Wuhan University
Journal of Natural Sciences

2019, Vol.24 No.4, 277-282

Article ID 1007-1202(2019)04-0277-06
DOI https://doi.org/10.1007/s11859-019-1397-3

Nonparametric Regression Combining Linear Structure

☐ **ZHANG Yanli[1], SONG Yunquan[2], LIN Lu[3], WANG Xiuli[4]**

1. School of Statistics, Shandong University of Finance and Economics, Jinan 250014, Shandong, China;

2. College of Science, China University of Petroleum, Qingdao 266580, Shandong, China;

3. Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, Shandong, China;

4. School of Mathematics and Statistics, Shandong Normal University, Jinan 250014, Shandong, China

Abstract: Nonparametric models are popular owing to their flexibility in model building and optimality in estimation. However nonparametric models have the curse of dimensionality and do not use any of the prior information. How to sufficiently mine structure information hidden in the data is still a challenging issue in model building. In this paper, we propose a parametric family of estimators which allows for penalizing deviation from linear structure. The new estimator can automatically capture the linear information underlying regressions function to avoid the curse of dimensionality and offers a smooth choice between the full nonparametric models and parametric models. Besides, the new estimator is the linear estimator when the model has linear structure, and it is the local linear estimator when the model has no linear structure. Compared with the complete nonparametric models, our estimator has smaller bias due to using linear structure information of the data. The new estimator is useful in higher dimensions; the usual nonparametric methods have the curse of dimensionality. Based on the projection framework, the theoretical results give the structure of the new estimator and simulation studies demonstrate the advantages of the new approach.

Key words: nonparametric; full model; nonlinear
CLC number: O 212.4

Received date: 2018-07-12
Foundation item: Supported by the Project of Humanities and Social Science of Ministry of Education of China (16YJA910003), the "Financial Statistics and Risk Management" Fostering Team of Shandong University of Finance and Economics and the Project of Shandong Province Higher Educational Science and Technology (J16LI56 and J17KA163)
Biography: ZHANG Yanli, female, Ph. D., Associate professor, research direction: multivariate statistics. E-mail: zhangylhappy@ 163.com

0 Introduction

Let $(\boldsymbol{X}_i, Y_i), (i = 1, 2, \cdots, n)$ be independent and identically distributed observations from $(\boldsymbol{X}, Y)$ with smooth joint density $p(\boldsymbol{x}, y) = f(\boldsymbol{x})g(y \mid \boldsymbol{x})$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{id})^{\mathrm{T}} \in \mathbf{R}^d, Y_i \in \mathbf{R}$. Consider the following regression:

$$Y = m(\boldsymbol{x}) + \varepsilon \tag{1}$$

where $m(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x})$ is the conditional mean function, the error ε with mean zero and variance σ^2 is independent of $\boldsymbol{X}$. In this work, our aim is to estimate $m(\boldsymbol{x})$ given the data $(\boldsymbol{X}_i, Y_i)$.

In typical regression models we assume that $m(\boldsymbol{x})$ is linear. For linear models, there exist fruitful results (see example Ref.[1]). Various models with given special structure have been extensively investigated in the literatures and widely applied in practice, such as parametric models (linear or nonlinear), semi-parametric models (partial linear model and general linear model), additive model and so on (Refs. [2-6]). If the underlying assumptions are correct, the fitted models can be easily interpreted and estimated. If they are incorrect, the special estimators may be inconsistent and give a misleading picture of the regression relationship.

Models without restrictions fall into the broad class of nonparametric regression models which are called full models. In the full model, we only assume that $m(\boldsymbol{x})$ is a smooth function. Some nonparametric estimators have been well studied, among which the local linear estimator is minimax optimal in case of more than one dimension. The convergence rate of mean squared error of a common nonparametric estimator is of the order $O(n^{-\frac{4}{4+d}})$. The convergence rate is optimal and cannot be improved in some sense. Nonparametric regression esti-

mators are very flexible but their statistical accuracy decreases greatly with the growth of the dimension of explanatory variables in the model. The latter caveat has been appropriately called the curse of dimensionality (Refs. [7-12]).

Therefore, researchers have tried to develop models and estimators to offer more flexibility than special cases and to overcome the curse of dimensionality. Whether or how to combine advantages of the stability of the special estimators and the optimality of the nonparametric estimators is the key issue of the methods.

We note that an adaptive combination between the full model and the additive model has been studied (see Refs. [13-15]). These methods solve the curse of dimensionality and do not need any limitation for the regression function. At the same time, the methods mentioned above are combinations of two nonparametric estimators, hence the estimators can not obtain the optimal convergence rate when the true regression function has parametric structure.

Example: $m(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$.

Choosing the linear or additive model will lead to serious bias owing to neglecting the nonparametric component of the regression function $m(\boldsymbol{x})$. Fitting the full model also causes large bias because a large bandwidth must be used in order to achieve the same rate for variance and the squared bias. If the structure information in the data can be automatically captured and sufficiently used, the accuracy of estimator will be greatly improved.

Based on the arguments above and motivated by Ref. [13], in this paper, we propose a continuous parametric family of estimators $\hat{m}_\lambda(\lambda > 0)$ which can automatically capture the linear structure information of regression function. The method has following advantages:

1) It offers a continuous model choice via the tuning parametric λ, including the full model $(\lambda = 0)$ and the linear model $(\lambda = \infty)$ as special cases;

2) It is an adaptive combination between the local linear estimator and the global linear estimator;

3) It overcomes the curse of dimensionality and the convergence rate is the parametric rate when the true regression function is linear.

1 Definition of the Penalized Estimator

This section is divided into two subsections. In the first subsection, we briefly give a review of the local linear estimator and the linear estimator for model (1) which can be viewed as a projection of the data with respect to appropriate norms. The second subsection develops the definition of our estimator.

1.1 Simple Smoothers Viewed as Projections

For a fixed $\boldsymbol{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}} \in \mathbf{R}^d$, the local linear estimator for $m(\boldsymbol{x})$ in the full model (1) is defined as $\hat{m}_{\mathrm{ll}}(\boldsymbol{x}) = \hat{r}_0(\boldsymbol{x})$ by minimizing

$$\mathrm{SSR}(r(\boldsymbol{x}), \boldsymbol{x}) = \sum_{i=1}^{n}(Y_i - r_0(\boldsymbol{x}) - \sum_{k=1}^{d} r_k(\boldsymbol{x})\frac{X_{ik} - x_k}{h_k})^2 K_h(\boldsymbol{X}_i, \boldsymbol{x})$$

where $K_h(\boldsymbol{X}_i, \boldsymbol{x})$ is the d-dimensional kernel function of the observation $\boldsymbol{X}_i$ for the output point $\boldsymbol{x}$. We assume the bandwidths $h_1, \cdots, h_k$ are of the same order.

According to Ref. [16], many of smoothing methods are projections with respect to a particular norm.

Define a normalized vector space of $n \times (d+1)$ functions

$$\boldsymbol{F} = \left\{\boldsymbol{r} = (r^{i,l} \mid i = 1, \cdots, n; l = 0, 1, \cdots, d), r^{i,l} : \mathbf{R}^d \to \mathbf{R}\right\}$$

that contains the space of data vectors and the space of candidate regression functions. Let

$$\boldsymbol{r}_Y = (Y_1, 0, \cdots, 0, Y_2, 0, \cdots, 0, Y_n, 0, \cdots, 0)$$

that is, within blocks of d+1, only the first entries may be nonzero, then $\boldsymbol{r}_Y \in F$. Define

$$\boldsymbol{F}_{\mathrm{full}} = \left\{\boldsymbol{r} = (r^0, r^1, \cdots, r^d, r^0, r^1, \cdots, r^d, \cdots, r^d) \mid r^l : \mathbf{R}^d \to \mathbf{R}\right\}$$

that is, within blocks of d+1, $r^{i,l}(\boldsymbol{x})$ are independent of i. $\boldsymbol{F}_{\mathrm{full}}$ is a subspace of $\boldsymbol{F}$.

Define the seminorm $\|.\|_*^2$ on $\boldsymbol{F}$ by

$$\|r\|_*^2 = \int \frac{1}{n}\sum_{i=1}^{n}\left[r^{i,0}(\boldsymbol{x}) + \sum_{k=0}^{d} r^{i,k}(\boldsymbol{x})\frac{X_{i,k} - x_k}{h_k}\right]^2 K_h(\boldsymbol{X}_i, \boldsymbol{x}) v(\mathrm{d}\boldsymbol{x})$$

Hence, for $\boldsymbol{r} \in \boldsymbol{F}_{\mathrm{full}}$, we have

$$\|\boldsymbol{r}_Y - \boldsymbol{r}\|_*^2 = \int \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - r^0(\boldsymbol{x}) - \sum_{k=0}^{d} r^k(\boldsymbol{x})\frac{X_{i,k} - x_k}{h_k}\right]^2 K_h(\boldsymbol{X}_i, \boldsymbol{x}) v(\mathrm{d}\boldsymbol{x}) \tag{2}$$

The $\int$ can be removed because the minimum can be found for each $\boldsymbol{x}$ individually. The integration corresponds to the minimization problem for the local linear estimator. Consequently, $\hat{m}_{\mathrm{ll}}$ is the projection of the response Y onto the subspace $\boldsymbol{F}_{\mathrm{full}}$ under the seminorm $\|.\|_*^2$.

The centralized linear regression model assumes that

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i = \sum_{j=1}^{d} \boldsymbol{\beta}_j X_{ij} + \varepsilon_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i + \varepsilon_i,\ i = 1, \cdots, n$$

where $E\varepsilon_i = 0$ and $D\varepsilon_i = \sigma^2$. The least-squares estimation for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} X_i)^2 = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n)^{\mathrm{T}}$. Then the linear estimator of regression function $m(\boldsymbol{x})$ is $\hat{m}_l = \hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}$. The least- squares estimation has the parametric convergence rate $O_p(n^{-1/2})$. Set

$$\boldsymbol{F}_{\text{linear}} = \left\{ \boldsymbol{r} = (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}, 0, \cdots, 0, \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}, 0, \cdots, 0, \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}, 0, \cdots, 0) \mid \boldsymbol{\beta} \in \mathbf{R}^d \right\}$$

it is obvious that $\boldsymbol{F}_{\text{linear}}$ is a subspace of $\boldsymbol{F}_{\text{full}}$.

Letting $K_h(\boldsymbol{X}_i, \boldsymbol{x}) \equiv C$ and $v(\mathrm{d}\boldsymbol{x}) = \mathrm{d}(F_n(\boldsymbol{x}))$ in formula (2), we have, for $\boldsymbol{r} \in \boldsymbol{F}_{\text{linear}}$,

$$\left\| \boldsymbol{r}_Y - \boldsymbol{r} \right\|_*^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2 \triangleq \left\| \boldsymbol{r}_Y - \boldsymbol{r} \right\|_2^2$$

The formula is the minimization problem for the linear estimator $\hat{m}_l$. Hence, it is the projection of the response Y to subspace $\boldsymbol{F}_{\text{linear}}$ under the seminorm $\|\cdot\|_2^2$. We use P_l and P_* to denote the $\|\cdot\|_2^2$ orthogonal projection from $\boldsymbol{F}_{\text{full}}$ onto $\boldsymbol{F}_{\text{linear}}$ and the $\|\cdot\|_*^2$ orthogonal projection from $\boldsymbol{F}$ onto $\boldsymbol{F}_{\text{full}}$, respectively.

1.2 Definition of the Penalized Estimator

In this subsection, we construct a family of estimators connecting linearity with nonlinearity components. The approach offers us more flexibility in case of highly nonlinear functions and chooses a fit between the linear and the full model. The new method decomposes $m(\boldsymbol{x})$ into the linear part and the orthogonal components. Set the seminorm

$$\left\| \cdot \right\|_\lambda^2 = \left\| \boldsymbol{r}(\boldsymbol{x}) \right\|_*^2 + \lambda \left\| (\boldsymbol{I} - P_l) \boldsymbol{r}(\boldsymbol{x}) \right\|_2^2$$

define the penalty estimator by

$$\hat{m}_\lambda(\boldsymbol{x}) = \arg\min_{r \in \boldsymbol{F}_{\text{full}}} \left\| \boldsymbol{r}_Y - \boldsymbol{r}(\boldsymbol{x}) \right\|_\lambda^2 = \arg\min_{\boldsymbol{r} \in \boldsymbol{F}_{\text{full}}}$$

$$\int \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{r}^{i,0}(\boldsymbol{x}) + \sum_{k=0}^{d} r^{i,k}(\boldsymbol{x}) \frac{X_{i,k} - x_k}{h_k} \right]^2 K_h(\boldsymbol{X}_i, \boldsymbol{x}) v(\mathrm{d}\boldsymbol{x})$$

$$+ \lambda \left\| (I - P_l) r(\boldsymbol{x}) \right\|_2^2$$

Let $P_{l,\lambda}$ be the orthogonal projection from $\boldsymbol{F}_{\text{full}}$ onto $\boldsymbol{F}_{\text{linear}}$ under $\|\cdot\|_\lambda^2$. Note that $\hat{m}_\lambda(\boldsymbol{x})$ ($\lambda > 0$) is a parametric family of estimators which includes asymptotically optimal estimators for the full model ($\lambda = 0$) and the linear model ($\lambda = \infty$) as special cases. It offers a continuous model via the tuning parameter λ. For general λ, we get a family of estimators connecting the local linear estimator with the linear estimator. We call $\hat{m}_\lambda(\boldsymbol{x})$ as a local linear-linear estimation (LLLE).

2 Properties of the Estimator

In this section, we study the properties of the estimator given in Section 1. In order to investigate the asymptotic properties, we give the following conditions:

C1: The kernel K is bounded, and it has compact support $[-a, a]$. The kernel K is symmetric about zero and Lipschitz continuity;

C2: The d-dimensional vector $\boldsymbol{X}$ has compact support $[-a, a]^d$ and its density f is bounded away from zero;

C3: $h_k = O(n^{-1/5}), k = 1, \cdots, d$;

C4: for some $k > 5/2$, $E \mid Y^k \mid < \infty$;

C5: $m(\boldsymbol{x})$ is twice continuously differentiable and f is once continuously differentiable.

Let

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & \frac{X_{1,1} - x_1}{h_1} & \frac{X_{1,2} - x_2}{h_2} & \cdots & \frac{X_{1,d} - x_d}{h_d} \\ 1 & \frac{X_{2,1} - x_1}{h_1} & \frac{X_{2,2} - x_2}{h_2} & \cdots & \frac{X_{2,d} - x_d}{h_d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{X_{n,1} - x_1}{h_1} & \frac{X_{n,2} - x_2}{h_2} & \cdots & \frac{X_{n,d} - x_d}{h_d} \end{pmatrix},$$

$$\boldsymbol{W} = \begin{pmatrix} K_h(\boldsymbol{X}_1 \quad \boldsymbol{x}) & 0 & \cdots & 0 \\ 0 & K_h(\boldsymbol{X}_2 \quad \boldsymbol{x}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_h(\boldsymbol{X}_n \quad \boldsymbol{x}) \end{pmatrix}$$

Then we have the following theorems.

Theorem 1 Under the conditions C1-C3, we have

$$\hat{m}_\lambda(\boldsymbol{x}) = P_{l,\lambda} \hat{m}_{\text{ll}} + (\tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \tilde{\boldsymbol{X}} \left(\boldsymbol{I} - P_l \right) \hat{m}_{\text{ll}} \tag{3}$$

$$P_l \hat{m}_\lambda(\boldsymbol{x}) = P_{l,\lambda} \hat{m}_{\text{ll}}(\boldsymbol{x}) \tag{4}$$

and

$$\left(\boldsymbol{I} - P_l \right) \hat{m}_\lambda(\boldsymbol{x}) = (\tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \tilde{\boldsymbol{X}} \left(\boldsymbol{I} - P_l \right) \hat{m}_{\text{ll}} \tag{5}$$

Proof Let $\boldsymbol{S}(\boldsymbol{x}) = \tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \tilde{\boldsymbol{X}}$ and $\boldsymbol{L}(\boldsymbol{x}) = \tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{Y}$, then the normal equations for $\hat{m}_{\text{ll}}$ and $\hat{m}_\lambda$ respectively are

$$\boldsymbol{S}(\boldsymbol{x}) \hat{m}_{\text{ll}} = L(\boldsymbol{x})$$

and

$$(\boldsymbol{S}+\lambda(\boldsymbol{I}-P_l))\hat{m}_\lambda(\boldsymbol{x})=\boldsymbol{L}(\boldsymbol{x})=\boldsymbol{S}(\boldsymbol{x})\hat{m}_{\text{ll}},$$

We have

$$\hat{m}_\lambda(\boldsymbol{x})=(\boldsymbol{S}+\lambda(\boldsymbol{I}-P_l))^{-1}\boldsymbol{S}(\boldsymbol{x})\hat{m}_{\text{ll}}.$$

The above equation is equivalent to

$$\hat{m}_\lambda(\boldsymbol{x})=\{(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\lambda P_{l,\lambda}+(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}S\}\hat{m}_{\text{ll}} \tag{6}$$

Since $(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\lambda+(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}=\boldsymbol{I}$, we obtain

$$\hat{m}_\lambda(\boldsymbol{x})=\{P_{l,\lambda}+(\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{X}}+\lambda\boldsymbol{I})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{X}}\left(\boldsymbol{I}-P_l\right)\}\hat{m}_{\text{ll}}$$

Then equation (3) holds. Because $P_{l,\lambda}\hat{m}_{\text{ll}}(\boldsymbol{x})\in\boldsymbol{F}_{\text{linear}}$, so formula (4) holds. Let $\boldsymbol{S}_l=P_l\boldsymbol{S}P_l$, $\Lambda_{*,\lambda}=P_l(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}P_l$. If S_l is invertible, then we define the projection

$$P_{*,\lambda}=P_l\Lambda_{*,\lambda}P_l(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}$$

Then we have $P_lP_{*,\lambda}=P_{*,\lambda}$ and

$$P_l(\boldsymbol{I}-(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\lambda)P_{*,\lambda}=$$
$$\Lambda_{*,\lambda}\Lambda^{-1}_{*,\lambda}P_l(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}=P_l(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S} \tag{7}$$

By the definition of $P_{*,\lambda}$, $P_{*,\lambda}\hat{r}_{\text{ll}}\in\boldsymbol{F}_{\text{linear}}$, $(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}=\boldsymbol{I}-(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\lambda$ and equation (7), we obtain

$$P_l(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}(\boldsymbol{I}-P_{*,\lambda})=0$$

The above formula shows that the reminder is orthogonal to $\boldsymbol{F}_{\text{linear}}$.

Remark 1 Formula (3) and (4) show that $\hat{r}_\lambda(\boldsymbol{x})$ is decomposed into the linear part and the orthogonal remainder, and it is a weighted sum of the local linear estimator and the linear estimator. The linear part is just the projection of $\hat{m}_\lambda(\boldsymbol{x})$ with respect to the norm $\|\cdot\|_2^2$. It is obvious that in sparse regions, the local linear estimator is unstable and does not exist when $\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{X}}$ is singular. Formula (5) shows that our method solves this problem. When all eigenvalues of $\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{X}}$ are large, the effect of penalty vanishes.

Theorem 2 Under the conditions C1-C4 and $\lambda\to 0$, we have

$$\left\|\hat{m}_\lambda(\boldsymbol{x})-\hat{m}_{\text{ll}}(\boldsymbol{x})\right\|_2=O_P(\lambda) \tag{8}$$

Proof By Formula (6), we have

$$\hat{m}_\lambda(\boldsymbol{x})=(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\lambda P_{l,\lambda}\hat{m}_{\text{ll}}+(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\hat{m}_{\text{ll}}$$
$$\hat{=}W_1P_{l,\lambda}\hat{m}_{\text{ll}}+W_2\hat{m}_{\text{ll}}$$

which shows that $\hat{m}_\lambda(\boldsymbol{x})$ is some kind of convex combination of $\hat{m}_{\text{ll}}(\boldsymbol{x})$ and $P_{l,\lambda}\hat{r}_{\text{ll}}(\boldsymbol{x})$. Let $W_1=O_P(\lambda)$ and $W_2=1+O_P(\lambda)$, then the result follows directly.

Remark 2 The conclusion implies that when the regression function has no linear structure, the new estimator behaves as well as the local linear estimator and therefore they can achieve the optimal convergence rate, provided that λ is not very large.

Theorem 3 Under the conditions C1-C5, and $\lambda\to\infty$, we have

$$\left\|\hat{m}_\lambda(\boldsymbol{x})-\hat{m}_l(\boldsymbol{x})\right\|_2=O(\frac{1}{\lambda\sqrt{nh^d}}) \tag{9}$$

Proof If the regression function is linear, then the variance term is the nonlinear part. It indicates

$$\|(\boldsymbol{I}-P_{*,\lambda})\hat{m}_{\text{ll}}\|_2^2=O_p(\frac{1}{nh^d})$$

By Theorem 1, we have

$$\hat{m}_l(\boldsymbol{x})-\hat{m}_\lambda(\boldsymbol{x})=P_{*,\lambda}\hat{m}_\lambda(\boldsymbol{x})-P_l\hat{m}_\lambda(\boldsymbol{x})-\left(\boldsymbol{I}-P_l\right)\hat{m}_\lambda(\boldsymbol{x})$$
$$=P_*\left(\boldsymbol{I}-P_l\right)\hat{m}_\lambda(\boldsymbol{x})-\left(\boldsymbol{I}-P_l\right)\hat{m}_\lambda(\boldsymbol{x})$$
$$=(P_*-\boldsymbol{I})\left(\boldsymbol{I}-P_l\right)\hat{m}_\lambda(\boldsymbol{x})$$
$$=(P_*-\boldsymbol{I})(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\left(\boldsymbol{I}-P_{*,\lambda}\right)\hat{m}_{\text{ll}}(\boldsymbol{x})$$

Under the condition $\|(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\|_{2,\sup}\quad 1/\lambda\|\boldsymbol{S}\|_{2,\sup}$, we obtain Theorem 3.

Remark 3 Formula (9) indicates that $\hat{m}_\lambda(\boldsymbol{x})$ and $\hat{m}_l(\boldsymbol{x})$ are equivalent if the true regression function is linear for $h=O(n^{-1/5})$ and $1/\lambda=O(n^{-d/10})$. That is to say, the estimator performs as the linear estimator and has parametric convergence rate.

All the properties aforementioned reveal that the new estimator can adapt to the global and local linearity of the regression function by choosing regularization parameter λ.

3 Simulation Studies

3.1 The Choice of Tuning Parametric

In this section we consider the choice of the tuning parametric λ and the bandwidth *h*. It is observed from formula (3) that the first part of bias of $\hat{m}_\lambda(\boldsymbol{x})$ is irrelevant to λ if the study model is linear. Now the parameter selection is asymptotically equivalent to the classical variance/bias compromise, that is to say $h\propto n^{-1/5}$ and $\lambda\to\infty$. Next we investigate the rate of λ for the general case. Set

$$\hat{Y}=[(\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{X}}+\lambda(\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-}\tilde{\boldsymbol{X}}^{\mathrm{T}}-\boldsymbol{I})]^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y}\hat{=}\boldsymbol{M}_\lambda\boldsymbol{Y}$$

which is a linear function of $\boldsymbol{Y}$. $\boldsymbol{M}_\lambda$ is called the hat matrix. Consider the following criteria:

$$\mathrm{AIC}(\lambda,h)=\log(\sigma^2)+2\mathrm{tr}(\boldsymbol{M}_\lambda)/n$$
$$\mathrm{GCV}(\lambda,h)=\sigma^2/(1-\mathrm{tr}(\boldsymbol{M}_\lambda)/n)^2$$
$$\mathrm{AIC}_C(\lambda,h)=\log(\hat{\sigma}^2)+\frac{1+\mathrm{tr}(\boldsymbol{M}_\lambda)/n}{(1-(\mathrm{tr}(\boldsymbol{M}_\lambda)+2)/n}$$

where $\hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{M}_\lambda Y\|^2$. AIC and GCV criteria are classical model selection (see Ref. [6]). Here we use AIC_C criteria which was proposed by Ref. [17]. By the Taylor expansion,

$$\log(\hat{\sigma}^2) = \log(\sigma^2) + \frac{\hat{\sigma}^2}{\sigma^2} - 1 + \frac{2}{n}\mathrm{tr}(\boldsymbol{M}_\lambda),$$

we have

$$\mathrm{AIC}_C = \mathrm{AIC} - \log(\sigma^2) = \frac{\hat{\sigma}^2}{\sigma^2} - 1 + \frac{2}{n}\mathrm{tr}(\boldsymbol{M}_\lambda)$$

Under the AIC_C criteria, similar to Ref. [13], we obtain $1/\lambda = O(n^{-d/10})$.

3.2 Numerical Simulations

In this section, we conduct a simulation study to compare our method with the local linear estimation, linear estimation, additive estimation, the local additive estimation and locally linear-additive estimation. Consider the following regression function:

$$m(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$$

In this model, $\boldsymbol{x}$ is uniformly distributed in $[-1,1]\times[0,1]$ and the error is normally distributed with mean zero and variance 0.25. We generate 400 datasets from the model. The mean squared errors (MSE) for our estimator and the other five estimators are 5.6, 8.5, 7.4, 13.1, 11.2, 6.3, respectively. It is obvious that the MSE of our estimator is smaller than those of other five estimators and our estimator is more efficient than them.

Figures 1-7 are the true regression function surface contour and six estimation surface contours. It is observed that even though the true regression function is clearly nonlinear, penalizing the nonlinear part leads to a remarkable improvement in optimal MSE.

4 Conclusion

Semiparametric models have been widely used in practice, but how to consistently distinguish parametric and nonparametric terms for the full models is still a challenging problem. In this paper, motivated by Ref. [13],

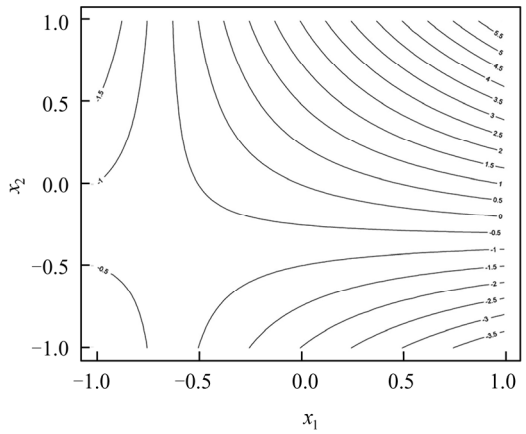

Fig. 1 Contour lines of true regression function

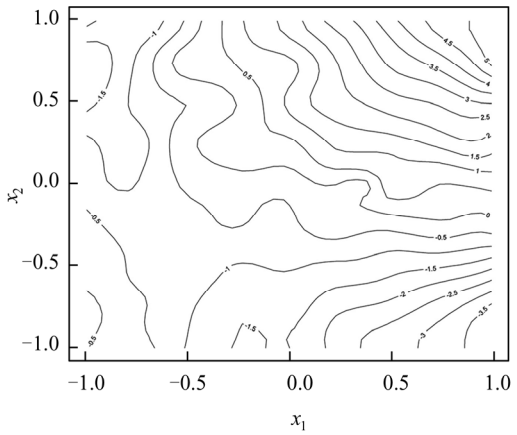

Fig. 2 Contour of local linear estimation

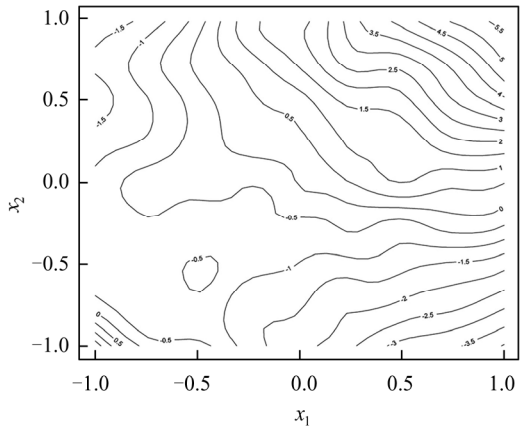

Fig. 3 Contour of linear estimation

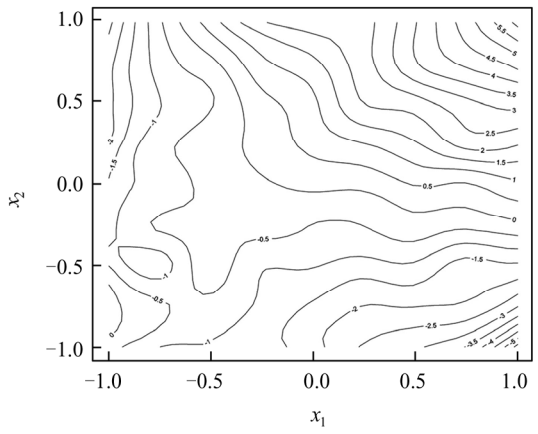

Fig. 4 Contour of additive estimation

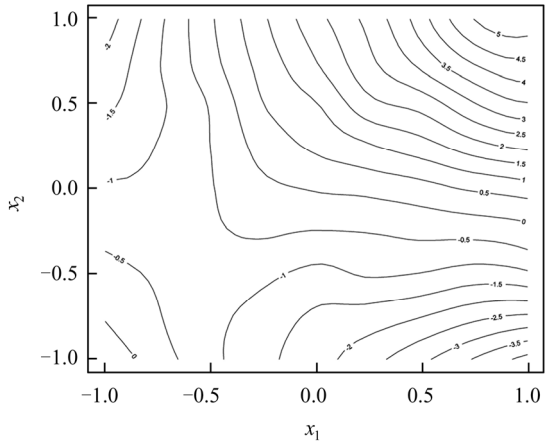

Fig. 5 Contour of local additive estimation

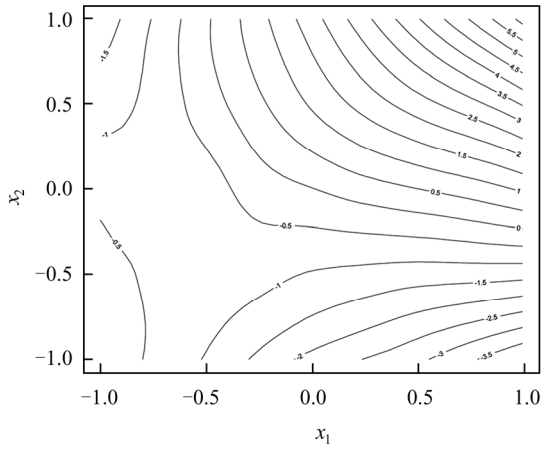

Fig. 6 Contour of local linear-additive estimation

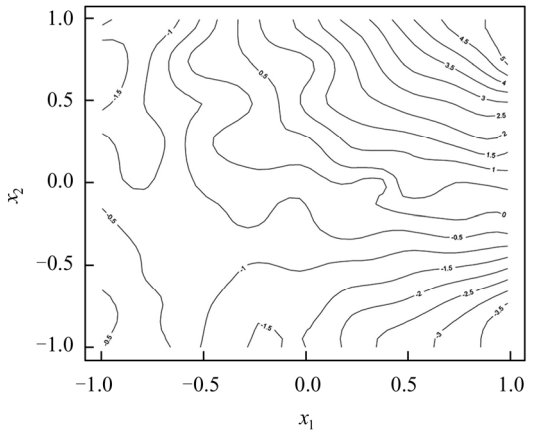

Fig. 7 Contour of local linear-linear estimation

we propose a continuous parametric family of estimators, which can automatically capture the linear structure information of regression function. The new method is an adaptive combination between the local linear estimator and the global linear estimator, including the full model and the linear model as special cases. The estimator avoids the curse of dimensionality and the convergence rate is the parametric rate when the true regression function is linear.

Note that the new method only captures the linear parametric term behind the data and the remainder fitted local linear estimator. It would be interesting to extend the approach to other parametric and nonparametric terms, such as generalized linear-local linear, linear-B-Spline, generalized linear-B-Spline, etc.

References

[1] Arnold S F. *The Theory of Linear Models and Multivariate Analysis* [M]. New York : Wiley, 1981.

[2] Carroll R J, Fan J, Gijbels W, *et al*. Generalized partially linear single-index models [J]. *Journal of the American Statistical Association*, 1997, **92**: 477-489.

[3] Härdle W, Müller M, Sperlich S, *et al*. *Nonparametric and Semiparametric Models* [M]. New York: Springer-Verlag, 2004.

[4] Buja A, Hastie T J, Tibshirani R J. Linear smoothers and additive models (with discussion) [J]. *Annals of Statistics*, 1989, **17**: 453-555.

[5] Nelder J A, Wedderburn R W M. Generalized linear models [J]. *Journal of Applied Econometrics*, 1972, **5**: 99-135.

[6] Hastie T, Tibshirani R. *Generalized Additive Models* [M]. London: Chapman and Hall, 1990.

[7] Fan J. Local linear regression smoothers and their minimax efficiencies [J]. *Annals of Statistics*, 1993, **21**: 196-216.

[8] Fan J, Gasser T, Gijbels I, *et al*. Local polynomial regression: Optimal kernels and asymptotic minimax efficiency [J]. *Ann Inst Statist Math*, 1997, **49**: 79-99.

[9] Stone C. Optimal rates of convergence for nonparametric estimators [J]. *Annals of Statistics*, 1980, **8:** 1348-1360.

[10] Chu C K, Marron J S. Choosing a kernel regression estimator (with discussion) [J]. *Statist Sci*, 1991, **6**: 404-436.

[11] Hardle W, Gasser T. Robust nonparametric function fitting [J]. *Roy Statist Soc Ser B*, 1984, **46**: 42-51.

[12] Stone C. Optimal global rates of convergence for nonparametric regression [J]. *Annals of Statistics*, 1982, **10**: 1040- 1053.

[13] Studer M, Seifert B, Gasser T. Nonparametric regression penalizing deviations from additivity [J]. *Annals of Statistics*, 2005, **33**: 1295-1329.

[14] Park J, Seifert B. Local additive estimation [J]. *Roy Statist Soc Ser B*, 2010, **72**: 171-191.

[15] Lin L, Song Y Q, Liu Z. Local linear-additive estimation for multiple nonparametric regressions [J]. *Journal of Multivariate Analysis*, 2014, **123**: 252-269.

[16] Mammen E, Marron J S, Turlach B, *et al*. A general projection framework for constrained smoothing [J]. *Statist Sci*, 2001, **16**: 232-248.

[17] Hurvich C, Simonoff J, Tsai C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion [J]. *Roy Statist Soc Ser B*, *Methodol*, 1998, **60**: 271-293.

□