Nonparametric regression with parametric help

Young K. Lee; Enno Mammen; Jens P. Nielsen; Byeong U. Park

Kangwon National University; Heidelberg University; City, University of London; Seoul National University

Date: February 5, 2020

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

Key Words: Regression function, bias, profiling technique, local linear estimation, cross-validatory bandwidth selectors.

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 $\mathcal{G} = \{g : g'' \text{ exists and } 0 < \mathrm{E}\, g''(X)^2 < \infty\}$, 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 $\theta g(X) + \eta(x) = 0$ 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 Hjort and Glad (1995), Glad (1998), Gozalo and Linton (2000), Rahman and Ullah (2002), Fan et al. (2009) and Talamakrouni et al. (2015, 2016). 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 (Severini and Wong, 1992) to estimate (1.3), which is known as a useful technique of fitting partially linear models. Our discussion in this paper can be generalized to more complex semiparametric models, such as generalized partially linear models and generalized partially linear additive models, with common

covariates in the parametric and nonparametric components. In these models one may also allow specifications of the parametric part $g(\theta, X)$ where the parameter θ does not enter linearly. In this paper, to avoid technical difficulties and to make the presentation transparent, we focus our discussion on the model (1.3) where $g(\theta, X)$ is linear in θ. For simplicity we also assume that the covariate X is univariate.

This paper is organized as follows. In the next section we discuss the estimation of m based on the decomposition (1.3), and develop its asymptotic theory. In Section 3 we present numerical evidences that support the theory. 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

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

which is equivalent to choosing g that minimizes

$$\mathrm{E}\left(\frac{g''(X)}{\sqrt{\mathrm{E} g''(X)^2}} - m''(X)\right)^2 = 1 + \mathrm{E} m''(X)^2 - 2\,\mathrm{E}\left(\frac{g''(X)}{\sqrt{\mathrm{E} g''(X)^2}} \cdot m''(X)\right). \tag{2.3}$$

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

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

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

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

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.

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 is done with the CV bandwidth selectors introduced also in the previous section. We generate (X_i, Y_i) according to the model

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

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 $\sigma = 0.5$. In the application of our approach, 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 made two choices for λ: $\lambda = 0, 0.5$, and three choices for ρ: $\rho = 0, 1, 2$.

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 $\tilde{m}^{\rm pa} = \tilde{\theta} g$ where $\tilde{\theta}$ minimizes $\sum_{i=1}^n (Y_i - \theta g(X_i))^2$. We denote the direct local linear smoother by $\tilde{m}^{\rm ll}_{\tilde{h}} = \mathcal{S}_{\tilde{h}}(Y)$, where $\tilde{h}$ is chosen to minimize $\sum_{i=1}^n (Y_i - \mathcal{S}_h^{(-i)}(Y)(X_i))^2$ with respect to h. The two-step procedure with $\tilde{m}^{\rm pa}$ as a parametric start is $\tilde{m}^{\rm ts}_{\tilde{b}} = \tilde{\theta} g + \tilde{m}_{\tilde{b}}(\cdot, \tilde{\theta})$, where $\tilde{b}$ is chosen by minimizing the CV criterion $\sum_{i=1}^n (Y_i - \tilde{m}_b^{{\rm ts}(-i)}(X_i))^2$. 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 $\hat{m}_{\hat{b}}$, $\tilde{m}^{\rm ts}_{\tilde{b}}$, $\tilde{m}^{\rm ll}_{\tilde{h}}$ and $\tilde{m}^{\rm pa}$. 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 $\tilde{m}^{\rm pa}$ 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 $\tilde{m}^{\rm ll}_{\tilde{h}}$ do not change as ρ varies. This stems basically from the property of the weight w_b that

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

Our theory in Section 2 tells that there is a larger reduction in the bias of our proposal in comparison with that of the direct local linear estimator if g'' is closer to m'', see (2.3). This is evident in the numerical results. We note that under the data generating model (3.1) g'' gets closer to m'' when $\lambda = 0$ than when $\lambda = 0.5$. The ISB values of $\hat{m}_{\hat{b}}$ in the tables are less than those of $\tilde{m}^{\rm ll}_{\tilde{h}}$ for both values of λ and the relative difference is larger when $\lambda = 0$. 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 of the latter. The smaller bias then gives our proposal some room for sacrificing bias to reduce variance by

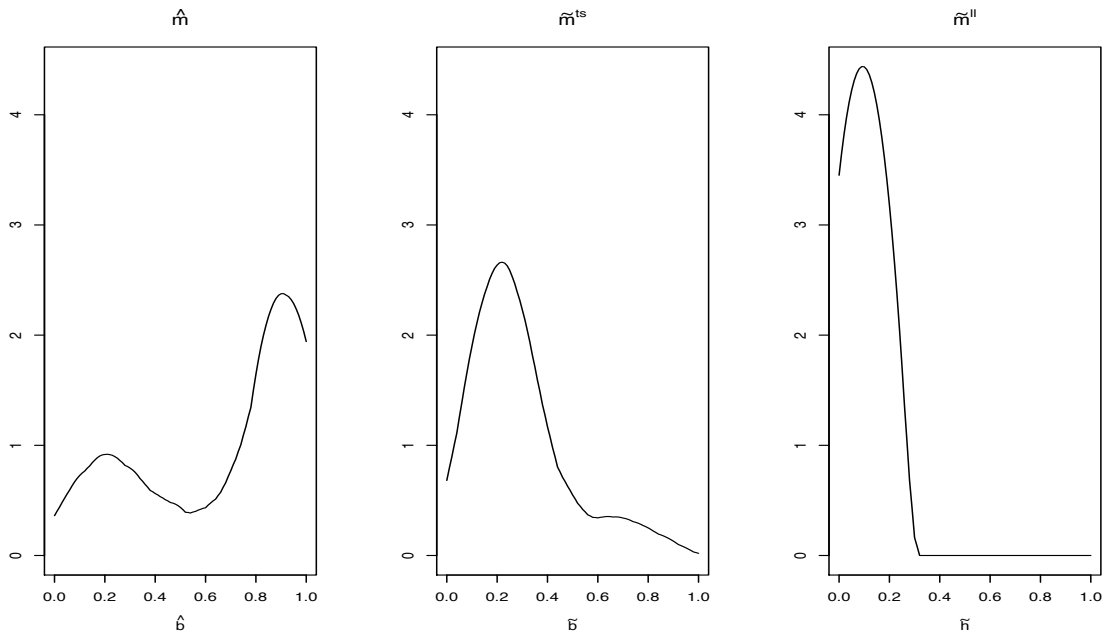

Figure 1: *The distributions of the CV bandwidth selectors. From left to right,* $\hat{b}$ *for our proposal* $\hat{m}_{\hat{b}}$*,* $\tilde{b}$ *for the two-step procedure* $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ *and* $\tilde{h}$ *for the direct local linear estimator* $\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$*.*

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 1, which depicts the distributions of the CV bandwidth choices $\hat{b}$ (left) for our proposal and $\tilde{h}$ for the direct local linear estimator (right).

Our proposal exhibits the best performance in all cases except $(\lambda = 0, \rho = 0)$, in which case the parametric method is the best as expected. For the two cases of $\rho = 0$ ($\lambda = 0$ and 0.5), our proposal and the two-step procedure show comparable performance. In these cases, the true regression function m is not far from the parametric function g. Indeed,

$$\int_0^1 (m(x) - g(x))^2 \, dx = \frac{1}{2}\rho^2 - \frac{4}{\pi^2}\rho\lambda + \frac{1}{2}\lambda^2, \tag{3.2}$$

so that the squared distances between m and g in the case $\rho = 0$ equal 0 and 1/8 for $\lambda = 0$ and $\lambda = 0.5$, respectively. However, m gets away from g as $\rho > 0$ increases and

is more distant from g when $\lambda = 0$ than when $\lambda = 0.5$ if $\rho > 0$. The main lesson from the results in the tables is that the existing two-step procedure $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ 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.

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. 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 parametric rate. This observation and our simulation results suggest that the success of the existing two-step procedure $\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$ depends highly on the choice of a pilot parametric model, while our approach does not as long as the chosen function g satisfies $\mathrm{E}\, g''(X)^2 > 0$.

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

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

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

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

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

Table 1: 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 $\sigma = 0.1$.

			$\lambda = 0$				$\lambda = 0.5$			
			$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	$n = 100$	MISE	0.35	0.32	0.91	0.09	0.71	0.70	0.94	126
		ISB	0.00	0.00	0.17	0.00	0.13	0.11	0.19	125
		IV	0.35	0.32	0.74	0.08	0.58	0.58	0.75	0.69
	$n = 400$	MISE	0.10	0.09	0.26	0.02	0.21	0.21	0.27	125
		ISB	0.00	0.00	0.04	0.00	0.03	0.03	0.04	125
		IV	0.10	0.09	0.22	0.02	0.18	0.18	0.22	0.22
$\rho = 1$	$n = 100$	MISE	0.35	6.29	0.91	126	0.71	1.21	0.94	53.3
		ISB	0.00	3.36	0.17	126	0.13	0.43	0.19	53.0
		IV	0.35	2.92	0.74	0.48	0.58	0.78	0.75	0.27
	$n = 400$	MISE	0.10	5.26	0.26	126	0.21	0.29	0.27	53.1
		ISB	0.00	2.88	0.04	126	0.03	0.10	0.04	53.0
		IV	0.10	2.38	0.22	0.15	0.18	0.19	0.22	0.07
$\rho = 2$	$n = 100$	MISE	0.35	15.8	0.91	505	0.71	2.33	0.94	233
		ISB	0.00	7.32	0.17	503	0.13	1.27	0.19	233
		IV	0.35	8.49	0.74	1.66	0.58	1.06	0.75	0.64
	$n = 400$	MISE	0.10	9.98	0.26	504	0.21	0.53	0.27	233
		ISB	0.00	4.82	0.04	503	0.03	0.30	0.04	233
		IV	0.10	5.16	0.22	0.58	0.18	0.23	0.22	0.22

Table 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 $\sigma = 0.5$.

			$\lambda = 0$				$\lambda = 0.5$			
			$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$	$\hat{m}_{\hat{b}}$	$\tilde{m}^{\mathrm{ts}}_{\tilde{b}}$	$\tilde{m}^{\mathrm{ll}}_{\tilde{h}}$	$\tilde{m}^{\mathrm{pa}}$
$\rho = 0$	$n = 100$	MISE	8.73	7.92	15.2	2.13	9.78	9.48	15.5	128
		ISB	0.04	0.06	2.27	0.02	0.74	0.74	2.54	125
		IV	8.69	7.86	12.9	2.11	9.04	8.74	13.0	2.68
	$n = 400$	MISE	2.55	2.30	4.24	0.53	3.49	3.38	4.38	126
		ISB	0.01	0.01	0.44	0.00	0.39	0.38	0.51	125
		IV	2.54	2.29	3.80	0.53	3.10	3.00	3.87	0.79
$\rho = 1$	$n = 100$	MISE	8.73	20.4	15.2	128	9.77	20.1	15.5	55.4
		ISB	0.04	8.35	2.27	126	0.74	8.35	2.54	53.1
		IV	8.69	12.0	12.9	2.53	9.03	11.8	13.0	2.29
	$n = 400$	MISE	2.55	14.1	4.24	126	3.49	9.30	4.38	53.7
		ISB	0.01	8.61	0.44	126	0.39	4.01	0.51	53.1
		IV	2.54	5.46	3.80	0.59	3.10	5.29	3.87	0.58
$\rho = 2$	$n = 100$	MISE	8.73	47.6	15.2	507	9.78	49.3	15.5	235
		ISB	0.04	25.2	2.27	503	0.74	28.9	2.54	232
		IV	8.69	22.4	12.9	3.73	9.04	20.4	13.0	2.68
	$n = 400$	MISE	2.55	38.8	4.24	504	3.49	26.2	4.38	233
		ISB	0.01	24.1	0.44	503	0.39	14.1	0.51	233
		IV	2.54	14.7	3.80	0.95	3.10	12.1	3.87	0.65

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,

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

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

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

Thus, it holds that

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

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). To prove the first assertion, put $\mu_j(x;b) = f(x)\int_0^1((u-x)/b)^j K_b(u-x)\,du$. 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

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

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

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

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

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

This completes the proof of the proposition. □

References

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

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

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

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

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.

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

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

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.