Electronic Journal of Statistics
Vol. 14 (2020) 3845–3868
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1760

Nonparametric regression with parametric help

Young K. Lee[*]

Kangwon National University
e-mail: youngklee@kangwon.ac.kr

Enno Mammen

Heidelberg University
e-mail: mammen@math.uni-heidelberg.de

Jens P. Nielsen[†]

City, University of London
e-mail: jens.nielsen.1@city.ac.uk

and

Byeong U. Park[‡]

Seoul National University
e-mail: bupark@stats.snu.ac.kr

Abstract: In this paper we propose a new nonparametric regression technique. Our proposal has common ground with existing two-step procedures in that it starts with a parametric model. However, our approach differs from others in the choice of parametric start within the parametric family. Our proposal chooses a function that is the projection of the unknown regression function onto the parametric family in a certain metric, while the existing methods select the best approximation in the usual L_2 metric. We find that the difference leads to substantial improvement in the performance of regression estimators in comparison with direct one-step estimation, irrespective of the choice of a parametric model. This is in contrast with the existing two-step methods, which fail if the chosen parametric model is largely misspecified. We demonstrate this with sound theory and numerical experiment.

AMS 2000 subject classifications: 62G08, 62G20.
Keywords and phrases: Regression function, bias, profiling technique, local linear estimation, cross-validatory bandwidth selectors.

Received February 2020.

[*] Research of Young K. Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2018R1A2B6001068).

[†] Research of Jens P. Nielsen was supported by the Institute and Faculty of Actuaries, London, UK.

[‡] Research of B. U. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2C3007355).

1. Introduction

We study a new approach to nonparametric regression. Let $m = \mathrm{E}\,(Y|X = \cdot)$ denote the true regression function and we assume that m is twice continuously differentiable with $\mathrm{E}\,m''(X)^2 < \infty$. Instead of estimating m directly by a local smoother, we choose a function g in a class of functions $\mathcal{G} = \{g : g''$ exists and $0 < \mathrm{E}\,g''(X)^2 < \infty\}$, and estimate a parameter θ_0 and a nonparametric function m_0 defined by

$$\theta_0 = \frac{\mathrm{E}g''(X)m''(X)}{\mathrm{E}g''(X)^2}, \quad m_0(x) = m(x) - \frac{\mathrm{E}g''(X)m''(X)}{\mathrm{E}g''(X)^2} \cdot g(x). \tag{1.1}$$

By definition m_0 satisfies

$$\mathrm{E}\,g''(X)m_0''(X) = 0 \tag{1.2}$$

and m is decomposed as

$$m(x) = \theta_0 g(x) + m_0(x). \tag{1.3}$$

For each given $g \in \mathcal{G}$, the decomposition (1.3) is unique under the constraint (1.2). To see this, suppose that $\theta g(\cdot) + \eta(\cdot) = 0$ and $\mathrm{E}\,g''(X)\eta''(X) = 0$. Then, $\theta^2 \mathrm{E}\,g''(X)^2 + \mathrm{E}\,\eta''(X)^2 = 0$ so that $\theta = 0$ and $\eta \equiv 0$.

The decomposition (1.3) with θ_0 and m_0 as given in (1.1) has a projection interpretation. For this, we consider an equivalence relation such that two functions f_1 and f_2 are equivalent if the difference is a linear function. The space of the equivalence classes forms a Hilbert space if we endow it with the inner product

$$\langle f_1, f_2 \rangle = \mathrm{E}f_1''(X)f_2''(X).$$

Let $\mathcal{H}_g$ be the space of equivalence classes spanned by g, i.e., $\mathcal{H}_g = \{c \cdot g(\cdot) : c \in \mathbb{R}\}$. Then, we get

$$\mathrm{Proj}(m|\mathcal{H}_g) = \frac{\mathrm{E}g''(X)m''(X)}{\mathrm{E}g''(X)^2}\, g = \theta_0\, g.$$

By estimating m through the decomposition (1.3), as described in the next section, we may afford a substantial room for reducing the bias. In this paper, we demonstrate the advantage with a local linear smoother, but the main idea can be extended to other local smoothers, see Remark 1 in Section 2. The conventional local linear estimator of m with a bandwidth b has the asymptotic bias $b^2 c_K m''(x)/2$ with a constant c_K depending on the kernel of the local linear smoother, while our new approach based on the decomposition (1.3) gives $b^2 c_K m_0''(x)/2$, see Proposition 1. This implies a reduction in the asymptotic average squared error since

$$\begin{aligned} \mathrm{E}\,m''(X)^2 &= \mathrm{E}\,(\theta_0 g''(X) + m_0''(X))^2 \\ &= \theta_0^2\, \mathrm{E}\,g''(X)^2 + \mathrm{E}\,m_0''(X)^2 \\ &> \mathrm{E}\,m_0''(X)^2. \end{aligned} \tag{1.4}$$

Our approach is related to the existing literature where two-step procedures have been proposed that consist of a parametric and a nonparametric fit of the data. These include [7, 5, 6, 10, 3, 12, 13]. All these papers considered the approach that finds a pilot estimator of a parametric model assuming that the chosen parametric model is correct, and then updates the parametric fit by a nonparametric adjustment. This was done by an additive, multiplicative or a more general adjustment based on nonparametric fits of the data or of the residuals from a parametric fit. The success of these two-step procedures turns out to depend highly on the choice of a pilot parametric model, which we illustrate in Section 3. Our approach is differentiated from these in that we do not fit a parametric model in the first step, but estimate θ_0 such that $\mathrm{E}g''(X)(m''(X) - \theta_0 g''(X)) = 0$. By doing this we can always reduce the bias for any choice of g with $\mathrm{E}\, g''(X)^2 > 0$, as is seen from (1.4).

The estimation of the model (1.3) is also of independent interest as it answers the question of what happens in the estimation of partially linear models $Y = \theta_0 g(Z) + m_0(X) + \varepsilon$ if the two covariates X and Z are identical or if they nearly coincide. Indeed, we use the profiling technique [11] to estimate (1.3), which is known as a useful technique of fitting partially linear models. We conjecture that our findings in this paper can be generalized to more complex semiparametric models, such as partially linear additive models [14] and partially linear single index models [2], with common covariates in the parametric and nonparametric components.

Furthermore, our idea of bias reduction by introducing parametric components in nonparametric regression functions may be extended to various structured nonparametric regression problems, such as in (generalized) additive models [9, 15], in (generalized) varying coefficient models [1, 8] and in single index models [4]. In these models one may also specify a parametric part $g(\theta, X)$ in a way that the parameter θ does not enter linearly, or allow for multivariate $\mathbf{X}$ or multi-dimensional $\boldsymbol{\theta}$. In this paper, to avoid technical complication and to make the presentation transparent, we first focus our discussion on the model where $g(\theta, X)$ is linear in $\theta \in \mathbb{R}$ for univariate X. Then, we discuss some extensions including nonlinear parametrisation, multivariate $\mathbf{X}$ and multi-dimensional $\boldsymbol{\theta}$ later in this paper.

This paper is organized as follows. In the next section we discuss two-step estimators based on a pilot estimator $\hat{\theta}$ that converges to θ_0 and develop asymptotic theory for such estimators. We also demonstrate that profiling for θ gives a consistent pilot estimator $\hat{\theta}$. In Section 3 we present numerical evidences that support the theory. Section 4 contains extensions of our approach to the cases mentioned above. Proofs are deferred to the Appendix.

2. Methodology and theory

Our estimation procedure consists of two steps. In the first step, the parameter θ_0 is estimated by an estimator $\hat{\theta}$. A choice of $\hat{\theta}$ will be discussed below. In the second step, a local smoother is applied to regress $Y - \hat{\theta} g(X)$ onto X. The result of the second step is our estimator of m_0. We take a local linear regression estimator as the local smoother.

Specifically, let $\mathcal{S}_b U$ denote the local linear kernel smoother with a baseline kernel function K and a bandwidth b taking X as the predictor and U as the response. It can be written as $\mathcal{S}_b U(x) = n^{-1} \sum_{i=1}^n w_b(x, X_i) U_i$, where

$$w_b(x,u) = \frac{\hat{\mu}_2(x;b) - \hat{\mu}_1(x;b)(u-x)/b}{\hat{\mu}_0(x;b)\hat{\mu}_2(x;b) - \hat{\mu}_1(x;b)^2} \cdot K_b(u-x),$$

$K_b(v) = K(v/b)/b$ and $\hat{\mu}_k(x;b) = n^{-1} \sum_{i=1}^n ((X_i - x)/b)^k K((X_i - x)/b)/b$ for integers $k \geq 0$. Define

$$\tilde{m}_b(x, \theta) = \mathcal{S}_b(Y - \theta g(X))(x)$$

for each θ. We propose

$$\hat{m} = \hat{\theta} g + \tilde{m}_b(\cdot, \hat{\theta}) \tag{2.1}$$

as an estimator of $m = \theta_0 g + m_0$.

The difference between our proposal and the existing two-step procedures is in the first step. For a direct comparison between the two approaches, suppose that one chooses a parametric model of the form $\{\theta g(\cdot) : \theta \in \mathbb{R}\}$. Then, the existing two-step procedures estimate θ_* where $\theta_* g$ is the best approximation of the true regression function m in the usual L_2 metric so that $\theta_* = \mathrm{E} m(X) g(X) / \mathrm{E} g(X)^2$, while ours estimates θ_0 as defined in (1.1).

We discuss the statistical properties of $\hat{m}$ at (2.1). Our first result states that $\hat{m}$ as an estimator of $m = \theta_0 g + m_0$ behaves like $\tilde{m}_b(\cdot, \theta_0)$ as an estimator of m_0 that utilizes the knowledge of θ_0 and for this it suffices to have a consistent estimator $\hat{\theta}$ of θ_0:

$$\hat{\theta} \ \to \ \theta_0 \quad \text{in probability.} \tag{2.2}$$

In particular, it is not required that $\hat{\theta}$ approximates θ_0 with a certain rate of convergence. For stating this result we make use of the following assumptions.

(A1) We observe i.i.d. copies (X_i, Y_i), $i = 1, \ldots, n$, of (X, Y), where X is supported on $[a_L, a_U]$ for some $-\infty < a_L < a_U < \infty$ and has a continuous strictly positive density f on $[a_L, a_U]$. For the error variable $\varepsilon = Y - m(X)$, it holds that $\mathrm{E}(\varepsilon | X) = 0$ and $\sigma^2(\cdot) = \mathrm{Var}(\varepsilon | X = \cdot)$ is continuous on $[a_L, a_U]$.

(A2) The function g and the true regression function m have continuous second-order derivatives and fulfill $0 < \mathrm{E}\, g''(X)^2 < \infty$ and $\mathrm{E}\, m_0''(X)^2 < \infty$.

(A3) The kernel K is a probability density function with compact support, say $[-1, 1]$.

(A4) For the bandwidth b it holds that $b \to 0$ and $nb \to \infty$.

Proposition 1. *Assume (A1)–(A4) and that an estimator $\hat{\theta}$ fulfills (2.2). Then, it holds that*

$$\hat{m}(x) - m(x) = \mathcal{S}_b \varepsilon(x) + \mathcal{S}_b(m_0(X))(x) - m_0(x) + o_P(b^2),$$

uniformly for $x \in [a_L, a_U]$.

We note that $\mathcal{S}_b\varepsilon + \mathcal{S}_b(m_0(X))$ is the local linear estimator $\tilde{m}_b(\cdot, \theta_0)$ of m_0 that is based on $(X_i, Y_i - \theta_0 g(X_i))$. The proposition demonstrates that the asymptotic variance and bias of $\hat{m}$ as an estimator of m are the same as those of $\tilde{m}_b(\cdot, \theta_0)$ as an estimator of m_0. The asymptotic variance equals that of the direct estimator $\mathcal{S}_b Y$. However, the asymptotic bias of $\hat{m}$ is $b^2\beta(x)m_0''(x)$, in contrast with $b^2\beta(x)m''(x)$ of the direct estimator $\mathcal{S}_b Y$, where $\beta(x)$ is a function of $\mu_k(x) = \int_{a_L}^{a_U} ((u-x)/b)^k K_b(u-x)\, du$. Thus, the average squared bias of $\hat{m}$ is smaller than that of $\mathcal{S}_b Y$, see (1.4). To maximize the reduction of the bias, one may choose $g \in \mathcal{G}$ that maximizes

$$\theta_0^2 \mathrm{E} g''(X)^2 = \left[\mathrm{E} \left(\frac{g''(X)}{\sqrt{\mathrm{E} g''(X)^2}} \cdot m''(X) \right) \right]^2, \tag{2.3}$$

which is equivalent to choosing g that minimizes

$$\mathrm{E} \left(\frac{g''(X)}{\sqrt{\mathrm{E} g''(X)^2}} - m''(X) \right)^2 = 1 + \mathrm{E} m''(X)^2 - 2\, \mathrm{E} \left(\frac{g''(X)}{\sqrt{\mathrm{E} g''(X)^2}} \cdot m''(X) \right).$$

Remark 1. *The main idea behind the bias reduction implied by Proposition 1 can be applied to other local smoothers. For example, in the case of the pth order local polynomial smoother with an odd p, we choose a function g such that $0 < \mathrm{E}\, g^{(p+1)}(X)^2 < \infty$, where $\eta^{(k)}$ for a function η denotes its kth derivative. Then, there is a unique decomposition $m = \theta_0 g + m_0$ under the constraint $\mathrm{E}\, g^{(p+1)}(X) m_0^{(p+1)}(X) = 0$, where θ_0 and m_0 are redefined in an obvious way. The estimator $\hat{m}$ as defined in (2.1), with a consistent estimator $\hat{\theta}$ of θ_0 and $\tilde{m}_b(\cdot, \hat{\theta})$ now obtained by applying the pth order local polynomial smoother, admits the uniform expansion in Proposition 1 with a remainder of order $o_P(b^{p+1})$. The leading bias of the local polynomial estimator applied directly to Y_i equals $b^{p+1}\beta(x)m^{(p+1)}(x)$ for some function β, while the estimator based on the decomposition gives $b^{p+1}\beta(x)m_0^{(p+1)}(x)$. In this case,*

$$\mathrm{E}\, m^{(p+1)}(X)^2 - \mathrm{E}\, m_0^{(p+1)}(X)^2 = \frac{\left(\mathrm{E}\, g^{(p+1)}(X) m^{(p+1)}(X)\right)^2}{\mathrm{E}\, g^{(p+1)}(X)^2}.$$

It remains to find a consistent estimator of θ_0. Recall that θ_0 we need to estimate is the one that fulfills $\mathrm{E}\, g''(X) m''(X, \theta) = 0$, among all θ in the decompositions $m = \theta g + m(\cdot, \theta)$, where $m(x, \theta) = m(x) - \theta g(x)$. We achieve this by using the profiling technique. The profiling technique has been proposed for the partially linear model $Y = \theta_0 g(Z) + m_0(X) + \varepsilon$ with $Z \neq X$. The profile least squares estimator of θ_0 is given by

$$\hat{\theta}_h = \arg\min_\theta \sum_{i=1}^n \left(Y_i - \theta g(X_i) - \tilde{m}_h(X_i, \theta)\right)^2, \tag{2.4}$$

where h is a second bandwidth, which may be chosen to be the same as b in (2.1). The next proposition demonstrates that $\hat{\theta}_h$ is a consistent estimator of θ_0. We need the following additional assumption for the statement of this proposition.

(A5) For the bandwidth h it holds that $h \to 0$ and $nh^4 \to \infty$.

Proposition 2. *Assume (A1)–(A3) and (A5). Then, $\hat{\theta}_h \to \theta_0$ in probability.*

Remark 2. *The condition $nh^4 \to \infty$ in (A5) is needed to take care of the properties of the local linear estimator at the boundary of the interval $[a_L, a_U]$. We note that, although the local linear smoother $\mathcal{S}_h$ affords the same order of biases $O(h^2)$ at the boundary and in the interior, their constant factors are still different. The condition can be relaxed if we remove boundary regions in the definitions of $\mathcal{S}_h$ and the profile estimator of θ_0 and if the pilot model g and the density f are sufficiently smooth. In such a case the leading stochastic terms of the magnitude $n^{-1/2}h^{-2}$ in an expansion of $\hat{\theta}_h - \theta_0$ cancel each other, which may be deduced from our asymptotic analysis presented in the Appendix.*

From our propositions we get the following corollary.

Corollary 1. *Assume (A1)–(A5). Then, we have for $\hat{m} = \hat{\theta}_h g + \tilde{m}_b(\cdot, \hat{\theta}_h)$ that*

$$\hat{m}(x) - m(x) = \mathcal{S}_b \varepsilon(x) + \mathcal{S}_b(m_0(X))(x) - m_0(x) + o_P(b^2),$$

uniformly for $x \in [a_L, a_U]$.

We have again the interpretation that we already formulated after the statement of Proposition 1. Also by profile estimation we get an estimator of $m = \theta_0 g + m_0$ that optimally chooses one from a class of local linear estimators. Thus, profile estimation works quite well also in the degenerate case $X = Z$ of the partially linear model $Y = \theta_0 g(Z) + m_0(X) + \varepsilon$.

The estimator $\hat{m} = \hat{\theta}_h g + \tilde{m}_b(\cdot, \hat{\theta}_h)$ depends on the bandwidths b and h. We may take $h = b$ for simplicity and choose a common bandwidth by cross validation. We employed this strategy in our simulation and found that it worked quite well, see Section 3. To indicate its dependence on b we write $\hat{m}_b$ for $\hat{m}$ with $h = b$. Let $\hat{m}_b^{(-i)}$ denote the leave-one-out version of $\hat{m}_b$ that makes use of only the observations $\{(X_{i'}, Y_{i'}) : i' \neq i\}$. We choose the bandwidth b by minimizing a CV criterion. The CV bandwidth $\hat{b}$ is defined by

$$\hat{b} = \arg\min_{b \in B_n} \sum_{i=1}^{n} \left(Y_i - \hat{m}_b^{(-i)}(X_i) \right)^2. \tag{2.5}$$

Our estimator of m is then given by $\hat{m}_{\hat{b}}$. We will check whether the cross validation approach works in the next section by simulation.

As discussed before the statement of Proposition 1, we only need a consistent estimator of θ_0. Clearly, there are alternatives to the profiling approach. One example would be to start with a pilot estimator of m'' and then plug the estimator into the definition of θ_0 at (1.1).

3. Simulation results

The purpose of this simulation study is to support the asymptotic theory we demonstrated in Section 2 and to compare our approach with other competitors.

This was done with the CV bandwidth selectors introduced also in the previous section. We considered three models that generated (X_i, Y_i). The first model was

$$Y_i = \sin(\pi X_i) + \rho X_i + \lambda \cos(\pi X_i) + \varepsilon_i \tag{3.1}$$

with X_i being generated from the uniform distribution on $[a_L, a_U]$ with $a_L = 0$ and $a_U = 1$, and ε_i from $N(0, \sigma^2)$ independent of X_i. For noise level we made two choices, $\sigma = 0.1$ and $\sigma = 0.5$. We made three choices for λ: $\lambda = 0, 0.5, 1$, and three choices for ρ: $\rho = 0, 1, 2$. The true regression curves are depicted in the three panels in the top row of Figure 1. In the case where $\sigma = 0.1$, the values of the noise-to-signal ratio $\mathrm{NSR} = \mathrm{Var}(\varepsilon)/\mathrm{Var}(m(X))$ are 0.106, 0.056, 0.023 for $\rho = 0, 1, 2$, respectively, when $\lambda = 0$; 0.046, 0.100, 0.068, respectively, when $\lambda = 0.5$; 0.017, 0.037, 0.085, respectively, when $\lambda = 1$. The values in the case where $\sigma = 0.5$ may be obtained by multiplying these values by 25. In the first application of our approach to the model (3.1), we took $g(x) = \sin(\pi x)$. According to (1.1), this choice gives $\theta_0 = 1$ and $m_0(x) = \rho x + \lambda \cos(\pi x)$.

We compared our approach with a parametric fit, the direct local linear fit and the two-step procedure starting with a parametric fit to the model $\mathrm{E}(Y_i|X_i) = \theta g(X_i)$ and then making a nonparametric adjustment. The parametric fit we considered in this comparison is $\tilde{m}^{\mathrm{pa}} = \tilde{\theta} g$ where $\tilde{\theta}$ minimizes $\sum_{i=1}^n (Y_i - \theta g(X_i))^2$. We denote the direct local linear smoother by $\tilde{m}^{\mathrm{ll}}_{\tilde{h}} = \mathcal{S}_{\tilde{h}}(Y)$, where

$$\tilde{h} = \arg\min_{h \in H_n} \sum_{i=1}^n \left(Y_i - \mathcal{S}_h^{(-i)}(Y)(X_i) \right)^2 . \tag{3.2}$$

The two-step procedure with $\tilde{m}^{\mathrm{pa}}$ as a parametric start is $\tilde{m}^{\mathrm{ts}}_{\tilde{b}} = \tilde{\theta} g + \tilde{m}_{\tilde{b}}(\cdot, \tilde{\theta})$, where $\tilde{b}$ is chosen by minimizing the CV criterion $\sum_{i=1}^n (Y_i - \tilde{m}_b^{\mathrm{ts}(-i)}(X_i))^2$. For comparison of these estimators, we computed

$$\mathrm{MISE}(\bar{m}) := \mathrm{E} \int_{a_L}^{a_U} (\bar{m}(x) - m(x))^2 \, dx$$

for each $\bar{m}$ of $\hat{m}_{\hat{b}}$, $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$, $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ and $\tilde{m}^{\mathrm{pa}}$. Tables 1 and 2 give the Monte Carlo approximations of the MISE values. They also contain the Monte Carlo approximations of the values of $\mathrm{ISB}(\bar{m}) := \int_{a_L}^{a_U} (\mathrm{E}\, \bar{m}(x) - m(x))^2 \, dx$ and $\mathrm{IV}(\bar{m}) := \int_{a_L}^{a_U} \mathrm{Var}(\bar{m}(x)) \, dx$.

From the tables we note that the bias of $\tilde{m}^{\mathrm{pa}}$ does not change as n or the noise level σ varies, which is well expected. We also note that the properties of our proposal $\hat{m}_{\hat{b}}$ and the direct local linear estimator $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ do not change as ρ varies. This stems basically from the property of the weight w_b that

$$n^{-1} \sum_{i=1}^n w_b(x, X_i) X_i = n^{-1} \sum_{i=1}^n w_b(x, X_i) x = x \tag{3.3}$$

so that $\mathcal{S}_b(a + bX)(x) = a + bx$ for any real numbers a and b. Because of (3.3) we see that, for the direct local linear smoother, $\tilde{m}^{\mathrm{ll}}_h(x) - m(x) = \mathcal{S}_h(Y)(x) - m(x)$

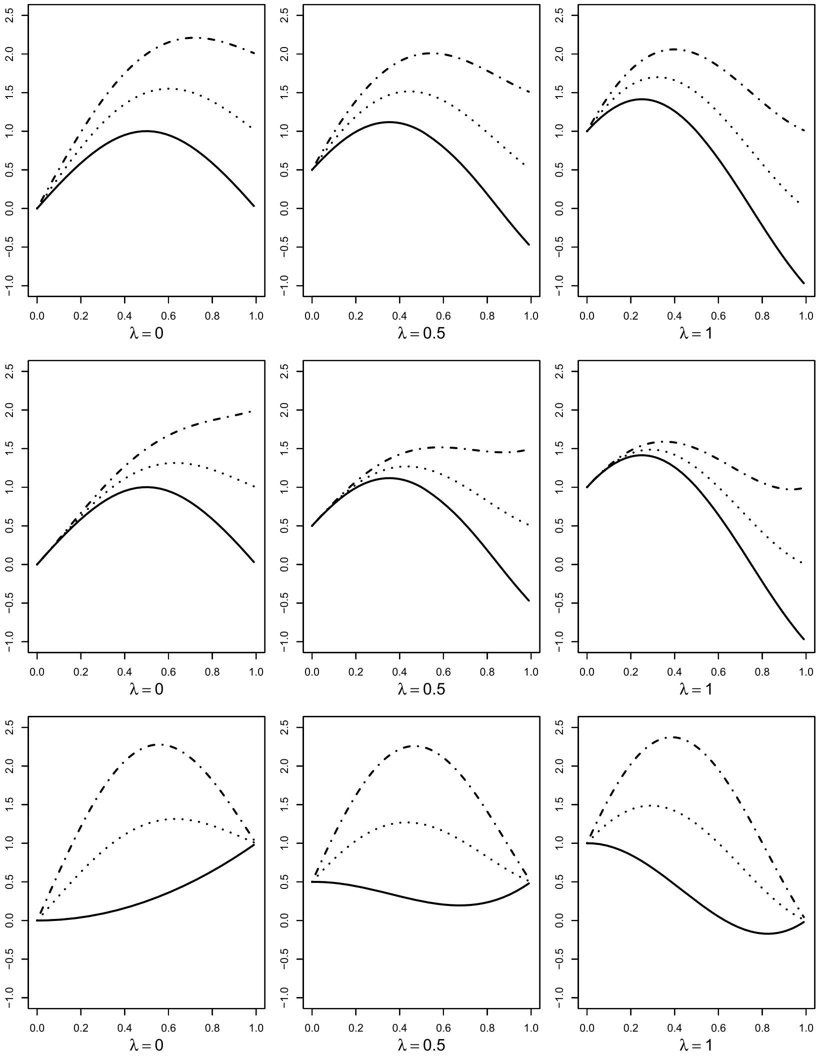

Fig 1. *True regression curves: top row for the model (3.1), middle row for the model (3.6) and bottom row for the model (3.7). Solid curves correspond to $\rho = 0$, dotted to $\rho = 1$ and dot-dashed to $\rho = 2$.*

does not involve ρ. As for our estimator $\hat{m}_{\hat{b}} = \hat{\theta} g + \tilde{m}_b(\cdot, \hat{\theta})$, it holds that

$$\begin{aligned} \hat{m}_b(x) - m(x) &= \hat{\theta}_b g(x) + \mathcal{S}_b(Y - \hat{\theta}_b g(X))(x) - m(x) \\ &= \lambda \left[\mathcal{S}_b(\cos(\pi X))(x) - \cos(\pi x) \right] \\ &\qquad + (1 - \hat{\theta}_b) \left[\mathcal{S}_b(g(X))(x) - g(x) \right] + \mathcal{S}_b(\varepsilon)(x), \end{aligned} \tag{3.4}$$

which does not depend on ρ. Furthermore, the profiling estimator $\hat{\theta}_b$ is also

TABLE 1

For the model (3.1) with $g(x) = \sin(\pi x)$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.1$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	0.35	0.32	0.91	0.09	0.71	0.70	0.94	126	0.86	0.87	1.00	502
		ISB	0.00	0.00	0.17	0.00	0.13	0.11	0.19	125	0.16	0.15	0.21	500
		IV	0.35	0.32	0.74	0.08	0.58	0.58	0.75	0.69	0.70	0.72	0.79	2.50
	400	MISE	0.10	0.09	0.26	0.02	0.21	0.21	0.27	125	0.25	0.25	0.28	501
		ISB	0.00	0.00	0.04	0.00	0.03	0.03	0.04	125	0.04	0.04	0.05	500
		IV	0.10	0.09	0.22	0.02	0.18	0.18	0.22	0.22	0.21	0.21	0.23	0.76
$\rho = 1$	100	MISE	0.35	6.29	0.91	126	0.71	1.21	0.94	53.3	0.86	0.93	1.00	231
		ISB	0.00	3.36	0.17	126	0.13	0.43	0.19	53.0	0.16	0.21	0.21	230
		IV	0.35	2.92	0.74	0.48	0.58	0.78	0.75	0.27	0.70	0.72	0.79	1.28
	400	MISE	0.10	5.26	0.26	126	0.21	0.29	0.27	53.1	0.25	0.27	0.28	230
		ISB	0.00	2.88	0.04	126	0.03	0.10	0.04	53.0	0.04	0.06	0.05	230
		IV	0.10	2.38	0.22	0.15	0.18	0.19	0.22	0.07	0.21	0.21	0.23	0.33
$\rho = 2$	100	MISE	0.35	15.8	0.91	505	0.71	2.33	0.94	233	0.86	1.20	1.00	213
		ISB	0.00	7.32	0.17	503	0.13	1.27	0.19	233	0.16	0.42	0.21	212
		IV	0.35	8.49	0.74	1.66	0.58	1.06	0.75	0.64	0.70	0.78	0.79	0.84
	400	MISE	0.10	9.98	0.26	504	0.21	0.53	0.27	233	0.25	0.35	0.29	212
		ISB	0.00	4.82	0.04	503	0.03	0.30	0.04	233	0.04	0.13	0.05	212
		IV	0.10	5.16	0.22	0.58	0.18	0.23	0.22	0.22	0.21	0.22	0.23	0.21

invariant to the change of ρ since $Y_i - \theta g(X_i) - \tilde{m}_b(X_i, \theta)$ for all θ and $1 \le i \le n$ do not involve ρ, see (2.4). Similarly, the CV criteria at (3.2) and (2.5) do not depend on ρ. This explains why $\hat{m}_{\hat{b}}$ and $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ do not change as ρ varies. This is not the case with $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$, however, since $\tilde{m}^{\mathrm{pa}}$ depends on ρ, so does $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ that has $\tilde{m}^{\mathrm{pa}}$ as a parametric start.

Our theory in Section 2 tells that there is a relatively larger reduction in the bias of our proposal in comparison with that of the direct local linear estimator if

$$\left[\frac{\mathrm{E}\,(g''(X)m''(X))}{\sqrt{\mathrm{E}\,(g''(X)^2)}\sqrt{\mathrm{E}\,(m''(X)^2)}}\right]^2 = \frac{1}{1+\lambda^2}$$

is larger, see (2.3). This is evident in the numerical results. The ISB values of $\hat{m}_{\hat{b}}$ in the tables are less than those of $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ for the three values of λ and the relative difference is the largest when $\lambda = 0$ and decreases as λ increases. We also find that $\hat{m}_{\hat{b}}$ has smaller variance as well. The smaller variance achieved by our proposal is due to the reduced bias and the CV bandwidth choice $\hat{b}$ that trades off the bias and the variance. Theoretically, with a fixed bandwidth applied to both methods, the variance of our proposal is asymptotically the same as that of the direct local linear estimator while the bias of the first is smaller than that

Table 2

For the model (3.1) with $g(x) = \sin(\pi x)$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.5$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	8.73	7.92	15.2	2.13	9.78	9.48	15.5	128	12.4	14.0	16.1	505
		ISB	0.04	0.06	2.27	0.02	0.74	0.74	2.54	125	1.87	1.79	2.74	500
		IV	8.69	7.86	12.9	2.11	9.04	8.74	13.0	2.68	10.5	12.2	13.4	4.46
	400	MISE	2.55	2.30	4.24	0.53	3.49	3.38	4.38	126	4.14	4.11	4.58	501
		ISB	0.01	0.01	0.44	0.00	0.39	0.38	0.51	125	0.54	0.51	0.54	500
		IV	2.54	2.29	3.80	0.53	3.10	3.00	3.87	0.79	3.60	3.60	4.04	1.41
$\rho = 1$	100	MISE	8.73	20.4	15.2	128	9.77	20.1	15.5	55.4	12.4	17.3	16.1	233
		ISB	0.04	8.35	2.27	126	0.74	8.35	2.54	53.1	1.87	4.86	2.74	230
		IV	8.69	12.0	12.9	2.53	9.03	11.8	13.0	2.29	10.5	12.4	13.4	3.26
	400	MISE	2.55	14.1	4.24	126	3.49	9.30	4.38	53.7	4.14	4.79	4.58	231
		ISB	0.01	8.61	0.44	126	0.39	4.01	0.51	53.1	0.54	1.03	0.54	230
		IV	2.54	5.46	3.80	0.59	3.10	5.29	3.87	0.58	3.60	3.76	4.04	0.91
$\rho = 2$	100	MISE	8.73	47.6	15.2	507	9.78	49.3	15.5	235	12.4	37.8	16.1	215
		ISB	0.04	25.2	2.27	503	0.74	28.9	2.54	232	1.87	17.1	2.74	212
		IV	8.69	22.4	12.9	3.73	9.04	20.4	13.0	2.68	10.5	20.7	13.4	2.84
	400	MISE	2.55	38.8	4.24	504	3.49	26.2	4.38	233	4.14	7.57	4.58	213
		ISB	0.01	24.1	0.44	503	0.39	14.1	0.51	233	0.54	2.82	0.54	212
		IV	2.54	14.7	3.80	0.95	3.10	12.1	3.87	0.65	3.60	4.75	4.04	0.71

of the latter. The smaller bias then gives our proposal some room for sacrificing bias to reduce variance by increasing bandwidth in trading off the bias and the variance. Thus, the CV criteria tend to choose $\hat{b} > \tilde{h}$, which results in the smaller variance as well as the smaller bias. This is well demonstrated in Figure 2 for the case where $\lambda = 0$, $\rho = 2$ and $\sigma = 0.1$, which depicts the distributions of the CV bandwidth choices $\hat{b}$ (left) for our proposal, $\tilde{b}$ for the two-step estimator (middle) and $\tilde{h}$ for the direct local linear estimator (right). Recall that the CV bandwidth selectors $\hat{b}$ and $\tilde{h}$ do not depend on ρ as we discussed above.

Our proposal exhibits the best performance in all cases except the case $(\lambda, \rho) = (0, 0)$ where the parametric method is the best as expected. For the three cases of $\rho = 0$ ($\lambda = 0$, 0.5 and 1), our proposal and the two-step procedure show comparable performance. The success of $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ when $\rho = 0$ is mainly due to the fact that $g(X)$ is orthogonal to $m_0(X)$ in the space of square-integrable random variables, i.e., $\mathrm{E}(g(X)m_0(X)) = \rho/\pi = 0$. In this case, the estimation of θ_0 and m_0 in $m = \theta_0 g + m_0$ may be done by marginal regression. The marginal regression for θ_0 is simply the parametric fit that minimizes $\sum_{i=1}^{n}(Y_i - \theta g(X_i))^2$ with respect to θ. Thus, in this case the minimizer $\tilde{\theta}$, which is the parametric start of the two-step estimator $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$, approximates well the true $\theta_0 = 1$ at the

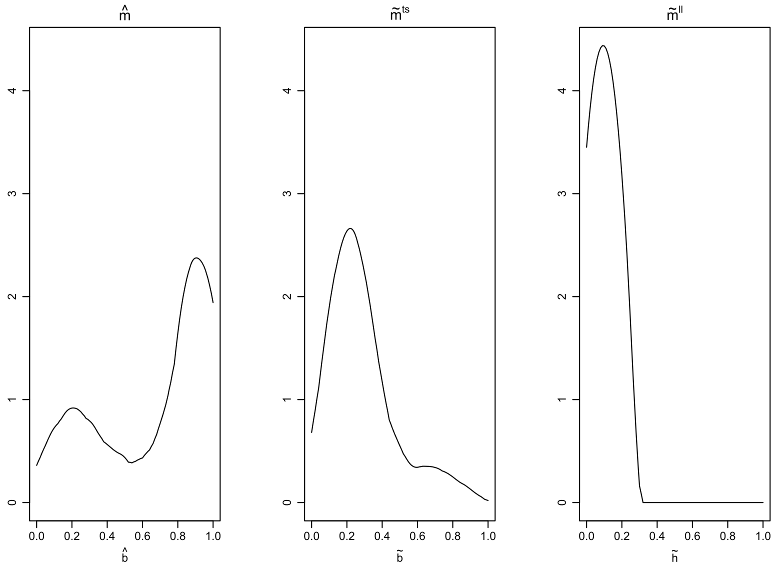

FIG 2. *The distributions of the CV bandwidth selectors for* $n = 400$ *in the case where* $\lambda = 0$, $\rho = 2$ *and* $\sigma = 0.1$ *in the model (3.1). From left to right,* $\hat{b}$ *for our proposal* $\hat{m}_{\hat{b}}$, $\tilde{b}$ *for the two-step procedure* $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ *and* $\tilde{h}$ *for the direct local linear estimator* $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$.

parametric rate. However, the two-step estimator $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ with the CV choice $\tilde{b}$ deteriorates very fast as ρ departs from $\rho = 0$. The performance of $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ is even worse than the direct local linear $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ when $\rho > 0$. This is in contrast with our proposal $\hat{m}_{\hat{b}}$ whose performance does not change as ρ varies. Another point to note is that the 'cosine similarity' between g and m_0 in the L^2 space, which is given by

$$\frac{\mathrm{E}\left(g(X)m_0(X)\right)}{\sqrt{\mathrm{E}g(X)^2 \cdot \mathrm{E}m_0(X)^2}} = \frac{\rho}{\sqrt{\pi^2(\rho^2/6 + \lambda^2/4 - 2\rho\lambda/\pi^2)}},$$

does not change as ρ increases on $(0, \infty)$ when $\lambda = 0$. Nevertheless, the performance of $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ gets quite worse as $\rho > 0$ increases. In fact, the L^2 distance between the parametric model and the true function m is given by

$$\min_{\theta} \int_0^1 \left(m(x) - \theta g(x)\right)^2 \, dx = \left(\frac{1}{3} - \frac{2}{\pi^2}\right)\rho^2 - \frac{4}{\pi^2}\rho\lambda + \frac{1}{2}\lambda^2, \qquad (3.5)$$

so that the L^2 distance increases when ρ increases from 0, in case $\lambda = 0$. This suggests that the performance of $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ depends on the departure of g from m, not only in terms of the cosine similarity but also in terms of the L^2 distance. In practice, one may plot the data, choose a good parametric model and then use the fitted model to improve the parametric estimator by nonparametric estimation at the second stage. In doing so, it may be helpful to include more

TABLE 3

For the model (3.1) with $g(x) = x^2$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\rm ts}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\rm ll}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\rm pa}$), for the error level $\sigma = 0.1$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\rm ts}_{\tilde{b}}$	$\tilde{m}^{\rm ll}_{\tilde{h}}$	$\tilde{m}^{\rm pa}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\rm ts}_{\tilde{b}}$	$\tilde{m}^{\rm ll}_{\tilde{h}}$	$\tilde{m}^{\rm pa}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\rm ts}_{\tilde{b}}$	$\tilde{m}^{\rm ll}_{\tilde{h}}$	$\tilde{m}^{\rm pa}$
$\rho = 0$	100	MISE	0.63	1.01	0.91	320	0.76	1.08	0.94	587	0.89	1.08	1.00	1004
		ISB	0.10	0.27	0.17	316	0.13	0.33	0.19	583	0.16	0.32	0.21	999
		IV	0.53	0.74	0.74	3.80	0.63	0.75	0.75	4.26	0.73	0.76	0.79	4.80
	400	MISE	0.20	0.29	0.26	317	0.23	0.29	0.27	584	0.26	0.30	0.28	1001
		ISB	0.03	0.07	0.04	316	0.04	0.08	0.04	583	0.05	0.08	0.05	1000
		IV	0.17	0.22	0.22	0.84	0.19	0.21	0.22	0.98	0.21	0.22	0.23	1.16
$\rho = 1$	100	MISE	0.63	1.10	0.91	504	0.76	1.14	0.94	821	0.89	1.16	1.00	1290
		ISB	0.10	0.32	0.17	498	0.13	0.38	0.19	815	0.16	0.38	0.21	1283
		IV	0.53	0.78	0.74	5.92	0.63	0.76	0.75	6.48	0.73	0.78	0.79	7.12
	400	MISE	0.20	0.31	0.26	499	0.23	0.32	0.27	817	0.26	0.32	0.28	1284
		ISB	0.03	0.08	0.04	498	0.04	0.10	0.04	815	0.05	0.10	0.05	1282
		IV	0.17	0.23	0.22	1.30	0.19	0.22	0.22	1.48	0.21	0.22	0.23	1.69
$\rho = 2$	100	MISE	0.63	1.18	0.91	729	0.76	1.20	0.94	1098	0.89	1.24	1.00	1616
		ISB	0.10	0.36	0.17	720	0.13	0.40	0.19	1088	0.16	0.43	0.21	1606
		IV	0.53	0.82	0.74	8.52	0.63	0.79	0.75	9.18	0.73	0.81	0.79	9.92
	400	MISE	0.20	0.33	0.26	722	0.23	0.34	0.27	1091	0.26	0.34	0.29	1609
		ISB	0.03	0.09	0.04	720	0.04	0.11	0.04	1088	0.05	0.11	0.05	1606
		IV	0.17	0.24	0.22	1.87	0.19	0.23	0.22	2.08	0.21	0.23	0.23	2.32

parameters in candidate parametric models. The flexible parametric modeling may be also beneficial to our approach. We give a short discussion on this extension at the end of Section 4.

In the second application to the model (3.1) we considered $g(x) = x^2$. This was to see how our approach works when one picks out a parametric model that is not a part of the underlying regression function. We note that the quadratic function is absent in the true regression function $m(x) = \sin(\pi x) + \rho x + \lambda \cos(\pi x)$. With this choice, $\theta_0 = -\pi$. Tables 3 and 4 summarize the results of the second application. They show that our approach exhibits the best performance in all cases. As in the first application, our proposal $\hat{m}_{\hat{b}}$ and the direct local linear estimator $\tilde{m}^{\rm ll}_{\tilde{h}}$ are invariant to the change of ρ. The results also confirms our theory that $\hat{m}_{\hat{b}}$ gets better and the gap in the performance between $\hat{m}_{\hat{b}}$ and $\tilde{m}^{\rm ll}_{\tilde{h}}$ gets larger as

$$\left[\frac{\mathrm{E}\left(g''(X) m''(X)\right)}{\sqrt{\mathrm{E}\left(g''(X)^2\right)} \sqrt{\mathrm{E}\left(m''(X)^2\right)}} \right]^2 = \frac{8}{\pi^2 (1 + \lambda^2)}$$

increases, i.e., as λ decreases. As for the two-step estimator $\tilde{m}^{\rm ts}_{\tilde{b}}$ with $\tilde{m}^{\rm pa}$ as a parametric start, its performance does not change much as ρ or λ varies,

TABLE 4

For the model (3.1) with $g(x) = x^2$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$), local linear estimator ($\tilde{m}_{\tilde{h}}^{\mathrm{ll}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.5$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$	$\tilde{m}_{\tilde{h}}^{\mathrm{ll}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$	$\tilde{m}_{\tilde{h}}^{\mathrm{ll}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$	$\tilde{m}_{\tilde{h}}^{\mathrm{ll}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	8.98	16.6	15.2	322	9.86	17.9	15.5	589	12.3	24.3	16.1	1006
		ISB	0.21	3.45	2.27	316	0.83	4.78	2.54	583	2.02	8.35	2.74	1000
		IV	8.77	13.1	12.9	5.80	9.03	13.3	13.0	6.27	10.3	16.0	13.4	6.81
	400	MISE	2.83	4.64	4.24	318	3.64	4.96	4.38	585	4.35	5.24	4.58	1001
		ISB	0.12	0.74	0.44	316	0.43	1.07	0.51	583	0.61	1.20	0.54	999
		IV	2.71	3.90	3.80	1.35	3.21	3.89	3.87	1.52	3.74	4.04	4.04	1.71
$\rho = 1$	100	MISE	8.98	17.3	15.2	506	9.86	18.2	15.5	824	12.3	21.2	16.1	1292
		ISB	0.21	3.43	2.27	498	0.83	4.85	2.54	815	2.02	6.92	2.74	1283
		IV	8.77	13.9	12.9	7.98	9.03	13.4	13.0	8.54	10.3	14.3	13.4	9.19
	400	MISE	2.83	4.93	4.24	500	3.64	5.08	4.38	817	4.35	5.41	4.58	1285
		ISB	0.12	0.80	0.44	498	0.43	1.05	0.51	815	0.61	1.26	0.54	1283
		IV	2.71	4.13	3.80	1.81	3.21	4.03	3.87	2.01	3.74	4.15	4.04	2.24
$\rho = 2$	100	MISE	8.98	17.8	15.2	731	9.86	18.8	15.5	1100	12.3	20.7	16.1	1618
		ISB	0.21	3.57	2.27	720	0.83	4.91	2.54	1089	2.02	6.45	2.74	1606
		IV	8.77	14.2	12.9	10.6	9.03	13.9	13.0	11.3	10.3	14.3	13.4	12.1
	400	MISE	2.83	5.10	4.24	723	3.64	5.26	4.38	1091	4.35	5.52	4.58	1609
		ISB	0.12	0.86	0.44	720	0.43	1.07	0.51	1088	0.61	1.31	0.54	1606
		IV	2.71	4.24	3.80	2.38	3.21	4.19	3.87	2.61	3.74	4.21	4.04	2.87

contrary to the first application. We see that, although $\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$ starts from $\tilde{m}^{\mathrm{pa}}$ and the latter gives the worst performance in all cases, $\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$ makes drastic recover at the second stage. Its performance gets slightly worse as ρ increases, which may be explained by the distance of the chosen parametric model from m, $\min_\theta \int_0^1 \left(m(x) - \theta g(x)\right)^2 \, dx$. Comparing the columns for $\hat{m}_{\hat{b}}$ and $\tilde{m}_{\tilde{b}}^{\mathrm{ts}}$ in Tables 1 and 2, with the corresponding ones in Tables 3 and 4, respectively, we find that $\hat{m}_{\hat{b}}$ is much less affected by the choice of g.

Now, we present the simulation results for the other two models. The second model we considered was

$$Y_i = \sin(\pi X_i) + \rho X_i^2 + \lambda \cos(\pi X_i) + \varepsilon_i, \tag{3.6}$$

where the distributions of X_i and ε_i are the same as for the first model. The true regression curves are depicted in the three panels in the middle row of Figure 1. In the case where $\sigma = 0.1$, the values of the noise-to-signal ratio $\mathrm{NSR} = \mathrm{Var}(\varepsilon)/\mathrm{Var}(m(X))$ are 0.106, 0.073, 0.028 for $\rho = 0, 1, 2$, respectively, when $\lambda = 0$; 0.046, 0.166, 0.127, respectively, when $\lambda = 0.5$; 0.017, 0.043, 0.208, respectively, when $\lambda = 1$. For this model, we exercised $g(x) = \sin(\pi x)$ as in the first application to the first model (3.1). The last model was

$$Y_i = \rho \sin(\pi X_i) + X_i^2 + \lambda \cos(\pi X_i) + \varepsilon_i, \tag{3.7}$$

where we tried $g(x) = x^2$. The true regression curves are depicted in the three panels in the bottom row of Figure 1. In the case where $\sigma = 0.1$, the values of NSR are 0.113, 0.073, 0.027 for $\rho = 0, 1, 2$, respectively, when $\lambda = 0$; 0.889, 0.166, 0.034, respectively, when $\lambda = 0.5$; 0.054, 0.043, 0.021, respectively, when $\lambda = 1$.

The simulation results with the two models at (3.6) and (3.7) are contained in Tables 5–8. In both models, the parametric model $\theta g(\cdot)$ with the corresponding choice of g is correct when $\rho = \lambda = 0$. Thus, $\tilde{m}^{\mathrm{pa}}$ performs the best when $\rho = \lambda = 0$ in both models. The lessons from Tables 5 and 6 for the second model (3.6) are much the same as those from the first application to the first model (3.1), except slight changes in the performances of $\hat{m}_{\hat{b}}$ and $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ as ρ varies due to the fact that they are no more invariant to the change of ρ. In particular, our proposal and the two-step procedure show comparable performance when $\rho = 0$. Regarding these results, we note that, as in the first application to the model (3.1), g is perpendicular to m_0 when $\rho = 0$ regardless of the values of λ so that $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ works well in this case. As for the last model (3.7), the results contained in Tables 7 and 8 give similar lessons for $\hat{m}_{\hat{b}}$ and $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$ as the results for the second model (3.6). The two-step estimator $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ works still well when $\rho = 0$ although the corresponding g and m_0 are not perpendicular. Also, it does not deteriorate much as ρ increases, contrary to the case of the second model (3.6) and the first application to the first model (3.1).

To summarise, our approach outperforms the direct local linear estimator in all scenarios. Its performance is less affected by the choice of a parametric model than the two-step procedure $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ that starts from fitting the chosen parametric model. For the two-step estimator a parametric start needs to be chosen very carefully. It works very well if a parametric model is well chosen, such as $g(x) = \sin(\pi x)$ in the models (3.1) and (3.6) with $\rho = 0$, but largely fails if the choice is inadequate, such as $g(x) = \sin(\pi x)$ in the models (3.1) and (3.6) with $\rho > 0$.

4. Extensions

In this section we will discuss three extensions of our approach. First, we consider the case of a nonlinear parametric component $g(\theta, \cdot)$ for θ in a subset Θ of $\mathbb{R}$. We define θ_0 as the minimizer of

$$S(\theta) = \mathrm{E}[(m''(X) - g''(\theta, X))^2], \tag{4.1}$$

where g'' denotes the second derivative with respect to the second argument. We decompose the function m as

$$m(x) = g(\theta_0, x) + m_0(x).$$

As in Section 2, suppose that there exists a consistent estimator $\hat{\theta}$ of θ_0. Again, below we will discuss an estimator $\hat{\theta}$ based on profiling. We now regress $Y_i - g(\hat{\theta}, X_i)$ onto X_i by local linear smoothing. Define

$$\hat{m} = g(\hat{\theta}, \cdot) + \tilde{m}_b(\cdot, \hat{\theta}) \tag{4.2}$$

TABLE 5

For the model (3.6) with $g(x) = \sin(\pi x)$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.1$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	0.35	0.32	0.91	0.09	0.71	0.70	0.94	126	0.86	0.87	1.00	503
		ISB	0.00	0.00	0.17	0.00	0.13	0.11	0.19	125	0.16	0.15	0.21	500
		IV	0.35	0.32	0.74	0.08	0.58	0.58	0.75	0.69	0.70	0.72	0.79	2.50
	400	MISE	0.10	0.09	0.26	0.02	0.21	0.21	0.27	125	0.25	0.25	0.29	501
		ISB	0.00	0.00	0.04	0.00	0.03	0.03	0.04	125	0.04	0.04	0.05	500
		IV	0.10	0.09	0.22	0.02	0.18	0.18	0.22	0.22	0.21	0.21	0.23	0.76
$\rho = 1$	100	MISE	0.37	6.40	0.86	124	0.71	1.18	0.89	51.0	0.88	0.94	0.96	229
		ISB	0.01	3.42	0.15	123	0.12	0.40	0.16	50.7	0.15	0.21	0.19	228
		IV	0.36	2.98	0.71	0.47	0.59	0.78	0.73	0.26	0.72	0.73	0.76	1.26
	400	MISE	0.12	4.40	0.24	123	0.21	0.29	0.25	50.8	0.25	0.27	0.27	228
		ISB	0.01	2.31	0.04	123	0.03	0.09	0.04	50.7	0.04	0.06	0.05	228
		IV	0.11	2.09	0.20	0.15	0.18	0.19	0.21	0.06	0.21	0.21	0.22	0.33
$\rho = 2$	100	MISE	0.44	13.9	0.78	495	0.72	2.15	0.85	224	0.88	1.19	0.91	204
		ISB	0.04	6.42	0.13	493	0.13	1.22	0.16	223	0.16	0.40	0.18	203
		IV	0.40	7.47	0.65	1.62	0.59	0.93	0.68	0.59	0.72	0.78	0.73	0.79
	400	MISE	0.16	6.14	0.22	494	0.22	0.49	0.24	223	0.25	0.35	0.26	203
		ISB	0.02	2.80	0.03	493	0.03	0.26	0.04	223	0.04	0.13	0.05	203
		IV	0.14	3.34	0.19	0.57	0.18	0.23	0.20	0.20	0.21	0.22	0.21	0.19

as an estimator of $m = g(\theta_0, \cdot) + m_0$, where

$$\tilde{m}_b(x, \theta) = \mathcal{S}_b(Y - g(\theta, X))(x)$$

for each θ with the smoothing operator $\mathcal{S}_b U$ as defined in Section 2.

For the discussion of the statistical properties of $\hat{m}$ at (4.2) we need the following additional assumptions.

(A6) The function $S(\theta)$ defined at (4.1) has a unique global minimizer θ_0.

(A7) The parameter space Θ is a finite interval in $\mathbb{R}$. The function m and the functions $g(\theta, \cdot)$ for $\theta \in \Theta$ are twice continuously differentiable. The functions $g'(\theta, x)$ and $g''(\theta, x)$ are continuous functions of (θ, x) for $(\theta, x) \in \Theta \times [a_L, a_U]$.

The following proposition demonstrates that the conclusion of Proposition 1 remains to hold for this more general setting.

Proposition 3. *Assume (A1), (A3)–(A4), (A6), (A7) and that an estimator $\hat{\theta}$ fulfills (2.2). Then, it holds that*

$$\hat{m}(x) - m(x) = \mathcal{S}_b \varepsilon(x) + \mathcal{S}_b(m_0(X))(x) - m_0(x) + o_P(b^2),$$

uniformly for $x \in [a_L, a_U]$.

TABLE 6

For the model ([3.6]) with $g(x) = \sin(\pi x)$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.5$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	8.73	7.92	15.2	2.13	9.78	9.48	15.5	128	12.4	14.0	16.1	504
		ISB	0.04	0.06	2.27	0.02	0.74	0.74	2.54	125	1.87	1.79	2.75	500
		IV	8.69	7.86	12.9	2.11	9.04	8.74	13.0	2.68	10.5	12.2	13.3	4.46
	400	MISE	2.55	2.30	4.24	0.53	3.49	3.38	4.38	126	4.14	4.11	4.94	501
		ISB	0.01	0.01	0.44	0.00	0.39	0.38	0.51	125	0.54	0.51	1.01	500
		IV	2.54	2.29	3.80	0.53	3.10	3.00	3.87	0.79	3.60	3.60	3.93	1.41
$\rho = 1$	100	MISE	8.70	19.9	13.3	126	9.75	19.5	13.7	53.0	12.3	16.8	15.6	231
		ISB	0.04	8.17	2.02	123	0.75	7.56	2.27	50.7	1.91	4.65	2.68	228
		IV	8.66	11.7	11.3	2.52	9.00	11.9	11.4	2.28	10.4	12.1	12.9	3.25
	400	MISE	2.55	13.7	4.06	124	3.50	9.29	4.20	51.3	4.16	4.89	4.47	229
		ISB	0.01	8.43	0.40	123	0.39	4.01	0.46	50.7	0.56	1.07	0.56	228
		IV	2.54	5.33	3.66	0.59	3.11	5.28	3.73	0.57	3.60	3.82	3.91	0.91
$\rho = 2$	100	MISE	8.79	46.7	12.2	497	9.74	48.6	12.7	226	12.3	36.6	14.4	206
		ISB	0.07	24.5	2.00	493	0.75	28.4	2.42	223	1.90	16.8	2.78	203
		IV	8.72	22.2	10.2	3.69	8.99	20.2	10.3	2.64	10.4	19.8	11.6	2.80
	400	MISE	2.58	37.9	3.65	494	3.53	24.2	3.94	224	4.18	7.51	4.35	204
		ISB	0.04	23.4	0.38	493	0.42	13.0	0.46	223	0.58	2.78	0.55	203
		IV	2.54	14.5	3.27	0.93	3.11	11.2	3.48	0.64	3.60	4.73	3.80	0.70

We now discuss the estimation of θ_0. Define the profiling estimator $\hat{\theta}_h$ by

$$\hat{\theta}_h = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left(Y_i - g(\theta, X_i) - \tilde{m}_h(X_i, \theta)\right)^2 . \tag{4.3}$$

The following proposition shows that also in this more general setting θ_0 can be consistently estimated by profiling.

Proposition 4. *Assume (A1), (A3), (A5)–(A7). Then, for the estimator $\hat{\theta}_h$ defined at ([4.3]) it holds that $\hat{\theta}_h \to \theta_0$ in probability.*

Our result can be also extended to the case of multivariate covariates $\mathbf{X}_i$. For simplicity, assume that we use a product kernel $K_{\mathbf{b}}(\cdot - \mathbf{x})$ for smoothing around a point $\mathbf{x} \equiv (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ of the covariate domain such that $K_{\mathbf{b}}(\mathbf{u} - \mathbf{x}) = \prod_{j=1}^{d} K^{\mathrm{u}}_{b_j}(u_j - x_j)$ for some univariate kernel function K^{u}, where $\mathbf{b} = (b_1, \ldots, b_d)^\top$ is a bandwidth vector. For an estimator $\hat{\theta}$ that converges in probability to a limit θ_0 one can prove, under appropriate conditions, an expansion similar to the one stated in Proposition [1] for the one-dimensional case: $\hat{m}(\mathbf{x}) - m(\mathbf{x}) = \mathcal{S}_b \varepsilon(\mathbf{x}) + \mathcal{S}_b(m - \theta_0 g)(\mathbf{X})(\mathbf{x}) - (m - \theta_0 g)(\mathbf{x}) + o_P(b^2_{max})$, where $b_{max} = \max_{1 \le j \le d} b_j$. Define $\mu_0(\mathbf{x}) = \int K_{\mathbf{b}}(\mathbf{u} - \mathbf{x})\, d\mathbf{u} \in \mathbb{R}$. Also, define a

Table 7

For the model ([3.7]) with $g(x) = x^2$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.1$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	0.35	0.29	0.63	0.07	0.72	0.70	0.72	75.0	0.86	0.86	0.87	300
		ISB	0.00	0.00	0.07	0.00	0.13	0.09	0.12	75.0	0.16	0.12	0.15	300
		IV	0.35	0.29	0.56	0.07	0.59	0.61	0.60	0.12	0.70	0.74	0.72	0.25
	400	MISE	0.10	0.09	0.18	0.02	0.21	0.21	0.22	75.0	0.25	0.26	0.25	300
		ISB	0.00	0.00	0.02	0.00	0.03	0.03	0.03	74.9	0.04	0.04	0.04	300
		IV	0.10	0.09	0.16	0.02	0.18	0.18	0.18	0.04	0.21	0.22	0.21	0.09
$\rho = 1$	100	MISE	0.63	1.01	0.86	320	0.76	1.09	0.89	587	0.89	1.08	0.95	1004
		ISB	0.10	0.27	0.15	316	0.13	0.33	0.16	583	0.16	0.32	0.19	999
		IV	0.53	0.74	0.71	3.80	0.63	0.75	0.73	4.26	0.73	0.76	0.76	4.80
	400	MISE	0.20	0.29	0.24	317	0.23	0.29	0.25	584	0.26	0.30	0.27	1001
		ISB	0.03	0.07	0.04	316	0.04	0.08	0.04	583	0.05	0.08	0.05	1000
		IV	0.17	0.22	0.20	0.84	0.19	0.21	0.21	0.98	0.21	0.22	0.22	1.16
$\rho = 2$	100	MISE	0.79	1.32	1.08	1280	0.85	1.34	1.10	1740	0.93	1.36	1.13	2349
		ISB	0.13	0.47	0.26	1265	0.16	0.50	0.27	1724	0.17	0.51	0.30	2332
		IV	0.66	0.85	0.82	14.9	0.69	0.84	0.83	15.7	0.76	0.85	0.83	16.7
	400	MISE	0.24	0.36	0.31	1268	0.25	0.37	0.31	1278	0.27	0.38	0.32	2336
		ISB	0.04	0.10	0.06	1265	0.04	0.11	0.06	1724	0.05	0.13	0.06	2332
		IV	0.20	0.26	0.25	3.28	0.21	0.26	0.25	3.55	0.22	0.25	0.26	3.85

d-vector function $\boldsymbol{\mu}_1$ and a $d \times d$ matrix function $\boldsymbol{\mu}_2$ by

$$\boldsymbol{\mu}_1(\mathbf{x}) = \int \left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right) K_{\mathbf{b}}(\mathbf{u}-\mathbf{x})\, d\mathbf{u},$$

$$\boldsymbol{\mu}_2(\mathbf{x}) = \int \left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right) \left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right)^{\top} K_{\mathbf{b}}(\mathbf{u}-\mathbf{x})\, d\mathbf{u},$$

where $\mathbf{u}/\mathbf{b} = (u_1/b_1, \ldots, u_d/b_d)^{\top}$. Write $\mathbf{B} = \mathrm{diag}(b_j)$. Then, the asymptotic bias of $\hat{m}(\mathbf{x})$ is equal to

$$\frac{1}{2}\left[\mu_0(\mathbf{x}) - \boldsymbol{\mu}_1(\mathbf{x})^{\top}\boldsymbol{\mu}_2(\mathbf{x})^{-1}\boldsymbol{\mu}_1(\mathbf{x})\right]^{-1} \cdot \int \left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right)^{\top} \mathbf{B} \cdot D^2(m - \theta_0 g)(\mathbf{x})$$
$$\cdot\, \mathbf{B}\left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right) \cdot \left[1 - \boldsymbol{\mu}_1(\mathbf{x})^{\top}\boldsymbol{\mu}_2(\mathbf{x})^{-1}\left(\frac{\mathbf{u}-\mathbf{x}}{\mathbf{b}}\right)\right] \cdot K_{\mathbf{b}}(\mathbf{u}-\mathbf{x})\, d\mathbf{u}$$

where $D^2 f$ for a multivariate function f denotes the Hessian matrix consisting of the second-order partial derivatives of f. If one uses a symmetric kernel K^{u},

Table 8

For the model (3.7) with $g(x) = x^2$. Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV), multiplied by 10^3, of the four methods: our proposal ($\hat{m}_{\hat{b}}$), two-step estimator ($\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$), local linear estimator ($\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$) and parametric method ($\tilde{m}^{\mathrm{pa}}$), for the error level $\sigma = 0.5$.

			$\lambda = 0$				$\lambda = 0.5$				$\lambda = 1$			
	n		$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	100	MISE	8.83	7.19	9.88	1.86	9.78	8.88	10.6	76.8	12.4	12.0	12.6	302
		ISB	0.04	0.04	0.70	0.01	0.76	0.75	1.32	74.9	1.25	1.43	2.31	300
		IV	8.79	7.15	9.18	1.85	9.02	8.13	9.30	1.91	10.4	10.6	10.3	2.04
	400	MISE	2.59	2.13	3.26	0.53	3.52	3.29	3.58	75.5	4.17	4.09	4.00	300
		ISB	0.01	0.01	0.32	0.00	0.39	0.30	0.47	74.9	0.54	0.41	0.60	299
		IV	2.58	2.12	2.94	0.53	3.13	2.98	3.11	0.57	3.63	3.68	3.40	0.65
$\rho = 1$	100	MISE	8.98	16.6	13.3	322	9.86	17.9	13.7	589	12.3	24.3	15.6	1006
		ISB	0.21	3.45	2.02	316	0.83	4.78	2.27	583	2.02	8.35	2.67	999
		IV	8.77	13.1	11.3	5.80	9.03	13.1	11.4	6.27	10.3	16.0	12.9	6.81
	400	MISE	2.83	4.64	4.06	317	3.64	4.96	4.20	585	4.35	5.24	4.47	1001
		ISB	0.12	0.74	0.40	316	0.43	1.07	0.46	583	0.61	1.20	0.56	999
		IV	2.71	3.90	3.66	1.35	3.21	3.89	3.73	1.52	3.74	4.04	3.91	1.72
$\rho = 2$	100	MISE	10.4	19.5	17.1	1282	10.4	19.5	17.1	1742	12.9	20.1	17.4	2351
		ISB	0.66	4.54	2.54	1265	1.27	4.89	2.76	1724	2.19	5.51	3.04	2332
		IV	9.76	15.0	14.6	17.1	9.14	14.6	14.3	18.0	10.7	14.6	14.4	18.9
	400	MISE	3.54	5.48	4.88	1269	3.98	5.57	4.97	1728	4.53	5.71	5.03	2337
		ISB	0.38	0.99	0.57	1265	0.53	1.13	0.59	1724	0.64	1.25	0.62	2332
		IV	3.16	4.49	4.31	3.80	3.45	4.44	4.38	4.09	3.89	4.46	4.41	4.42

then for an interior point $\mathbf{x}$ the asymptotic bias is simplified to

$$(1/2)\kappa_2 \sum_{j=1}^{d} b_j^2 \bigl[\partial^2 m(\mathbf{x})/\partial x_j^2 - \theta_0 \partial^2 g(\mathbf{x})/\partial x_j^2\bigr],$$

where $\kappa_2 = \int v^2 K^{\mathrm{u}}(v)\, dv$. In this case, neglecting boundary regions, the expected squared bias is given by

$$\frac{\kappa_2^2}{4}\mathrm{E}\left[\left(\sum_{j=1}^{d} b_j^2 \bigl(m_{jj}(\mathbf{X}) - \theta_0 g_{jj}(\mathbf{X})\bigr)\right)^2\right],$$

where $m_{jj}(\mathbf{x}) = \partial^2 m(\mathbf{x})/\partial x_j^2$ and $g_{jj}(\mathbf{x}) = \partial^2 g(\mathbf{x})/\partial x_j^2$. The expected squared bias is minimized for the choice

$$\theta_0 = \frac{\sum_{j=1}^{d} b_j^2 \mathrm{E}[m_{jj}(\mathbf{X}) g_{jj}(\mathbf{X})]}{\sum_{j=1}^{d} b_j^2 \mathrm{E}[g_{jj}(\mathbf{X})^2]}.$$

We conjecture that the minimizing value θ_0 is again consistently picked out by the corresponding profiling estimator. We see that in the multivariate case the minimizing θ_0 depends on the ratios $(b_2/b_1, \ldots, b_d/b_1)$. Only in the one-dimensional case the dependence on the bandwidth disappears.

We finish this section by mentioning shortly another extension of the method in Section 2 to the case of choosing a multi-dimensional parametric model, say $\boldsymbol{\theta}^\top \mathbf{g}(\cdot)$ for $\boldsymbol{\theta} \in \mathbb{R}^d$ with $\mathbf{g} = (g_1, \ldots, g_d)^\top$. In the latter case, assume that $\mathrm{E}[\mathbf{g}''(X)\mathbf{g}''(X)^\top]$ is invertible and $\mathrm{E}\left(g_j''(X)^2\right) < \infty$ for $1 \le j \le d$. Define

$$\boldsymbol{\theta}_0 = \left(\mathrm{E}[\mathbf{g}''(X)\mathbf{g}''(X)^\top]\right)^{-1} \mathrm{E}[\mathbf{g}''(X)m''(X)] \in \mathbb{R}^d,$$

where $\mathbf{g}''(x) = (g_1''(x), \ldots, g_d''(x))^\top$. Then, one can easily verify that Proposition 1 with $m_0 = m - \boldsymbol{\theta}_0^\top \mathbf{g}$ remains to hold for this extension.

Appendix

A.1. Proof of Proposition 1

From the standard kernel smoothing theory, the condition (A1) gives that, if a function η is twice continuously differentiable on $[a_L, a_U]$, then

$$\mathcal{S}_b\eta(X)(x) - \eta(x) = \frac{1}{2} \cdot \frac{\hat{\mu}_2(x;b)^2 - \hat{\mu}_1(x;b)\hat{\mu}_3(x;b)}{\hat{\mu}_0(x;b)\hat{\mu}_2(x;b) - \hat{\mu}_1(x;b)^2} \cdot b^2 \cdot \eta''(x) + o_P(b^2), \quad \text{(A.1)}$$

uniformly for $x \in [a_L, a_U]$. We also note that there exists an absolute constant $0 < C < \infty$ such that

$$\sup_{x \in [a_L, a_U]} \left| \frac{\hat{\mu}_2(x;b)^2 - \hat{\mu}_1(x;b)\hat{\mu}_3(x;b)}{\hat{\mu}_0(x;b)\hat{\mu}_2(x;b) - \hat{\mu}_1(x;b)^2} \right| \le C \quad \text{(A.2)}$$

with probability tending to one. For (A.2) what we need is that the support of the baseline kernel K contains a nontrivial interval in both of the half intervals $[-1, 0]$ and $[0, 1]$, which is ensured by the condition (A3). Note that $\tilde{m}_b(\cdot, \theta) = \mathcal{S}_b\varepsilon + \mathcal{S}_b(m_0(X)) - (\theta - \theta_0)\mathcal{S}_b(g(X))$. Thus,

$$\begin{aligned}
&\hat{m}(x) - m(x) \\
&\qquad = \hat{\theta}g(x) + \tilde{m}_b(x, \hat{\theta}) - m(x) \\
&\qquad = \mathcal{S}_b\varepsilon(x) + [\mathcal{S}_b(m_0(X))(x) - m_0(x)] - (\hat{\theta} - \theta_0)\,[\mathcal{S}_b(g(X))(x) - g(x)] \\
&\qquad = \mathcal{S}_b\varepsilon(x) + [\mathcal{S}_b(m_0(X)) - m_0]\,(x) + o_P(b^2)
\end{aligned}$$

uniformly for $x \in [a_L, a_U]$. Here, we used (A.1) and (A.2). □

A.2. Proof of Proposition 2

From the definition of $\hat{\theta}_h$ in Section 2 and writing simply $\mathcal{S}_h\eta$ for $\mathcal{S}_h(\eta(X))$, we get

$$\hat{\theta}_h = \arg\min_{\theta} \sum_{i=1}^{n} \left[\varepsilon_i - \mathcal{S}_h\varepsilon(X_i) - (\mathcal{S}_h m_0 - m_0)(X_i) + (\theta - \theta_0)(\mathcal{S}_h g - g)(X_i)\right]^2.$$

Thus, it holds that

$$\begin{aligned}\hat{\theta}_h - \theta_0 = &\left[n^{-1}\sum_{i=1}^{n}(\mathcal{S}_h g - g)^2(X_i)\right]^{-1} \\ &\cdot \left[n^{-1}\sum_{i=1}^{n}\big(\mathcal{S}_h\varepsilon(X_i) - \varepsilon_i\big)\cdot(\mathcal{S}_h g - g)(X_i)\right. \\ &\quad \left. + n^{-1}\sum_{i=1}^{n}(\mathcal{S}_h m_0 - m_0)(X_i)\cdot(\mathcal{S}_h g - g)(X_i)\right].\end{aligned} \tag{A.3}$$

We now argue that with $\mu_2 = \int u^2 K(u)\, du$

$$\begin{aligned}T_1 &:= n^{-1}\sum_{i=1}^{n}(\mathcal{S}_h g - g)^2(X_i) - \frac{1}{4}h^4\,\mu_2^2\,\mathrm{E}\,g''(X)^2 = o_P(h^4), \\ T_2 &:= n^{-1}\sum_{i=1}^{n}(\mathcal{S}_h m_0 - m_0)(X_i)\cdot(\mathcal{S}_h g - g)(X_i) = o_P(h^4), \\ T_3 &:= n^{-1}\sum_{i=1}^{n}\big(\mathcal{S}_h\varepsilon(X_i) - \varepsilon_i\big)\cdot(\mathcal{S}_h g - g)(X_i) = O_P(h^2/\sqrt{n}).\end{aligned} \tag{A.4}$$

From (A.3) and (A.4) we get $\hat{\theta}_h - \theta_0 = O_P(n^{-1/2}h^{-2}) + o_P(1)$. The statement of the proposition now follows because of (A5).

It remains to prove (A.4). Put $\mu_j(x;b) = f(x)\int_{a_L}^{a_U}((u-x)/b)^j K_b(u-x)\,du$. For $j \geq 0$, we get $\hat{\mu}_j(x;b) = \mu_j(x;b) + o_P(1)$ uniformly for $x \in [a_L, a_U]$. Let

$$c(x;h) = \frac{\mu_2(x;b)^2 - \mu_1(x;b)\mu_3(x;b)}{\mu_0(x;b)\mu_2(x;b) - \mu_1(x;b)^2}.$$

Note that $c(x;h) = \mu_2$ for all $x \in [a_L + h, a_U - h]$. This and a version of (A.1) for $(\mathcal{S}_h g - g)(x)$ give

$$\begin{aligned}T_1 &= \frac{1}{4}h^4\,n^{-1}\sum_{i=1}^{n}c(X_i;h)^2 g''(X_i)^2 - \frac{1}{4}h^4\,\mu_2^2\,\mathrm{E}\,g''(X)^2 + o_P(h^4) \\ &= \frac{1}{4}h^4\int_{\mathcal{I}_B}\big(c(x;h)^2 - \mu_2^2\big)\,g''(x)^2 f(x)\,dx + o_P(h^4) \\ &= o_P(h^4),\end{aligned}$$

where $\mathcal{I}_B = [a_L, a_U]\setminus[a_L + h, a_U - h]$. Similarly, for the second assertion it holds that

$$\begin{aligned}T_2 &= \frac{1}{4}h^4\,n^{-1}\sum_{i=1}^{n}c(X_i;h)^2 m_0''(X_i)g''(X_i) + o_P(h^4) \\ &= \frac{1}{4}h^4\,\mu_2^2\,\mathrm{E}\,m_0''(X)g''(X) + o_P(h^4) \\ &= o_P(h^4),\end{aligned}$$

where the last equality follows from the definition of m_0 at (1.1). For the last assertion at (A.4), let

$$D_h(x) = (\mathcal{S}_h g - g)(x), \quad J_h(x) = n^{-1} \sum_{i=1}^n w_h(X_i, x) D_h(X_i).$$

Then, $T_3 = n^{-1} \sum_{i=1}^n (J_h(X_i) - D_h(X_i)) \varepsilon_i$. From the versions of (A.1) and (A.2) for the bandwidth h, we have $\sup_{x \in [a_L, a_U]} |D_h(x)| = O_P(h^2)$. Also, similarly as in (A.2) there exists an absolute constant $0 < C' < \infty$ such that

$$n^{-1} \sum_{i=1}^n |w_h(X_i, x)| \le C' \, n^{-1} \sum_{i=1}^n K_h(x - X_i),$$

so that $\sup_{x \in [a_L, a_U]} |J_h(x)| = O_P(h^2)$. Thus,

$$\sup_{x \in [a_L, a_U]} |J_h(x) - D_h(x)| = O_P(h^2). \tag{A.5}$$

At this point we remark that the difference $|J_h(x) - D_h(x)|$ is of smaller order than $O_P(h^2)$ uniformly in $[a_L + 2h, a_U - 2h]$ under additional smoothness assumptions on g and f. The continuity of $\sigma^2(\cdot)$ in the assumption (A1) and the result (A.5) give

$$\mathrm{Var}(T_3 | X_1, \ldots, X_n) = n^{-2} \sum_{i=1}^n (J_h(X_i) - D_h(X_i))^2 \, \sigma^2(X_i) = O_P(n^{-1} h^4).$$

This completes the proof of the proposition. □

A.3. Proof of Proposition 3

By proceeding as in the proof of Proposition 1 it only remains to show that

$$\mathcal{S}_b g(\hat{\theta}, X)(x) - \mathcal{S}_b g(\theta_0, X)(x) - g(\hat{\theta}, x) + g(\theta_0, x) = o_P(b^2),$$

uniformly for $x \in [a_L, a_U]$. For a proof of this claim note that

$$\begin{aligned}
&\mathcal{S}_b g(\hat{\theta}, X)(x) - \mathcal{S}_b g(\theta_0, X)(x) - g(\hat{\theta}, x) + g(\theta_0, x) \\
&\quad = \frac{1}{n} \sum_{i=1}^n w_b(x, X_i) \Big[(g'(\hat{\theta}, x) - g'(\theta_0, x))(X_i - x) \\
&\qquad\qquad + \frac{1}{2} \big(g''(\hat{\theta}, \tilde{X}_i) - g''(\theta_0, \tilde{X}_i) \big) (X_i - x)^2 \Big] \\
&\quad = \frac{1}{n} \sum_{i=1}^n w_b(x, X_i) \Big[\frac{1}{2} \big(g''(\hat{\theta}, \tilde{X}_i) - g''(\theta_0, \tilde{X}_i) \big) (X_i - x)^2 \Big] \\
&\quad = \max_{v \in [a_L, a_U]} |g''(\hat{\theta}, v) - g''(\theta_0, v)| \cdot O(b^2) \\
&\quad = o_P(b^2)
\end{aligned}$$

with $\tilde{X}_i$ such that $|\tilde{X}_i - x| \le b$. □

A.4. Proof of Proposition 4

Define

$$S_n(\theta) = \frac{4}{n}\sum_{i=1}^{n}(Y_i - g(\theta, X_i) - \tilde{m}_h(X_i, \theta))^2 - \frac{4}{n}\sum_{i=1}^{n}(\varepsilon_i - \mathcal{S}_h\varepsilon(X_i))^2.$$

We will show that

$$S_n(\theta) = h^4 S(\theta) + o_P(h^4), \tag{A.6}$$

uniformly over $\theta \in \Theta$. This implies the statement of the proposition because $\hat{\theta}_h$ minimizes $S_n(\theta)$, θ_0 minimizes $S(\theta)$, and $S(\theta)$ is a continuous function on Θ with a unique minimum.

For the proof of (A.6) note first that

$$\begin{aligned} S_n(\theta) = \frac{4}{n}\sum_{i=1}^{n}\Big(&\varepsilon_i - \mathcal{S}_h\varepsilon(X_i) - \big(g(\theta, X_i) - g(\theta_0, X_i)\big) \\ &+ \mathcal{S}_h\big(g(\theta, X) - g(\theta_0, X)\big)(X_i) + m_0(X_i) - \mathcal{S}_h m_0(X)(X_i)\Big)^2 \\ &- \frac{4}{n}\sum_{i=1}^{n}(\varepsilon_i - \mathcal{S}_h\varepsilon(X_i))^2. \end{aligned}$$

We now argue that

$$\begin{aligned} &\frac{4}{n}\sum_{i=1}^{n}\Big(\varepsilon_i - \mathcal{S}_h\varepsilon(X_i)\Big)\Big(- \big(g(\theta, X_i) - g(\theta_0, X_i)\big) \\ &\qquad + \mathcal{S}_h\big(g(\theta, X) - g(\theta_0, X)\big)(X_i) + m_0(X_i) - \mathcal{S}_h m_0(X)(X_i)\Big) \\ &\quad = O_P(h^2/\sqrt{n}) = o_P(h^4). \end{aligned}$$

This can be shown by an extension of the arguments used in the treatment of T_4 in the proof of Proposition 2. Here, one has to make use of the smoothness properties of $g(\theta, x)$ as a function of (θ, x). Using this bound we get that

$$\begin{aligned} S_n(\theta) = \frac{4}{n}\sum_{i=1}^{n}\Big(&- \big(g(\theta, X_i) - g(\theta_0, X_i)\big) + \mathcal{S}_h\big(g(\theta, X) - g(\theta_0, X)\big)(X_i) \\ &+ m_0(X_i) - \mathcal{S}_h m_0(X)(X_i)\Big)^2 + o_P(h^4). \end{aligned}$$

The claim now follows from the application of (A.1) with $\eta = m_0$, $b = h$, (A.3) and

$$\begin{aligned} &\mathcal{S}_h g(\theta, X)(x) - g(\theta, x) \\ &\quad = \frac{1}{2}\cdot\frac{\hat{\mu}_2(x;h)^2 - \hat{\mu}_1(x;h)\hat{\mu}_3(x;h)}{\hat{\mu}_0(x;h)\hat{\mu}_2(x;h) - \hat{\mu}_1(x;h)^2}\cdot h^2 \cdot g''(\theta, x) + o_P(h^2), \end{aligned} \tag{A.7}$$

uniformly for $x \in [a_L, a_U]$ and $\theta \in \Theta$. The expansion (A.7) follows similarly as (A.1). In particular, one makes use of the assumption that $g''(\theta, x)$ is a continuous function of (θ, x) for $(\theta, x) \in \Theta \times [a_L, a_U]$, see the assumption (A7), for seeing that (A.7) holds uniformly over $\theta \in \Theta$ and $x \in [a_L, a_U]$. □

References

[1] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95**, 888–902. MR1804446

[2] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489. MR1467842

[3] Fan, J., Wu, Y. and Feng, Y. (2009). Local quasi-likelihood with a parametric guide. *Annals of Statistics* **37**, 4153–4183. MR2572456

[4] Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society, Series B* **65**, 57–80. MR1959093

[5] Glad, I. K. (1998). Parametrically guided nonparametric regression. *Scandinavian Journal of Statistics* **25**, 649–668. MR1666776

[6] Gozalo, P. and Linton, O. (2000). Local nonlinear least squares: using parametric information in nonparametric regression. *Journal of Econometrics* **99**, 63–106. MR1793389

[7] Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics* **23**, 882–904. MR1345205

[8] Lee, Y. K., Mammen, E. and Park, B. U. (2012). Flexible generalized varying coefficient regression models. *Annals of Statistics* **40**, 1906–1933. MR3015048

[9] Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**, 1443–1490. MR1742496

[10] Rahman, M. and Ullah, A. (2002). Improved combined parametric and non-parametric regression: estimation and hypothesis testing. In *Handbook of Applied Econometrics and Statistical Inference*, Edited by Ullah, A., Wan, A. and Chaturvedi, A. Marcel Dekker, New York. MR1893335

[11] Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768–1802. MR1193312

[12] Talamakrouni, M., El Ghouch, A. and van Keilegom, I. (2015). Guided censored regression. *Scandinavian Journal of Statistics* **42**, 214–233. MR3318033

[13] Talamakrouni, M., van Keilegom, I. and El Ghouch, A. (2016). Parametrically guided nonparametric density and hazard estimation with censored data. *Computational Statistics and Data Analysis* **93**, 308–323. MR3406214

[14] Yu, K., Mammen, E. and Park, B. U. (2011). Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli* **17**, 736–748. MR2787613

[15] Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Annals of Statistics* **36**, 228–260. MR2387970