IRTG 1792 Discussion Paper 2018-002

Nonparametric Variable Selection and Its Application to Additive Models

Zheng-Hui Feng *
Lu Lin *[2]
Ruo-Qing Zhu *[3]
Li-Xing Zhu *[4]

* Xiamen University, China
*[2] Shandong University, China
*[3] Yale University, United States of America
*[4] Hong Kong Baptist University, China

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Nonparametric Variable Selection and Its Application to Additive Models

ZHENG-HUI FENG, LU LIN, RUO-QING ZHU and LI-XING ZHU

Abstract

For multivariate nonparametric regression models, existing variable selection methods with penalization require high-dimensional nonparametric approximations in objective functions. When the dimension is high, none of methods with penalization in the literature are readily available. Also, ranking and screening approaches cannot have selection consistency when iterative algorithms cannot be used due to inefficient nonparametric approximation. In this paper, a novel and easily implemented approach is proposed to make existing methods feasible for selection with no need of nonparametric approximation. Selection consistency can be achieved. As an application to additive regression models, we then suggest a two-stage procedure that separates selection and estimation steps. An adaptive estimation to the smoothness of underlying components can be constructed such that the consistency can be even at parametric rate if the underlying model is really parametric. Simulations are carried out to examine the performance of our method, and a real data example is analyzed for illustration.

[2] *Address for correspondence: lzhu@hkbu.edu.hk.* Zhenghui Feng is an associate professor, School of Economics & Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China. Lu Lin is a professor, School of Mathematical Sciences, Shandong University, China. Ruoqing Zhu is a postdoctoral fellow, Department of Biostatistics, Yale University. Lixing Zhu is a chair professor, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China. (Email: lzhu@hkbu.edu.hk). He 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.

KEY WORDS: Adaptive estimation; non-parametric additive model; purely non-parametric regression; variable selection.

1 INTRODUCTION

It is well known that for multivariate nonparametric regression models with many predictors, even with moderate number of predictors, estimation could be very inefficient, see Härdle (1990). Therefore, when the model is sparse, it is necessary to select active predictors into and rule out inactive ones from a parsimonious working model such that further statistical analysis can be performed efficiently. For parametric models, the most promising methodology in the literature is with use of various penalized objective functions for simultaneous selection and estimation. Among them, the LASSO (Tibshirani 1996), the SCAD (Fan and Li 2001), and the Dantzig selector (Candés and Tao 2007) are the proven powerful methods. Several efforts have been devoted to apply or extend these methods to handle multivariate nonparametric models. However, when the number of predictors is large, usually these methods may not work efficiently because of an important feature of these methods, that is, these methods work on variable selection and estimation simultaneously. Therefore, using these methods, "residuals" in objective functions have to involve approximations to underlying nonparametric regression functions. As is known, any approximation is a parametrization for nonparametric function and thus its approximation accuracy merely depends on the extent of data denseness in the space and the smoothness of regression functions. For instance, Lin and Zhang (2006) investigated variable selection for nonparametric regression, Storlie et al.(2011)

refined the algorithm proposed in Lin and Zhang (2006), and used smoothing splines to approximate nonparametric regression function. However, in high-dimensional space with a sample of moderate size, meaningful nonparametric approximation is often not possible. This means that inaccurate fitting to the true nonparametric regression function would seriously affect the accuracy of further variable selection and estimation. Thus, both Lin and Zhang (2006) and Storlie et al.(2011) in effect mainly 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 high-dimensionality to a relative low-dimensionality. There are several nonparametric sure screening approaches available in the literature, which are based on different correlations between response and every predictor, see Zhu et al.(2011), Li, Zhong and Zhu (2012), and Lin et al. (2013). Nevertheless, sure screening cannot ensure selection consistency, and selected models would still contain many inactive predictors. So iterative algorithms are often necessary via combining existing penalty-based selection methods. However, for purely multivariate nonparametric models, iterative algorithms cannot be efficiently implemented because of the same nonparametric approximation difficulty as discussed above even when screening can reduce the number of predictors down to a number much less than the sample size. Although the robust rank correlation screening approach developed by Li et al. (2012) can implement iteration to select predictors, it is still a question for it to apply to purely multivariate nonparametric models. The motivation of this paper is to propose an efficient variable selection and estimation for the purely multivariate nonparametric model, and apply it to the additive model.

As an application, we will consider the nonparametric additive model (Hastie and Tibshirani 1986). In the literature, all available methods have a common feature: using a nonparametric smoothing approach to locally linearize the components to define

their estimates and then using an objective function with penalty, such as the group LASSO, to select groups of variables as the corresponding estimates of selected components. The examples of references include the following. Lin and Zhang (2006) proposed the component selection and smoothing operator(COSSO) method when p is fixed. p is the dimension of predictor $\mathbf{X} = (X_1, \cdots, X_p)^T$. It is an extension of the group LASSO (see Yuan and Lin 2006) and is applicable for the cases where p is smaller than n (n is the sample size). Meier et al. (2009) investigated variable selection in the additive model with $p \gg n$ with a "sparsity-smoothness penalty", then again it is a group LASSO after parametrization. Huang et al. (2010) and Peng et al. (2013) similarly used the above idea of nonparametric approximation and group variable selection. That is, in the above methods, the components are approximated by groups of variables, and are selected through an *all-in-all-out* fashion, the original p-dimensional space is enlarged to be $\tilde{p} := \sum_{j=1}^{p} k_j$-dimensional space when the corresponding approximation of each function has k_j unknown parameters. To guarantee the consistency of estimates, k_j are necessary to go to infinity as the sample size n goes to infinity. This is a must in nonparametric estimation, see Härdle (1990). In other words, it increases the difficulty to handle the large p scenarios. Ravikumar et al. (2009) proposed the sparse additive models(SpAM) using the same penalty as the COSSO. Their backfitting algorithm for the SpAM allows use of arbitrary nonparametric smoothing techniques, but does not give the convergence rate of the nonparametric estimates. Applying the method we propose in this paper to the additive model, the algorithm without any nonparametric approximation is thus very different from all the above. After components are selected, an adaptive estimation procedure is recommended. When the underlying model is really parametric, the convergence rate of the estimate can achieve parametric rate. The details are in Section 3.

On the other hand, this very simple approach reasonably has its limitations. The

main cost is at fairly strong conditions on the distribution of the predictors and the shape of the regression function. These can be seen in Theorem 1 in Section 2. In the simulations, we can see that the method has difficulty to select the active predictors who are in symmetric component functions. Therefore, an ad-hoc approach is suggested to deal with this issue when the designed conditions are violated. Theoretically, how to weaken those conditions while to maintain the implementation simplicity and selection efficiency of the method is an interesting but challenging topic, deserving a further study.

The paper is organized as follows. The selection procedure is described in Section 2. The application to additive models is described in Section 3. Simulations are carried out in Section 4. A real data analysis is presented in Section 5. A brief proof of Theorem 1 is postponed in the Appendix.

2 SELECTION PROCEDURE

For the response Y and the column predictor vector $\mathbf{X} = (X_1, \cdots, X_p)^T$, assume that

$$Y \perp\!\!\!\perp \mathbf{X}_{A^c} | \mathbf{X}_A, \tag{2.1}$$

where $\mathbf{X}_A = \{X_i : i \in A\}$ is the set of the relevant X_i's such that A is the index set. $\mathbf{X}_{A^c}$ is the compliment of $\mathbf{X}_A$ in $\mathbf{X}$. 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 regression function, say, $G(\mathbf{X}) = E(Y|\mathbf{X})$. 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 $A = \{1, \cdots, d\}$ and $\mathbf{X}_A = \{X_1, \cdots, X_d\}$. From the above description, the strategy that performs simultaneous selection and estimation is not a necessary way. To make selection without any nonparametric approximation

realistic, we describe the following linear least squares sparse solution.

2.1 Linear Least Squares Sparse Solution

Let I_i be the p-dimensional column vector whose ith element is 1 and all others are zero. For any index $l_i \in A$, we have a vector I_{l_i} to indicate it. Denote a $p \times d$ matrix by $\mathbf{A}_d = (I_{l_1}, \cdots, I_{l_i}, \cdots, I_{l_d})$. Then the conditional independence in (2.1) can be rewritten as $Y \perp\!\!\!\perp \mathbf{x} | \mathbf{A}_d^T \mathbf{x}$. Note that $\mathbf{A}_d$ is not unique because for any $d \times d$ orthogonal matrix $\mathbf{C}$, $\mathbf{A}_d\mathbf{C}$ can also make the conditional independence hold. For this general estimation problem, sufficient dimension reduction (Li 1991, Cook 1998) is often employed to deal with. What sufficient dimension reduction approaches can estimate is the column subspace of $\mathbf{A}_d$ with minimum dimension, which is denoted by $\mathcal{S}_{\mathcal{Y}|\mathbf{x}}$. The space $\mathcal{S}_{\mathcal{Y}|\mathcal{X}}$ is called the central subspace (CS, Cook 1998). The dimension d of $\mathcal{S}_{\mathcal{Y}|\mathcal{X}}$ is called the structural dimension. There are a number of methods available in the literature to identify and estimate the central subspace $\mathcal{S}_{\mathcal{Y}|\mathcal{X}}$. For instance, sliced inverse regression (SIR, Li 1991), sliced average variance estimation (SAVE, Cook and Weisberg 1991), directional regression (DR, Li and Wang 2007) and discretization-expectation estimation (DEE, Zhu et al. 2010).

On the other hand, the problem under study is more specific. We are not interested in the central subspace $\mathcal{S}_{\mathcal{Y}|\mathcal{X}}$, while the indices I_{l_i} themselves. Any aforementioned sufficient dimension reduction technique cannot do this directly. Therefore, we cannot use the column vectors in the central subspace $\mathcal{S}_{\mathcal{Y}|\mathbf{x}}$ to identify $I_{l_i}^T \mathbf{x} = x_{l_i}$. To attack this difficulty, we suggest the following method.

Without loss of generality, assume $\mu = 0$. The predictor vector $\mathbf{x} = (x_1, ..., x_p)^T$ is centered, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ is the $n \times p$ design matrix. Σ is the covariance matrix of $\mathbf{X}$.

Write $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ as the first d and last $p-d$ columns of $\mathbf{X}$ respectively, and then we can express the sample covariance matrix $C = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ in a block-wise form

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where $C_{ij} = \frac{1}{n}\mathbf{X}_{(i)}^T\mathbf{X}_{(j)}, i, j = 1, 2$. Denote $\mathbf{x} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)})^T$, where $\mathbf{x}_{(1)} = (x_1, \cdots, x_d)^T$ and $\mathbf{x}_{(2)} = (x_{(d+1),\cdots,x_{(p)}})^T$. Then $\mathbf{X}_{(1)} = (\mathbf{x}_{(1)1}, \cdots, \mathbf{x}_{(1)n})^T$, $\mathbf{X}_{(2)} = (\mathbf{x}_{(2)1}, \cdots, \mathbf{x}_{(2)n})^T$. From the above description about sufficient dimension reduction, $\mathbf{x}_{(1)} = (x_1, \cdots, x_d)^T$ are relevant to Y. We now work on identifying $\mathbf{x}_{(1)}$.

By the above notations, $A_1 = \sum_{i=1}^{d} I_i$ is a $p \times 1$ vector whose first d components are 1, otherwise 0. This is a very useful index for us to identify the active predictors and then the corresponding components. In other words, to select the active predictors, it is enough for us to identify the vector A_1. To this end, we let $\mathbf{Z} = \Sigma^{-1/2}\mathbf{X}$, and $\eta = \Sigma^{1/2}\mathbf{A}_d$. It is easy to see that η consists of the columns of $\Sigma^{1/2}$ corresponding to $\mathbf{A}_d$ and $\mathbf{A}_d^T\mathbf{X} = \eta^T\mathbf{Z}$. Further, define $\eta_1 = \Sigma^{1/2}A_1$, $\mathbf{B}_1$ as a $p \times (p-1)$ matrix orthogonal to $\eta_1/\|\eta_1\|$ and $\mathbf{B} = (\mathbf{B}_1, \eta_1/\|\eta_1\|)$ an orthogonal matrix, where $\|\eta_1\|$ is $A_1^T\Sigma A_1$. The following theorem provides a sparse solution of $\mathbf{x}$ in the least squares formulation.

Theorem 1 (Sparse Solution) *Assume that Σ_x is positive definite. Then, almost surely*

$$E(\mathbf{B}_1^T\mathbf{Z}|Y) = 0 \tag{2.2}$$

is necessary and sufficient for any function $h(\cdot)$ on the response Y, there exists some constant c_h such that

$$\Sigma_x^{-1}\mathrm{Cov}\big(\mathbf{X}, h(Y)\big) = c_h A_1 =: \gamma_h \tag{2.3}$$

provided that it is finite where c_h *depends on* h *and* $c_h = A_1^T \mathrm{E}\big(\mathbf{X}h(Y)\big)/\|\eta_1\|^2 = E(\sum_{i=1}^d x_i h(Y))/\|\eta_1\|^2$. *A sufficient condition for the above linear least squares formulation to hold is that the distribution of* $\mathbf{X}$ *is elliptically symmetric.*

Note that γ_h is proportional to A_1 which takes value 1 in the locations of x_i's. As such, this result provides us a very simple, but efficient way to identify the active predictors x_i through those nonzero elements of γ_h. This is a sparse least squares solution. From the sufficient condition (2.3) in Theorem 1, when $c_h \neq 0$, we can simply use the identity function as h to establish the following model:

$$Y = c + \gamma^T \mathbf{x} + e, \tag{2.4}$$

where $E(e\mathbf{x}) = 0$. Therefore, we transfer variable selection of the nonparametric regression model (2.1) to variable selection of the linear model (2.4). By selecting the "active elements" of γ, we can identify the corresponding active predictors x_i's. Thus, our method is rather simple and efficient, but very different from all existing methods which usually select active predictors and estimate the corresponding regression function simultaneously. It is obvious that this sparse solution of γ in model (2.4) makes any successful variable selection approach for linear models feasible, for example, the classical LASSO.

For given data points $\{\mathbf{x}_i, y_i\}_{i=1}^n$, the LASSO estimate is defined as

$$\hat{\gamma}(\lambda) = \arg\min_{\gamma} \{\sum_{i=1}^{n} (y_i - \gamma^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\gamma_j|\}, \tag{2.5}$$

where $\lambda \geq 0$ controls the amount of regularization applied to the estimate. $\lambda = 0$ changes the LASSO to the ordinary least squares. Because the selection is exactly the same as that for linear models, the selection consistency can hold. Therefore, we will not give the 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 we have*

$$\lim_{n\to\infty} P(sgn(\hat{\gamma}) = sgn(\gamma)) = 1, \tag{2.6}$$

where $sgn(A)$ *is the sign function componentwise. Let* $\hat{d} = \#\{k : \hat{\gamma}_k \neq 0\}$, d *is the true number of nonzero components in model* (3.1). *Then*

$$\lim_{n\to\infty} P(\hat{d} = d) = 1. \tag{2.7}$$

Remark 1 *These two theorems provide that, the LASSO can select the true indices* $\{I_{l_1}, \cdots, I_{l_d}\}$ *of the active predictors* $\mathbf{X}_A$ *in model* (2.1) *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). The details are skipped here.*

3 Application to the additive model (3.1)

3.1 Estimation

Suppose that a sample $\{(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)\}$ is available. The model takes the form as

$$y_i = \mu + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, ..., n, \tag{3.1}$$

where μ is an intercept term, x_{ij} is the j-th component of $\mathbf{x}_i$, $f_j(x_{\cdot j})$ 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. Furthermore, the function $f_j(x_{\cdot j})$ is normalized so that $\int_0^1 f_j(u)du = 0$ to make model identification possible. Assume there are d nonzero components in model (3.1) with $d \ll p$.

The result that $\hat{d}$ is consistent to d with a probability going to 1 is applicable to model (3.1) as well. To efficiently estimate the $\hat{d}$ nonzero components selected, we suggest the following adaptive method. The estimation could be adaptive to the smoothness of underlying function such that the convergence rate could be faster than the usual optimal nonparametric rate when the function is smooth enough. The basic idea is to adjust initial estimates to adapt the smoothness. The resulting estimates are of optimal non-parametric rate generally, and when the model is actually parametric, it can achieve the parametric convergence rate $O(n^{-1})$ in mean squared error (MSE). In the following, we describe it briefly by assuming the true nonzero number d is given.

Define an initial estimate first. 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 of 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's. More about the RKHS and their reproducing kernels are given in Wahba(1990). The second order Sobolev Hilbert space S_2 is the 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

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

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

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

According to the above approximation, denote $\{q_l(t)\}$ as the group of the $\{\sin, \cos\}$ orthogonal basis $\{\sqrt{2}\cos 2\pi t, \sqrt{2}\sin 2\pi t, ..., \sqrt{2}\cos \pi M t\}$ with coefficients a_ν, b_ν being denoted as β. Using this orthogonal decomposition, the initial estimates for μ and $f_j(x_j)$, denoted respectively, by $\tilde{\mu}$ an $\tilde{f}_j(x_j) = \sum_{l=1}^{M} \tilde{\beta}_{jl} q_l(x_j)$, can be obtained by minimizing,

over μ, β_{jl} and $\{x_{j_1}, \cdots, x_{j_d}\}$:

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

Here $q_l(u)$ are the basis functions taken as stated above and also satisfying

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

and M depends on n and tends to infinity as n tends to infinity. The initial estimation used as plug-in in the following step is obtained by the least squares. Especially, the estimation can be solved by (5) in Lin and Zhang (2006) with $\lambda = 0$.

Now we are in the position to adjust each initial component estimate, say $\tilde{f}_1(x_1)$, 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 respectively adjustment factor and adjustment shift which will be specified later. To determine $\xi(x_1)$ and $\zeta(x_1)$, we use the following steps. Motivated by Lin et al. (2009), in this paper a local L_2-fitting criterion is defined as

$$r_1(t_1, \xi) = \frac{1}{h}E\left(K\left(\frac{x_1 - t_1}{h}\right)\left[f_1(x_1) - \tilde{f}_1(x_1)\xi\right]^2\right), \tag{3.2}$$

where $K(\cdot)$ is a kernel function satisfying some 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 $\zeta(t_1)$:

$$r_2(t_1, \zeta) = \frac{1}{h}E\left(K\left(\frac{x_1 - t_1}{h}\right)\left[f_1(x_1) - (\tilde{f}_1(x_1) + \zeta)\right]^2\right). \tag{3.3}$$

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

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

$\xi(\cdot)$ and $\zeta(\cdot)$ can be estimated via, 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 $j \geq 2$:

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

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

$$\hat{f}_1(x_1) = \tilde{f}_1(x_1)\frac{\sum_{i=1}^{n}\{Y_i - \tilde{\mu} - \sum_{j=2}^{d}\tilde{f}_j(x_{ij})\}\tilde{f}_1(x_{i1})K(\frac{x_{i1}-x_1}{h})}{\sum_{i=1}^{n}\tilde{f}_1^2(x_{i1})K(\frac{x_{i1}-x_1}{h})}, \tag{3.4}$$

$$\check{f}_1(x_1) = \tilde{f}_1(x_1) + \frac{\sum_{i=1}^{n}\{Y_i - \tilde{\mu} - \sum_{j=1}^{d}\tilde{f}_j(x_{ij})\}K(\frac{x_{i1}-x_1}{h})}{\sum_{i=1}^{n}K(\frac{x_{i1}-x_1}{h})}. \tag{3.5}$$

For the other additive components $f_j(\cdot), j = 2, \cdots, d$, the construction scheme is similar; the details are omitted here.

3.2 Asymptotics

In this part, we discuss the adaptivity property for our proposals in (3.4) and (3.5). As the result is similar to that in Lin et al. (2009), we only present a brief description and explanation. Theorem 2 guarantees that $\mathbf{X}_A$ can be selected into the working model with a probability going to one and also d is consistently estimated by $\hat{d}$ with a probability tending to one as well. Without loss of generality, we assume that the first d components of model (3.1) are nonzero, and with a probability going to one, the working model can be written as

$$y_i = \mu + \sum_{j=1}^{\hat{d}} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, ..., n. \tag{3.6}$$

Assume that the regression function $f_j \in \bar{T}$, then 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 by $r''_{jM}(x_j)$ the second order derivative of $r_{jM}(x_j)$. Write $\sigma_K^2 = \int_{-1}^{1} u^2K(u)du$, $J_K = \int_{-1}^{1} K^2(u)du$ and let $p_j(x_j)$

be the density function of x_j. Assume that $K(\cdot)$ on the support $[-1,1]$ is Lipschitz continuous and $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} uK(u)du = 0$, and without loss of generality that $x_j \in [0,1]$ and $p_j(x_j) > 0$ for $x_j \in [0,1]$. To get the adaptivity given in the following theorem, we need the following condition:

C1 There exist nonzero functions $e_{jk}(x_j), k = 0,1,2, j = 1,\cdots,d$, such that

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

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

This condition requests the convergence rates of the remainder terms and their derivatives, which are also related to the smoothness of f_j. The decreasing relationship between the rates described by $\gamma_{j2} \le \gamma_{j1} \le \gamma_{j0}$ is also common. For example, if the basis functions are chosen to be trigonometric functions or polynomial functions, the remainder term has this property. The following theorem gives the details.

Theorem 3 (Adaptivity) *Assume that Condition C1 holds as $n \to \infty$. For $x_1 \in (0,1)$, the bias and variance of the second stage estimates in (3.4) and (3.5) have the following representations:*

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

The proof of the theorem is similar to that of Theorem 1 of Lin et al. (2009). We omit the detail here. The theorem shows that although the variance is the same as that of the common kernel estimation, the bias can adapt to the smoothness of the underlying function f_1. More precisely, since the value of $|r''_{1M}(x_1)|$ can describe the smoothness of f_1, the more smooth the function f_1 is, the smaller the value of $|r''_{1M}(x_1)|$ is, and consequently, the smaller bias the estimates in (3.4) and (3.5) have. Furthermore, when f_1 is smooth enough, $|r''_{1M}(x_1)| \to 0$ as $n \to \infty$ where M is dependent on n, the biases of the estimates are of the order smaller than h^2. In this case, the estimates are super-consistent in the sense that the convergence rate in mean squared error is faster than the standard order of $n^{-4/5}$. Particularly, if f_1 satisfies $h^2|r''_{1M}(x_1)| = O(n^{-1/2})$, the estimates can achieve the convergence rate n^{-1} of parametric estimation.

3.3 Bandwidth Selection

For the adaptive estimation procedure, cross-validation (CV) is applied. For the component $f_1(x_1)$, we describe the selection procedure. First assume that the parameter μ and functions $f_j(x_j), j = 2, ..., d$ are known. Then model (3.1) can be rewritten as a one-dimensional non-parametric regression:

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

Denoted by $\hat{f}_{i1}(x_1)$, the leave-one-out form is defined as

$$\mathbf{CV}(h) = n^{-1} \sum_{i=1}^{n} \left\{ \left(y_i - \mu - \sum_{j=2}^{d} f_j(x_{ij}) \right) - \hat{f}_{i1}(x_{i1}) \right\}^2 w(x_{i1}), \tag{3.7}$$

where $w(\cdot)$ is a weight function. 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) that

the choice is based on the following criterion:

$$\lim_{n\to\infty} \frac{d(\hat{m}_h, m)}{\inf_{h\in H_n} d(\hat{m} - h, m)} = 1, \tag{3.8}$$

where m is a non-parametric 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 $f_j(x_j)$, $j = 2, ..., d$, which are in fact unknown. We replace the unknown parameter μ and the function $f_j(x_j)$, $j = 2, ..., d$, respectively by the leave-one-out forms $\tilde{\mu}_i$ and $\tilde{f}_{ij}(x_j)$ of the first-stage estimates $\tilde{\mu}$ and $\tilde{f}_j(x_j)$ and define

$$\tilde{\mathrm{CV}}(h) = n^{-1} \sum_{i=1}^{n} \Big\{ \big(y_i - \tilde{\mu} - \sum_{j=2}^{d} \tilde{f}_{ij}(x_{ij})\big) - \hat{f}_{i1}(x_{i1}) \Big\}^2 w(x_{i1}), \tag{3.9}$$

and

$$\tilde{h}_c = \arg \inf_{h\in H_n} \tilde{\mathrm{CV}}(h). \tag{3.10}$$

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

3.4 The Algorithm

For our two-stage estimation, the algorithm can be summarized the following steps.

1. Use the LASSO for model (2.4) 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 the BIC or the 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. For the additive model, estimate each $f_j(\cdot), j \in \Omega$ by the adaptive method we provided above. 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.

For convenience, in the simulations below, the initial estimates are computed by the COSSO without penalty (the tuning parameter $\lambda = 0$ in the COSSO), and the bandwidth selection is based on the 5-fold CV.

4 NUMERICAL STUDIES

In this section, 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 in parentheses.

4.1 Nonparametric Regression Models

In this subsection, we examine the performance of our selection method for nonparametric models.

Example 1 Consider the following models:

$$Y = \exp\{\frac{X_1 + \cdots + X_5}{\sqrt{5}}\} + \varepsilon, \tag{4.1}$$

$$Y = \left(5[X_1^3 + X_2^3 + \frac{1}{4}(X_1 + X_2)^2 - \frac{1}{4}(X_1 - X_2)^2]\right)^{3/5} + \varepsilon. \tag{4.2}$$

$$Y = \left(5X_1 + 5X_2 + 10X_3^2 + 10X_4^2\right)^{3/5} + \varepsilon, \tag{4.3}$$

Models in this example are regarded as nonparametric ones. Model (4.1) has $d = 5$ significant predictors, model (4.2) has an interaction term with $d = 2$, because it has another expression $Y = \left(5[X_1^3 + X_2^3 + X_1X_2]\right)^{3/5} + \varepsilon$. Model (4.3) is with $d = 4$ and contains two square functions that are symmetric about 0. $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$, $i =$

$1, ..., n$, $X_{i1}, ..., X_{ip}$ are independently distributed with U(-1,1). The errors ε_i, $i = 1, ..., n$, follow the normal distribution $N(0, \sigma^2)$ with the variance σ^2 such that the signal-to-noise ratio is $3 : 1$. Sample size is $n = 200$. The LASSO is applied and the results are reported in Table 4.1.

Table 4.1: Performance of Example 1

	$p = 10$			$p = 30$			$p = 50$		
Model	MS	TP	FP	MS	TP	FP	MS	TP	FP
(4.1)	5.05(0.26)	5(0)	0.05(0.26)	5.08(0.34)	5(0)	0.08(0.34)	5.04(0.24)	5(0)	0.04(0.24)
(4.2)	2.01(0.10)	2(0)	0.01(0.10)	2(0)	2(0)	0(0)	2(0)	2(0)	0(0)
(4.3)	2(0)	2(0)	0(0)	1.99(0.10)	1.99(0.10)	0(0)	1.98(0.14)	1.98(0.14)	0(0)

From the results in Table 4.1, we can see that our method can well identify the active predictors in the first two models (4.1) and (4.2). However, it can only identify X_1 and X_2 in model (4.3), while is not able to find X_3 and X_4 in the square functions. This confirms that our method heavily relies on the asymmetry of the predictor distribution. This is the main limitation of our method.

To be a remedy, an ad-hoc approach may be considered to take care of symmetry issue. Note that our method transfers the original model to be a linear model $Y = a + b_1^\tau X + e$ and see if any component X_i significantly affects Y. From this idea, we may consider to check whether a polynomial of X_i significantly affects Y. For instance, a second order polynomial $Y = a + b_1^\tau X + b_2^\tau X^2 + e$ could be considered. In other words, for any X_i, we will examine the relationship between Y and $a + b_1^\tau X_i + b_2^\tau X_i^2$. Higher order polynomial such as of third order or fourth order could also be considered. This selection approach may be regarded as a higher order "approximation" to the underlying model between Y and X. The predictor X_i is selected if any coefficient of X_i^j, either $j = 1$ or 2, is nonzero. In Table 4.2, "Square" means $a + b_1 X + b_2 X^2$ is used. Similarly, "Cubic" and "Fourth"

mean that the third and fourth order polynomials are used.

Table 4.2: Performance of Example 1 by Modified Method

	$p = 10$			$p = 30$			$p = 50$		
				Model (4.1)					
Method	MS	TP	FP	MS	TP	FP	MS	TP	FP
Square	5.09(0.29)	5(0)	0.09(0.29)	5.04(0.20)	5(0)	0.04(0.20)	5.01(0.22)	5(0)	0.01(0.22)
Cubic	5.10(0.41)	5(0)	0.10(0.41)	5.01(0.10)	5(0)	0.01(0.10)	4.97(0.17)	4.97(0.17)	0(0)
Fourth	5.22(0.82)	5(0)	0.22(0.28)	5.01(0.10)	5(0)	0.01(0.10)	49.02(2.0)	5(0)	44.02(2.30)
				Model (4.2)					
Square	2.02(0.14)	2(0)	0.02(0.14)	2(0)	2(0)	0(0)	2(0)	2(0)	0(0)
Cubic	2.10(0.54)	2(0)	0.10(0.54)	2(0)	2(0)	0(0)	2(0)	2(0)	0(0)
Fourth	2.31(1.14)	2(0)	0.31(1.14)	3.08(2.86)	2(0)	1.08(2.86)	49.96(0.20)	2(0)	47.96(0.20)
				Model (4.3)					
Square	4.14(0.40)	4(0)	0.14(0.40)	4.40(0.96)	4(0)	0.40(0.96)	4.55(2.29)	4(0)	0.55(2.29)
Cubic	4.12(0.36)	4(0)	0.12(0.36)	4.49(2.58)	3.98(0.14)	0.51(2.57)	3.22(3.11)	2.87(1.05)	0.35(2.78)
Fourth	4.34(0.98)	4(0)	0.34(0.98)	4.11(2.46)	3.77(0.58)	0.34(2.36)	12.08(15.49)	3.32(1.09)	8.76(15.06)

Results in Table 4.2 tells the "Square" performs best. Through Models (4.1)-(4.3), it could select the true active components 100% correctly and the false positive number is very small. This method also works for model (4.3) with small FP values. The "Cubic" and "Fourth" seems to choose, either too less or too more, than the true values when p is large. Overall, the second order polynomial is worthy of recommendation.

4.2 Application to Additive Models

Here, the COSSO (Lin and Zhang 2006) (for $p < n$), and the Boosting (Bülhmann and Yu 2003) (for $p \geq n$) are taken for comparison because they have been proved to be powerful. In the following example, we design two scenarios with $p = 10$, $n = 100$, and $p = 100$, $n = 100$.

Example 2 The model is as follows:

$$Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + \sum_{i=4}^{p} f_i(X_i) + \varepsilon, \tag{4.4}$$

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 $f_i = 0, i = 4..., p$, $p = 10$. The data $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$, $i = 1, ..., n$, $X_{i1}, ..., X_{ip}$ are independent having the following two distributions:

(1) Trimmed AR(1): $W_1, ..., W_p \sim$ N(0,1) i.i.d., and $X_1 = W_1$, $X_j = \rho X_{j-1} + (1-\rho^2)^{1/2} W_j$, $j = 2, ...d$. 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, $\text{corr}(X_j, X_k) = t^2/(1+t^2), j \neq k$.

Also the errors ε_i, $i = 1, ..., n$, follow the normal distribution $N(0, \sigma^2)$, where σ^2 is chosen according to the standard deviation of signal-to-noise ratio (SNR) at around 3 : 1. In the simulations, $\rho = 0, 0.5$, and $t = 0, 1$. As estimation is also involved, we then report ISE as well to measure the estimation accuracy. Here $ISE = E[\hat{f}(\mathbf{X}_i) - f(\mathbf{X}_i)]^2$, which is estimated by 10000 testing points from the same distribution with the training points, where $f(\mathbf{X}_i) = \sum_{j=1}^{p} f_j(X_{ij})$, is the true conditional mean function. For the comparison with the COSSO and the Boosting, we use the codes of these two methods available online. For the Boosting, the default values $\nu = 0.1, df = 4$ are for the shrinkage factor ν and the degrees of freedom df.

We consider the following two cases.

Case I: Results reported in Table 4.3 are based on $p = 10, n = 100$. We compare our proposed two-stage estimation for additive models with the COSSO. Our two-stage method is abbreviated to be "LLSS".

It can be seen from Table 4.3 that, no matter the predictors have correlation or not, our LLSS method wins in this case. Both COSSO and LLSS can select all the positive

Table 4.3: Measurements for model (4.4) case I, $n = 100, p = 10$.

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)

variables, however, LLSS has smaller ISE, MS and FP. The adaptive approach works better.

Case II: In this case, the setting is the same as that in Case I, except that $n = 100, p = 100$. As the COSSO may not be able to handle the $p = n$ case, we thus make a comparison with the Boosting (Bülhmann and Yu 2003). To efficiently apply our proposed two-stage method for additive models when $p = n$ in Case II below, we use the SIS and ISIS (Fan and Lv 2008) to first reduce the dimensionality by using the codes provided Fan and Lv (2008) available online. For the fair play, we also check the performance of the Boosting when the SIS and the ISIS are used to help on reducing the dimensionality before performing the Boosting. The results are reported in Table 4.4.

From the results in Table we can have the following observations. First, although the Boosting can handle the $p \geq n$ case, the selection is not very efficient. Second, the SIS can help on reducing the dimensionality, but the models could be too parsimonious to loss some important variables as we can see smaller TP than the true number of the important variables in the table. In this situation, the ISIS can help on rescuing them. Third, although the models are in favor of the competitors, our method only has slight

Table 4.4: Measurements for model (4.4), case II, $n = p = 100$.

	Trimmed AR(1), $\rho = 0$				Trimmed AR(1), $\rho = 0.5$			
method	ISE(sd)	MS(sd)	TP(sd)	FP(sd)	ISE(sd)	MS(sd)	TP(sd)	FP(sd)
Boosting	0.228(0.090)	16.490(2.176)	3(0)	13.490(2.176)	0.383(0.148)	16.360(2.272)	3(0)	13.360(2.272)
SIS + Boosting	0.218(0.324)	7.640(1.133)	2.960(0.197)	4.680(1.205)	1.279(0.840)	8.120(1.365)	2.390(0.490)	5.730(1.728)
ISIS + Boosting	0.108(0.056)	4.30(1.573)	3(0)	1.30(1.573)	0.170(0.093)	4.270(1.254)	3(0)	1.270(1.254)
SIS + LLSS	0.233(0.345)	7.40(1.675)	2.960(0.197)	4.440(1.684)	1.290(0.909)	6.770(1.958)	2.390(0.490)	4.380(1.927)
ISIS + LLSS	0.127(0.098)	4.420(1.776)	3(0)	1.420(1.776)	0.152(0.104)	4.420(1.464)	3(0)	1.420(1.464)
	Compound Symmetry, $t = 0$				Compound Symmetry, $t = 1$			
method	ISE(sd)	MS(sd)	TP(sd)	FP(sd)	ISE(sd)	MS(sd)	TP(sd)	FP(sd)
Boosting	0.135(0.049)	14.870(2.372)	3(0)	11.870(2.372)	1.117(0.705)	14.920(2.187)	3(0)	11.920(2.187)
SIS + Boosting	0.158(0.330)	6.840(1.261)	2.960(0.197)	3.880(1.335)	2.423(0.809)	8.150(1.009)	2(0)	6.150(1.009)
ISIS + Boosting	0.085(0.037)	5.060(1.699)	3(0)	2.060(1.699)	0.932(0.80)	3.70(1)	3(0)	0.70(1)
SIS + LLSS	0.153(0.313)	6.770(1.746)	2.960(0.197)	3.810(1.756)	1.741(0.2)	3.930(1.81)	2(0)	1.930(1.81)
ISIS + LLSS	0.103(0.066)	5.590(2.261)	3(0)	2.590(2.261)	0.132(0.098)	3.430(0.967)	3(0)	0.430(0.967)

loss in selection efficiency when compared with ISIS+Boosting.

Overall, the newly proposed method tends to be not conservative, and competitive to existing powerful approaches in the literature.

5 Real Data Example

We apply our method to the Hitters' salary data which was firstly given in 1988 ASA Garphics Poster Session. Its main interest "why they make what they make" was a main

topic of this session organized by the American Statistical Association. 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, put-outs (x_{14}), assistances (x_{15}) and errors (x_{16}). Let $\mathbf{X} = (x_1, ..., x_{16})^T$. The size of the data is $n = 263$.

In a nonparametric regression structure, $p = 16$ is too large for an efficient nonparametric estimation with a size of $n = 263$ in the sample. Therefore, there are several attempts to work on estimation. Sufficient dimension reduction (Li, 1991; Cook 1998) is a promising way to handle it via selecting some representative predictors or the linear combinations of the predictors to establish the underlying model. When sliced inverse regression (SIR, Li, 1991) with a BIC type structural dimension determination (Zhu, Miao and Peng, 2006) is applied, 2 linear combinations of the 16 predictors are determined.

As there is no specific prior information about the model structure, we first fit the data nonparametrically. For this purpose, our linear least squares sparse method in Section 2 (denoted by LLSS here) is used to select predictors. For comparison, sliced inverse regression (SIR, Li 1991) is also used to select projection indices to achieve the purpose of sufficient dimension reduction. After that, nonparametric regression models are fitted with the predictors or indices selected by these two methods. The fitting is made by a matlab package named "Multivariant Kernel Regression and Smoothing". To compare the performance, we list the selected predictors and the indices, and the regression R^2 in Table 5.5.

Table 5.5: The indices selected by the SIR, the predictors selected by the LLSS and the R^2 values

Method	SIR $R^2 = 0.72$		LLSS $R^2 = 0.76$	
Index	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
x_1	**0.38**	0.0	1	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	0	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

The LLSS selected two predictors into the working model. We note that the covariance matrix shows that there are 3 groups to separate all predictors: $\{x_1, \cdots, x_6\}$, $\{x_8, \cdots, x_{13}\}$, and $\{x_{14}, x_{15}, x_{16}\}$. Within the groups, the predictors are highly positively correlated with the correlation coefficients around 0.8, whereas between the groups, they are positively, but weakly correlated with the correlation coefficients around 0.2. Thus, x_1 and x_{11} could be regarded as the representatives of the first two sets, respectively, and they are respectively the strength in 1986, and comprehensive strength of a player. The number of selected predictors coincides with the dimension of central subspace determined by the SIR. Further, both the SIR and the LLSS show that the set $\{x_{14}, x_{15}, x_{16}\}$ has very little contribution to the response Y. We note that when the SIR is applied, and looking at the coefficients with large loadings, $\hat{\gamma}_1$ seems a contrast between x_1 and x_2, while $\hat{\gamma}_2$ would be a contrast between x_8 and x_9. However, these two pairs are respectively highly positively correlated with the correlation coefficients $\rho \approx 0.8$. Thus, it

is hard to explain their meanings. In contrast, our method provides a better fitted model with larger R^2 value, which is much simpler and more interpretable.

To further explore the regression relationship between the response Y and the two predictors, x_1 and x_{11}, we draw the scatter plots in Figure 5.1. Both show monotonicity.

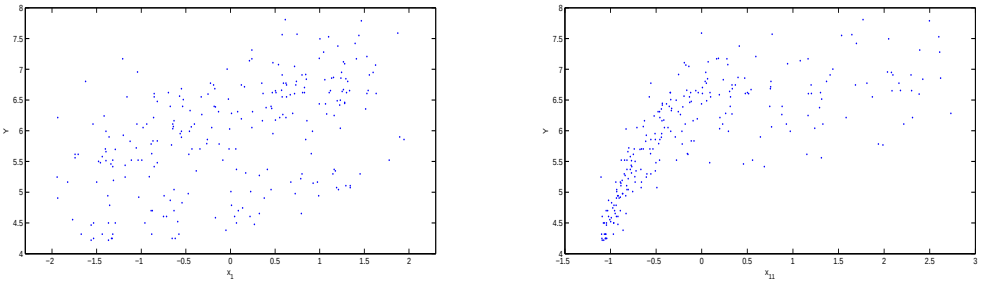


Figure 5.1: Scatter Plot

This enlightens us to further try a nonparametric additive model:

$$y = \mu + f_1(x_1) + f_{11}(x_{11}) + \varepsilon \tag{5.1}$$

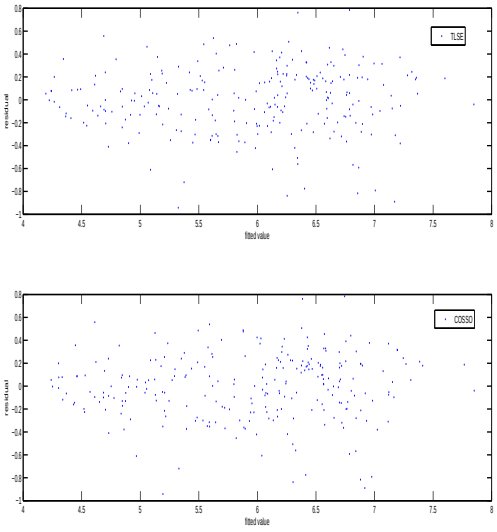
to fit this dataset. This means that x_1 and x_{11} are regarded as the two active predictors selected by the LLSS, and then the corresponding functions are estimated adaptively. In Table 5.6, the results are listed. As the COSSO is particularly designed for additive models, we also use it to select predictors to fit an additive model $y = \mu + \sum_{j=1}^{d} f_{i_j}(x_{i_j}) + \varepsilon$ with a $d \leq p$. Measurements are: RSS, the squared sum of residuals divided by the sample size; R^2, the coefficient of determination in regression; $\hat{d}$, the number of active predictors selected; Index, indices selected.

In Table 5.6, this further estimation makes a slightly larger R^2 value when the LLSS and the adaptive estimation are applied. The COSSO gets $R^2 = 0.9$ larger than that of

Table 5.6: Model fitting of the two methods

method	RSS	R^2	$\hat{d}$	Index
LLSS	0.155	0.797	2	$\{1,\ 11\}$
COSSO	0.077	0.900	8	$\{2, 6, 7, 8, 9, 11, 12, 14\}$

the LLSS. However, it is inefficient in selection with half out of all the predictors being included in the working model. Again the our two-stage estimation proposed in Section 3 owns a much clearer and interpretable result at a cost, but not much, of losing a certain regression fitting R^2. Residual plots tell that the working models by both the methods well fit the data. Thus, our two-stage estimation is worthy of recommendation.



*: TLSE stands for Adaptive method

Figure 5.2: The residual plot

Acknowledgement Financial support from the German Research Foundation (DFG) via the International Research Training Group 1792 "High Dimensional Nonstationary Time Series, Humboldt-University zu Berlin, is gratefully acknowledged.

Appendix

Proof of Theorem 1. Recall the definition of η and $\mathbf{Z}$ and ${\mathbf{A}_d}^T\mathbf{X} = \eta^T\mathbf{Z}$. Also, let $\eta_1 = A_1^T\Sigma^{1/2}$ recalling that A_1 is a p-dimensional vector whose first d elements are 1, otherwise 0. We have

$$\begin{aligned}
\Sigma_x^{-1}\mathrm{E}(\mathbf{X}h(Y)) &= \Sigma_x^{-1/2}(\mathbf{B}_1, \eta_1/\|\eta_1\|)(\mathbf{B}_1, \eta_1/\|\eta_1\|)^T\mathrm{E}\big(\mathbf{Z}h(Y)\big) \\
&= \Sigma_x^{-1/2}\mathbf{B}_1\mathbf{B}_1^T\mathrm{E}\big(\mathbf{Z}h(Y)\big) + \Sigma_x^{-1/2}\eta_1\eta_1^T\mathrm{E}\big(\mathbf{Z}h(Y)\big)/\|\eta_1\|^2 \\
&= \Sigma_x^{-1/2}\mathbf{B}_1\mathbf{B}_1^T\mathrm{E}\big(\mathbf{Z}h(Y)\big) + \Sigma_x^{-1/2}\eta_1\eta_1^T\mathrm{E}\big(\mathbf{Z}h(Y)\big)/\|\eta_1\|^2 \\
&= \Sigma_x^{-1/2}\mathbf{B}_1\mathbf{B}_1^T\mathrm{E}\big(\mathbf{Z}h(Y)\big) + A_1A_1^T\mathrm{E}\big(\mathbf{X}h(Y)\big)/\|\eta_1\|^2 \\
&=: \Sigma_x^{-1/2}\mathbf{B}_1\mathrm{E}\big(\mathrm{E}(\mathbf{B}_1^T\mathbf{Z}|Y)h(Y)\big) + c_hA_1. \qquad \text{(A.1)}
\end{aligned}$$

It is obvious that the first term is equal to zero when the condition $\mathrm{E}\big(\mathrm{E}(\mathbf{B}_1^T\mathbf{Z}|Y)\big) = 0$ almost surely. Thus (2.2) implies (2.3). On the other hand, when (2.3) holds, for any transformation $h(\cdot)$, then for every component $a_j(Y)$ of $\mathrm{E}\big(\mathrm{E}(\mathbf{B}_1^T\mathbf{Z}|Y)\big)$, $j = 1, \cdots, p-1$, we choose a function $h(\cdot)$ of y to be $a_j(Y)$ so that every component of $\mathrm{E}\big(\mathrm{E}(\mathbf{B}_1^T\mathbf{Z}|Y)h(Y)\big)$ is equal to $\mathrm{E}\big(a_j(Y)h(Y)\big) = \mathrm{E}\big(a_j^2(Y)\big) = 0$ implying that $a_j(Y) = 0$ almost surely. (2.3) implies (2.2). The necessary and sufficient condition is then proved. When the distribution of $\mathbf{Z}$ is elliptically symmetric, the equation (2.3) can be proved similarly by following the argument in Li (1991), we then omit the detail. □

References

[1] Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324-339.

[2] Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, **35**, 2313-2404.

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

[4] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley, New York.

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

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

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

[8] Härdle W. (1990). *Applied Nonparametric Regression*, Econometric Society Monograph Series, 19, Cambridge University Press.

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

[10] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **1**, 297-318.

[11] Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric aditive models. *The Annals of Statistics*, **38**, 2282-2313.

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

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

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

[15] Li, G. R., Peng, H, Zhang, J, and Zhu, L. X. (2012). Robust rank correlation based screening. *Annals of Statistics*. **40**, 1846 - 1877

[16] Li, R., Zhong, W. and Zhu, L. P.(2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, in press.

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

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

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

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

[21] Peng, H., Cui, X., Wen, S. Q. and Zhu, L. X. (2013). Component selection in an additive regression model. *Scand. Journal of Statistics*, to appear.

[22] Ravikumar, P., Lafferty, J.,Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society. Series B.*, **71**, 1009-1030.

[23] Storlie, C.B., Bonedll, H.D., Reich, B.J. and Zhang, H. H. (2011). Surface estimation, variance selection, and the nonparametric oracle property. *Statistica Sinica*, to appear.

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

[25] Wahba, G. (1990). *Spline models for observational data*, **59**, SIAM. CBMSNSF Regional Conference Series in Applied Mathematics.

[26] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variable. *Journal of the Royal Statistical Society, Series B*, **68**, 49-67.

[27] Zhao, P. and Yu, B. (2006). On model selection consisitency of Lasso. *Journal of Machine larning Research*, **7**, 2541-2563.

[28] Zhu, L. P., Li, L. X., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, **106**, 1464-1474.

[29] Zhu, L., W P., ang T., Zhu, L. and Ferré, L. X. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295-304.

[30] Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On Sliced Inverse Regression with High Dimensional Covariates. *Journal of the American Statistical Association*, **101**, 630-643.

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

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu And Li-Xing Zhu, January 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.