Journal of Multivariate Analysis 123 (2014) 252–269

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Local linear–additive estimation for multiple nonparametric regressions

Lu Lin [a,*], Yunquan Song [a,b], Zhao Liu [c]

[a] Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan, China
[b] College of Science, China University of Petroleum, Qingdao, China
[c] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 16 April 2013
Available online 7 October 2013

AMS 2000 subject classifications:
62G08
62H99

Keywords:
Multiple nonparametric regression
Local linear estimation
Local additive estimation
Local linear–additive estimation
Double nonadditivity penalty

ABSTRACT

How to sufficiently use the structure information behind the data is still a challenging issue. In this paper, a local linear–additive estimation and its relevant version are proposed to automatically capture the additive information for general multiple nonparametric regressions. Our method connects two types of local estimators, the local linear (or the local constant) estimator and the local additive estimator. Thus the new estimators can achieve an adaptive fitting between the full model and the local (additive) model, and can adapt to the double additivity: local additivity and global additivity. On the other hand, like the local linear estimator, the new estimators can obtain the optimal convergence rate when the model has no additive structure. Moreover, the new estimators have closed representations and thus make the computation easy and accurate. The theoretical results and simulation studies show that the new approach has a low computational complexity and can significantly improve the estimation accuracy. Also a new theoretical framework is introduced as a foundation of locally and globally connected statistical inference. Based on this framework, the newly defined estimator can be regarded as a projection of the response variable onto full function space with respect to the locally and globally connected norms.

1. Introduction

Consider the following nonparametric regression:

$$Y = r(X) + \varepsilon, \tag{1.1}$$

where Y is the real-valued response variable, $X = (X^{(1)}, \ldots, X^{(p)})'$ is the p-dimensional covariate and ε is the error with conditional mean zero and variance $\sigma^2(X)$ given X. In the full model, we only assume that the unknown regression function

$$r(x) = E(Y|X = x) \tag{1.2}$$

is smooth. Such a full model is usually regarded as to be most flexible. Under regularity conditions (e.g., twice continuous differentiability of $r(x)$), the convergence rate in mean squared error of a common nonparametric estimator (e.g., kernel estimator) is of the order $O(n^{-4/(4+p)})$ [13,14]. This convergence rate is optimal and thus cannot be improved in some sense.

* Corresponding author.
E-mail addresses: linlu@sdu.edu.cn, linlu@eyou.com (L. Lin).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.09.012

It implies that the common nonparametric estimators suffer from the curse of dimensionality when the dimension p is high.

To avoid the problem, one usually considers the model with a specified structure as an approximation to the full model. Additive model, for instance, is a popular approximation to the full model, which has the special form:

$$r(x) = \mu + \sum_{k=1}^{p} r_j(x^{(j)}), \tag{1.3}$$

where regression functions satisfy $E(r_j(X^{(j)})) = 0, j = 1, \ldots, p$. In this paper we prefer to call the above form "the global additive model" because we will use this terminology to discriminate (1.3) from the local additive model [12]. Benefiting from the additive assumption of (1.3), the convergence rate in mean squared error of an estimator is of the order $O(n^{-4/5})$ as that for $p = 1$ [15,16,1]. However, the global additive assumption leads to serious bias when the regression function $r(\cdot)$ is far away from such a global additive form.

Various models with given structure have been extensively investigated in the literature and widely applied in practice. However, there has been little previous literature on the model structures that can adapt automatically to the data from the underlying model. How to sufficiently use the structure information behind the data is still a challenging issue. Studer, Seifert and Gasser [17] proposed a globally penalized estimator (GPE) by penalizing the nonadditivity of local linear estimator. Such an innovative estimator offers an adaptive combination between the full model (1.2) and the global additive model (1.3), instead of the single use of either the full model or the global additive model. Thus the method is automatically adaptive to the global additivity. However, the estimator is based only on a global penalty and thus the information of local additivity is not taken into account. It can be shown that this method is not adaptive to local additivity. For illustration, we consider the following regression function:

$$r(x) = x_1^2 + x_2^2 + x_1^2 x_2^2. \tag{1.4}$$

When $x_1^2 x_2^2$ is small enough, the function is approximately additive. But the local additivity in the local region $\{(x_1, x_2) : x_1^2 x_2^2 < C\}$ cannot be captured by the global nonadditivity penalty, where $C > 0$ is a small constant. In Section 5, we will further illustrate this point of view by simulation. Moreover, the method is computationally intensive and involves the dimensional problem because $m(p+1)$ values of some functions at grids have to be estimated in the calculation procedure, where $m = m_1 \times \cdots \times m_p$ and $m_k, k = 1, \ldots, p$, should tend to infinity to guarantee the accuracy of the calculation. Park and Seifert [12] employed a local additive estimator to capture the underlying local additivity of regression function. This estimator behaves well in the sense that it outperforms the local linear estimator when the true model is close to the additive model locally and is better than the global additive estimator when the true model is far away from the global additive structure. However, when the model is globally nonadditive, unlike the local linear estimator, the local additive estimator cannot achieve the optimal convergence rate because in the estimation procedure it only uses the estimation method designed for the additive model.

Motivated by the ideas of Studer, Seifert and Gasser [17] and Park and Seifert [12], in this paper we suggest a local linear–additive estimator and its relevant version to capture the local and global additive information for general nonparametric regression. The main idea of the new method is that an adaptive combination between the local linear estimation and the local additive estimator (rather than global additive estimator) is built by a local–global nonadditive penalty. For example, consider the regression function in (1.4). In local region $\{(x_1, x_2) : x_1^2 x_2^2 < C\}$, we use an additive model (local additive estimation) to fit the regression function and in local region $\{(x_1, x_2) : x_1^2 x_2^2 \geq C\}$, we use the full model (local linear estimation) to fit the regression function. If we are able to explore the local additivity by a data driven method, an adaptive combination between the local model and full model could be achieved. Our estimation therefore has the following distinguishing features:

(1) the new method can achieve an adaptive fitting between the full model and local additive model and therefore can adapt to the double additivity: local additivity and global additivity. More precisely, when the true model is approximately or completely additive in a local region, the estimation procedure is automatically adjusted to fit the corresponding additive model in this local region; when the true model is globally additive, the estimator is adjusted automatically to be an additive estimator; when the true model is completely nonadditive, the estimator can approach the local linear estimator and the optimal convergence rate is attained as well;
(2) although new idea is motivated by Studer, Seifert and Gasser [17], the new one connects two types of local estimators. Hence, the resultant estimators have closed representations and make the computation easy and accurate;
(3) a new theoretical framework is introduced as a foundation of locally and globally connected statistical inference. In the framework, locally and globally connected norms and relevant projections are defined, and based on the newly defined norms and projections, the new estimator can be regarded as a projection of response variable onto the full function space with respect to the locally and globally connected norms.

The remainder of the paper is organized in the following way. After a brief review of the local linear and the local additive estimators, the local linear–additive estimation and relevant version are defined and their properties are investigated in Section 2. Adaptive selections for regularization parameters and the computational steps are presented in Section 3 and in

the computational aspect, the main feasibility and limitation of such type of estimators are discussed. The general theoretical framework is established in Section 4. Simulation studies are given in Section 5 and the proof for theorem is postponed to Appendix.

2. Local linear–additive estimation

2.1. Brief reviews

To get the motivation for new theoretical and methodological developments, we first briefly review the relevant backgrounds. It is known that the local linear estimation for nonparametric regression function $r(x)$ in full model (1.2) is defined as $\tilde{r}_{ll}(x) = \tilde{\beta}^{(0)}(x)$ with $\tilde{\beta}^{(0)}(x)$ being the first component of the following solution:

$$(\tilde{\beta}^{(0)}(x), \tilde{\beta}^{(1)}(x), \ldots, \tilde{\beta}^{(p)}(x))' = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - \beta^{(0)} - \sum_{j=1}^{p} \beta^{(j)} \left(\frac{X_i^{(j)} - x^{(j)}}{h_j} \right) \right\}^2 K_h(X_i, x), \tag{2.1}$$

where $\beta = (\beta^{(0)}, \beta^{(1)}, \ldots, \beta^{(p)})'$, and $(X_i, Y_i), i = 1, \ldots, n$, are i.i.d. observations from the model (1.1). Here $K_h(X, x)$ is a p-dimensional kernel function; for example, it can be chosen as the product form: $K_h(X, x) = h_1^{-1} K(\frac{X^{(1)} - x^{(1)}}{h_1}) \cdots h_p^{-1} K(\frac{X^{(p)} - x^{(p)}}{h_p})$, where $K(\cdot)$ is a 1-dimensional kernel function and h_j are the corresponding bandwidths. It is known that this estimator asymptotically achieves the linear minimax risk when spherically symmetric Epanechnikov kernels are used; see [2] for the case of $p = 1$, and [3] for the case of $p > 1$. As was mentioned in Introduction, the optimal rate of convergence in mean squared error is of order $O(n^{-4/(4+p)})$, thus it suffers from the curse of dimensionality when the dimension p is high. This leads to the consideration of less general models. Global additive model, for instance, is a common approximation to the full model $r(x)$, which has the special form as in (1.3). The convergence rate of the estimator based on the global additive assumption is of the order $O(n^{-4/5})$ as for $p = 1$. But the global additive assumption results in serious bias due to neglecting the nonadditive component of the regression function. Studer, Seifert and Gasser [17] introduced a global penalty to combine the local linear and global additive estimators. Such an idea can achieve smooth choice between the two estimators and thus can adapt to the global additivity. As was stated in the Introduction, however, the estimator is not adaptive to the local additivity, and is computationally intensive and involves the dimensional problem in the procedure of calculation.

Park and Seifert [12] employed a local additive estimator to partially avoid the above problems. Assume without loss of generality that X has compact support $[-1, 1]^p$. For $w = (w^{(1)}, \ldots, w^{(p)})'$ and a given point $x \in (-1, 1)^p$, consider a p-dimensional rectangular region $[x \pm w] = \{X : X \in [x - w, x + w]\}$. Suppose that all $w^{(j)}, j = 1, \ldots, p$, are of the same order and let $w^{(j)} = \eta$ for the simplicity of notation. Here η may depend on $x : \eta = \eta(x)$. Write

$$U = \frac{X - x}{\eta}. \tag{2.2}$$

Then U is a re-scaled random vector defined on $[-1, 1]^p$. Denote by n_x the number of data $U_i = (X_i - x)/\eta$ in $[-1, 1]^p$. By the new variable U, the corresponding regression function can be rewritten as

$$r_x(u) = r(x + \eta u). \tag{2.3}$$

Consequently, an estimator of $r(\cdot)$ at x is defined as

$$\tilde{r}_{lad,\eta}(x) = \tilde{r}_{ad,x}(u)|_{u=0}, \tag{2.4}$$

where $\tilde{r}_{ad,x}(u)$ is a global additive estimator of $r_x(u)$ based on data U_i in $[-1, 1]^p$. More precisely, it is first assumed that

$$r_x(u) = \mu + \sum_{j=1}^{p} r_j(u^{(j)}), \tag{2.5}$$

and then $r_x(u)$ is estimated by data U_i in $[-1, 1]^p$ and the existing methods designed for global additive model, such as backfitting [1,10], marginal integration [9,8] and two-stage estimation [6,7]. If the resultant estimator is denoted by $\tilde{r}_{ad,x}(u)$, then its value at $u = 0$ is the local additive estimator as in (2.4). Particularly, when $\eta \to \infty$, the resultant estimator $\tilde{r}_{lad,\infty}(x)$ is indeed the global additive estimator of $r(x)$, which is obtained as the model could be exactly expressed as a global additive model as in (1.3). If η is small, the estimator $\tilde{r}_{lad,\eta}(x)$ is related only to the data in the local region $[x \pm w]$ and depends on the local additive assumption on $[x \pm w]$. Then one calls it the local additive estimator. This estimator behaves well in the sense that it outperforms the local linear estimator when the true model is close to the additive model locally or globally, and is better than the global additive estimator when the true model is far away from the global additive structure.

However, unlike the local linear estimator, the local additive estimator does not have the optimal property for general models, i.e., when the model is (globally) nonadditive, unlike the local linear estimator, the local additive estimator cannot achieve the optimal convergence rate of order $O(n^{-4/(4+p)})$. It is because, as was shown above, the approach only uses the estimation method designed for (locally) additive model (2.5) to estimate the regression function.

Thus, it is desirable to develop a new methodology that can adapt to the local and global additivity automatically when the model has the additivity, and can achieve the optimal convergence rate of order $O(n^{-4/(4+p)})$ if the model is (globally) nonadditive.

2.2. Local linear–additive estimation

To construct a new estimator, here we briefly reexamine the local additive estimator defined by (2.4). In fact it can be obtained by minimizing

$$\frac{1}{n_x}\sum_{i=1}^{n_x}\left\{Y(U_i)-\beta^{(0)}(u)-\sum_{j=1}^{p}\beta^{(j)}(u^{(j)})\frac{U_i^{(j)}-u^{(j)}}{h_j(x)}\right\}^2 K_n(U_i-u) \tag{2.6}$$

with respect to $\beta^{(0)}(u)$ and $\beta^{(j)}(u^{(j)}), j=1,\ldots,p$, respectively, where n_x is the number of data U_i in $[-1,1]^p$ and $Y(U_i)$ are the corresponding responses. Note that here $\beta^{(j)}(u^{(j)}), j=1,\ldots,p$, depend only on the components $u^{(j)}$ of u, respectively. Denote by $\tilde{r}_{ad,x}(u)$ and $\tilde{r}^{(j)}_{ad,x}(u^{(j)}), j=1,\ldots,p$, the solutions of $\beta^{(0)}$ and $\beta^{(j)}, j=1,\ldots,p$, respectively. Then $\tilde{r}_{ad,x}(u)$ evaluated at $u=0$ is just the local additive estimator defined by (2.4).

Although local additive estimator $\tilde{r}_{ad,x}(u)$ obtained by minimizing (2.6) and local linear estimator $\tilde{r}_{ll}(x)$ given by (2.1) are respectively local estimators of the regression function $r(x)$ that only use the data around x, the corresponding models do not involve any structural constraint, essentially. This property provides us a reason to combine (2.6) and (2.1) to define the local linear–additive estimator by minimizing

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\beta^{(0)}-\sum_{j=1}^{p}\beta^{(j)}\frac{X_i^{(j)}-x^{(j)}}{h_j}\right\}^2 K_h(X_i-x)\\&+\lambda\eta(x)\frac{1}{n_x}\sum_{i=1}^{n_x}\left\{\beta^{(0)}-\tilde{r}_{ad,x}(u)+\sum_{j=1}^{p}(\beta^{(j)}-\tilde{r}^{(j)}_{ad,x}(u^{(j)}))\frac{U_i^{(j)}-u^{(j)}}{h_j(u)}\right\}^2 K_h(U_i-u)\end{aligned} \tag{2.7}$$

with respect to $\beta^{(0)}$ and $\beta^{(j)}, j=1,\ldots,p$. In (2.7), h_j and $h_j(u)$ may be different; the former is the global bandwidth used for the local linear estimator and the latter is the local bandwidth used for the local additive estimator. Note that here $\beta^{(j)}, j=1,\ldots,p$, are functions of the vector u, thus they are different from those used in (2.6), in which each $\beta_x^{(j)}(u^{(j)})$ depends only on the corresponding component $u^{(j)}$. In (2.7), $\lambda\geq 0$ and $\eta(x)\geq 0$ are the global and the local penalty parameters, respectively. We use them to penalize the global and the local nonadditivity of the local linear estimator. The adaptive selections for λ and $\eta(x)$ will be given in Section 4. Denote by $\hat{r}^{(0)}_{\lambda,\eta,x}(u)$ and $\hat{r}^{(j)}_{\lambda,\eta,x}(u), j=1,\ldots,p$, the solutions respectively of $\beta^{(0)}$ and $\beta^{(j)}, j=1,\ldots,p$, in the optimization problem above. Then

$$\hat{r}^{(0)}_{\lambda,\eta}(x)=\hat{r}^{(0)}_{\lambda,\eta,x}(u)|_{u=0} \tag{2.8}$$

is defined as the local linear–additive estimator of $r(x)$.

The objective function (2.7) combines the local linear estimator with the local additive estimator, instead of the global additive estimator. A reason why we do so is that the local additive estimator can capture the local additivity by use of a small η, and can capture the global additivity by use of a large η. Intuitionally, the objective function (2.7) allows penalizing the global and local nonadditivity of regression function. The resulting estimator (2.8) can achieve an adaptive fitting between the full model and local additive model by choosing λ and $\eta(x)$. More particularly, for $\lambda\eta(x)=0$, we get the local linear estimator and for $\lambda\eta(x)=\infty$ we obtain the local additive estimator; for general λ and $\eta(x)$ we get a family of estimators connecting the local linear estimator with the local additive estimator. We therefore call the estimator in (2.8) the local linear–additive estimation. In Section 4 we will propose a theoretical foundation to guarantee that the local linear–additive estimator is the projection of the response variable onto the full function space with respect to a new locally and globally connected norm.

To simplify the representation, we use matrices and vectors to represent the above estimator. Denote

$$\beta=(\beta^{(0)},\beta^{(1)},\ldots,\beta^{(p)})',\qquad \tilde{\boldsymbol{r}}_{ad,x}(u)=(\tilde{r}_{ad,x}(u),\tilde{r}^{(1)}_{ad,x}(u^{(1)}),\ldots,\tilde{r}^{(p)}_{ad,x}(u^{(p)}))',$$
$$\mathscr{W}_x=\mathrm{diag}(K_h(X_1,x),\ldots,K_h(X_n,x)),\qquad \mathscr{W}_u=\mathrm{diag}(K_h(U_1,u),\ldots,K_h(U_{n_x},u)),$$
$$\tilde{\mathbf{X}}=\begin{pmatrix}1 & \frac{X_1^{(1)}-x^{(1)}}{h_1} & \cdots & \frac{X_1^{(p)}-x^{(p)}}{h_p}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & \frac{X_n^{(1)}-x^{(1)}}{h_1} & \cdots & \frac{X_n^{(p)}-x^{(p)}}{h_p}\end{pmatrix},\qquad \tilde{\mathbf{U}}(u)=\begin{pmatrix}1 & \frac{U_1^{(1)}-u^{(1)}}{h_1(u)} & \cdots & \frac{U_1^{(p)}-u^{(p)}}{h_p(u)}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & \frac{U_{n_x}^{(1)}-u^{(1)}}{h_1(u)} & \cdots & \frac{U_{n_x}^{(p)}-u^{(p)}}{h_p(u)}\end{pmatrix}.$$

Thus the new estimator can be expressed as $\hat{r}^{(0)}_{\lambda,\eta}(x) = \hat{\beta}^{(0)}_x(u)|_{u=0}$ with $\hat{\beta}^{(0)}_x(u)$ being the first component of the following solution:

$$(\hat{\beta}^{(0)}_x(u), \hat{\beta}^{(1)}_x(u), \ldots, \hat{\beta}^{(p)}_x(u))' = \arg\min_{\beta} \left\{ \frac{1}{n}(\mathbf{Y} - \tilde{\mathbf{X}}\beta)' \mathscr{W}_x (\mathbf{Y} - \tilde{\mathbf{X}}\beta) + \lambda\eta(x) \frac{1}{n_x} (\beta - \tilde{\mathbf{r}}_{ad,x}(u))' \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) (\beta - \tilde{\mathbf{r}}_{ad,x}(u)) \right\}. \tag{2.9}$$

Solving the above optimalization problem yields

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = \mathbf{e}_1' \left\{ \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \tilde{\mathbf{X}} + \frac{\lambda\eta(x)}{n_x} \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) \right\}^{-1} \left\{ \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \mathbf{Y} + \frac{\lambda\eta(x)}{n_x} \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) \tilde{\mathbf{r}}_{ad,x}(u) \right\} \Big|_{u=0}, \tag{2.10}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$. Such an estimator can be further expressed as a weighted sum of the local linear estimator and the local additive estimator as

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = W_1(x; \lambda, \eta)\tilde{\beta}(x) + W_2(x; \lambda, \eta)\tilde{\mathbf{r}}_{ad,x}(u)|_{u=0}, \tag{2.11}$$

where $\tilde{\beta}(x)$ is the local linear estimator of β defined by

$$\tilde{\beta}(x) = \left\{ \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \tilde{\mathbf{X}} \right\}^{-1} \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \mathbf{Y},$$

$\tilde{\mathbf{r}}_{ad,x}(u)$ is the local additive estimator that can be expressed as

$$\tilde{\mathbf{r}}_{ad,x}(u) = \left\{ \frac{1}{n}\tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) \right\}^{-1} \frac{1}{n}\tilde{\mathbf{U}}'(u) \mathscr{W}_u \mathbf{Y}(u)$$

with $\mathbf{Y}(u) = (Y(U_1), \ldots, Y(U_{n_x}))'$ being given in (2.6), and weights

$$W_1(x; \lambda, \eta) = \mathbf{e}_1' \left\{ \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \tilde{\mathbf{X}} + \frac{\lambda\eta(x)}{n_x} \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) \right\}^{-1} \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \tilde{\mathbf{X}},$$

$$W_2(x; \lambda, \eta) = \mathbf{e}_1' \left\{ \frac{1}{n}\tilde{\mathbf{X}}' \mathscr{W}_x \tilde{\mathbf{X}} + \frac{\lambda\eta(x)}{n_x} \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u) \right\}^{-1} \frac{\lambda\eta(x)}{n_x} \tilde{\mathbf{U}}'(u) \mathscr{W}_u \tilde{\mathbf{U}}(u).$$

The new estimator has the explicit representation (2.10) and (2.11). Thus the computation procedure can be easily implemented. The existing method of Studer, Seifert and Gasser [17], however, cannot achieve this goal because it is involved with both local and global estimators: local linear estimator and global additive estimator. Moreover, the estimator is an adaptive combination between the full model and the local additive model by selecting regularization parameters λ and η.

2.3. Local N-W-additive estimator

To further facilitate the estimation procedure, we now consider a special version. In the procedure above, set $\beta^{(j)} = r^{(j)}_{ad,x} = 0$ for $j = 1, \ldots, p$. We then get a special estimator as

$$\tilde{r}^{NW}_{\lambda,\eta}(x) = \arg\min_{\beta^{(0)}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \beta^{(0)} \right)^2 K_h(X_i, x) + \lambda\eta(x)(\beta^{(0)} - \tilde{r}_{lad,\eta}(x))^2,$$

where $\tilde{r}_{lad,\eta}(x)$ is the local additive estimator defined in (2.4). We call it local N-W-additive estimation (or local constant–additive estimation). Obviously, the resulting estimator has the following simple form:

$$\tilde{r}^{NW}_{\lambda,\eta}(x) = \frac{n^{-1} \sum_{i=1}^{n} Y_i K_h(X_i, x) + \lambda\eta(x)\tilde{r}_{lad,\eta}(x)}{n^{-1} \sum_{i=1}^{n} K_h(X_i, x) + \lambda\eta(x)}. \tag{2.12}$$

This estimator is a weighted sum of the standard N-W estimator and the local additive estimator; more precisely,

$$\tilde{r}^{NW}_{\lambda,\eta}(x) = \tilde{W}_1(x; \lambda, \eta)\tilde{r}^{NW}(x) + \tilde{W}_2(x; \lambda, \eta)\tilde{r}_{lad,\eta}(x),$$

where $\tilde{r}^{NW}(x)$ is the standard N-W estimator defined by

$$\tilde{r}^{NW}(x) = \frac{n^{-1}\sum_{i=1}^{n} Y_i K_h(X_i, x)}{n^{-1}\sum_{i=1}^{n} K_h(X_i, x)},$$

and weights $\tilde{W}_1$ and $\tilde{W}_2$ are expressed as

$$\tilde{W}_1(x;\lambda,\eta) = \frac{n^{-1}\sum_{i=1}^{n} K_h(X_i, x)}{n^{-1}\sum_{i=1}^{n} K_h(X_i, x) + \lambda\eta},$$

$$\tilde{W}_2(x;\lambda,\eta) = \frac{\lambda\eta}{n^{-1}\sum_{i=1}^{n} K_h(X_i, x) + \lambda\eta}.$$

This estimator also offers an adaptive fitting between the full model and the local additive model by selecting regularization parameters λ and η, and has a more simple explicit representation.

2.4. Asymptotic properties

To get the asymptotic conclusion we need the following conditions:

(C1) : the regression function r and the design density f of X are twice continuously differentiable, and furthermore f is bounded away from zero on $[-1, 1]^p$;
(C2) : the kernel K is bounded and Lipschitz continuous, has compact support and is symmetric around 0;
(C3) : for some $\delta > 5/2$, $E[|Y|^\delta] < \infty$;
(C4) : $\tilde{h}_j \to 0$ such that $n_x \tilde{h}_j^p / \log n_x \to \infty$ as $n_x \to \infty$, where $\tilde{h}_j = h_j/\eta$.

Then we have the following main theorem.

Theorem 1. *For model (1.1) with conditions* (C1)–(C4)*, the local linear–additive estimator and local N-W-additive estimator have the following properties:*

(1) *if* $\lambda\eta \to 0$*, then for* $x \in [-1, 1]^p$*,*

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = \tilde{r}_{ll}(x) + O_p(\lambda\eta) \quad \text{and} \quad r^{NW}_{\lambda,\eta}(x) = \tilde{r}^{NW}(x) + O_p(\lambda\eta).$$

(2) *if* $p \le 8$ *and* $\lambda\eta \le c$ *for a positive constant* c*, then for* $x \in (-1, 1)^p$*,*

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = r(x) + O_p\left(h^2 + \frac{1}{\sqrt{nh^p}}\right) \quad \text{and} \quad \tilde{r}^{NW}_{\lambda,\eta}(x) = r(x) + O_p\left(h^2 + \frac{1}{\sqrt{nh^p}}\right).$$

(3) *if* $\lambda\eta \to \infty$*, then for* $x \in [-1, 1]^p$*,*

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = \tilde{r}_{lad,\eta}(x) + O_p((\eta\lambda)^{-1}) \quad \text{and} \quad \tilde{r}^{NW}_{\lambda,\eta}(x) = \tilde{r}_{lad,\eta}(x) + O_p((\eta\lambda)^{-1}).$$

(4) *if model (1.1) is globally additive and* $\lambda\eta \to \infty$*, then for* $x \in (-1, 1)^p$*,*

$$\hat{r}^{(0)}_{\lambda,\eta}(x) = r(x) + O_p((\eta\lambda)^{-1}) + O_p\left(h^2 + \frac{1}{\sqrt{n\eta^{p-1}h}}\right)$$

$$\text{and} \quad \tilde{r}^{NW}_{\lambda,\eta}(x) = r(x) + O_p((\eta\lambda)^{-1}) + O_p\left(h^2 + \frac{1}{\sqrt{n\eta^{p-1}h}}\right).$$

By the theorem, we have the following findings:

(a) the conclusions in Theorem 1(1) and (2) together imply that for the case without global additive structure, the new estimators perform like the local linear estimator (or N-W estimator) and therefore they can achieve the optimal (or the standard) convergence rate provided that λ and η are not very large. For achieving this goal, the theoretical constraints on λ and η are quite mild, for example $\lambda, \eta \in [0, c]$ with c being an arbitrary fixed positive constant. By comparison, the new estimators outperform the additive estimators (e.g., the backfitting estimator and marginal integral estimator) when the model under study is nonadditive;

(b) the results of Theorem 1(3) indicate that when λ and η is large enough, the new estimators perform like the local additive estimator and therefore they can adapt to the local additivity of the underlying regression function;

(c) the results of Theorem 1(4) show that if the model is globally additive, the convergence rate of the new estimators is the same as for $p = 1$ provided that λ and η are chosen to be large enough. In this case they can avoid the curse of dimensionality and outperform the estimators (e.g., the local linear estimator) designed for the full model.

In short, all the properties aforementioned reveal that, by choosing regularization parameters λ and η, the new estimators can adapt to the global and local additivity of the regression function, and moreover, if the model is not additive, they still behave like the local linear estimator (or N-W estimator) and therefore retain the optimal convergence rate. The next section will discuss how to obtain the data-driven choices for λ and η.

3. Algorithm aspects

3.1. Regularization parameters and bandwidths selections

Under the transformation (2.2), if the regression function is twice continuous differentiable, then by Taylor expansion, it can be written as

$$\begin{aligned} r(u) &\approx r(x) + \eta(x)\sum_{j=1}^{p} r'_j(x)u^{(j)} + 2\eta^2(x)\sum_{j<k} r''_{j,k}(x)u^{(j)}u^{(k)} \\ &= \text{additive} + 2\eta^2(x)\sum_{j<k} r''_{j,k}(x)u^{(j)}u^{(k)}, \end{aligned}$$

where $r'_j(x) = \partial r(x)/\partial x^{(j)}$ and $r''_{j,k}(x) = \partial^2 r(x)/\partial x^{(j)}\partial x^{(k)}$. Then $\eta(x)$ can possess different degrees of additivity at different locations. It shows that we can use $\eta(x)$ to measure the local additivity. As stated in the previous section we need to construct adaptive choices for λ, $h_j\eta(x)$ and $h_j(x)$. For simplicity we assume that $h_j = h$ and $h_j(x) = h_x$ for $j = 1, \ldots, p$.

We first select the local regularization parameter $\eta(x)$ and local bandwidth $h(x)$. Motivated by the variable bandwidth selection [4], we split up the support $[-1, 1]^p$ of X into m rectangular regions R_k, $k = 1, \ldots, m$, satisfying $\bigcup_{k=1}^{m} R_k = [-1, 1]^p$ and $R_k \bigcap R_j = \phi$ for $k \neq j$. Suppose $x \in R_{k_x}$, where R_{k_x} is an element of $\{R_k, k = 1, \ldots, m\}$, denote by m_x the number of data in R_{k_x}, and let $X_i(x)$, $i = 1, \ldots, m_x$, be the covariates in R_{k_x} and $Y_i(x)$, $i = 1, \ldots, m_x$, be the corresponding responses. By the definition of local additive estimator defined in the previous section, the fitted values of the local responses $Y_i(x)$, $i = 1, \ldots, m_x$, can be expressed as

$$\tilde{\mathbf{Y}}(x) = (\tilde{Y}_1(x), \ldots, \tilde{Y}_{m_x}(x))' = (\tilde{r}_{lad,\eta}(X_1), \ldots, \tilde{r}_{lad,\eta}(X_{m_x}))' = H(x)\mathbf{Y}(x),$$

where $\mathbf{Y}(x) = (Y_1(x), \ldots, Y_{m_x}(x))'$ and $H(x)$ is the corresponding hat matrix. According to the AIC-type model selection criterion used for selecting regularization parameter in local additive estimator [12], here we define a local AIC-type model selection criterion

$$AIC(h_x, \eta(x)) = \log(\hat{\sigma}^2(x)) + 2\,\mathrm{tr}(H(x))/m_x, \tag{3.1}$$

where $\hat{\sigma}^2(x) = \|\mathbf{Y}(x) - \tilde{\mathbf{Y}}(x)\|^2/m_x$. Such a criterion is related only to the data around x and the corresponding responses. Thus, we choose h_x and $\eta(x)$ by minimizing the local AIC-type model selection criterion $AIC(h_x, \eta(x))$. Denote by $\tilde{h}_x$ and $\tilde{\eta}(x)$ the choices of h_x and $\eta(x)$ via the above method, respectively. Note that $\tilde{h}_x$ and $\tilde{\eta}(x)$ are step functions in the sense that they are constant in each region R_k and their values in different regions R_k and R_j may be different. We therefore smooth them by averaging locally. More precisely, we define the smoothed versions of $\tilde{h}_x$ and $\tilde{\eta}(x)$ respectively by

$$\hat{h}_x = \frac{\sum_{i=1}^{n} \tilde{h}_{X_i} w(X_i - x)}{\sum_{i=1}^{n} w(X_i - x)} \quad \text{and} \quad \hat{\eta}(x) = \frac{\sum_{i=1}^{n} \tilde{\eta}(X_i) w(X_i - x)}{\sum_{i=1}^{n} w(X_i - x)}, \tag{3.2}$$

where $w(\cdot)$ is a weigh function. We use $\hat{h}_x$ and $\hat{\eta}(x)$ as the final choices of h_x and $\eta(x)$, respectively.

Further, by a newly defined estimator, say the local N-W-additive estimator $\tilde{r}^{NW}_{\lambda,\eta}(x)$, after $h(x)$ and $\eta(x)$ being estimated by $\hat{h}_x$ and $\hat{\eta}(x)$ respectively, the fitted value of the global response $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ can be expressed as

$$\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)' = (\tilde{r}^{NW}_{\hat{\eta}(X_1),\lambda}(X_1), \ldots, \tilde{r}^{NW}_{\hat{\eta}(X_n),\lambda}(X_n))' = H\mathbf{Y}, \tag{3.3}$$

where H is the corresponding hat matrix. According to the suggestion of Park and Seifert [12], we can choose h and λ by minimizing the AIC-type model selection criterion

$$AIC(h, \lambda) = \log(\hat{\sigma}^2) + 2\,\mathrm{tr}(H)/n, \tag{3.4}$$

where $\hat{\sigma}^2 = \|\mathbf{Y} - \tilde{\mathbf{Y}}\|^2/n$. Denote by $\hat{h}$ and $\hat{\lambda}$ the solutions of h and λ, respectively. This leads to the final choices of h and λ.

3.2. Computation steps

Now we briefly sum up all we have done from the aspect of algorithm. To truly realize the estimation procedures together with the selections for regularization parameters and bandwidths, we divide the algorithm into the following steps:

Step 1: use the algorithm of calculating local additive estimator $\tilde{r}_{lad,\eta}(x)$ in [12] and the AIC-type model selection criterion to choose the local bandwidth h_x and local regularization parameter $\eta(x)$. As in (3.1), the selected ones can expressed as

$$(\tilde{h}_x, \tilde{\eta}(x)) = \arg\min_{(h_x,\eta(x))} AIC(h_x, \eta(x)).$$

Step 2: smooth $\tilde{h}_x$ and $\tilde{\eta}(x)$ by averaging locally and then get the smoothed versions as $\hat{h}_x$ and $\hat{\beta}(x)$ as in (3.2).

Step 3: after h_x and η_x being estimated by $\hat{h}_x$ and $\hat{\eta}_x$ respectively, employ the newly defined estimator, say the local N-W-additive estimator $\tilde{r}^{NW}_{\lambda,\eta}(x)$, to express the global response vector $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$ as in (3.3), and then choose the global bandwidth $\hat{h}$ and global regularization parameter $\hat{\lambda}$ by minimizing the AIC-type model selection criterion. As in (3.4), the selected ones can be written as

$$(\hat{h}, \hat{\eta}) = \arg\min_{(h,\eta)} AIC(h, \eta).$$

Step 4: use the regularization parameters and bandwidths obtained from *Step* 1 to *Step* 3, together with formulas in Sections 2.2 and 2.3, to construct the weights $W_1(x; \hat{\lambda}, \hat{\eta}(x))$ and $W_2(x; \hat{\lambda}, \hat{\eta}(x))$ (or $\tilde{W}_1(x; \hat{\lambda}, \hat{\eta}(x))$ and $\tilde{W}_2(x; \hat{\lambda}, \hat{\eta}(x))$).

Step 5: finally, combine the local linear estimator $\tilde{\beta}(x)$ (or N-W estimator $\tilde{r}^{NW}(x)$) and the local additive estimator $\tilde{r}_{lad,\hat{\eta}(x)}(x)$ through the weights obtained in *Step* 4 to conduct the local linear–additive estimator $\hat{r}^{(0)}_{\hat{\lambda},\hat{\eta}(x)}(x)$ as in (2.11) or local N-W-additive estimator $\tilde{r}^{NW}_{\hat{\lambda},\hat{\eta}(x)}(x)$ as in (2.12).

As mentioned in the previous sections and the computation steps above, the new estimators have closed expressions and moreover, can be written as weighted sums of known estimators: local linear estimator and local additive estimator. The existing algorithms for local linear estimator and local additive estimator can be employed in our computation procedures. Thus, the computational complexity is relatively low and the estimation procedures can be easily implemented. Nevertheless, unlike the earlier estimators, such as local linear and local additive estimators, and globally penalized estimator, the new estimators involve two regularization parameters (local and global regularization parameters), the choices for them are relatively complex. This is the main limitation of such type of methods. However, to adapt to the double additivity, such a computational cost is worthwhile to pay.

4. Theoretical framework

Now we turn to establishing the theoretical framework for the globally and locally connected inference. We will use the theory to prove that the new estimation performs like a projection, i.e., it is a projection of the response variable onto the full function space with respect to the newly defined norm.

4.1. Globally design-dependent norm and projection

According to Studer, Seifert and Gasser [17], and Mammen, Linton and Nielsen [10], we now outline some design-dependent definitions. Denote the vector space of $(n+1)(p+1)$ functions by

$$\mathscr{F} = \left\{ r = (r^{(i,l)} | \, i = 0, \ldots, n; l = 0, \ldots, p) | \, r^{(i,l)} : [-1, 1]^p \to \mathbb{R} \right\}.$$

Set the projection $\mathscr{P}_0$ on $\mathscr{F}$ by replacing $r^{(i,l)}$ with $r^{(0,l)}$, i.e., if $\check{r} = \mathscr{P}_0 r$, then $\check{r}^{(i,l)} = r^{(0,l)}$. The image of $\mathscr{P}_0$ is denoted by $\mathscr{F}_{full}$, that is

$$\mathscr{F}_{full} = \left\{ r = (r^{(0)}, \ldots, r^{(p)}) | \, r^{(l)} : [-1, 1]^p \to \mathbb{R}, l = 0, \ldots, p \right\},$$

in which the index i is omitted for the simplicity of notation. By the definitions given above, the observations Y_i, $i = 1, \ldots, n$, can be coded as $r \in \mathscr{F}$ by

$$r_Y^{(i,l)} = \begin{cases} Y_i, & \text{for } i > 0 \text{ and } l = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Define the design-dependent seminorm on $\mathscr{F}$ by

$$\|r\|^2 = \int_{[-1,1]^p} \frac{1}{n} \sum_{i=1}^{n} \left\{ r^{(i,0)}(x) + \sum_{j=1}^{p} r^{(i,j)}(x) \frac{X_i^{(j)} - x^{(j)}}{h_j} \right\}^2 K_h(X_i, x) dx, \tag{4.1}$$

where the bandwidths h_j may be different from those used before, but for the simplicity of representation, we use the same notations without confusion. Then for $r \in \mathscr{F}_{full}$ we have

$$\|r_Y - r\|^2 = \int_{[-1,1]^p} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - r^{(0)}(x) - \sum_{j=1}^{p} r^{(j)}(x) \frac{X_i^{(j)} - x^{(j)}}{h_j} \right\}^2 K_h(X_i, x) dx,$$

and consequently, the local linear estimator $\tilde{r}_{ll}(x)$ can be thought of as the projection of $r_Y = (Y_1, \ldots, Y_n)'$ to $\mathscr{F}_{full}$ with respect to the seminorm above [11].

4.2. Locally design-dependent norm and projection

Now we introduce new local norm and projection. For the re-scaled variable U given in (2.2) and the corresponding local function $r_x(u)$ in (2.3), similar to the definitions of $\mathscr{F}$ and $\mathscr{F}_{full}$, we define local functional spaces

$$\mathscr{F}_x = \left\{ r = \left(r_x^{(i,l)} \,|\, i = 0, \ldots, n_x; l = 0, \ldots, p \right) \,|\, r_x^{(i,l)} : [-1, 1]^p \to \mathbb{R} \right\}$$

with n_x being the number of data U_i in $[-1, 1]^p$, and

$$\mathscr{F}_{full,x} = \left\{ r_x = \left(r_x^{(0)}, \ldots, r_x^{(p)} \right) \,|\, r_x^{(l)} : [-1, 1]^p \to \mathbb{R}, l = 0, \ldots, p \right\}.$$

Correspondingly, the locally design-dependent seminorm on $\mathscr{F}_x$ is given by

$$\|r\|_x^2 = \int_{[-1,1]^p} \frac{1}{n_x} \sum_{i=1}^{n_x} \left\{ r_x^{(i,0)}(u) + \sum_{j=1}^{p} r_x^{(i,j)}(u) \frac{U_i^{(j)} - u^{(j)}}{h_j(x)} \right\}^2 K_h(U_i, u) du. \tag{4.2}$$

Furthermore, we define the local additive subspace by

$$\mathscr{F}_{ad,x} = \left\{ r_x \in \mathscr{F}_{full,x} |\, r_x^{(0)}(u) = r_{ad,x}^{(0)}(u) \text{ is additive and } r_x^{(j)}(u) = r_{ad,x}^{(j)}(u^{(j)}) \text{ depends only on } u^{(j)} \text{ for } j = 1, \ldots, p \right\}.$$

Under this situation, the local additive estimator $\tilde{r}_{lad,\eta}$ in (2.4) could be regarded as the projection of $r_Y = (Y_1, \ldots, Y_n)'$ onto the local additive subspace $\mathscr{F}_{ad,x}$ with respect to the seminorm in (4.2). Moreover, we denote by $r_{ad,x} = \mathscr{P}_x r_x$ the local additive projection $\mathscr{P}_x$ from $\mathscr{F}_{full,x}$ to $\mathscr{F}_{ad,x}$, where $r_x \in \mathscr{F}_{full,x}$ and $r_{ad,x} \in \mathscr{F}_{ad,x}$. More precisely, $r_{ad,x}(u) = \sum_{j=1}^{p} \int_{[-1,1]^{p-1}} r_x(u) du^{(-j)} - (p-1) \int_{[-1,1]^p} r_x(u) du$ and $r_{ad,x}^{(j)}(u^{(j)}) = \int_{[-1,1]^{p-1}} r_x^{(j)}(u) du^{(-j)}$, where $\int_{[-1,1]^{p-1}} \cdots du^{(-j)}$ denotes the integral with respect to all components of u except $u^{(j)}$.

4.3. Locally and globally connected projection and estimation

Instead of the use of the global additive estimator, now we use the local additive estimator to build a new penalized local linear estimator. To this end, we further introduce a penalized (or locally and globally connected) seminorm on $\mathscr{F}_{full}$ as

$$\|r\|_{\lambda,\eta}^2 = \|r\|^2 + \lambda \eta(x) \|(\mathscr{J}_x - \mathscr{P}_x) \mathscr{P}_{0,x} \mathscr{T}_x r_x\|_x^2, \tag{4.3}$$

where $\mathscr{J}_x$ is the identity on $\mathscr{F}_x$, $\mathscr{P}_{0,x}$ is the projection on $\mathscr{F}_x$ by replacing $r_x^{(i,l)}$ with $r_x^{(0,l)}$, $\mathscr{T}_x : r_x = \mathscr{T}_x r$ is the re-scale transformation defined by (2.3). Based on the new seminorm (4.3), we obtain a new estimator $\hat{r}_{\lambda,\eta}^{(0)}(x, u)$ of $r(x)$ as the minimizer of $\|r_Y - r\|_{\lambda,\eta}^2$ for $r \in \mathscr{F}_{full}$. Thus $\hat{r}_{\lambda,\eta}^{(0)}(x, u)$ evaluated at $u = 0$ is the final estimator of $r(x)$. Note that $\mathscr{P}_{0,x} r_Y = 0$ and for $r \in \mathscr{F}_{full}$, $\mathscr{P}_{0,x}$ is the identity. Thus $\|r_Y - r\|_{\lambda,\eta}^2$ can be represented in detail as

$$\|r_Y - r\|_{\lambda,\eta}^2 = \|r_Y - r\|^2 + \lambda \eta(x) \|r_x - r_{lad,\eta}\|_x^2$$

with

$$\|r_Y - r\|^2 = \frac{1}{n} \int_{[-1,1]^p} \sum_{i=1}^{n} \left\{ Y_i - r^{(0)}(x) - \sum_{j=1}^{p} r^{(j)}(x)(X_i^{(j)} - x^{(j)})/h_j \right\}^2 K_h(X_i - x) dx,$$

$$\|r_x - r_{lad,\eta}\|_x^2 = \frac{1}{n_x} \int_{[-1,1]^p} \sum_{i=1}^{n_*} \left\{ r_x^{(0)}(u) - r_{ad,x}(u) + \sum_{j=1}^{p} (r_x^{(j)}(u) - r_{ad,x}^{(j)}(u^{(j)}))(U_i^{(j)} - u^{(j)})/h_j(x) \right\}^2 K_h(U_i - u) du.$$

It shows that the new estimator $\hat{r}_{\lambda,\eta,x}(u)$ is the projection of $r_Y = (Y_1, \ldots, Y_n)$ onto the full space $\mathscr{F}_{full}$ with respect to the seminorm (4.3).

Then the remainder is to find the relation between the estimator defined here and the estimator defined in Section 2.2. Since our interest is to estimate $r^{(0)}(x)$ (or $(r^{(0)}(x), r^{(1)}(x), \ldots, r^{(p)}(x))$), $r_{ad,x}(u)$ and $r_{ad,x}^{(j)}(u^{(j)}), j = 1, \ldots, p$, in (4.3) can

be replaced respectively by their estimators $\tilde{r}_{ad,x}(u)$ and $\tilde{r}^{(j)}_{ad,x}(u^{(j)}), j = 1, \ldots, p$, which are obtained based on the (local) additive model (2.5) with local data $U_i \in [-1, 1]$; for details see Section 2.2. Therefore, the new estimator can be expressed as $\hat{r}^{(0)}_{\lambda,\eta}(x, u) = \hat{\beta}^{(0)}(x, u)$ with $\hat{\beta}^{(0)}(x, u)$ being the first component of the following solution:

$$(\hat{\beta}^{(0)}(x, u), \hat{\beta}^{(1)}(x, u), \ldots, \hat{\beta}^{(p)}(x))' = \arg \min_{\beta^{(0)}, \beta^{(1)}, \ldots, \beta^{(p)}} \{\|r_Y - r\|^2 + \lambda\eta(x)\|r_x - \tilde{r}_{lad,\eta}\|^2_x\}, \tag{4.4}$$

where

$$\|r_Y - r\|^2 = \int_{[-1,1]^p} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \beta^{(0)} - \sum_{j=1}^{p} \beta^{(j)}(X_i^{(j)} - x^{(j)})/h_j \right\}^2 K_h(X_i - x)dx,$$

$$\|r_x - \tilde{r}_{lad,\eta}\|^2_x = \frac{1}{n_x} \int_{[-1,1]^p} \sum_{i=1}^{n_*} \left\{ \beta^{(0)} - \tilde{r}_{ad,x}(u) + \sum_{j=1}^{p} (\beta^{(j)} - \tilde{r}^{(j)}_{ad,x}(u^{(j)}))(U_i^{(j)} - u^{(j)})/h_j(x) \right\}^2 K_h(U_i - u)du.$$

However, the above integrations with respect to dx and du have no effect because the minimum can be found for each x and u, individually. By this reason and the same argument as in [10], we can delete the integrations in (4.4) and consequently, the minimizer in (4.4) can be rewritten by the minimizer in (2.7). We therefore have the conclusion:

Theorem 2. *The new estimation defined in (2.8) is exactly the projection of $r_Y = (Y_1, \ldots, Y_n)$ onto the full space $\mathscr{F}_{full}$ with respect to the seminorm (4.3).*

We can see that the above framework formulates a theoretical foundation for the globally and locally connected inference.

5. Simulation studies

In this section, we examine our method by simulation studies. In the following, the figures of the estimated regression surfaces, the corresponding integrated squared errors (ISE) and mean integrated squared errors (MISE) are used to compare our local N-W-additive estimate (LNWADE) and local linear–additive estimate (LLADE) with the common competitors including the local linear estimate (LLE), the local additive estimate (LADE), the additive estimate (AE) and the globally penalized estimator (GPE). To obtain comprehensive comparisons, we investigate the behaviors of these estimations in several models such as global additive model, local additive model and completely nonadditive model with dimensions $p = 2$ and $p > 2$. As shown in computation steps, in the simulation procedure, the data-driven method is employed to select h, η and λ, and the product Epanechnikov kernel is used to construct nonparametric estimation. Also, for other nonparametric estimators, the bandwidths are chosen by data-driven methods such as CV and GCV.

In the following Sections 5.1–5.3, we first consider various types of models with dimension $p = 2$, and in Section 5.4, we then study three models having dimension $p = 10$. Most models below are the same as in [17] or [12]. Thus, to obtain comparable conclusions, we choose the same model conditions and the same criterion ISE or MISE as used respectively by them to evaluate the estimators.

5.1. Nonadditive regression

Example 1. Consider the following nonadditive regression function:

$$r(x) = 15e^{-32\|x-(1/4)\mathbf{1}\|^2} + 35e^{-128\|x-(3/4)\mathbf{1}\|^2} + 25e^{-2\|x-(1/2)\mathbf{1}\|^2},$$

where $x = (x_1, x_2)'$ and $\mathbf{1} = (1, 1)'$. This model was investigated by Studer, Seifert and Gasser [17]. To get the true regression surface, 200 random points of X are generated uniformly on $[0, 1]^2$; see Fig. 1(a). Then, by regression function $r(x)$ given above, the figure of true regression surface is obtained, which is reported in Fig. 1(b). Furthermore, in the corresponding nonadditive regression model:

$$Y = r(X) + \varepsilon,$$

X is uniformly distributed on $[0, 1]^2$ and ε is normally distributed with mean zero and variance $\sigma^2 = 0.5^2$. The sample size is chosen as $n = 200$. The simulated regression surfaces for $r(x)$ are reported in Fig. 2 and the ISEs of the new estimators and the competitors are listed in Table 1.

By Fig. 2 and the ISEs given in Table 1, we have the following findings: (1) when the regression function is completely nonadditive, the local linear estimator, the local-additive estimator, the globally penalized estimator and the two new estimators, the local N-W-additive estimator and the local linear–additive estimator, perform almost equally well, but the two new estimators are slightly better than the local linear estimator, the local-additive estimator and the globally penalized estimator in under the criterion of the ISE. More precisely, the five simulated regression surfaces are close to the true one (see

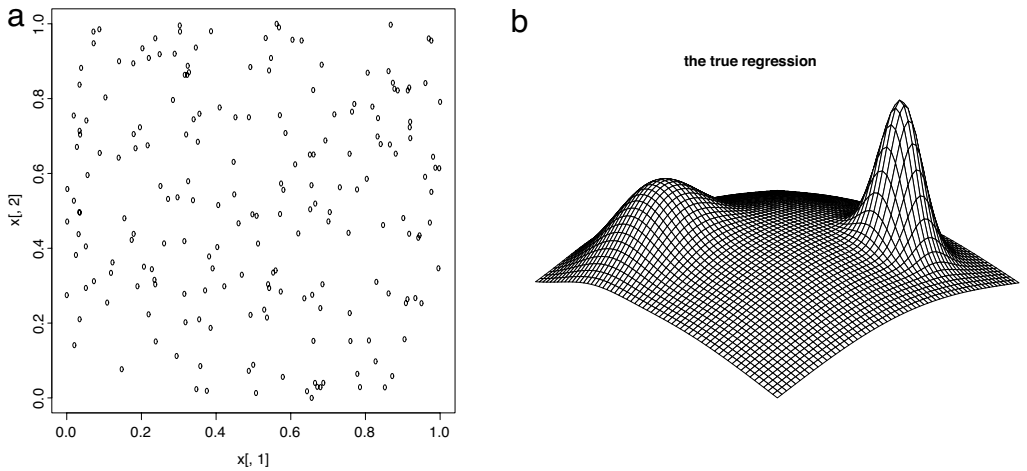

Fig. 1. Figures (a) and (b) are respectively the 200 random points of X and the true regression surface in Example 1.

Table 1
ISEs of six estimators in Example 1.

Estimators	LLE	AE	LADE	GPE	LNWADE	LLADE
ISE	8.4	17.2	6.7	6.4	6.3	6.1

Fig. 2(a), (c), (d), (e) and (f)), and the corresponding ISEs are relatively small. It is known that for nonparametric regression without any structure assumption, the local linear estimator has the optimal property in the sense of convergence rate. The simulation shows that in this case the two new estimators are comparable to the local linear estimator and thus have the same optimal property. (2) Because the model is completely nonadditive, the local additive estimator is somewhat wiggly and the additive estimation is the worst among the six estimators.

5.2. Additive regression

Example 2. Consider the following additive regression function:

$$r(x) = \sum_{k=1}^{2} \left(\frac{15}{2} e^{-32(x_k - 1/4)^2} + \frac{35}{2} e^{-128(x_k - 3/4)^2} + \frac{25}{2} e^{-128(x_k - 1/2)^2} \right).$$

This model was investigated by Studer, Seifert and Gasser [17] as well. In the corresponding regression model, X is distributed as the same as in Example 1, and the error ε is normally distributed with mean zero and variance $\sigma^2 = 0.5^2$. We choose $n = 200$ in the simulation procedure. The true regression surface and the estimated regression surfaces are reported in Figs. 3 and 4, respectively. The ISEs of the six estimators are listed in Table 2.

The Fig. 4 and the ISEs given Table 2 lead to the following conclusions: (1) when the regression function is completely additive, the additive estimator and the two new estimators have the similar behavior, i.e., the three simulated regression surfaces are close to the true one and the corresponding ISEs are approximately equal to each other. When the model is additive, the additive estimator has the optimal property in the sense of convergence rate. In this case, our estimators are comparable to the additive estimator; (2) the global penalized estimator and the local additive estimator are slightly biased because the estimated regression surfaces are smooth but the corresponding ISEs are slightly large. The local linear estimator is the worst among the six estimators.

5.3. Local additive model

Example 3. We first consider the following local additive regression function:

$$r(x) = x_1^2 + x_2^2 + \frac{\alpha}{1-\alpha} x_1^2 x_2^2,$$

where constant α controls the amount of nonadditive structure in the function. This model was considered in [12]. In the model, when $x_1^2 x_2^2$ is small enough, the function is approximately additive. In the corresponding regression model, X is

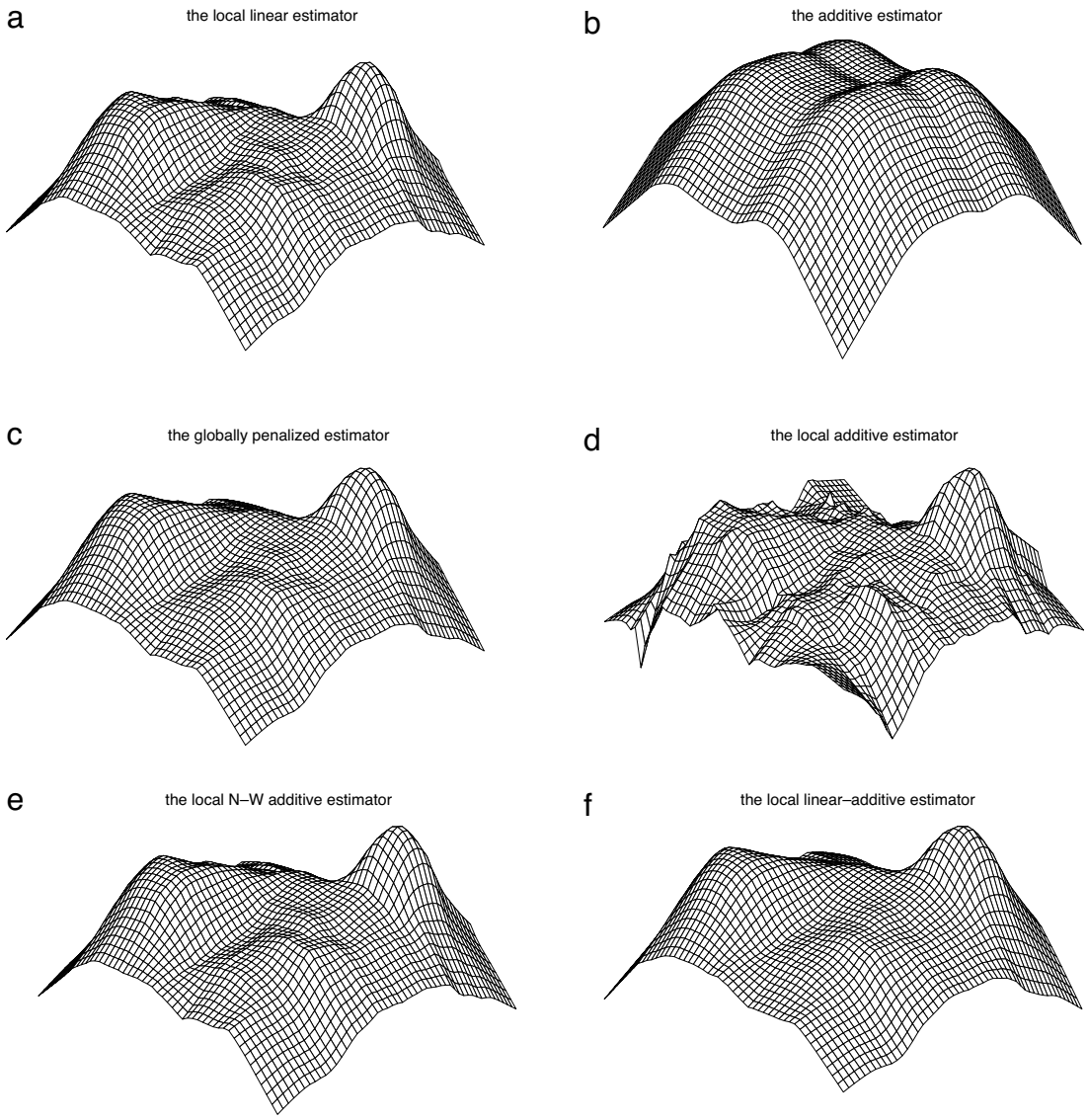

Fig. 2. Figures (a), (b), (c), (d), (e) and (f) are respectively the local linear estimator surface, the additive estimator surface, the globally penalized estimator surface, the local additive estimator surface, the local N-W-additive estimator surface and the local linear–additive estimator surface in Example 1.

Table 2
ISEs of six estimators in Example 2.

Estimators	LLE	AE	LADE	GPE	LNWADE	LLADE
ISE	11.2	6.4	9.6	8.1	7.3	6.9

Table 3
MISEs of six estimators in Example 3.

Estimators	LLE	AE	LADE	GPE	LNWADE	LLADE
MISE	8.3	15	6.4	7.2	6.1	5.8

uniformly distributed on $[-1, 1]^2$ and the error is normally distributed as $N(0, \sigma^2)$ with $\sigma^2 = 0.5^2$. In the simulation procedure, we choose $n = 400$ and $\alpha = 0.4$. The contour plots of the true regression surface and the estimated regression surfaces are presented in Fig. 5. The MISEs of the six estimators are given in Table 3.

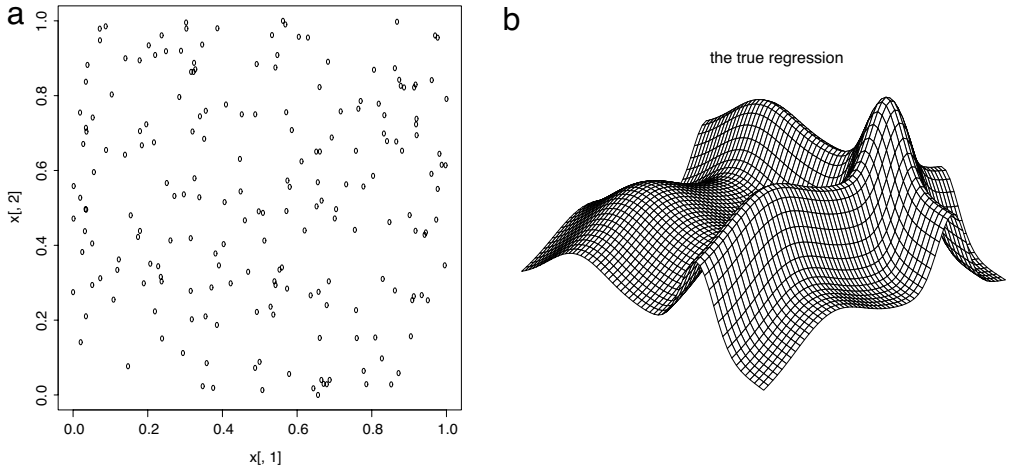

Fig. 3. Figures (a) and (b) are respectively the 200 random points of X and the true regression surface in Example 2.

Table 4
ISEs of six estimators in Example 4.

Estimators	LLE	AE	LADE	GPE	LNWADE	LLADE
ISE	9.1	16	7.2	7.9	6.5	5.3

By comparing the contour plots and the MISEs, we have the following conclusions: (1) the local additive estimator and the two new estimators can capture the local additive information of the true regression surface such that the corresponding contour plots are similar to the true one in the sense that three contours have similar shape and tendency, and the MISEs of the local additive estimator and the two new estimators are relatively small; (2) the MISEs of the globally penalized estimator is larger than those of the local additive estimator and the new estimators. This illustrates that the globally penalized estimator cannot capture the local additive information. The behaviors of the local linear estimator and the additive estimator behave poorly and the additive estimator is the worst one among the six estimators.

Example 4. To further examine if the aforementioned estimators can capture the local additive information, we consider the following local additive regression function:

$$r(x) = \sum_{k=1}^{2} \left(\frac{25}{2} e^{-32(x_k - 1/2)^2} \right) + \frac{25}{2}(x_1 - 1/2)^3 (x_2 - 1/2)^3.$$

Similar to the regression function in Example 3, the above regression function is local additive, i.e., when x_1 and x_2 are close to 1/2, the regression function is additive approximately. The distribution conditions on X and the regression error are the same as in Example 2. The main difference from Example 3 is that here the regression surface is more wiggly. The ISEs of the six estimators are reported in Table 4. The true regression surface and the estimated regression surfaces are presented in Figs. 6 and 7, respectively. Similar to the conclusions in Example 3, the local additive and our new estimators can capture the local additive information; the details are omitted here.

In summary, all the simulation conclusions are coincident with the theoretical properties mentioned in the previous sections. In other words, the new estimators (the local N-W-additive estimator and the local linear–additive estimator) are adaptive to local and global additivity. More precisely, when the model is global additive, the new ones can capture global additive information and have the same behavior as that of the additive estimator; when the model is local additive, our estimators can adapt to local additivity and behave as the local additive does; if the model is completely nonadditive, the new estimators have the same optimal property (the same convergence rate) as the local linear estimator (or the N-W estimator) has.

5.4. *Models with* $p = 10$

Example 5. We consider the regression function with dimension $p = 10$ as follows:

$$r(x) = \sum_{j=1}^{10} x_j^2 + \alpha x_1 \sum_{j=2}^{10} x_j, \quad \alpha = 0,\ 0.5,\ 1.$$

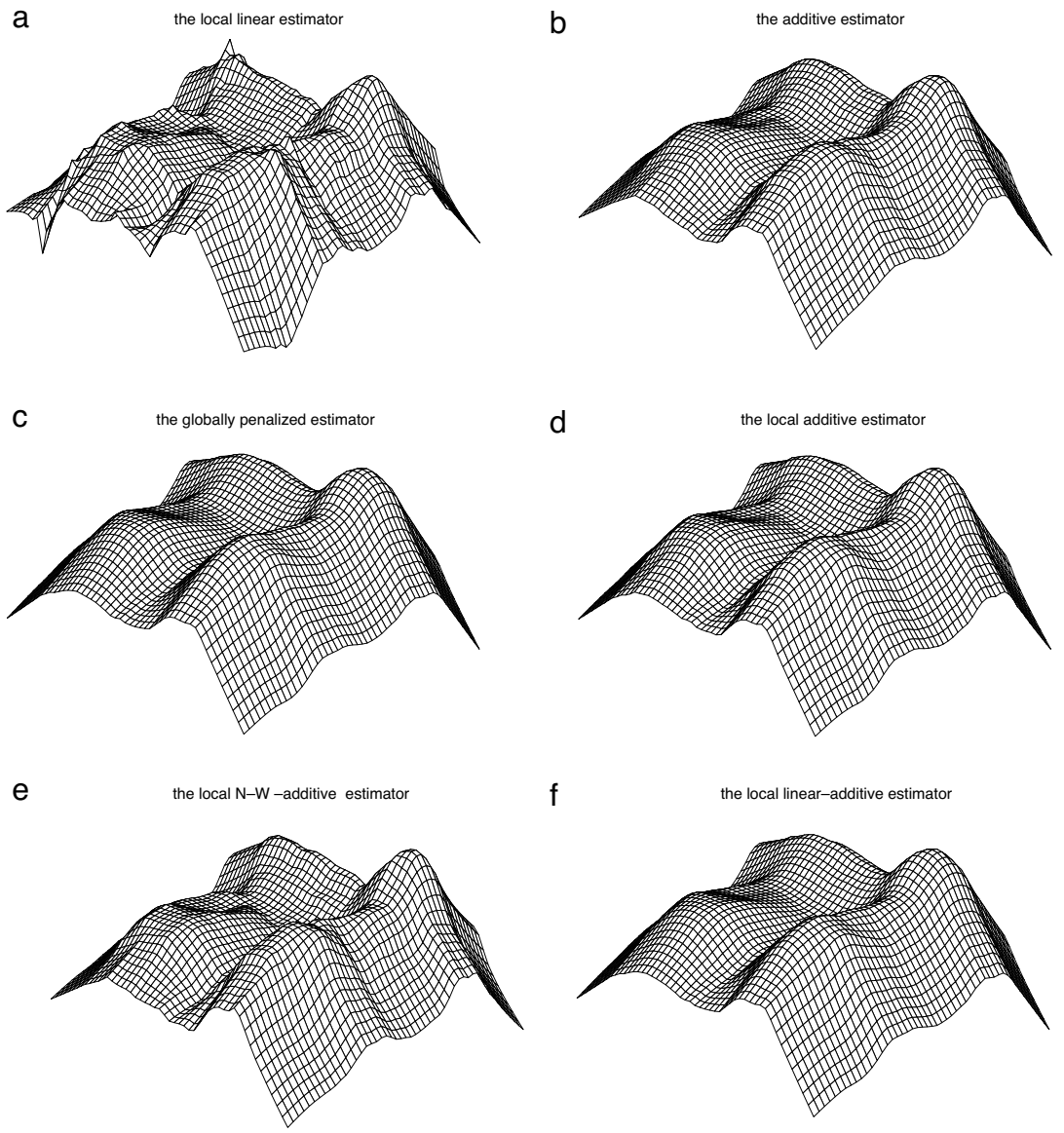

Fig. 4. Figures (a), (b), (c), (d), (e) and (f) are respectively the local linear estimator surface, the additive estimator surface, the globally penalized estimator surface, the local additive estimator surface, the local N-W-additive estimator surface and the local linear–additive estimator surface in Example 2.

For $\alpha = 0, 0.5$ and 1, the corresponding models are additive, approximately additive and nonadditive, respectively. The models were studied by [12] as well. In the corresponding regression models, the covariates are uniformly distributed on $[-1, 1]$ and the error is distributed as $N(0, 0.2^2)$, and then 2000 observations are obtained. The unconditional mean averaged squared error (MASE) is approximated with 20 runs of simulation, where the MASE is defined as MASE $= \frac{1}{n}\sum_{i=1}^{n}\{\hat{r}(x_i) - r(x_i)\}^2$. To reduce the computational burden, the estimators are evaluated at 50 design points randomly chosen at each simulation. The resulting relative standard error of the MASE-estimators is about 3%–5%.

Fig. 8 reports the performances of the estimators for the three different values of α, in which the broken curve, the full curve and the dot curve stand for the MASEs of the local linear estimator, the local additive estimator and the local linear–additive estimator, respectively. Since the curve of the MASE for the additive estimator is almost the same as that of the local additive estimator (the full curve), we then use the same curve (the full curve) together with a letter '*a*' at the end of the curve to denote the MASE of the additive estimator. The *x*-axis represents the smoothing parameter; for the local linear estimator, it is the bandwidth h and for the local additive estimator, it is η, where η is the size of the local region defined in Section 2.1. From Fig. 8, we see (1) the performance of the local linear estimator is not insensitive to the choices of α but it has relatively large MASE when the true model is equal to or approximate to an additive model (for $\alpha = 0, 0.5$); (2) the local additive estimator adapts to additivity, more precisely, the MASEs are relatively small when the true model is equal

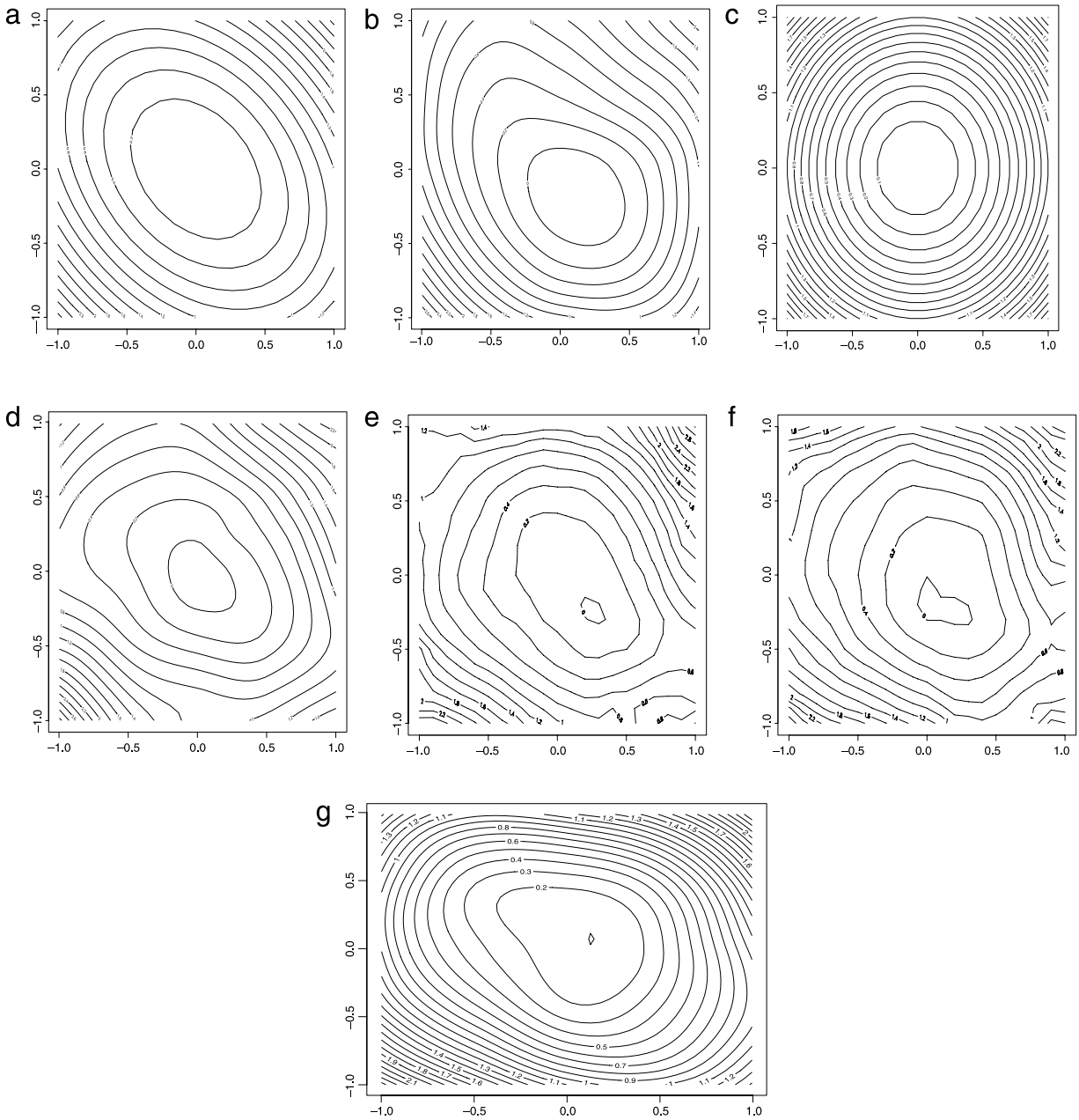

Fig. 5. Figure (a) is the contour plot of the true regression function. Figures (b), (c), (d), (e), (f) and (g) are respectively the contour plots of the local linear estimator surface, the additive estimator surface, the globally penalized estimator surface, the local additive estimator surface, the local N-W-additive estimator surface and the local linear–additive estimator surface in Example 3.

to or approximate to an additive model (for $\alpha = 0, 0.5$); (3) the local linear–additive estimator exceeds all the others and adapts to the additivity and nonadditivity (for all the choices of α).

Acknowledgments

The first author's research was supported by NNSF project (11171188 and 11231005) of China, Mathematical Finance-Backward Stochastic Analysis and Computations in Financial Risk Control of China (11221061), NSF and SRRF projects (ZR2010AZ001 and BS2011SF006) of Shandong Province of China and K C Wong-HKBU Fellowship Programme for Mainland China Scholars 2010–11.

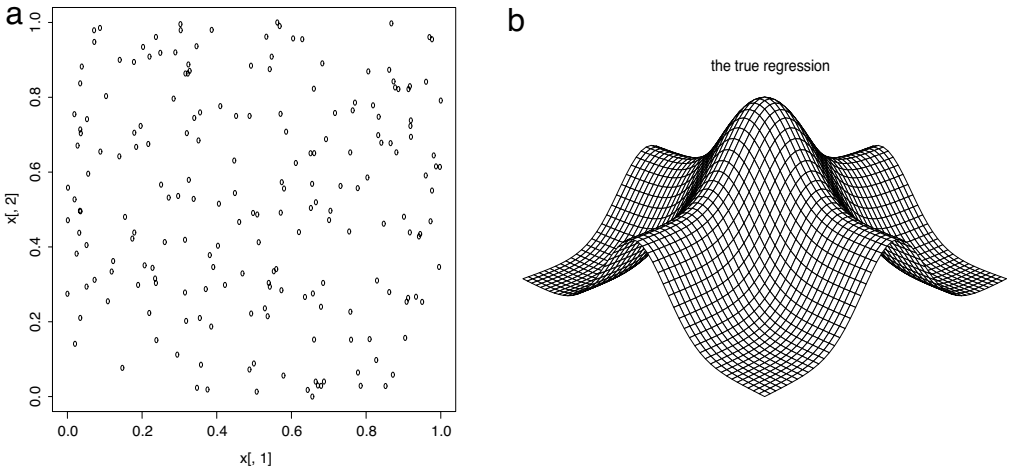

Fig. 6. Figures (a) and (b) are respectively the 200 random points of X and the true regression surface in Example 4.

Appendix. Proof

Proof of Theorem 1. *Part* 1: we first prove the conclusions for local linear–additive estimator $\hat{r}^{(0)}_{\lambda,\eta}(x)$.

(1) In this case, the weights satisfy $W_1 = 1 + O_p(\lambda\eta)$ and $W_2 = O_p(\lambda\eta)$. Then the conclusion follows directly.

(2) By the methods in [5,12], we have

$$\frac{1}{n}\tilde{\mathbf{X}}'\mathscr{W}_x\tilde{\mathbf{X}} = \Sigma + O_p(\tau_n) \quad \text{and} \quad \frac{1}{n_x}\tilde{\mathbf{U}}'(u)\mathscr{W}_u\tilde{\mathbf{U}}(u) = \Sigma + o_p(\tau_n),$$

where Σ is a known positive definite matrix and $\tau_n = h^2 + \frac{1}{\sqrt{nh^p}}$. Then

$$W_1 = \mathbf{e}_1'\left\{I + \frac{\lambda\eta(x)}{n_x}I\right\}^{-1} + O_p(\tau_n), \qquad W_1 = \mathbf{e}_1'\left\{I + \frac{\lambda\eta(x)}{n_x}I\right\}^{-1}\frac{\lambda\eta(x)}{n_x} + O_p(\tau_n). \tag{A.1}$$

By the properties of local linear estimator and local additive estimator, we have

$$\begin{cases}\tilde{\beta}(x) = (r(x), \dot{r}_1(x), \ldots, \dot{r}_p(x))' + (O_p(\tau_n), o_p(1), \ldots, o_p(1))', \\ \tilde{\boldsymbol{r}}_{ad,x}(u)|_{u=0} = (r(x), \dot{r}_1(x), \ldots, \dot{r}_p(x))' + (o_p(\tau_n), o_p(1), \ldots, o_p(1))',\end{cases} \tag{A.2}$$

where $\dot{r}_j(x)$ is the derivative of $r(x)$ with respect to $x^{(j)}$. Then, (A.1), (A.2) and (2.11) together complete the proof for Theorem 1(2).

(3) It is a direct result of (2.11).

(4) When $\lambda\eta \to \infty$, the weight functions satisfy

$$W_1 = O_p((\lambda\eta)^{-1}), \qquad W_2 = 1 + O_p((\lambda\eta)^{-1}). \tag{A.3}$$

Park and Seifert [12] shows

$$\tilde{r}_{lad,\eta}(x) = r(x) + O_p(\varsigma_n), \tag{A.4}$$

where $\varsigma_n = h^2 + \frac{1}{\sqrt{n\eta^{p-1}h}}$, provided that the model is globally additive. Then, (2.11), (A.3) and (A.4) lead to the result of Theorem 1(4).

Part 2: we now prove the conclusions for local-N-W additive estimator $r^{NW}_{\lambda,\eta}(x)$.

(1) The result directly follows from (2.12).

(2) It is known that under the conditions of theorem, we have

$$n^{-1}\sum_{i=1}^{n} Y_iK_h(X_i, x) = r(x)f(x) + O_p(\tau_n), \qquad n^{-1}\sum_{i=1}^{n} K_h(X_i, x) = f(x) + O_p(\tau_n),$$

where $f(x)$ is the density function of X and $\tau_n = h^2 + \frac{1}{\sqrt{nh^p}}$. On the other hand, Park and Seifert [12] show that if $p \le 8$, then

$$\tilde{r}_{lad,\eta}(x) = r(x) + O_p(\varsigma_n),$$

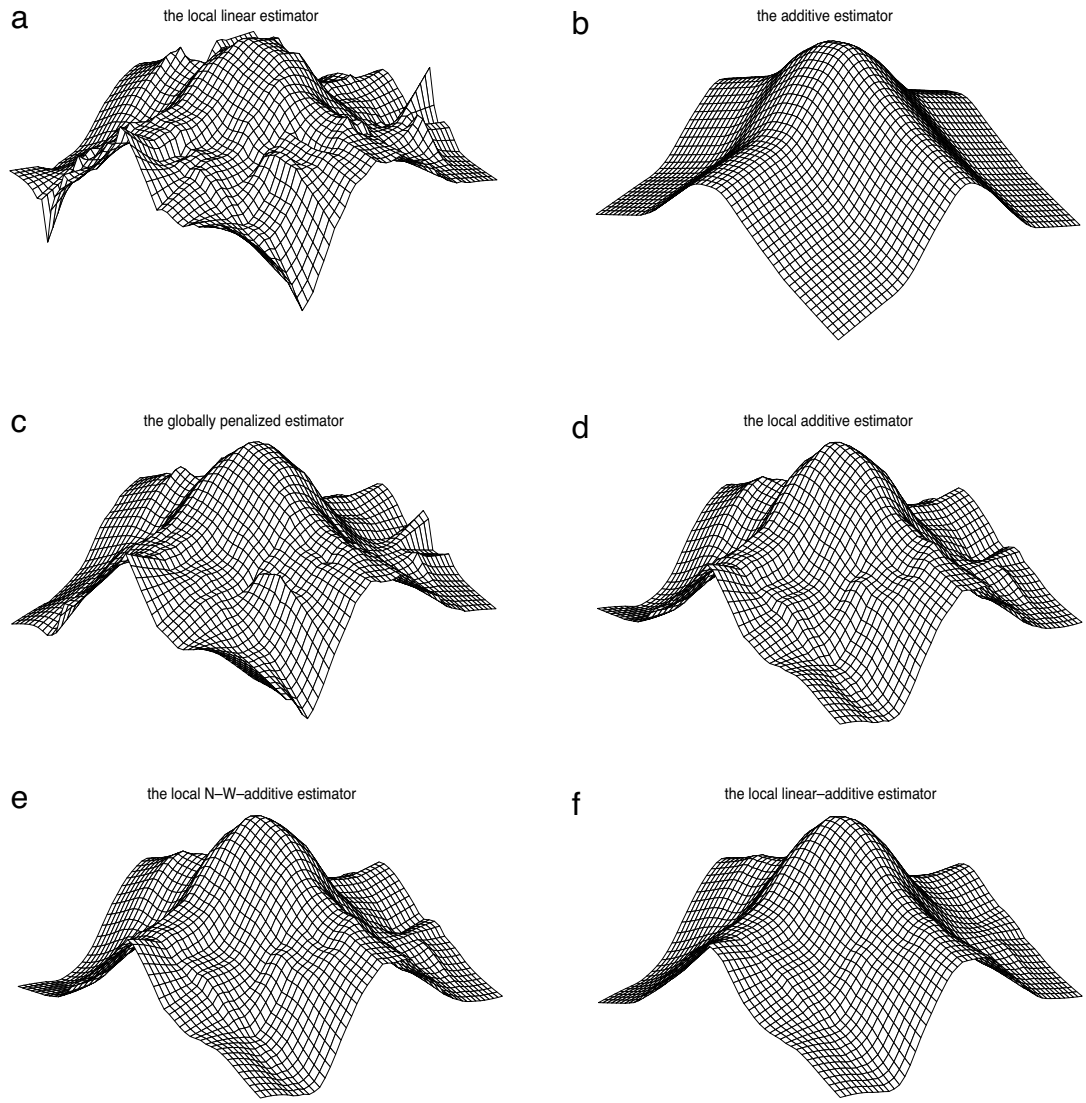

Fig. 7. Figures (a), (b), (c), (d), (e) and (f) are respectively the local linear estimator surface, the additive estimator surface, the globally penalized estimator surface, the local additive estimator surface, the local N-W-additive estimator regression surface and the local linear–additive estimator surface in Example 4.

where $\varsigma_n = o(\tau_n)$. Thus, for the case of $\lambda\eta$ being bounded, the above results and (2.12) lead to

$$\begin{aligned}\tilde{r}^{NW}_{\lambda,\eta}(x) &= \frac{rf + O_p(\tau_n) + \lambda\eta(r + O_p(\varsigma_n))}{f + O_p(\tau_n) + \lambda\eta} \\ &= \frac{r(f + n^{-1}\lambda\eta) + O_p(\lambda\eta\varsigma_n) + O_p(\tau_n)}{(f + \lambda\eta) + O_p(\tau_n)} \\ &= (f + n^{-1}\lambda\eta)^{-1}(1 + O_p(\tau_n))\{r(f + \lambda\eta) + O_p(\lambda\eta\varsigma_n) + O_p(\tau_n)\} \\ &= r + O_p(\tau_n), \end{aligned} \tag{A.5}$$

as required.

(3) The result directly follows from (2.12).

(4) When $\lambda\eta \to \infty$, by (A.5) and (A.4), we have

$$\tilde{r}^{NW}_{\lambda,\eta}(x) = r + O_p((\eta\lambda)^{-1}) + O_p(\varsigma_n).$$

Then we complete the proof. □

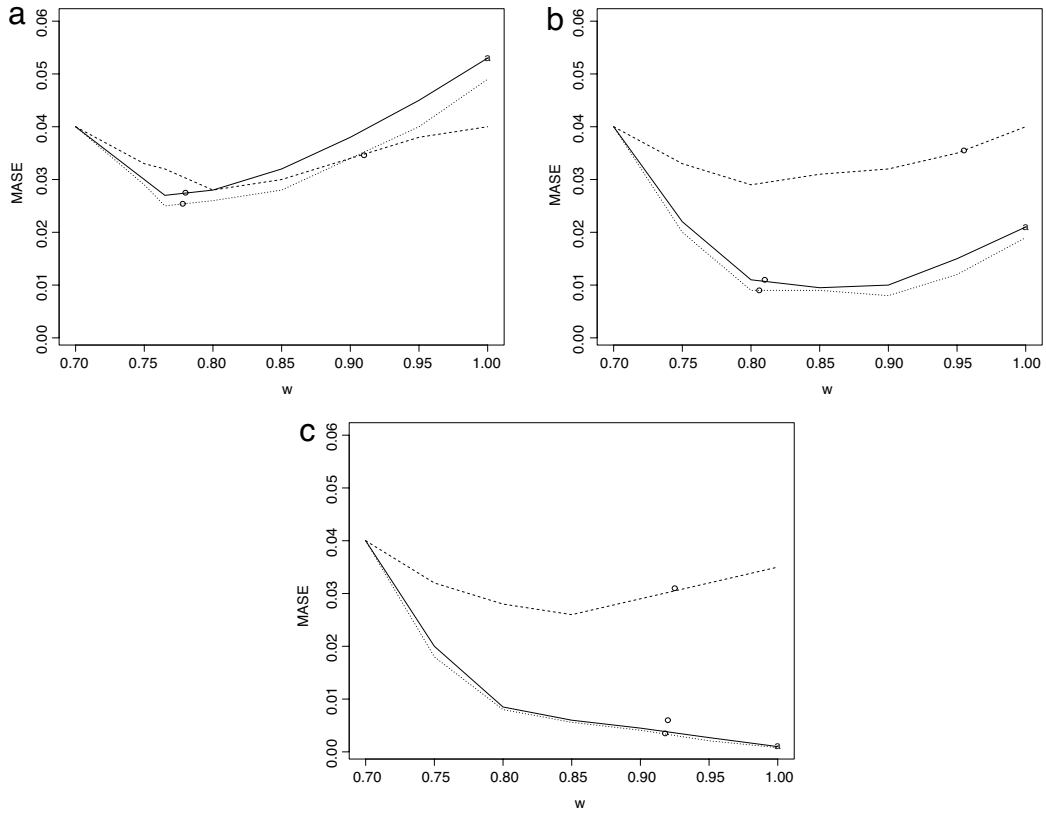

Fig. 8. Figures (a), (b) and (c) are the curves of MASEs according respectively to the models with $\alpha = 1, 0.5$ and 0. "- - -" is the MASE of the local linear estimator, "—" is the MASE of the local additive estimator, "—" with "a" is the MASE of the additive estimator and "· · ·" stands for the MASE the local linear–additive estimator. The positions "○" are the MASEs valued at the optimal smoothing parameter.

References

[1] A. Buja, T. Hastie, R. Tibshirani, Linear smoothers and additive models, Ann. Statist. 17 (1989) 435–555.
[2] J. Fan, Local linear regression smoothers and their minimax efficiencies, Ann. Statist. 21 (1993) 196–216.
[3] J. Fan, T. Gasser, I. Gijbels, M. Brockmann, J. Engel, Local polynomial regression: optimal kernels and asymptotic minimax efficiency, Ann. Inst. Statist. Math. 49 (1997) 79–99.
[4] J. Fan, I. Gijbels, Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation, J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1995) 371–394.
[5] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall, 1996.
[6] L.J. Horowitz, E. Mammen, Nonparametric estimation of an additive model with a link function, Ann. Statist. 32 (2004) 2412–2443.
[7] L. Lin, X. Cui, L. Zhu, An adaptive two-stage estimation method for additive models, Scand. J. Statist. 36 (2009) 248–269.
[8] O.B. Linton, W. Härdle, Estimating additive regression with known link functions, Biometrika 83 (1996) 529–540.
[9] O.B. Linton, J.P. Nielsen, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82 (1995) 93–101.
[10] E. Mammen, O. Linton, J. Nielsen, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, Ann. Statist. 27 (1999) 1443–1490.
[11] E. Mammen, J.S. Marron, B. Turlach, M. Wang, A general projection framework for constrained smoothing, Statist. Sci. 16 (2001) 232–248.
[12] J. Park, B. Seifert, Local additive estimation, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (2010) 171–191.
[13] C. Stone, Optimal rates of convergence for nonparametric estimators, Ann. Statist. 8 (1980) 1348–1360.
[14] C. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040–1053.
[15] C. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[16] C. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. 14 (1986) 590–606.
[17] M. Studer, B. Seifert, T. Gasser, Nonparametric regression penalizing deviations from additivity, Ann. Statist. 33 (2005) 1295–1329.