensionality to a relatively low level. Several nonparametric sure screening approaches based on different correlations between the response and every predictor have been reported (Zhu et al. 2011; Li et al. 2012; Lin et al. 2013). These selected models still contain many inactive predictors, and so iterative algorithms combined with existing penalty-based selection methods are required. Fan et al. (2011) proposed the nonparametric independent screening (NIS) method for sparse ultra-high-dimensional additive models. In this method, a B-spline is used to parameterize the nonparametric component functions before screening, and only marginal regression is considered.

All methods described in the literature have a common feature: A nonparametric smoothing approach is used to linearize the component functions $f_j(\cdot)$, and then an objective function with a penalty, such as the group LASSO (Yuan and Lin 2006), is applied to select and estimate groups of variables. For example, Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) method when p is fixed where p is the dimension of the predictors $\mathbf{X} = (X_1, \ldots, X_p)^\top$. This is an extension of the group LASSO and is applicable in cases where p is less than the sample size n. Meier et al. (2009) investigated variable selection in additive models for which $p \gg n$ with a "sparsity-smoothness penalty." This is another group LASSO after model parameterization. Cui et al. (2013) similarly used the idea of nonparametric approximation and group variable selection. Fan et al. (2011) extended the sure independent screening method to sparse ultra-high-dimensional additive models. In all these methods, the nonparametric $f_j(\cdot)$ are approximated by groups of variables and are selected through an *all-in-all-out* fashion. The original p-dimensional space is enlarged to an $\tilde{p} := \sum_{j=1}^{p} k_j$-dimensional space when the corresponding approximation of each function has k_j unknown parameters. To guarantee the consistency of estimates, the k_j must go to infinity as the sample size n goes to infinity, which is required by the consistency of nonparametric estimation (Härdle 1990). In other words, it becomes more difficult to handle large p scenarios.

To attack this difficulty, we propose direct variable selection method for multivariate nonparametric models. Enlightened by the sufficient dimension reduction theory (see, e.g., Li 1991; Cook 1998, and so on), a selection algorithm without any nonparametric approximation is recommended. After "active" variables have been selected, any estimation method for low-dimensional nonparametric models can be easily applied. As an application, we study the variable selection problem for nonparametric additive models (Hastie and Tibshirani 1986). A two-step approach is proposed for direct nonparametric variable selection, and then an adaptive estimation followed. The impact of violating some of the conditions is discussed, and an ad hoc method is proposed for practical implementation.

The remainder of this paper is organized as follows. The selection procedure is introduced in Sect. 2, and its application to additive models is studied in Sect. 3. Simulations are described in Sect. 4, and real data analysis is presented in Sect. 5. We give some discussions on the limitations of our method and also raise some issues that deserve further studies in Sect. 6. The proofs of the theorems are given in the appendix.

2 Selection procedure

For the response Y and the column predictor vector $\mathbf{X} = (X_1, \ldots, X_p)^\top$ with fixed $p < n$, a general multivariate nonparametric model has the form

$$E(Y|\mathbf{X}) = G(\mathbf{X}). \tag{1}$$

Assume that

$$Y \perp\!\!\!\perp \mathbf{X}_{A^c} | \mathbf{X}_A, \tag{2}$$

where $\mathbf{X}_A = \{X_i : i \in A\}$ is the set of relevant X_i such that A is the index set. $\mathbf{X}_{A^c}$ is the complement of $\mathbf{X}_A$ in $\mathbf{X}$. That is, $G(\mathbf{X}) = G(\mathbf{X}_A)$. Only $\mathbf{X}_A$ have effect on Y through $G(\cdot)$; thus, they are defined as "active," and $\mathbf{X}_{A^c}$ are "inactive." Let $d = |A|$ be the cardinality of A. When d is relatively small, and $\mathbf{X}_A$ can be identified, we can then efficiently estimate the regression function $G(\cdot)$. This model is very general, including $Y = G(\mathbf{X}) + G_1(\mathbf{X})\varepsilon$ and $Y = G(\mathbf{X}, \varepsilon)$ as special cases, where ε is independent of $\mathbf{X}$. Throughout this paper, we assume, without loss of generality, that $A = \{1, \ldots, d\}$ and $\mathbf{X}_A = \{X_1, \ldots, X_d\}$. When there is no possibility of confusion, we denote $\mathbf{X}_A = (X_1, \ldots, X_d)^\top$. As long as $\mathbf{X}_A$ can be identified and selected correctly, the subsequent estimation for the regression function is relatively easy. The key is to make the selection without any nonparametric approximation. In this paper, we propose the following linear least-squares sparse (LLSS) solution to identify $\mathbf{X}_A$.

2.1 Linear least-squares sparse solution

Assume $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is the sample from (1). Without loss of generality, assume $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is centered. $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is the $n \times p$ design matrix. $\mathbf{\Sigma}$ is the

covariance matrix of $\mathbf{X}$. Let I_i be the p-dimensional column vector whose ith element is 1 and all others are zero. For each $l_i \in A$, there is a corresponding vector I_{l_i}. Denote a $p \times d$ matrix by $\mathbf{A}_d = (I_{l_1}, \ldots, I_{l_i}, \ldots, I_{l_d}) = (I_1, \ldots, I_d)$. Then, $I_{l_i}^\top \mathbf{X} = X_{l_i}$, and $\mathbf{A}_d^\top \mathbf{X} = \mathbf{X}_A$. The conditional independence in (2) can be rewritten as $Y \perp\!\!\!\perp \mathbf{X} | \mathbf{A}_d^\top \mathbf{X}$.

Note that $\mathbf{A}_d$ is not unique because for any $d \times d$ orthogonal matrix $\mathbf{D}$, $\mathbf{A}_d\mathbf{D}$ can also make the conditional independence hold. This is because $G(\mathbf{A}_d^\top \mathbf{X})$ can also be written as $\tilde{G}(\mathbf{D}^\top \mathbf{A}_d^\top \mathbf{X})$. Thus, it is sufficient to estimate $\mathbf{A}_d\mathbf{D}$. To achieve this, we use the sufficient dimension reduction (Li 1991; Cook 1998) technique. Sufficient dimension reduction approaches estimate the column subspace of $\mathbf{A}_d$ with the minimum dimension, which is denoted by $\mathcal{S}_{Y|\mathbf{X}}$. The space $\mathcal{S}_{Y|\mathbf{X}}$ is called the central subspace (CS, Cook 1998). The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ is the structural dimension. There are a number of methods of identifying and estimating $\mathcal{S}_{Y|\mathbf{X}}$, such as the sliced inverse regression (SIR, Li 1991), the sliced average variance estimation (SAVE, Cook and Weisberg 1991), the directional regression (DR, Li and Wang 2007) and the discretization-expectation estimation (DEE; Zhu et al. 2010), and so on. We are also interested in identifying the indices I_{l_i} themselves correctly. The column vectors in the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ cannot be used directly. To solve this problem, we suggest the following method.

Before proceeding the discussion, let us define some notations. $A_1 = \sum_{i=1}^d I_i$ is a $p \times 1$ vector whose first d components are 1, and other components are 0. In other words, to select the active predictors, it is sufficient to identify A_1. Let $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, and $\eta = \boldsymbol{\Sigma}^{1/2}\mathbf{A}_d$. It is easy to see that η consists of columns of $\boldsymbol{\Sigma}^{1/2}$ corresponding to $\mathbf{A}_d$, and $\mathbf{A}_d^T\mathbf{X} = \eta^T\mathbf{Z}$. Further, define $\eta_1 = \boldsymbol{\Sigma}^{1/2}A_1$, $\mathbf{B}_1$ is a $p \times (p-1)$ matrix orthogonal to $\eta_1/\|\eta_1\|$ and $\mathbf{B} = (\mathbf{B}_1, \eta_1/\|\eta_1\|)$ is an orthogonal matrix, where $\|\eta_1\|^2$ is $A_1^\top \boldsymbol{\Sigma} A_1$. The following theorem provides a sparse solution of $\mathbf{X}$ in the least squares formulation.

Theorem 1 (Sparse solution) *Assume that* $\boldsymbol{\Sigma}$ *is positive definite. Then, almost surely*

$$E(\mathbf{B}_1^\top \mathbf{Z} | Y) = 0 \tag{3}$$

is necessary and sufficient for any function $h(\cdot)$ *on the response* Y*, and* $E(\mathbf{X}h(Y)) \neq 0$*, there exists some constant* c_h *such that*

$$\boldsymbol{\Sigma}^{-1}\mathrm{Cov}\big(\mathbf{X}, h(Y)\big) = c_h A_1 =: \gamma_h \tag{4}$$

provided that it is finite where c_h *depends on* h *and* $c_h = A_1^\top \mathrm{E}\big(\mathbf{X}h(Y)\big)/\|\eta_1\|^2 = E(\sum_{i=1}^d X_i h(Y))/\|\eta_1\|^2$.

Remark 1 A sufficient condition is that the distribution of $\mathbf{X}$ is elliptically symmetric, which includes normality as a special case. This condition is widely used in sufficient dimension reduction theory (Li 1991). In this paper, it can be considered as a mild condition. Because Hall and Li (1993) proved a very useful result: when p tends to infinity, the projections $A_d^\top X$ approximately follow elliptical symmetry. Thus, the proposed method can be theoretically valid when condition (3), including elliptical

symmetry as a special case, holds, and is more applicable in practice when the dimension is high. Li and Duan (1989) provide some detailed discussions on the elliptically symmetric condition. Our result gives a partial solution to relax the elliptical symmetry constraint.

Remark 2 $E(\mathbf{X}h(Y)) \neq 0$ is an identification condition. That is, if $E(\mathbf{X}h(Y)) = 0$, γ_h is not identifiable. When $E(X|Y) = 0$, $E(\mathbf{X}h(Y)) = 0$. This is the problem SIR encounters so that the corresponding central subspace cannot be identified. SAVE is proposed to attack this problem using the conditional variance instead of the conditional mean. Inspired by SAVE, in practice, if $E(\mathbf{X}h(Y)) = 0$, we might use higher orders of covariance $E(\mathbf{X}^k h(Y))$. Thus, in Example 1 in Sect. 4.2, we suggest an ad hoc method for practical use.

Remark 3 For function $h(\cdot)$, there are a number of choices. The simplest option is the identity function $h(y) = y$, such that $E(\mathbf{X}h(Y))$ becomes a real least-squares solution for a pro forma linear model. That is, the formula (4) in Theorem 1, when $c_h \neq 0$, is the coefficient in the following ordinary linear model:

$$Y = c + \gamma^\top \mathbf{X} + e, \tag{5}$$

where $E(e\mathbf{x}) = 0$. In the following, we take $h(y) = y$ for sake of simplicity. In the simulations in Sect. 4.1, we also consider some monotonic transformations.

Note that γ_h is proportional to $A_1 = (1, \ldots, 1, 0, \ldots, 0)^\top$, which takes a value of 1 in the first d elements. Then, the first d components of γ_h should be nonzero, while the others are zero. Elements with nonzero values indicate the active predictors X_i. Thus, a very simple, but efficient, way to identify the active predictors X_i is through the nonzero elements of γ_h. By selecting the "active elements" (nonzero) of γ, we can identify the corresponding active predictors X_i. Therefore, we successfully transfer the variable selection for the nonparametric regression model (2) to variable selection of the linear model (5). This is a sparse least-squares solution. We call this the "Linear Least Squares Sparse" (LLSS) solution. LLSS is rather simple and efficient, and very different from existing methods that select the active predictors and estimate the corresponding regression functions simultaneously. It is obvious that this sparse solution of γ in model (5) makes any successful variable selection approach for linear models feasible, for example, LASSO. For given data points $\{\mathbf{x}_i, y_i\}_{i=1}^n$, the LASSO estimate is defined as

$$\hat{\gamma}(\lambda) = \arg\min_{\gamma} \left\{ \sum_{i=1}^{n} (y_i - c - \gamma^\top \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\gamma_j| \right\}, \tag{6}$$

where $\lambda \geq 0$ controls the amount of regularization applied to the estimate. $\lambda = 0$ changes LASSO to the ordinary least-squares method. Because the selection is exactly the same as that for linear models, the selection consistency holds. Therefore, we will not provide a proof of the following theorem.

Theorem 2 (Selection consistency) *In addition to the condition in Theorem 1, assume* $c_h \neq 0$ *and the conditions designed in* Zhao and Yu (2006) *hold. Then,*

$$\lim_{n\to\infty} P(sgn(\hat{\gamma}) = sgn(\gamma)) = 1, \tag{7}$$

where $sgn(\cdot)$ *is the sign function componentwise. Let* $\hat{d} = \#\{k : \hat{\gamma}_k \neq 0\}$ *that is an estimate of the true number* d *of nonzero components in model* (9)*. Then,*

$$\lim_{n\to\infty} P(\hat{d} = d) = 1. \tag{8}$$

Remark 4 Combined with Theorem 1, LASSO can select the true active predictors $\mathbf{X}_A$ in model (2) with a probability approaching 1. When the conditions in Zhao and Yu (2006) are not satisfied, the adaptive LASSO can be applied, see Zou (2006).

In practice, LLSS is ready to be applied to high-dimensional nonparametric models with $p \geq n$ and the sparse condition $d \ll n$. Note that LLSS can be realized by LASSO to choose the variables. When $p \geq n$, LLSS can first be conducted by sure independent screening (Fan and Lv 2008) to reduce the dimension to a value p' that is less than n, and then to select active variables by LLSS as above. We call this algorithm SLLSS.

3 Application to additive models

Consider the following additive model, which is a special case in the class of nonparametric models:

$$y_i = \mu + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{9}$$

where $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are the sample of $(\mathbf{X}, Y)$, p is the dimension of $\mathbf{X}$, μ is an intercept term, x_{ij} is the jth component of $\mathbf{x}_i$, and $f_j(x_{\cdot j})$ is the additive nonparametric component on $[0, 1]$. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with mean 0 and variance σ^2. Further, the function $f_j(x_{\cdot j})$ is normalized so that $\int_0^1 f_j(u)\mathrm{d}u = 0$ to make the model identification possible. Suppose that model (9) has d nonzero component functions, where $d \ll n$, and all the others are all zero functions. Identifying those "active" variable components that are involved in the nonzero component functions is then equivalent to identifying those nonzero function components. Based on this feature, we perform the model selection for the additive model through variable selection using LLSS and then estimating every nonzero component function. In this section, we propose a two-step method to do both.

For notational convenience, assume that the first d components $f_1(\cdot), \ldots, f_d(\cdot)$ are nonzero, and $f_{d+1}(\cdot) = \cdots = f_p(\cdot) \equiv 0$, $d < p$, and p ($p < n$) is fixed. That is, the parsimonious model is actually

$$y_i = \mu + \sum_{j=1}^{d} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{10}$$

To select and estimate model (10), we first use LLSS to select the "active" variables $\mathbf{X}_A = (X_1, \ldots, X_d)^\top$ and then use an adaptive method to estimate $f_j(\cdot)$ corresponding to $\mathbf{X}_A$. The estimation is adaptive to the smoothness of the underlying function such that the convergence rate is faster than the usual optimal nonparametric rate when the function is smooth enough. When the vector $\mathbf{X}_A$ is obtained, $\hat{d}$ is the number of the components in $\mathbf{X}_A$. By Theorem 2, $\hat{d}$ is consistent for d. In the following estimation procedure, we assume d is known.

3.1 Adaptive estimation

Define an initial estimate and consider the orthogonal decomposition of a reproducing kernel Hilbert space (RHKS) $\mathcal{F}$, as in Lin and Zhang (2006). Let H^j be a function space of functions on x_j over [0, 1] such that $H^j = \{1\} \oplus \bar{H}^j$. For additive models, the responses lie in the direct sum of d orthogonal subspaces H^j. Further details on the RKHS and their reproducing kernels are given in Wahba (1990). The second-order Sobolev-Hilbert space S_2 is most commonly used in practice. Following Lin and Zhang (2006), we use this in our implementation. A special case with the second-order Sobolev space of periodic functions can be written as $t = \{1\} \oplus \bar{T}$, where

$$\bar{T} = \left\{ f : f(t) = \sum_{\nu=1}^{\infty} a_\nu \sqrt{2} \cos 2\pi\nu t + \sum_{\nu=1}^{\infty} b_\nu \sqrt{2} \sin 2\pi\nu t, \right.$$
$$\left. \text{with} \sum_{\nu=1}^{\infty} (a_\nu^2 + b_\nu^2)(2\pi\nu)^4 < \infty \right\}.$$

For a sufficiently large value M, a good approximate subspace of T is $T_M = \{1\} \oplus \bar{T}_M$ with

$$\bar{T}_M = \left\{ f : f(t) = \sum_{\nu=1}^{M/2-1} a_\nu \sqrt{2} \cos 2\pi\nu t + \sum_{\nu=1}^{M/2-1} b_\nu \sqrt{2} \sin 2\pi\nu t + a_{M/2} \cos \pi M t \right\}.$$

M should depend on n and tend to infinity as n tends to infinity. In principle, different values of M can be used for different components $f_j(\cdot)$. For simplicity of implementation, we use the same M. According to the above approximation, denote $\{q_l(t)\}$ as the group with the {sin, cos} orthogonal basis $\{\sqrt{2}\cos 2\pi t, \sqrt{2}\sin 2\pi t, \ldots, \sqrt{2}\cos \pi M t\}$ with coefficients (a_ν, b_ν), denoted as β. Using this orthogonal decomposition, the initial estimates for μ and $f_j(x_j)$, denoted by $\tilde{\mu}$ an $\tilde{f}_j(x_j) = \sum_{l=1}^{M} \tilde{\beta}_{jl} q_l(x_j)$, respectively, can be obtained by minimizing, over μ and β_{jl}:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\left(\mu+\sum_{m=1}^{d}\sum_{l=1}^{M}\beta_{jl}q_l(x_{ij_m})\right)\right\}^2.$$

Here, $q_l(u)$ are the basis functions taken as stated above, which satisfy

$$\int_0^1 q_l(u)\mathrm{d}u=0,\quad \int_0^1 q_l(u)q_s(u)\mathrm{d}u=\begin{cases}1, & \text{for } l=s\\ 0, & \text{otherwise.}\end{cases}$$

The initial estimation used as plug-in in the following step is obtained by least squares. In particular, the estimation can be solved by (5) in Lin and Zhang (2006) with $\lambda=0$.

Plugging in the initial estimate, each initial component estimate, say $\tilde{f}_1(x_1)$, is adjusted by a semiparametric form $\tilde{f}_1(x_1)\xi(x_1)$ or $\tilde{f}_1(x_1)+\zeta(x_1)$, where $\xi(x_1)$ and $\zeta(x_1)$ are the adjustment factor and the adjustment shift, respectively, which will be specified later. To determine $\xi(x_1)$ and $\zeta(x_1)$, motivated by Lin et al. (2009), a local L_2-fitting criterion is defined as

$$r_1(t_1,\xi)=\frac{1}{h}E\left(K\left(\frac{x_1-t_1}{h}\right)\left[f_1(x_1)-\tilde{f}_1(x_1)\xi\right]^2\right), \tag{11}$$

where $K(\cdot)$ is a kernel function satisfying certain regularity conditions and h is a bandwidth depending on n. The minimizer over all ξ is defined as $\xi(t_1)$. We also use the minimizer of the following criterion to define $\zeta(t_1)$:

$$r_2(t_1,\zeta)=\frac{1}{h}E\left(K\left(\frac{x_1-t_1}{h}\right)\left[f_1(x_1)-(\tilde{f}_1(x_1)+\zeta)\right]^2\right). \tag{12}$$

It is easy to show that the minimizers have the following respective closed forms:

$$\xi(t_1)=\frac{E\left(K\left(\frac{x_1-t_1}{h}\right)f_1(x_1)\tilde{f}_1(x_1)\right)}{E\left(K\left(\frac{x_1-t_1}{h}\right)\tilde{f}_1^2(x_1)\right)},\quad \zeta(t_1)=\frac{E\left(K\left(\frac{x_1-t_1}{h}\right)[f_1(x_1)-\tilde{f}_1(x_1)]\right)}{E\left(K\left(\frac{x_1-t_1}{h}\right)\right)}.$$

$\xi(\cdot)$ and $\zeta(\cdot)$ can be estimated, respectively, using $Y-\tilde{\mu}-\sum_{j=2}^{d}\tilde{f}_j(x_j)$ to replace f_1, and the sample averages to the expectations, where $\tilde{\mu}$ and $\tilde{f}_j$ are the initial estimates of μ and f_j for $j\geq 2$:

$$\hat{\xi}(x_1)=\frac{\sum_{i=1}^{n}\left\{Y_i-\tilde{\mu}-\sum_{j=2}^{d}\tilde{f}_j(x_{ij})\right\}\tilde{f}_1(x_{i1})K\left(\frac{x_{i1}-x_1}{h}\right)}{\sum_{i=1}^{n}\tilde{f}_1^2(x_{i1})K\left(\frac{x_{i1}-x_1}{h}\right)},$$

$$\hat{\zeta}(x_1)=\frac{\sum_{i=1}^{n}\left\{Y_i-\tilde{\mu}-\sum_{j=2}^{d}\tilde{f}_j(x_{ij})\right\}K\left(\frac{x_{i1}-x_1}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x_{i1}-x_1}{h}\right)}.$$

Finally, the second-stage estimates of f_1 are respectively attained as

$$\hat{f}_1(x_1) = \tilde{f}_1(x_1) \frac{\sum_{i=1}^n \left\{ Y_i - \tilde{\mu} - \sum_{j=2}^d \tilde{f}_j(x_{ij}) \right\} \tilde{f}_1(x_{i1}) K\left(\frac{x_{i1}-x_1}{h}\right)}{\sum_{i=1}^n \tilde{f}_1^2(x_{i1}) K\left(\frac{x_{i1}-x_1}{h}\right)}, \tag{13}$$

$$\check{f}_1(x_1) = \tilde{f}_1(x_1) + \frac{\sum_{i=1}^n \left\{ Y_i - \tilde{\mu} - \sum_{j=2}^d \tilde{f}_j(x_{ij}) \right\} K\left(\frac{x_{i1}-x_1}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_{i1}-x_1}{h}\right)}. \tag{14}$$

For the other additive components $f_j(\cdot)$, $j = 2, \ldots, d$, the construction scheme is given by substituting $f_1(\cdot)$ for $f_j(\cdot)$, $j = 2, \ldots, d$; the details are omitted here.

3.2 Asymptotics

In this part, we discuss the adaptivity properties for $\hat{f}_j(\cdot)$ and $\check{f}_j(\cdot)$ in (13) and (14). As the result is similar to that in Lin et al. (2009), we only present a brief description and explanation. Assuming that the regression function $f_j \in \bar{T}$, it can be approximated by a function $f_{jM} \in \bar{T}_M$. Let $r_{jM}(x_j) = f_j(x_j) - f_{jM}(x_j)$ and denote the second-order derivative of $r_{jM}(x_j)$ by $r''_{jM}(x_j)$. Write $\sigma_K^2 = \int_{-1}^1 u^2 K(u)\mathrm{d}u$, $J_K = \int_{-1}^1 K^2(u)\mathrm{d}u$ and let $p_j(x_j)$ be the density function of x_j. $K(\cdot)$ on the support $[-1, 1]$ is Lipschitz continuous and $\int_{-1}^1 K(u)\mathrm{d}u = 1$ and $\int_{-1}^1 uK(u)\mathrm{d}u = 0$, and $p_j(x_j) > 0$ for $x_j \in [0, 1]$. We list two conditions below:

C1 For any $f_j(x_j)$, there are $\beta_{jl}^0 \in \Theta$ such that

$$f_j(x_j) = \sum_{l=1}^{\infty} \beta_{jl}^0 q_l(x_j).$$

C2 There exist nonzero functions $e_{jk}(x_j)$, $k = 0, 1, 2$, $j = 1, \ldots, d$ such that

$$\begin{aligned} &\lim_{M\to\infty} M^{\gamma_{j0}} r_{jM}(x_j) = e_{j0}(x_j), \quad \lim_{M\to\infty} M^{\gamma_{j1}} r'_{jM}(x_j) = e_{j1}(x_j), \\ &\lim_{M\to\infty} M^{\gamma_{j2}} r''_{jM}(x_j) = e_{j2}(x_j), \quad j = 1, \ldots, d, \end{aligned}$$

where $\gamma_{j2} \le \gamma_{j1} \le \gamma_{j0}$ and $\gamma_{j0} > 0$.

Condition C1 is almost necessary because any smooth function can admit an orthogonal decomposition (Lin et al. 2009). Condition C2 requests convergence rates of the remainder terms and their derivatives, which are also related to the smoothness of $f_j(\cdot)$. The decreasing relationship between the rates described by $\gamma_{j2} \le \gamma_{j1} \le \gamma_{j0}$ is mild. For example, if the basis functions are chosen to be trigonometric functions or polynomial functions, the remainder term has this property. Let $h = O(n^{-b})$ and $M = O(n^{\delta})$ for $0 < b < 1$, $0 < \delta < 1$. The following theorem gives the details.

Theorem 3 (Adaptivity) *Assume that Conditions C1 and C2 hold as* $n \to \infty$. *For* $x_1 \in (0, 1)$, *the bias and variance of the estimates in* (13) *and* (14) *have the following representations:*

$$bias(\hat{f}_1(x_1)) = \frac{1}{2}h^2\sigma_K^2 r''_{1M}(x_1) + o(h^2M^{-\gamma_{12}}) + O(M^{-\gamma_{10}}) + O(n^{-1}M),$$
$$bias(\check{f}_1(x_1)) = \frac{1}{2}h^2\sigma_K^2 r''_{1M}(x_1) + o(h^2M^{-\gamma_{12}}) + O(M^{-\gamma_{10}}) + O(n^{-1}M),$$
$$var(\hat{f}_1(x_1)) = \frac{\sigma^2 J_K}{nhp_1(x_1)} + O(n^{-1}) + O(n^{-2}h^{-2}),$$
$$var(\check{f}_1(x_1)) = \frac{\sigma^2 J_K}{nhp_1(x_1)} + O(n^{-1}) + O(n^{-2}h^{-2}).$$

The proof of the theorem is similar to that of Theorem 1 in Lin et al. (2009). We give a sketch of the proof in the appendix. $\hat{f}_j(\cdot)$ and $\check{f}_j(\cdot)$, $j = 2, \ldots, d$ have the same properties. For simplicity, we take $f_1(\cdot)$ as an example. The theorem shows that although the variance is the same as that of the common kernel estimation, the bias can be adaptive to the smoothness of the underlying function $f_1(\cdot)$. More precisely, as the value of $|r''_{1M}(x_1)|$ describes the smoothness of $f_1(\cdot)$, the smoother the function $f_1(\cdot)$ is, the smaller the value of $|r''_{1M}(x_1)|$ will be, and consequently, the smaller the bias in (13) and (14). Furthermore, when $f_1(\cdot)$ is sufficiently smooth, $|r''_{1M}(x_1)| \to 0$ as $n \to \infty$, where M is dependent on n and the biases of the estimates are of order smaller than h^2. In this case, the estimates are super-consistent in the sense that the convergence rate in mean squared error (MSE) is faster than the standard order of $n^{-4/5}$. In particular, if $f_1(\cdot)$ satisfies $h^2|r''_{1M}(x_1)| = O(n^{-1/2})$, the estimates can achieve the convergence rate n^{-1} of parametric estimation. Properties of MSE can further be obtained for $\hat{f}_1(x_1)$ and $\check{f}_1(x_1)$. For details, readers are referred to Lin et al. (2009) Corollary 1 and Remark 2.

3.3 Bandwidth selection

To select the bandwidth h, cross-validation (CV) is applied. Let us take the estimation procedure of $f_1(x_1)$ as an example. Assume that the parameter μ and functions $f_j(x_j)$, $j = 2, \ldots, d$ are known. Then, model (9) can be rewritten as a one-dimensional nonparametric regression:

$$y_i - \mu - \sum_{j=2}^{d} f_j(x_{ij}) = f_1(x_{i1}) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Denote $\hat{f}_{i1}(\cdot)$ as the leave-one-out estimator of $f_1(\cdot)$. That is, $\hat{f}_{i1}(\cdot)$ is obtained without the ith sample $(\mathbf{x}_i, y_i)$. Then, the CV function with bandwidth h is defined as

$$\mathbf{CV}(h) = n^{-1}\sum_{i=1}^{n}\left\{\left(y_i - \mu - \sum_{j=2}^{d} f_j(x_{ij})\right) - \hat{f}_{i1}(x_{i1})\right\}^2 w(x_{i1}), \tag{15}$$

where $w(\cdot)$ is a weight function, which is commonly taken as a Gaussian or constant function. For better choices of weight function and further details, please refer to Lin

et al. (2009). Let $h_c = \arg\inf_{h\in H_n} \mathrm{CV}(h)$, where the interval $H_n = (\underline{h}, \bar{h})$, and $\underline{h}$ and $\bar{h}$ satisfy the regularity conditions in Härdle and Marron (1985) so that the choice is based on the following criterion:

$$\lim_{n\to\infty} \frac{d(\hat{m}_h, m)}{\inf_{h\in H_n} d(\hat{m}_h, m)} = 1, \tag{16}$$

where m is a nonparametric function, $\hat{m}_h$ is the kernel estimate with bandwidth h, and d is the averaged squared error. The obtained h_c depends on the parameter μ and the functions $f_j(x_j)$, $j = 2, \ldots, d$, which are in fact unknown. We replace the unknown parameter μ and the function $f_j(x_j)$, $j = 2, \ldots, d$, respectively, by their leave-one-out forms $\tilde{\mu}_i$ and $\tilde{f}_{ij}(x_j)$ and define

$$\widetilde{\mathrm{CV}}(h) = n^{-1} \sum_{i=1}^{n} \left\{ \left(y_i - \tilde{\mu}_i - \sum_{j=2}^{d} \tilde{f}_{ij}(x_{ij}) \right) - \hat{f}_{i1}(x_{i1}) \right\}^2 w(x_{i1}), \tag{17}$$

$$\tilde{h}_c = \arg \inf_{h\in H_n} \widetilde{\mathrm{CV}}(h). \tag{18}$$

$\widetilde{\mathrm{CV}}(h) = \mathrm{CV}(h) + o(1), \quad a.s.$, see Lin et al. (2009).

3.4 Algorithm

The proposed two-step method can be conducted through the following algorithm:

1. Use LASSO for model (5) with the original dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$. $\hat{\gamma}_\lambda$ is the LASSO estimate according to λ which is chosen by BIC or CV.
2. Find the locations of the nonzero components $\{j : \hat{\gamma}_j \neq 0\} =: \Omega$ and the number of nonzero components $\hat{d} = \#\{j : \hat{\gamma}_j \neq 0\} =: |\Omega|$.
3. Corresponding to each $j \in \Omega$, estimate $f_j(\cdot)$, $j \in \Omega$ by the adaptive method. This step includes two substeps: (1) provide initial estimates $\tilde{\mu}$ and $\tilde{f}_j(x_j)$, $j \in \Omega$; (2) adjust them to be adaptive estimates using (13) and (14).

Remark 5 In step 1 of the algorithm, LASSO is not necessary. It can be substituted for another variable selection method for linear models that can shrink their coefficients for "inactive" variables to zero, such as SCAD (Fan and Li 2001) or adaptive LASSO (Zou 2006). When $p \geq n$ or $p \gg n$, sure independent screening (SIS, Fan and Lv 2008) and the iterated sure independent screening (ISIS) can also be used in step 1. The other steps are the same. We denote this method as "SLLSS."

4 Numerical studies

The following simulation studies are discussed. Example 1 considers variable selection for purely nonparametric models. An ad hoc method is proposed when there are symmetric terms in the models. Example 2 investigates variable selection for additive

models. In Case I, $p < n$ and the performance of the two-step method in Sect. 3 is examined. For simplicity, we still call it the "LLSS" method. In Case II, $p \geq n$ and screening is used before LLSS, i.e., "SLLSS." Specifically, comparisons are made with nonparametric independent screening (NIS) method for additive models (Fan et al. 2011), and the penalized method for additive models (penGAM; Meier et al. 2009). Example 3 concerns a classical additive model that has frequently been used in numerical studies (Lin and Zhang 2006). All the results are based on 100 replications. The following three quantities are used to measure the selection accuracy: (1) MS, the mean value of model size (the number of selected components); (2) TP, the mean value of the true positive variables selected; (3) FP, the mean value of the false positive variables missed. Their standard deviations are also reported. For convenience, in Example 2, the initial estimates are obtained by COSSO without any penalty (the tuning parameter $\lambda = 0$), and the bandwidth selection is based on fivefold CV.

4.1 Nonparametric regression models

In this subsection, we examine the performance of the LLSS method for nonparametric models without an additive structure.

Example 1 Consider the following models:

$$Y = \exp\left\{\frac{X_1 + \cdots + X_5}{\sqrt{5}}\right\} + \varepsilon, \tag{19}$$

$$Y = \left(5\left[X_1^3 + X_2^3 + \frac{1}{4}(X_1 + X_2)^2 - \frac{1}{4}(X_1 - X_2)^2\right]\right)^{3/5} + \varepsilon, \tag{20}$$

$$Y = \left(5X_1 + 5X_2 + 10X_3^2 + 10X_4^2\right)^{3/5} + \varepsilon, \tag{21}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$, the samples are $\{\mathbf{x}_i, y_i\}_{i=1}^n$. The errors ε_i, $i = 1, \ldots, n$ follow the normal distribution $N(0, \sigma^2)$ with variance σ^2 such that the signal-to-noise ratio is 3 : 1. The sample size is $n = 100$, $p = 10, 30$. Model (19) has $d = 5$ significant predictors; model (20) has an interaction term with $d = 2$, because it has another expression $Y = \left(5[X_1^3 + X_2^3 + X_1X_2]\right)^{3/5} + \varepsilon$. This model is designed to examine the model identification problem. Model (21) has $d = 4$ and contains two square functions that are symmetric about 0. To show the performance of distribution violations, $\mathbf{x}_i$ are generated independently from $N(0, var)$, $U(-1, 1)$, and $\chi^2(2)$ distributions; in each cases, var is set so as to make each component comparable. LASSO is used in LLSS with the tuning parameter λ selected by the CV method. The results are reported in Table 1, where the standard deviation is given in parentheses.

Various phenomena can be identified in Table 1. First, the distribution of $\mathbf{X}$ does indeed have an effect on the results. $\chi^2(2)$ does not satisfy the condition required for Theorem 1, whereas $U(-1, 1)$ gives comparable results to the normal distribution. Second, p does not have much of and effect on the results. Third, the LLSS method can select the true "active" variables in models (19) and (20) correctly (almost) with a very small false positive value. However, in model (21), X_3 and X_4 cannot be identified.

Table 1 Performance of Example 1

	$p = 10$			$p = 30$		
	MS	TP	FP	MS	TP	FP
Model (19)						
$N(0, 0.1)$	5.14 (0.377)	5 (0)	0.14 (0.377)	5.26 (0.824)	5 (0)	0.26 (0.824)
$U(-1, 1)$	5.07 (0.256)	5 (0)	0.07 (0.256)	5.13 (0.441)	4.99 (0.1)	0.14 (0.427)
$\chi^2(2)$	1.01 (0.1)	0.98 (0.2)	0.03 (0.171)	1.01 (0.1)	0.95 (0.621)	0.06 (0.239)
Model (20)						
$N(0, 1)$	2 (0.141)	1.99 (0.1)	0.01 (0.1)	2 (0.142)	1.99 (0.1)	0.01 (0.1)
$U(-1, 1)$	2.06 (0.312)	2 (0)	0.06 (0.312)	2.04 (0.243)	2 (0)	0.04 (0.243)
$\chi^2(2)$	1.95 (0.297)	1.93 (0.256)	0.02 (0.141)	1.84 (0.420)	1.82 (0.386)	0.02 (0.141)
Model (21)						
$N(0, 0.1)$	2.08 (0.367)	2.05 (0.261)	0.03 (0.171)	2.03 (0.171)	2.03 (0.171)	0 (0)
$U(-1, 1)$	1.87 (0.393)	1.85 (0.359)	0.02 (0.141)	1.84 (0.615)	1.76 (0.429)	0.08 (0.394)
$\chi^2(2)$	2.4 (0.898)	2.28 (0.587)	0.12 (0.456)	2.19 (0.506)	2.11 (0.345)	0.08 (0.339)

Table 2 Performance for Model (19) and (20)

	Model (19)			Model (20)		
	MS	TP	FP	MS	TP	FP
Identity	5.11 (0.373)	5 (0)	0.11 (0.373)	2.02 (0.245)	1.99 (0.1)	0.03 (0.222)
F_n	5.09 (0.351)	5 (0)	0.09 (0.351)	1.99 (0.1)	1.99 (0.1)	0 (0)
Logistic	5.07 (0.293)	5 (0)	0.07 (0.293)	2.03 (0.223)	2 (0)	0.03 (0.223)
Φ	5.16 (0.465)	5 (0)	0.16 (0.465)	2.03 (0.171)	2 (0)	0.03 (0.171)

Table 2 reports the results with different choices of $h(\cdot)$. Because LLSS failed to work for model (21), we only report the results for models (19) and (20). The settings are the same as described above, the predictor is normally distributed and $p = 10$. $h(y)$ is chosen to be identity function (Identity), the empirical cumulative distribution function($F_n(y)$), $\exp(y)/(1+\exp(y))$ (Logsitic), or the standard normal distribution's cumulative distribution function (Φ). We will discuss an ad hoc method in the next subsection.

4.2 An ad hoc method

As discussed in Remark 2, we suggest an ad hoc approach. Note that LLSS transfers the original model to a linear model $Y = a + b_1^\top \mathbf{X} + e$ and then checks whether any component X_i significantly affects Y. If $E(X|Y) = 0$, then $E(\mathbf{X}Y) = 0$, and LLSS fails to work. We then consider using higher-order covariances in practice. We consider determining whether a polynomial term of X_i has any effect on Y. For instance, a second-order polynomial $Y = a + b_1^\top \mathbf{X} + b_2^\top \mathbf{X}^2 + e$ is considered, where $\mathbf{X}^2 = (X_1^2, X_2^2, \ldots, X_p^2)^\top$. Higher-order polynomial terms such as the third and fourth

Table 3 Performance of the ad hoc approach for Model (21), $n = 100$, $p = 10$

Order	$N(0, 0.1)$			$U(-1, 1)$		
	MS	TP	FP	MS	TP	FP
Square	5.95 (1.546)	3.9 (0.362)	2.05 (1.431)	4.59 (0.9)	4 (0)	0.59 (0.9)
Cubic	5.69 (1.548)	3.85 (0.411)	1.84 (1.383)	4.42 (0.843)	4 (0)	0.42 (0.843)
Fourth	5.36 (1.618)	3.76 (0.553)	1.6 (1.363)	4.4 (0.752)	4 (0)	0.4 (0.752)

order could also be considered. The predictor component X_i is selected as "active" if, in model $Y = a + b_1^\top \mathbf{X} + b_2^\top \mathbf{X}^2 + b_3^\top \mathbf{X}^3 + b_4^\top \mathbf{X}^4 + e$, any coefficient corresponding to X_i^j, $j = 1, 2, 3, 4$, is nonzero under LASSO. This ad hoc selection is different from the group LASSO in both purpose and procedure although they look very similar. For example, X_1 is selected as long as any one of the coefficients of the terms X_1, X_1^2, X_1^3, X_1^4 is nonzero rather than all of these coefficients are nonzero, which is the procedure of the group LASSO in an all-in-all-out fashion to select X_1. This difference comes from the purpose of using higher-order covariances in our method to avoid the non-identification with $E(\mathbf{X}h(Y)) = 0$ when only the first order term of X_1 is used. The ad hoc method just aims at selecting variables, not estimation. The adaptive estimation in the second step can provide better estimation. Based on experience, this ad hoc method with order 2 or 3 in practice seems to work well. It is clear that the theoretical investigation deserves a further study.

In Table 3, we present the results of the ad hoc approach for model (21) with $p = 10$. "Square" means that $a + b_1^\top \mathbf{X} + b_2^\top \mathbf{X}^2$ is used. Similarly, "Cubic" and "Fourth" mean that the first third- and fourth-order polynomial terms of $\mathbf{X}$ are involved.

From Table 3, we find that the ad hoc method can determine X_3 and X_4 and select all four active components with almost 100% accuracy. For simplicity, the "square" case performs better. For the other two models in Example 1, the ad hoc method has a similar performance to the regular LLSS shown (Table 1), so the simulation results are omitted. When $p > n$, LLSS can be used after some screening methods. We present some examples of variable selection for additive models in the following subsection.

4.3 Application to additive models

The following examples consider $p = 10$, $p = 100$, and 500 as the dimensions of $\mathbf{X}$. The sample size is $n = 100$. In case 1, $p < n$ and LLSS is compared with COSSO (Lin and Zhang 2006). In case 2, $p \leq n$ and the sure independent screening (SIS; Fan and Lv 2008) is used before implementing LLSS, which is denoted as "SLLSS." As COSSO does not suit this case, we compare SLLSS with greedy NIS and iterative NIS (Fan et al. 2011), and penGAM (Meier et al. 2009). Example 3 is an additive model from Lin and Zhang (2006) for the use of COSSO. We compare the results of LLSS, LLSS with order 2, and COSSO.

Example 2 The model is

$$Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + \sum_{j=4}^{p} f_j(X_j) + \varepsilon, \tag{22}$$

where $f_1(x) = 5(x-1)$, $f_2(x) = 20(x-0.5)\Phi(-|x-0.5|)$, $f_3(x) = -4x^3 + 1$, and $f_j(X_j) = 0$, $j = 4\ldots, p$, $p = 10$. The sample data $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are independent and identically distributed according to the following two distributions: (1). Trimmed AR(1): $W_1, \ldots, W_p \overset{\text{def}}{=} \sim i.i.d.\ N(0,1)$, and $X_1 = W_1$, $X_j = \rho X_{j-1} + (1-\rho^2)^{1/2} W_j$, $j = 2, \ldots, p$. Trim X_j in $[-2.5, 2.5]$ and scale to [0, 1]. (2). Compound symmetry: $W_1, .., W_p$, $U \sim$ Uniform (0, 1) i.i.d., let $X_j = (W_j + tU)/(1+t)$. Therefore, $\text{corr}(X_j, X_k) = t^2/(1+t^2)$, $j \neq k$. Additionally, the errors ε_i, $i = 1, \ldots, n$, follow the normal distribution $N(0, \sigma^2)$, where σ^2 is chosen according to the standard deviation of the signal-to-noise ratio around 3:1. In the simulations, $\rho = 0, 0.5$, and $t = 0, 1$. As estimation is involved, we then report ISE to measure the estimation accuracy. Here, ISE $= E[\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)]^2$, which is estimated by 1000 testing points generated from the same distribution as the training points, where $f(\mathbf{x}_i) = \sum_{j=1}^{p} f_j(x_{ij})$ is the true conditional mean function.

Case I ($p < n$): The results reported in Table 4 are based on $p = 10$, $n = 100$. From the results, LLSS gives smaller ISE and FP than COSSO, which tends to select more variables than LLSS such that its MS is larger. Both methods can find X_1, X_2, X_3 with 100% accuracy and with MS close to the true value of 3. Weak correlation between the independent variables does not have a significant effect on the results.

When the condition (2.3) of LLSS is violated, the performance is not satisfactory. Table 5 presents the results from COSSO, LLSS, and LLSS with order 2 when X is $\chi^2(2)$ distributed or the error follows $\chi^2(2)$. Here, the performance of both COSSO and LLSS has deteriorated. The ad hoc method failed here because it is used for violation of $E(Xh(Y)) = 0$, not for the elliptical distribution condition. When the errors are not normally distributed, the performance of LLSS and LLSS with the ad hoc method is better than COSSO, with smaller FP value and closer to the true value of MS. We can see that, when using COSSO, MS is larger than the true value of 3, and FP is much larger than when using LLSS and LLSS with the ad hoc method. In other words, COSSO selects about four variables, but often wrongly selects variable that should not be active.

Case II ($p \geq n$): In this case, $p = 100, 500$. SIS is used before LLSS. Because this is a high-dimensional case, the greedy NIS(g-NIS), iterative NIS(INIS), and penGAM method are compared. Consider the case where $\mathbf{x}_i$ are independent, that is, trimmed AR(1) with $\rho = 0$. The results of these four methods are reported in Table 6. We also considered some other settings, but the details are omitted because they produce similar performance.

It is clear from Table 6 that all four methods can select X_1, X_2, and X_3 correctly. penGAM and INIS tend to select more variables, so they have larger MS and FP. Because LLSS is used after SIS, the dimension can be reduced to an appropriate

Table 4 Measurements for model (22) Case I

Method	ISE (sd)	MS (sd)	TP (sd)	FP (sd)	ISE (sd)	MS (sd)	TP (sd)	FP (sd)
	Trimmed AR(1), $\rho = 0$				Compound symmetry, $t = 0$			
COSSO	0.053 (0.025)	3.720 (1.147)	3.0 (0)	0.720 (1.147)	0.107 (0.057)	3.710 (1.192)	3.0 (0)	0.710 (1.192)
LLSS	0.041 (0.017)	3.080 (0.307)	3.0 (0)	0.080 (0.307)	0.091 (0.052)	3.070 (0.293)	3.0 (0)	0.070 (0.293)
	Trimmed AR (1), $\rho = 0.5$				Compound symmetry, $t = 1$			
COSSO	0.010 (0.004)	3.590 (1.055)	3.0 (0)	0.590 (1.055)	0.023 (0.011)	3.850 (1.445)	3.0 (0)	0.850 (1.445)
LLSS	0.008 (0)	3.030 (0.171)	3.0 (0)	0.030 (0.171)	0.018 (0.008)	3.030 (0.171)	3.0 (0)	0.030 (0.171)

Table 5 Measurements for model (22) Case I with different settings

Method	$X \sim \chi^2(2)$				$\varepsilon \sim \chi^2(2)$			
	ISE (sd)	MS (sd)	TP (sd)	FP (sd)	ISE (sd)	MS (sd)	TP (sd)	FP (sd)
COSSO	0.061 (0.032)	4.16 (1.973)	2.76 (0.429)	1.4 (1.848)	1.155 (0.219)	3.96 (1.645)	3 (0)	1.96 (1.645)
LLSS	0.095 (0.089)	2.43 (0.670)	2.34 (0.497)	0.09 (0.351)	1.243 (0.393)	2.93 (0.432)	2.88 (0.356)	0.05 (0.219)
LLSS (order 2)	0.135 (0.427)	2.87 (1.3231)	2.48 (0.502)	0.39 (1.063)	1.225 (0.382)	2.96 (0.448)	2.9 (0.333)	0.06 (0.278)

Table 6 Measurements for model (22) Case II

Method	$p = 100$				$p = 500$			
	ISE (sd)	MS (sd)	TP (sd)	FP (sd)	ISE (sd)	MS (sd)	TP (sd)	FP (sd)
SLLSS	0.094 (0.006)	3 (0)	3 (0)	0 (0)	0.064 (0.005)	3 (0)	3 (0)	0 (0)
g-INIS	6.632 (0.617)	3.99 (0.1)	3 (0)	0.99 (0.1)	0.817 (0.045)	3.01 (0.1)	3 (0)	0.01 (0.1)
INIS	9.608 (0.920)	7.95 (0.5)	3 (0)	4.95 (0.5)	1.366 (0.056)	4.99 (0.1)	3 (0)	1.99 (0.1)
penGAM	5.432 (0.426)	14.98 (0.2)	3 (0)	11.98 (0.2)	1.312 (0.024)	22.85 (1.5)	3 (0)	19.85 (1.5)

number, and then the two-stage selection and adaptive estimation can be performed. In this example, SLLSS works well.

Example 3 Following Lin and Zhang (2006), we generate data from the following additive model:

$$Y = 5g_1(X_1) + 2g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon, \tag{23}$$

where $g_1(x) = x, g_2(x) = (2x-1)^2, g_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)}$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. The covariates $X = (X_1, \ldots, X_p)^\top$ are simulated with the same settings as in Example 2: trimmed AR(1), $\rho = 0, 1$; compound symmetry, $t = 0, 1$, $\varepsilon \sim N(0, 1)$. $n = 100$, $p = 10$. We report the results of COSSO, LLSS, and LLSS with order 2 in Table 7.

From the results, we can see that LLSS does not perform well, especially when the predictor variable X is correlated. The ad hoc method can help to select more variables into the model, but it also wrongly selects some inactive variables. Specifically, LLSS tends to miss some active variables with TP smaller than 4, but with the help of the ad hoc method, most of the missing variables can be selected at the cost of some inactive variables being included. It is noticed that functions g_3 and g_4 contain symmetric patterns. Thus, overall LLSS is comparable in this example.

5 Real data example

We applied our method to the Hitters' salary data, which were first presented in the 1988 ASA Graphics Poster Session. The main topic of this session was "why they make what they make." Chaudhuri et al. (1994) considered a tree model, and Li et al. (2000) used a dimension reduction approach to fit a semiparametric model. In detail, the data set consists of the numbers of times at bat(X_1), hits (X_2), home runs (X_3), runs (X_4), runs batted in (X_5) and walks (X_6) in 1986, years in major leagues (X_7), times at bat (X_8), hits (X_9), home runs (X_{10}), runs (X_{11}), runs batted in (X_{12}), and walks (X_{13}) during their entire career up to 1986, annual salary (Y) in 1987, putouts (X_{14}), assistances (X_{15}), and errors (X_{16}). Let $\mathbf{X} = (X_1, \ldots, X_{16})^\top$.

In a nonparametric regression structure, $p = 16$ is too large to have an efficient nonparametric estimation with a sample of size $n = 255$. Therefore, there have been several attempts to work on estimation. Sufficient dimension reduction (Li 1991; Cook 1998) is a promising way to approach estimation via the selection of some representative predictors or the linear combinations of the predictors to establish the underlying model. When SIR (Li 1991) with a BIC-type structural dimension determination (Zhu et al. 2006) is applied, two linear combinations of the 16 predictors are determined.

As there is no specific prior information about the model structure, we first fitted the data nonparametrically. For this purpose, the LLSS method in Sect. 2 was used to select active predictors. To make the analysis robust, we repeated a bootstrap experiment 1000 times to examine the stability of the variable selection through the frequencies of the variables selected into the model. For each experiment, $n = 255$ independent data were sampled from the original dataset with equal probability. The frequencies of all

Table 7 Measurements for model (23)

Method	ISE (sd)	MS (sd)	TP (sd)	FP (sd)	ISE (sd)	MS (sd)	TP (sd)	FP (sd)
	Trimmed AR (1), $\rho = 0$				Compound symmetry, $t = 0$			
COSSO	0.825 (0.460)	4.25 (1.29)	3.76 (0.429)	0.49 (1.141)	0.657 (0.293)	4.09 (0.552)	3.97 (0.171)	0.12 (0.518)
LLSS	2.720 (1.788)	2.08 (0.929)	2.02 (0.828)	0.06 (0.278)	3.407 (2.625)	2.45 (0.903)	2.37 (0.08)	0.08 (0.307)
LLSS (order 2)	1.466 (0.870)	5.81 (2.521)	3.4 (0.778)	2.41 (1.954)	2.625 (0.293)	6.76 (2.275)	3.6 (0.532)	3.16 (1.911)
	Trimmed AR (1), $\rho = 0.5$				Compound symmetry, $t = 1$			
COSSO	0.823 (0.417)	4.04 (1.034)	3.71 (0.456)	0.33 (0.841)	0.806 (0.360)	3.96 (0.994)	3.62 (0.488)	0.34 (0.742)
LLSS	2.563 (1.194)	1.75 (0.642)	1.74 (0.630)	0.01 (0.1)	2.006 (1.346)	1.9 (0.798)	1.82 (0.520)	0.08 (0.464)
LLSS (order 2)	1.539 (0.829)	5.53 (2.619)	3.25 (0.925)	2.28 (1.870)	1.246 (0.494)	6.31 (2.394)	3.44 (0.770)	2.87 (1.807)

Table 8 Frequencies (out of 1000 times) selected by LLSS

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
LLSS	**659**	**736**	38	49	371	**680**	**754**	147
LLSS (order 2)	161	**987**	245	244	312	**981**	**1000**	**907**
Variable	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
LLSS	38	75	**986**	19	15	358	42	46
LLSS (order 2)	**722**	**724**	**969**	570	161	**985**	555	339

Numbers in bold have larger values than 600

variables are listed in Table 8. On average, LLSS identified five important variables, whereas the ad hoc method selected more. The sample covariance matrix shows that there are four separate groups of all predictors: $\{X_1, \ldots, X_6\}$, $\{X_7\}$, $\{X_8, \ldots, X_{13}\}$, and $\{X_{14}, X_{15}, X_{16}\}$. Within the groups, the predictors are highly positively correlated, with correlation coefficients of around 0.8, whereas between the groups, they are positively, but weakly, correlated, with correlation coefficients of around 0.2. Thus, X_1, X_2, X_6 are representatives of the first group, and X_{11} could be regarded as the representatives of the third group. It seems that X_{14} is only identified by the ad hoc method and belongs to group 4. Finally, according to the results, we decided to use variables X_1, X_2, X_6, X_7, X_{11}, X_{14} in the working model. These variables are related to a player's performance in the year 1986; his experience (years in major leagues); the comprehensive performance up to year 1986; and his bad performance.

For comparison, SIR was also used to select projection indices for the purpose of sufficient dimension reduction. To be fair, we select the first two dimension reduction directions corresponding to the first two largest eigenvalues of the SIR matrix and the two variables with the highest selected frequencies in LLSS, i.e., (X_7, X_{11}). Nonparametric regression models were fitted with these predictors (indices). To compare the performance, we list the selected predictors (indices), the regression coefficients, and R^2 in Table 9.

Both SIR and LLSS show that the set $\{X_{14}, X_{15}, X_{16}\}$ makes very little contribution to the response Y. However, LLSS with order 2 identifies X_{14}. This means that X_{14} would affect Y in a quadratic term. When SIR is applied, the large loadings of $\hat{\gamma}_1$ correspond to X_1 and X_2, X_{11}). This means the first direction corresponds to a contrast between X_1 and (X_2, X_{11}), whereas $\hat{\gamma}_2$ would be a contrast between X_8 and X_9. However, note that the pairs (X_1, X_2) and (X_8, X_9) have highly positive correlations, with correlation coefficients of around 0.8. It is difficult to explain why such contrasts could happen. The proposed LLSS method provides a better fitted model with a larger R^2 value and, more importantly, is more interpretable.

We now examine whether a better model could be fitted according to model fitting and interpretability. We consider fitting an additive model. The prediction accuracy can be checked by splitting the dataset into a training set with $n_1 = 200$ to estimate the component functions $f_j(\cdot)$ and a testing set with $n_2 = 55$. The residual sum of squares (RSS) defined as $\sum_{i=1}^{n_2}(y_i - \hat{y}_i)^2$, and the regression coefficient of determination R^2 is reported in Table 10. We list the results for COSSO, LLSS, and LLSS with order 2 for additive models.

Table 9 The indices selected by the SIR, the predictors selected by LLSS

Method	SIR $R^2 = 0.72$		LLSS $R^2 = 0.81$	
Index	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
X_1	**0.38**	0.0	0	0
X_2	**−0.67**	−0.06	0	0
X_3	−0.02	0.01	0	0
X_4	0.15	−0.04	0	0
X_5	0.07	−0.04	0	0
X_6	−0.21	0.00	0	0
X_7	−0.09	−0.17	1	0
X_8	−0.02	**−0.79**	0	0
X_9	0.03	**0.45**	0	0
X_{10}	−0.25	−0.04	0	0
X_{11}	**−0.45**	0.22	0	1
X_{12}	0.13	0.27	0	0
X_{13}	0.19	0.04	0	0
X_{14}	−0.10	0.03	0	0
X_{15}	−0.06	0.01	0	0
X_{16}	0.02	0.01	0	0

Numbers in bold have larger values than 0.3

Table 10 Model fitting of the two methods

Method	RSS	R^2	d	Variables
COSSO	3.74	0.89	8	$\{X_2, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{14}\}$
LLSS	5.35	0.85	5	$\{X_1, X_2, X_6, X_7, X_{11}\}$
LLSS (order 2)	3.81	0.89	8	$\{X_2, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{14}\}$
Final Model1	4.54	0.87	3	$\{X_2, X_7, X_{11}\}$
Final Model2	5.19	0.85	3	$\{X_1, X_7, X_{11}\}$
Final Model3	5.92	0.83	3	$\{X_6, X_7, X_{11}\}$

In Table 10, all three methods have larger R^2 values than those in Table 9. This suggests that additive models would be appropriate, while a purely nonparametric model suffers from the typical estimation problem although it is more general. We can also see that COSSO has a slightly larger R^2 and smaller RSS than LLSS and LLSS with order 2.

However, among these predictors, we find that X_1, X_2, X_6 are positively correlated, with large correlation coefficients of around 0.8, and thus, any one could be the representative of the group $X_1, \ldots, X_6$. Similarly, among $X_8, \ldots, X_{13}$, we may only need a single representative. Further, X_{14}, X_{15}, X_{16} make very little contribution to salary and, more importantly, their effect on salary is very difficult to explain. Thus, taking this information into consideration, it would be interesting to see whether we can use a more parsimonious model. Thus, we considered model fitting with one of X_1, X_2, X_6;

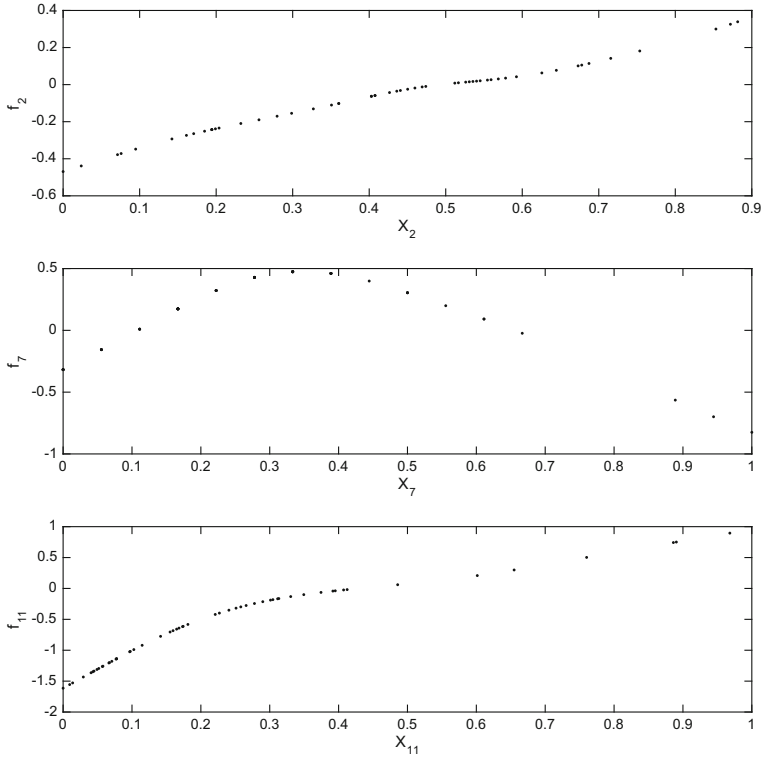

Fig. 1 Estimated components by LLSS

X_7; and X_{11}. The final model with X_2, X_7, X_{11} has the largest $R^2 = 0.87$ and the smallest RSS $= 4.54$ among the three models. Thus, this model is recommended, although the values of the two criteria are still not as large/small as those of the model obtained by COSSO and LLSS with order 2.

Figure 1 shows scatter plots of Y against X_2, X_7, X_{11}. It is clear that these predictors have positive effects for salary Y. That means that salary increases with better performance in year 1986 ($X_1, \ldots, X_6$) and overall performance to year 1986 ($X_8, \ldots, X_{13}$). Years in major leagues (X_7) can be viewed as a substitution for age of a player. Salary increases with experience at first, but decreases after a peak. This bell-shaped curve explains the salary change with age and coincides with common sense.

Figure 2 shows that the models given by COSSO and LLSS, combined with an interpretable manual selection, both fit the data well.

6 Discussion

In this paper, we have studied variable selection and estimation methods for purely nonparametric models that are applied to additive models. We proposed a least-squares-based variable selection method without any nonparametric approximation

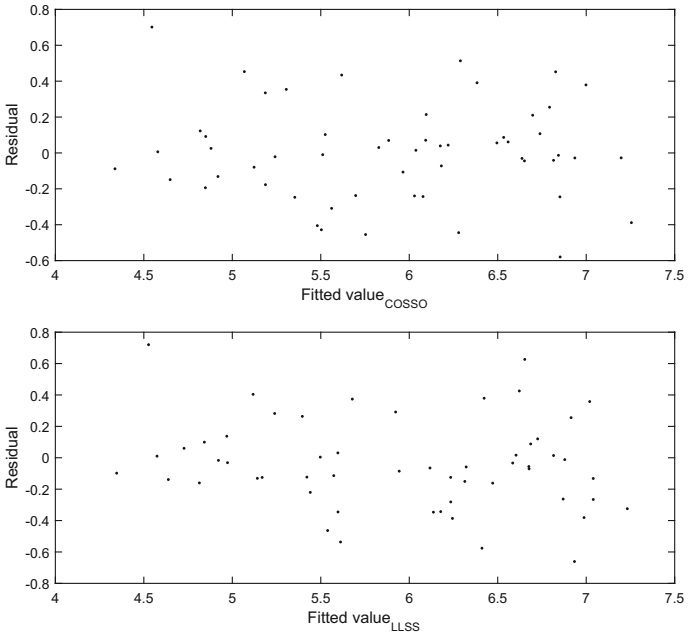

Fig. 2 The residual plot

for nonparametric regression functions. This method has the advantages of easy implementation and computational efficiency, with fewer problems caused by dimensionality. The cost is that the proposed method places some regularity conditions that may be regarded as restrictive conditions. A sufficient and special case of Condition (2.3) is the elliptical symmetric distribution assumption for predictor $\mathbf{X}$. This condition is often used in the sufficient dimension reduction and is approximately satisfied in practice. Recently, Guan et al. (2017) weakened these conditions and gave some systematic studies on the necessary and sufficient conditions. Further, another trade-off to use such a very simple method for such a general model is that in some cases, $E(\mathbf{X}h(Y))$ could be zero and then the identification of the direction in the central subspace becomes a problem. Because $E(\mathbf{X}h(Y))$ is the key to well identify the variables we want to select, so we propose an ad hoc method that can somehow improve the performance. The relevant theoretical investigations are ongoing.

Acknowledgements The authors thank Dr. Yang Feng in Columbia University for providing their NIS code. Dr. Lixing Zhu's work was supported by a Grant from the Research Grants Council of Hong Kong and a Faculty Research Grant (FRG) Grant from Hong Kong Baptist University and a Grant (NSFC11671042) from the National natural Science Foundation of China. Dr. Zhenghui Feng's work was supported by the Natural Science Foundation of Fujian Province of China, Grant No. 2017J01006, and German Research Foundation (DFG) via the International Research Training Group 1792 "High Dimensional Nonstationary Time Series," Humboldt-University zu Berlin. The authors thank the editor, associate editors, and referees for their constructive suggestions and comments that led to a significant improvement in an early manuscript.

Appendix

Proof of Theorem 1 Recall the definition of η, $\mathbf{Z}$, and $\mathbf{A}_d{}^T\mathbf{X} = \eta^T\mathbf{Z}$. A_1 is a p-dimensional vector whose first d elements are 1, otherwise 0. We have

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}h(Y)) &= \boldsymbol{\Sigma}^{-1/2}(\mathbf{B}_1, \eta_1/\|\eta_1\|)(\mathbf{B}_1, \eta_1/\|\eta_1\|)^\top \mathrm{E}\left(\mathbf{Z}h(Y)\right)\\ &= \boldsymbol{\Sigma}^{-1/2}\mathbf{B}_1\mathbf{B}_1^\top \mathrm{E}\left(\mathbf{Z}h(Y)\right) + \boldsymbol{\Sigma}^{-1/2}\eta_1\eta_1^\top \mathrm{E}\left(\mathbf{Z}h(Y)\right)/\|\eta_1\|^2\\ &= \boldsymbol{\Sigma}^{-1/2}\mathbf{B}_1\mathbf{B}_1^\top \mathrm{E}\left(\mathbf{Z}h(Y)\right) + A_1A_1^\top \mathrm{E}\left(\mathbf{X}h(Y)\right)/\|\eta_1\|^2\\ &=: \boldsymbol{\Sigma}^{-1/2}\mathbf{B}_1\mathrm{E}\left(\mathrm{E}(\mathbf{B}_1^\top\mathbf{Z}|Y)h(Y)\right) + c_hA_1. \end{aligned} \tag{24}$$

It is obvious that the first term is equal to zero when the condition $\mathrm{E}(\mathbf{B}_1^\top\mathbf{Z}|Y) = 0$ almost surely. Thus, (3) implies (4). On the other hand, when (4) holds, the first term in (24) is zero. For any transformation $h(\cdot)$, $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_1\mathrm{E}\big(\mathrm{E}(\mathbf{B}_1^\top\mathbf{Z}|Y)h(Y)\big) = 0$, implies $\mathrm{E}(\mathbf{B}_1^\top\mathbf{Z}|Y) = 0$ almost surely. (4) implies (3). When the distribution of $\mathbf{Z}$ is elliptically symmetric, the Eq. (4) can be proved similarly in Li and Duan (1989). □

Proof of Theorem 3 Here, we give the sketch of proof of Theorem 3, which is Theorem 1 in Lin et al. (2009). For details please refer to Lin et al. (2009).

To proof Theorem 3, two lemmas are needed.

Lemma 1 *Suppose the conditions of Theorem 3 hold and denote*

$$\hat{f}_{1M}(x_1) = f_{1M}(x_1)\frac{\sum_{i=1}^n\{Y_i - \mu - \sum_{j=2}^d f_{jM}(x_{ij})\}\int_{s_{i-1}}^{s_i} K\left(\frac{t-x_1}{h}\right) f_{1M}(t)\mathrm{d}t}{\int_0^1 K\left(\frac{t-x_1}{h}\right)(f_{1M}(t))^2\mathrm{d}t},$$

where $f_{1M}(x_j)$ are defined in Sect. 3.1. $s_i, i = 0, \ldots, n$ are defined as $s_0 = 0$, $s_i = (x_{i1} + x_{(i+1)1})/2, i = 1, \ldots, n-1$, $s_n = 1$, $0 \le x_{11} < x_{21} < \cdots < x_{n1} \le 1$ ordered. Then, as $h \to 0$ and $n \to \infty$, the bias and variance of $\hat{f}_{1M}(x_1)$ can be expressed, respectively, as

$$\begin{aligned} bias(\hat{f}_{1M}(x_1)) &= \frac{1}{2}h^2\sigma_K^2 M^{-\gamma_{12}}e_{21}(x_1) + O(n^{-1}) + o(h^2M^{-\gamma_{12}}) + O(M^{-\gamma_0}),\\ var(\hat{f}_{1M}(x_1)) &= \frac{\sigma^2 J_K}{nhp_1(x_1)} + O(n^{-1}) + O(n^{-2}h^{-2}), \end{aligned}$$

where $\gamma_0 = \min\{\gamma_{j0};\ j = 1, 2, \ldots, d\}$, $e_{21}(x_1)$ is defined in condition C2, and satisfying $\lim_{M\to\infty} M^{\gamma_{j2}}r''_{jM}(x_j) = e_{j2}(x_j)$, $j = 1, \ldots, d$.

Lemma 2 *Let the conditions of Theorem 3 hold and let*

$$\begin{aligned}\check{f}_{1M}(x_1) = f_{1M}(x_1)\\ + \frac{\sum_{i=1}^n\left\{Y_i - \mu - \sum_{j=2}^d f_{jM}(x_{ij})\right\}\int_{s_{i-1}}^{s_i} K\left(\frac{t-x_1}{h}\right)\mathrm{d}t - \int_0^1 K\left(\frac{t-x_1}{h}\right) f_{1M}(t)\mathrm{d}t}{\int_0^1 K\left(\frac{t-x_1}{h}\right)}.\end{aligned}$$

Then as $h \to 0$ and $n \to \infty$, the bias and variance of $\check{f}_{1M}(x_1)$ can be expressed, respectively, as

$$bias(\check{f}_{1M}(x_1)) = \frac{1}{2}h^2\sigma_K^2 M^{-\gamma_{12}} e_{21}(x_1) + O(n^{-1}) + o(h^2 M^{-\gamma_{12}}) + O(M^{-\gamma_0}),$$
$$var(\check{f}_{1M}(x_1)) = \frac{\sigma^2 J_K}{nhp_1(x_1)} + O(n^{-1}) + O(n^{-2}h^{-2}),$$

where $\gamma_0 = \min\{\gamma_{j0};\ j = 1, 2, \ldots, d\}$, $e_{21}(x_1)$ is defined in condition C2, and satisfying $\lim_{M\to\infty} M^{\gamma_{j2}} r''_{jM}(x_j) = e_{j2}(x_j)$, $j = 1, \ldots, d$.

To proof Theorem 3, in Sect. 3.2, we defined that $r_{jM}(x_j) = f_j(x_j) - f_{jM}(x_j)$; here, we define that

$$R_M(x) = \sum_{j=1}^{d} f_j(x_j) - \sum_{j=1}^{d}\sum_{l=1}^{M} \beta_{jl}^0 q_l(x_j).$$

Then, the first stage estimators of $f_j(x_j)$ can be expressed as

$$\tilde{f}_j(x_j) = f_{jM}(x_j) + \frac{1}{n}\sum_{i=1}^{n} PA(x_i)(\varepsilon_i + R_M(x_i)),\ j = 1, \ldots, d,$$

where $PA(x_i)$ is the summation components in equation (A1) in Lin et al. (2009). And using Taylor expansion, $\hat{f}_1(x_1) - \hat{f}_{1M}(x_1)$ can be expanded at $f_{1M}(x_1)$; then, we can get

$$\hat{f}_1(x_1) - \hat{f}_{1M}(x_1) = B_1(x_1)(\tilde{f}_1(x_1) - f_{1M}(x_1)) + B_2(x_1)\left\{\tilde{\mu} - \mu_0 + \sum_{j=2}^{d}(\tilde{f}_j(x_j) - f_{jM}(x_j))\right\} + o_p(hn^{-1}M),$$

where $B_1(x_1) = \eta_1/\eta_2 + f_1(x_1)\eta_3/\eta 2 - 2f_1(x_1)\eta_1/\eta_4^2$, and $B_2(x_1) = -f_1(x_1)\eta_5/\eta_2$ with $\eta_1 = \sum_{i=1}^{n}\{Y_i - \mu^0 - \sum_{j=2}^{d} f_{jM}(x_{ij})\}\int_{s_{i-1}}^{s_i} K(\frac{t-x_1}{h}) f_{1M}(t)\mathrm{d}t$, $\eta_2 = \int_0^1 K(\frac{t-x_1}{h}) f_{1M}^2(t)\mathrm{d}t$, $\eta_3 = \sum_{i=1}^{n}\{Y_i - \mu^0 - \sum_{j=2}^{d} f_{jM}(x_{ij})\}\int_{s_{i-1}}^{s_i} K(\frac{t-x_1}{h})\mathrm{d}t$, $\eta_4 = \int_0^1 K(\frac{t-x_1}{h}) f_{1M}(t)\mathrm{d}t$, $\eta_5 = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K(\frac{t-x_1}{h}) f_{1M}(t)\mathrm{d}t$.

From the results above, we have

$$E\{\eta_1(\tilde{f}_1(x_1) - f_{1M}(x_1))\} = O(hM^{-\gamma_0+1}) + O(hn^{-1}M).$$

And similarly $E\{\eta_3(\tilde{f}_1(x_1) - f_{1M}(x_1))\} = O(hM^{-\gamma_0+1}) + O(hn^{-1}M)$. So,

$$E\{B_1(x_1)(\tilde{f}_1(x_1) - f_{1M}(x_1))\} = O(M^{-\gamma_0+1}) + O(n^{-1}M).$$

And $E\{B_2(x_1)[\tilde{\mu} - \mu + \sum_{j=2}^{d}(\tilde{f}_j(x_j) - f_{jM}(x_j))]\} = O(M^{-\gamma_0+1}) + O(n^{-1}M)$. Combining these results with Lemma 1 and conditions C1 and C2 leads to the final expression of bias$(\hat{f}_1(x_1))$ and var$(\hat{f}_1(x_1))$ in Theorem 3.

The second part of the proof is similar, and so we omitted it here. □

References

Candés, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, *35*, 2313–2404.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y., Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, *4*, 143–167.

Cook, R. D. (1998). *Regression graphics*: *Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D., Weisberg, S. (1991). Discussion of 'Sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association*, *86*, 28–33.

Cui, X., Peng, H., Wen, S. Q., Zhu, L. X. (2013). Component selection in the additive regression model. *Scandinavian Journal of Statistics*, *40*(3), 491–510.

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

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.

Guan, Y., Xie, C., Zhu, L. (2017). Sufficient dimension reduction with mixture multivariate skew elliptical distributions. *Statistica Sinica*, *27*(1), 335–355.

Hall, P., Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, *21*, 867–889.

Härdle, W. (1990). *Applied nonparametric regression, econometric society monograph series, 19*. Cambridge: Cambridge University Press.

Härdle, W., Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, *13*, 1465–1481.

Hastie, T., Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–318.

Li, B., Wang, S. (2007). On directional regression for dimesnion reduction. *Journal of the American Statistical Association*, *102*, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–327.

Li, K.-C., Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, *17*(3), 1009–1052.

Li, K.-C., Lue, H. H., Chen, C. H. (2000). Interactive tree-truncated regression via principal Hessian directions. *Journal of the American Statistical Association*, *95*, 547–560.

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

Lin, L., Cui, X., Zhu, L. X. (2009). An adaptive two-stage estimation method for additive models. *Scandinavian Journal of Statistics*, *36*, 248–269.

Lin, L., Sun, J., Zhu, L. X. (2013). Nonparametric feature screening. *Computational Statistics and Data Analysis*, *36*, 162–174.

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

Meier, L., Van der Geer, S., Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, *37*, 3779–3821.

Storlie, C. B., Bonedll, H. D., Reich, B. J., Zhang, H. H. (2011). Surface estimation, variance selection, and the nonparametric oracle property. *Statistica Sinica*, *21*, 679–705.

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

Wahba, G. (1990). *Spline models for observational data*, vol. 59. SIAM. CBMSNSF Regional Conference Series in Applied Mathematics.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variable. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.
Zhao, P., Yu, B. (2006). On model selection consisitency of Lasso. *Journal of Machine learning Research*, *7*, 2541–2563.
Zhu, L. P., Wang, T., Zhu, L. X., Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, *97*, 295–304.
Zhu, L. P., Li, L. X., Li, R., Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*, 1464–1474.
Zhu, L. X., Miao, B. Q., Peng, H. (2006). On sliced inverse regression with high dimensional covariates. *Journal of the American Statistical Association*, *101*, 630–643.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Affiliations

Zhenghui Feng[1] · Lu Lin[2,3] · Ruoqing Zhu[4] · Lixing Zhu[5,6]

Zhenghui Feng
zhfengwise@xmu.edu.cn

Lu Lin
linlu@sdu.edu.cn

Ruoqing Zhu
rqzhu@illinois.edu

1 School of Economics, and the Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, Fujian, China

2 Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, Shandong, China

3 School of Statistics, Qufu Normal University, Qufu 273165, Shandong, China

4 Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

5 Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, China

6 School of Statistics, Beijing Normal University, Beijing 100875, China