Multiplicative Bias Corrected Nonparametric Smoothers

N. Hengartner, E. Matzner-Løber, L. Rouvière and T. Burr

Abstract The paper presents a general multiplicative bias reduction strategy for nonparametric regression. The approach is most effective when applied to an oversmooth pilot estimator, for which the bias dominates the standard error. The practical usefulness of the method was demonstrated in [2] in the context of estimating energy spectra. For such data sets, it was observed that the method could decrease significantly the bias with only negligible increase in variance. This paper presents the theoretical analysis of that estimator. In particular, we study the asymptotic properties of the bias corrected local linear regression smoother, and prove that it has zero asymptotic bias and the same asymptotic variance as the local linear smoother with with a suitably adjusted bandwidth. Simulations show that our asymptotic results are available for modest sample sizes.

1 Introduction

In nonparametric regression, the bias-variance tradeoff of linear smoothers such as kernel-based regression smoothers, wavelet based smoother or spline smoothers, is generally governed by a user-supplied parameter. This parameter is often called the bandwidth, which we will denote by h. As an example, assuming that the regression

N. Hengartner
Theoretical Biology and Biophysics group, Los Alamos National Laboratory, e-mail: nickh@lanl.gov

E. Matzner-Løber
Cepe-Ensae, e-mail: Eric.Matzner-Lober@ensae.fr

L. Rouvière
Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, e-mail: laurent.rouviere@univ-rennes2.fr

T. Burr
Statistical Sciences group, Los Alamos National Laboratory, e-mail: tburr@lanl.gov

function m is locally twice continuously differentiable at a point x, the local linear smoother with bandwidth h and kernel K has conditional bias at that point

$$\frac{h^2}{2} m''(x) \int u^2 K(u)\, \mathrm{d}u + \mathrm{o}_p(h^2)$$

and conditional variance

$$\frac{1}{nh} \frac{\sigma^2(x)}{f(x)} \int K^2(u)\, \mathrm{d}u + \mathrm{o}_p\left(\frac{1}{nh}\right)$$

where f stands for the density of the (one-dimensional) explanatory variable X and $\sigma^2(x)$ is the conditional variance of the response variable given $X = x$. See for example the book of [7]. Since the bias increases with the second order derivative of the regression function, the local linear smoother tends to under-estimate in the peaks and over-estimate in the valleys of the regression function. See for example [25, 26, 27].

The resulting bias in the estimated peaks and valleys is troublesome in some applications, such as the estimation of energy spectrum from nuclear decay. That example motivates the development of our multiplicative bias correction methodology. The interested reader is referred to [2] for a more detailed description and analysis.

All nonparametric smoothing methods are generally biased. There are a large number of methods to reduce the bias, but most of them do so at the cost of an increase in the variance of the estimator. For example, one may choose to undersmooth the energy spectrum. Undersmoothing will reduce the bias but will have a tendency of generating spurious peaks. One can also use higher order smoothers, such as local polynomial smoother with a polynomial of order larger than one. While again this will lead to a smaller bias, the smoother will have a larger variance. Another approach is to start with a pilot smoother and to estimate its bias by smoothing the residuals ([6, 3, 4]). Subtracting the estimated bias from the smoother produces a regression smoother with smaller bias and larger variance. For the estimation of an energy spectrum, the additive bias correction and the higher order smoothers have the unfortunate side effect of possibly generating a non-positive estimate.

An attractive alternative to the linear bias correction is the multiplicative bias correction pioneered by [20]. Because the multiplicative correction does not alter the sign of the regression function, this type of correction is particularly well suited for adjusting non-negative regression functions. [19] showed that if the true regression function has four continuous derivatives, then the multiplicative bias reduction is operationally equivalent to using an order four kernel. And while this does remove the bias, it also increases the variance because of the roughness of such a kernel.

Many authors have extended the work of [19]. [9, 10] propose to use a parametrically guided local linear smoother and Nadaraya-Watson smoother by starting with a parametric pilot. This approach is extended to a more general framework which includes both multiplicative and additive bias correction by [21] (see also [16, 28, 22] for an extension to time series conditional variance estimation and spectral estimation). For multiplicative bias correction in density estimation and hazard estimation, we refer the reader to the works of [11, 12, 17, 23, 24].

Although the bias-variance tradeoff for nonparametric smoothers is always present in finite samples, it is possible to construct smoothers whose *asymptotic bias* converges to zero while keeping the same asymptotic variance. [13] has exhibited a nonparametric density estimator based on multiplicative bias correction with that property, and has shown in simulations that his estimator also enjoys good finite sample properties. [2] adapts the estimator from [13] to nonparametric regression with aim to estimate energy spectra. They illustrate the benefits of their approach on real and simulated spectra. The goal of this paper is to study the asymptotic properties of that estimator. It is worth pointing out that these properties have already been studied by [20] for fixed design and further by [19]. We emphasize that there are two major differences between our work and that of [19].

- First, we do not add regularity assumptions on the target regression function. In particular, we do not assume that the regression function has four continuous derivatives as in [19].
- Second, we show that the multiplicative bias reduction procedure performs a bias reduction with no cost to the asymptotic variance. It is exactly the same as the asymptotic variance of the local linear estimate.

Finally, we note that we show a different asymptotic behavior under less restrictive assumptions than those found in [19]. Moreover our results and proofs are different from the above referenced works.

This paper is organized as follows. Section 2 introduces the notation and defines the estimator. Section 3 gives the asymptotic behavior of the proposed estimator. A brief simulation study on finite sample comparison is presented in Section 4. The interested reader is referred to Section 6 where we have gathered the technical proofs.

2 Preliminaries

2.1 Notations

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be n independent copies of the pair of random variables (X,Y) with values in $\mathbb{R}\times\mathbb{R}$. We suppose that the explanatory variable X has probability density f and model the dependence of the response variable Y to the explanatory variable X through the nonparametric regression model

$$Y = m(X) + \varepsilon. \tag{1}$$

We assume that the regression function $m(\cdot)$ is smooth and that the disturbance ε is a mean zero random variable with finite variance σ^2 that is independent of the covariate X. Consider the linear smoothers for the regression function $m(x)$ which we write as

$$\hat{m}(x) = \sum_{j=1}^{n} \omega_j(x;h) Y_j,$$

where the weight functions $\omega_j(x;h)$ depend on a bandwidth h. If the weight functions are such that $\sum_{j=1}^{n} \omega_j(x;h) = 1$ and $\sum_{j=1}^{n} \omega_j(x;h)^2 = (nh)^{-1}\tau^2$, and if the disturbances satisfy the Lindeberg's condition, then the linear smoother obeys the central limit theorem

$$\sqrt{nh}\left(\hat{m}(x) - \sum_{j=1}^{n} w_j(x;h) m(X_j)\right) \xrightarrow{d} \mathcal{N}(0,\tau^2) \quad \text{as} \quad n \to \infty. \tag{2}$$

We can use (2) to construct asymptotic pointwise confidence intervals for the unknown regression function $m(x)$. But unless the limit of the scaled bias

$$b(x) = \lim_{n \longrightarrow \infty} \sqrt{nh}\left(\sum_{j=1}^{n} w_j(x;h) m(X_j) - m(x)\right),$$

which we call the asymptotic bias, is zero, the confidence interval

$$\left[\hat{m}(x) - Z_{1-\alpha/2}\sqrt{nh}\tau, \hat{m}(x) + Z_{1-\alpha/2}\sqrt{nh}\tau\right]$$

will not cover asymptotically the true regression function $m(x)$ at the nominal $1-\alpha$ level ($Z_{1-\alpha/2}$ stands for the $(1-\alpha/2)$-quantile of the $\mathcal{N}(0,1)$ distribution). The construction of valid pointwise $1-\alpha$ confidence intervals for regression smoothers is another motivation for developing estimators with zero asymptotic bias.

2.2 Multiplicative bias reduction

Given a pilot smoother with bandwidth h_0 for the regression function $m(x)$,

$$\tilde{m}_n(x) = \sum_{j=1}^{n} \omega_j(x;h_0) Y_j,$$

consider the ratio $V_j = \frac{Y_j}{\tilde{m}_n(X_j)}$. That ratio is a noisy estimate of the inverse relative estimation error of the smoother $\tilde{m}_n$ at each of the observations, $m(X_j)/\tilde{m}_n(X_j)$. Smoothing V_j using a second linear smoother, say

$$\widehat{\alpha}_n(x) = \sum_{j=1}^{n} \omega_j(x;h_1) V_j,$$

produces an estimate for the inverse of the relative estimation error that can be used as a multiplicative correction of the pilot smoother. This leads to the (nonlinear) smoother

$$\widehat{m}_n(x) = \widehat{\alpha}_n(x)\tilde{m}_n(x). \tag{3}$$

The estimator (3) was first studied for fixed design by [20] and extended to the random design by [19]. In both cases, they assumed that the regression function had four continuous derivatives, and show an improvement in the convergence rate of the bias corrected Nadaraya-Watson kernel smoother. The idea of multiplicative bias reduction can be traced back to [9, 10], who proposed a parametrically guided local linear smoother that extended a parametric pilot regression estimate with a local polynomial smoother. It is showed that the resulting regression estimate improves on the naïve local polynomial estimate as soon as the pilot captures some of the features of the regression function.

3 Theoretical Analysis of Multiplicative Bias Reduction

In this section, we show that the multiplicative smoother has smaller bias with essentially no cost to the variance, assuming only two derivatives of the regression function. While the derivation of our results is for local linear smoothers, the technique used in the proofs can be easily adapted for other linear smoothers, and the conclusions remain essentially unchanged.

3.1 Assumptions

We make the following assumptions:

1. The regression function is bounded and strictly positive, that is, $b \geq m(x) \geq a > 0$ for all x.
2. The regression function is twice continuously differentiable everywhere.
3. The density of the covariate is strictly positive on the interior of its support in the sense that $f(x) \geq b(\mathscr{K}) > 0$ over every compact $\mathscr{K}$ contained in the support of f.
4. ε has finite fourth moments and has a symmetric distribution around zero.
5. Given a bounded symmetric probability density $K(\cdot)$, consider the weights $\omega_j(x;h)$ associated to the local linear smoother. That is, denote by $K_h(\cdot) = K(\cdot/h)/h$ the scaled kernel by the bandwidth h and define for $k = 0, 1, 2, 3$ the sums

   $$S_k(x) \equiv S_k(x;h) = \sum_{j=1}^{n} (X_j - x)^k K_h(X_j - x).$$

   Then

$$\omega_j(x;h) = \frac{S_2(x;h) - (X_j - x)S_1(x;h)}{S_2(x;h)S_0(x;h) - S_1^2(x;h)} K_h(X_j - x).$$

We set

$$\omega_{0j}(x) = \omega_j(x;h_0) \quad \text{and} \quad \omega_{1j}(x) = \omega_j(x;h_1).$$

6. The bandwidths h_0 and h_1 are such that

$$h_0 \to 0, \quad h_1 \to 0, \quad nh_0 \to \infty, \quad nh_1^3 \to \infty, \quad \frac{h_1}{h_0} \to 0 \quad \text{as} \quad n \to \infty.$$

The positivity assumption (assumption 1) on $m(x)$ is classical when we perform a multiplicative bias correction. It allows to avoid that the terms $Y_j/\tilde{m}_n(X_j)$ blows up. Of course, the regression function might cross the x-axis. For such a situation, [10] proposes to shift all response data Y_i a distance a, so that the new regression function $m(x)+a$ does not any more intersect with the x-axis. Such a method can also be performed here. Assumptions 2–4 are standard to obtain rate of convergence for nonparametric estimators. Assumption 5 means that we conduct the theory for the local linear estimate. The results can be generalized to other linear smoothers. Assumption 6 is not restrictive since it is satisfied for a wide range of values of h_0 and h_1.

3.2 A technical aside

The proof of the main results rests on establishing a stochastic approximation of estimator (3) in which each term can be directly analyzed.

Proposition 3.1 *We have*

$$\widehat{m}_n(x) = \mu_n(x) + \sum_{j=1}^{n} \omega_{1j}(x)A_j(x) + \sum_{j=1}^{n} \omega_{1j}(x)B_j(x) + \sum_{j=1}^{n} \omega_{1j}(x)\xi_j,$$

where $\mu_n(x)$, conditionally on $X_1,\ldots,X_n$ is a deterministic function, A_j, B_j and ξ_j are random variables. Under condition $nh_0 \to \infty$, the remainder ξ_j converges to 0 in probability and we have

$$\widehat{m}_n(x) = \mu_n(x) + \sum_{j=1}^{n} \omega_{1j}(x)A_j(x) + \sum_{j=1}^{n} \omega_{1j}(x)B_j(x) + \mathrm{O}_P\left(\frac{1}{nh_0}\right).$$

Remark 1. A technical difficulty arises because even though ξ_j may be small in probability, its expectation may not be small. We resolve this problem by showing that we only need to modify ξ_j on a set of vanishingly small probability to guarantee that its expectation is also small.

Definition 3.1 *Given a sequence of real numbers a_n, we say that a sequence of random variables $\xi_n = \mathrm{o}_p(a_n)$ if for all fixed $t > 0$,*

$$\limsup_{n \longrightarrow \infty} \mathbb{P}[|\xi_n| > t a_n] = 0.$$

We will need the following Lemma.

Lemma 3.1 *If $\xi_n = \mathrm{o}_p(a_n)$, then there exists a sequence of random variables $\xi_n^\star$ such that*

$$\limsup_{n \longrightarrow \infty} \mathbb{P}[\xi_n^\star \neq \xi_n] = 0 \quad \textit{and} \quad \mathbb{E}[\xi_n^\star] = \mathrm{o}(a_n).$$

We shall use the following notation

$$\mathbb{E}_\star[\xi_n] = \mathbb{E}[\xi_n^\star].$$

3.3 Asymptotic behavior

We deduce from Proposition 3.1 and Lemma 3.1 the following Theorem.

Theorem 3.1 *Under the assumptions (1)-(6), the estimator $\widehat{m}_n$ satisfies:*

$$\mathbb{E}_\star\left(\widehat{m}_n(x)|X_1,\ldots,X_n\right) = \mu_n(x) + \mathrm{O}_p\left(\frac{1}{n\sqrt{h_0 h_1}}\right) + \mathrm{O}_p\left(\frac{1}{nh_0}\right)$$

and

$$\mathbb{V}_\star(\widehat{m}_n(x)|X_1,\ldots,X_n) = \sigma^2 \sum_{j=1}^{n} w_{1j}^2(x) + \mathrm{O}_p\left(\frac{1}{nh_0}\right) + \mathrm{o}_p\left(\frac{1}{nh_1}\right).$$

We deduce from Theorem 3.1 that if the bandwidth h_0 of the pilot estimator converges to zero much slower than h_1, then $\widehat{m}_n$ has exactly the same asymptotic variance as the local linear smoother of the original data with bandwidth h_1. However, for finite samples, the two step local linear smoother can have a slightly larger variance depending on the choice of h_0. For the bias term, a limited Taylor expansion of $\mu_n(x)$ leads to the following result.

Theorem 3.2 *Under the assumptions (1)-(6), the estimator $\widehat{m}_n$ satisfies:*

$$\mathbb{E}_\star\left(\widehat{m}_n(x)|X_1,\ldots,X_n\right) = m(x) + \mathrm{o}_p(h_1^2).$$

Remark 2. Note that we only assume that the regression function is twice continuously differentiable. We do not add smoothness assumptions to improve the convergence rate from $\mathrm{O}_p(h_1^2)$ to $\mathrm{o}_p(h_1^2)$. In that manner, our analysis differs from that of [19] who assumed m to be four times continuously differentiable to conclude that the bias corrected smoother converged at the $\mathrm{O}_p(h_1^4)$ rate. For a study of the local linear estimate in the presence of jumps in the derivative, we refer the reader to [5].

Remark 3. Under similar smoothness assumptions, [8, 10, 21] have provided a comprehensive asymptotic behavior for the multiplicative bias corrected estimator with a parametric guide. They obtain the same asymtptotic variance as the local linear estimate and a bias reduction provided the parametric guide captures some of the features of the regression function. We obtain a similar result when the rate of decay of the bandwidth of the pilot estimate is carefully chosen.

Combining Theorem 3.1 and Theorem 3.2, we conclude that the multiplicative adjustment performs a bias reduction on the pilot estimator without increasing the asymptotic variance. The asymptotic behavior of the bandwidths h_0 and h_1 is constrained by assumption 6. However, it is easily seen that this assumption is satisfied for a large set of values of h_0 and h_1. For example, the choice $h_1 = c_1 n^{-1/5}$ and $h_0 = c_0 n^{-\alpha}$ for $0 < \alpha < 1/5$ leads to

$$\mathbb{E}_\star\left(\widehat{m}_n(x)|X_1,\ldots,X_n\right) - m(x) = \mathrm{o}_p(n^{-2/5})$$

and

$$\mathbb{V}_\star(\widehat{m}_n(x)|X_1,\ldots,X_n) = \mathrm{O}_p\left(n^{-4/5}\right).$$

Remark 4. Estimators with bandwidths of order $\mathrm{O}(n^{-\alpha})$ for $0 < \alpha < 1/5$ are oversmoothing the true regression function, and as a result, the magnitude of their biases are of larger than the magnitude of their standard deviations. We conclude that the multiplicative adjustment performs a bias reduction on the pilot estimator.

4 Numerical examples

Results presented in the previous sections show that our procedure allows to reduce the bias of nonparametric smoothers at no cost for the asymptotic variance. The simulation study in this section shows that the practical benefits of this asymptotic behavior already emerge at modest sample sizes.

4.1 Local study

To illustrate numerically the reduction in the bias and associate (limited) increase of the variance achieved by the multiplicative bias correction, consider estimating the regression function

$$m(x) = 5 + 3|x|^{5/2} + x^2 + 4\cos(10x)$$

at $x = 0$ (see Figure 1).
The local linear smoother is known to under-estimate the regression function at local maxima and over-estimate local minima, and hence, this example provides a

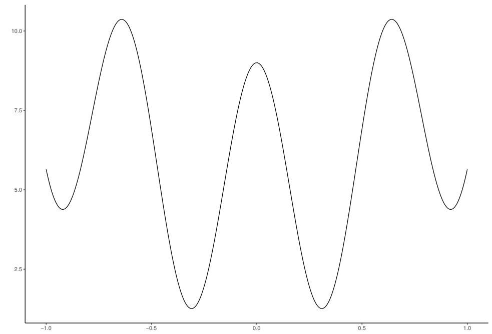

Fig. 1 The regression function to be estimated.

good example to explore bias-reduction variance-increase trade-off. Furthermore, because the second derivative of this regression function is continuous but not differentiable at the origin, the results previously obtained by [20] do not apply.

For our Monte-Carlo simulation, the data are generated according to the model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, 100,$$

where ε_i are independent $\mathscr{N}(0,1)$ variables, and the covariates X_i are independent Uniform random variables on the interval $[-1,1]$.

We first consider the local linear estimate and we study its performances over a grid of bandwidths $\mathscr{H} = [0.005, 0.1]$. For the new estimate, the theory recommends to start with an oversmooth pilot estimate. In this regard, we take $h_0 = 0.1$ and study the performance of the multiplicative bias corrected estimate for $h_1 \in \mathscr{H}_1 = [0.005, 0.12]$. To explore the stability of our two stages estimator with respect to h_0, we also consider the choice $h_0 = 0.02$. For such a choice, the pilot estimate clearly undersmoothes the regression function. For both estimates, we take the Gaussian kernel $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$.

We conduct a Monte Carlo study to estimate bias and variance of each estimate at $x = 0$. To this end, we compute the estimate at $x = 0$ for 1 000 samples $(X_i, Y_i), i = 1, \ldots, 100$. The same design $X_i, i = 1, \ldots, 100$ is used for all the sample. The bias at point $x = 0$ is estimated by subtracting $m(0)$ at the mean value of the estimate at $x = 0$ (the mean value is computed over the 1 000 replications). Similarly we estimate the variance at $x = 0$ by the variance of the values of the estimate at this point. Figure 2 presents squared bias, variance and mean square error of each estimate for different values of bandwidths h for the local linear smoother and h_1 for our estimate.
Comparing panel (a) and (c) in Figure 2, we see that if the pilot smoother underestimates the regression function, then the bias is small but the variance is large. For such a pilot smoother, applying a bias correction does not provide any benefit, and the resulting estimator can be worse than a good local linear smoother. Intuitively, the bias of the pilot smoother is already small at the cost of a larger variance, and

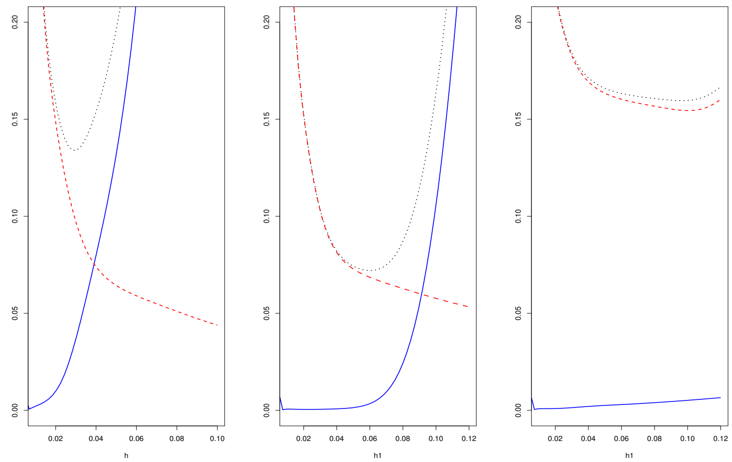

Fig. 2 Mean square error (dotted line), squared bias (solid line) and variance (dashed line) of the local linear estimate (left) and multiplicative bias corrected estimate with $h_0 = 0.1$ (center) and $h_0 = 0.02$ (right) at point $x = 0$.

operating a bias reduction provides little benefit to the bias and can only make the variance worse, leading to a suboptimal smoother.
Comparing panel (a) and (b) in Figure 2, we note that the squared bias is smaller for the bias corrected smoother over the standard local linear smoother, while the variance of both smoothers are essentially the same. As a result, the mean squared error for the bias corrected smoother is smaller than that of the local linear smoother. This shows that the asymptotic properties outlined in theorems 3.1 and 3.2 emerge for moderate sample sizes. Table 1 quantifies the benefits of the bias corrected smoother over the classical local linear smoother.

	MSE	Bias^2	Variance
LLE	0.130	0.031	0.098
MBCE	0.068	0.003	0.065

Table 1 Optimal mean square error (MSE) for the local linear estimate (LLE) and the multiplicative bias corrected estimate (MBCE) with $h_0 = 0.1$ at point $x = 0$.

We conclude our local study by comparing the multiplicative bias correction smoother starting from a nonparametric pilot with the multiplicative bias correction smoother starting from a parametric model, as suggested by Glad [10]. Specifically, we compare our smoother to multiplicative bias smoothers starting with the following three parametric models:

- first, the guide is chosen correctly and belong to the true parametric family:

$$\tilde{m}_n^1(x) = \hat{\beta}_0 + \hat{\beta}_1 |x|^{5/2} + \hat{\beta}_2 x^2 + \hat{\beta}_3 \cos(10x);$$

- second, we consider a linear parametric guide (which is obviously wrong):

$$\tilde{m}_n^2(x) = \hat{\beta}_0 + \hat{\beta}_1 x;$$

- finally, we use a more reasonable guide, not correct, but that can reflect some a priori idea on the regression function

$$\tilde{m}_n^3(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \ldots + \hat{\beta}_8 x^8.$$

All the estimates $\hat{\beta}_j$ stands for the classical least square estimates.

The multiplicative bias correction is performed on these parametric starts using the local linear estimate. The performance of the resulting estimates is measured over a grid of bandwidths $\mathscr{H}_2 = [0.005; 0.4]$. Bias and variance of each estimate are still estimated at $x = 0$. We keep the same setting as above and all the results are averaged over the same 1 000 replications. We display in Table 2 the optimal MSE calculated over the grid $\mathscr{H}_2$.

	MSE	Bias^2	Variance
start $\tilde{m}_n^1$	0.052	0.000	0.052
start $\tilde{m}_n^2$	0.129	0.031	0.098
start $\tilde{m}_n^3$	0.090	0.019	0.071
MBCE	0.068	0.003	0.065

Table 2 Pointwise optimal mean square error at $x = 0$ for the multiplicative bias corrected estimates with parametric starts $\tilde{m}_n^j$, $j = 1,2,3$, compared to a multiplicative bias corrected smoother starting with initial bandwidth $h_0 = 0.1$.

As expected, the performance depends on the choice of the parametric start. Unsurprisingly, the performance of the smoother starting with the parametric guide $\tilde{m}_n^1$ (which belongs to the true model) is best. Table 2 shows that (in term of MSE) the estimate studied in this paper is better than the corrected estimated with parametric start $\tilde{m}_n^2$ and $\tilde{m}_n^3$. This suggests that in practice, when little priori information on the target regression function is available, the method proposed in the present paper is preferable.

4.2 Global study

The theory in Section 3 does not address the practical issue of bandwidths selection for both the pilot smoother and the multiplicative adjustment. [2] suggests adapting existing automatic bandwidth selection procedures to this problem. There is a large literature on automatic bandwidth selection, including [14, 15]. In this section, we present a numerical investigation of the leave-one-out cross-validation method to select both bandwidths h_0 and h_1 as to minimize the integrated square error of the estimator. The resulting bias smoother is compared with a local polynomial smoother, whose bandwidth is selected in a similar manner.

Our selection of test functions for our investigation relies on the comprehensive numerical study of [18]. We will only compare our multiplicative bias corrected smoother with the classical local linear smoother. In all our examples, we use a Gaussian kernel to construct nonparametric smoothers to estimate the following regression functions (see Figure 3):

(1) $m_1(x) = \sin(5\pi x)$
(2) $m_2(x) = \sin(15\pi x)$
(3) $m_3(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$
(4) $m_4(x) = 0.3\exp[-64(x-.25)^2] + 0.7\exp[-256(x-.75)^2]$.

from data $Y_{ji} = m_j(X_i) + \varepsilon_{ji}$, with disturbances $\varepsilon_{j1},\ldots,\varepsilon_{jn}$ i.i.d. Normal with mean zero and standard deviation $\sigma_j = 0.25\|m_j\|_2,\ j = 1,\ldots,4$, and $X_1,\ldots,X_n$ i.i.d. Uniform on $[-0.2, 1.2]$.

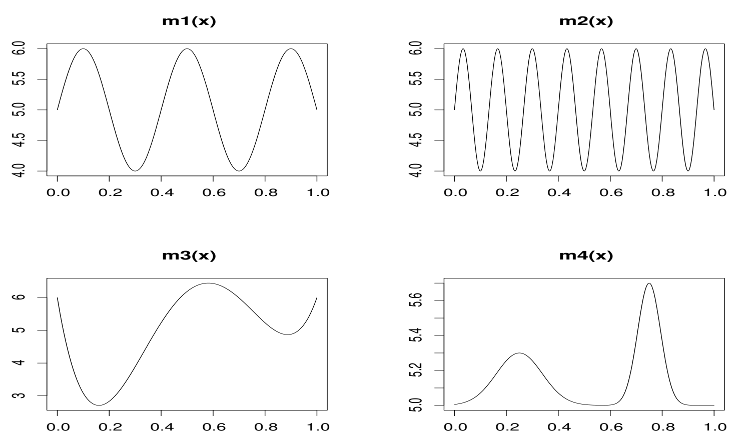

Fig. 3 Regression functions to be estimated.

We use a cross validation device to select both h_0 and h_1 by minimizing simultaneously over a finite grid $\mathscr{H}$ of bandwidths h_0 and h_1 the leave-one-out prediction error. That is, given a grid $\mathscr{H}$, we choose the pair $(\hat{h}_0, \hat{h}_1)$ defined by

$$(\hat{h}_0, \hat{h}_1) = \underset{(h_0,h_1)\in\mathscr{H}\times\mathscr{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_n^i(X_i))^2.$$

Here $\widehat{m}_n^i$ stands for the prediction of the bias corrected smoother at X_i, estimated without the observation (X_i, Y_i). We use the Integrated Square Error (ISE)

$$ISE(\widehat{m}) = \int_0^1 (m(x) - \widehat{m}(x))^2\, \mathrm{d}x,$$

to measure the performance of an estimator $\widehat{m}$. Note that even though our estimators are defined on the interval $[-0.2, 1.2]$ (the support of the explanatory variable), we evaluate the integral on the interval $[0, 1]$ to avoid boundary effects.
Table 3 compares the median ISE over 1 000 replication, of a standard local linear smoother and our bias corrected smoother from a samples of size $n = 100$. This table further presents the median selected bandwidth, and the ratio of the ISE.

	LLE		MBCE			
	h	ISE ($\times 100$)	h_0	h_1	ISE ($\times 100$)	R_{ISE}
m_1	0.023	0.957	0.050	0.032	0.735	1.316
m_2	0.011	6.094	0.028	0.012	4.771	1.286
m_3	0.028	2.022	0.071	0.054	1.281	1.591
m_4	0.018	0.087	0.034	0.024	0.074	1.187

Table 3 Median over 1 000 replications of the selected bandwidths and of the integrated square error of the selected estimates. LLE and MBCE stands for local linear estimate and multiplicative bias corrected estimate.

First, in all four cases, the ISE for the MBCE is smaller than that of the LLE. Second, we note that both bandwidths for the multiplicative bias corrected are larger than the optimal bandwidth of the classical local linear smoother. That h_0 is larger is supported by the theory, as the pilot smoother needs to oversmooth. We surmise that larger bandwidth h_1 reflects the fact that the pilot is reasonably close to the true regression function, and hence the multiplicative correction is quite smooth and thus can accomodate a larger bandwidth. Figure 4 displays the boxplots of the integrated square error for each estimate.
Figure 5 presents, for the regression function m_1 with $n = 100$ and 1 000 iterations, different estimators on a grid of points. In lines is the true regression function which is unknown. For every point on a fixed grid, we plot, side by side, the mean over 1 000 replications of our estimator at that point (left side) and on the right side of that point the mean over 1 000 replications of the local polynomial estimator. Leave-one-out cross validation is applied to select the bandwidths h_0 and h_1 for our estimator and the bandwidth h for the local polynomial estimator. We add also the interquartile interval in order to see the fluctuations of the different estimators. In this example, our estimator reduces the bias by increasing the peak and decreasing the valleys. Moreover, the interquartile intervals look similar for both estimator, as predicted by the theory.

5 Conclusion

This paper revisits the idea of multiplicative bias reduction under minimal conditions and shows that it is possible to reduce the bias with little effect to the variance. Our theory proves that our proposed estimator has zero asymptotic bias while main-

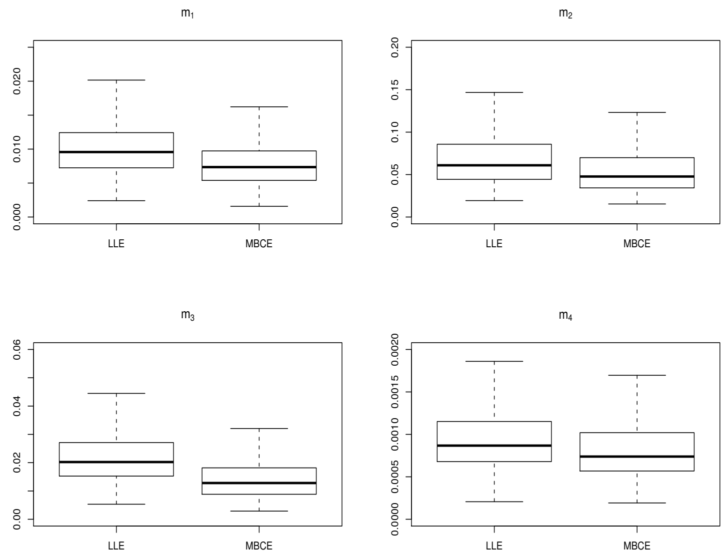

Fig. 4 Boxplot of the integrated square error over the 1 000 replications.

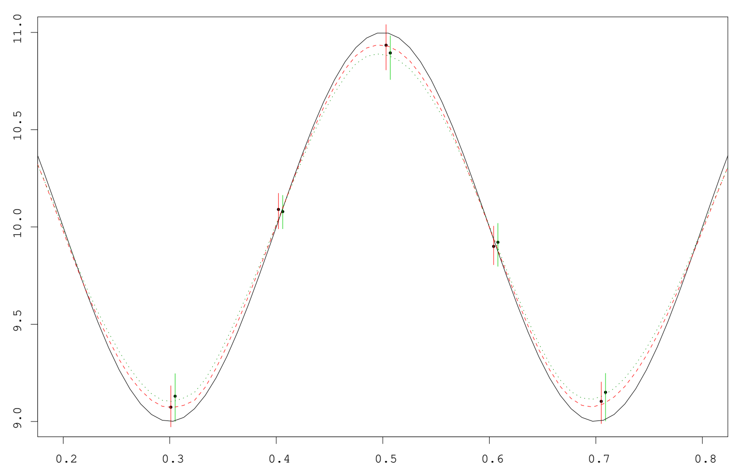

Fig. 5 The solid curve represents the true regression function, our estimator is in dashed line and local linear smoother is dotted.

taining the same asymptotic variance than the original smoother. The simulation study in this paper show that this desirable property emerges for even modest sample sizes. The one downside of our estimator is that the computation of data driven ”optimal” bandwidths is computationally expensive.

6 Proofs

6.1 Proof of Proposition 3.1

Write the bias corrected estimator

$$\widehat{m}_n(x) = \sum_{j=1}^{n} \omega_{1j}(x)\frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)}Y_j = \sum_{j=1}^{n} \omega_{1j}(x)R_j(x)Y_j,$$

and let us approximate the quantity $R_j(x)$. Define

$$\bar{m}_n(x) = \sum_{j=1}^{n} \omega_{0j}(x)m(X_j) = \mathbb{E}\left(\tilde{m}_n(x)|X_1,\ldots,X_n\right),$$

and observe that

$$\begin{aligned}
R_j(x) =& \frac{\tilde{m}_n(x)}{\tilde{m}_n(X_j)} \\
=& \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \times \left(1+\frac{\tilde{m}_n(x)-\bar{m}_n(x)}{\bar{m}_n(x)}\right) \times \left(1+\frac{\tilde{m}_n(X_j)-\bar{m}_n(X_j)}{\bar{m}_n(X_j)}\right)^{-1} \\
=& \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \times [1+\Delta_n(x)] \times \frac{1}{1+\Delta_n(X_j)},
\end{aligned}$$

where

$$\Delta_n(x) = \frac{\tilde{m}_n(x)-\bar{m}_n(x)}{\bar{m}_n(x)} = \frac{\sum_{l\le n}\omega_{0l}(x)\varepsilon_l}{\sum_{l\le n}\omega_{0l}(x)m(X_l)}.$$

Write now $R_j(x)$ as

$$R_j(x) = \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}\left[1+\Delta_n(x)-\Delta_n(X_j)+r_j(x,X_j)\right]$$

where $r_j(x,X_j)$ is a random variable converging to 0 to be define latter on. Given the last expression and model (1), estimator (3) could be written as

$$\begin{aligned}
\widehat{m}_n(x) =& \sum_{j=1}^{n} \omega_{1j}(x)R_j(x)Y_j \\
=& \sum_{j=1}^{n} \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}m(X_j) + \sum_{j=1}^{n} \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}\left[\varepsilon_j + m(X_j)\left(\Delta_n(x)-\Delta_n(X_j)\right)\right] \\
&+ \sum_{j=1}^{n} \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}\left(\Delta_n(x)-\Delta_n(X_j)\right)\varepsilon_j + \sum_{j=1}^{n} \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}r_j(x,X_j)Y_j \\
=& \mu_n(x) + \sum_{j=1}^{n} \omega_{1j}(x)A_j(x) + \sum_{j=1}^{n} \omega_{1j}(x)B_j(x) + \sum_{j=1}^{n} \omega_{1j}(x)\xi_j.
\end{aligned}$$

which is the first part of the proposition. Under assumption set forth in Section 3.1, the pilot smoother $\tilde{m}_n$ converges to the true regression function $m(x)$. [1] show that this convergence is uniform over compact sets $\mathscr{K}$ contained in the support of the density of the covariate X. As a result, for n large enough $\sup_{x\in\mathscr{K}} |\tilde{m}_n(x) - \bar{m}_n(x)| \leq \frac{1}{2}$ with probability 1. So a limited expansion of $(1+u)^{-1}$ yields for $x \in \mathscr{K}$

$$R_j(x) = \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} \left[1 + \Delta_n(x) - \Delta_n(X_j) + \mathrm{O}_p\left(|\Delta_n(x)\Delta_n(X_j)| + \Delta_n^2(X_j)\right)\right],$$

thus

$$\xi_j = \mathrm{O}_p\left(|\Delta_n(x)\Delta_n(X_j)| + \Delta_n^2(X_j)\right).$$

Under the stated regularity assumptions, we deduce that $\xi_j = \mathrm{O}_p\left(\frac{1}{nh_0}\right)$, leading to the announced result. Proposition 3.1 is proved.

6.2 Proof of Lemma 3.1

By definition $\limsup_{n\longrightarrow\infty} \mathbb{P}[|\xi_n| > ta_n] = 0$ for all $t > 0$, so that a triangular array argument shows that there exists an increasing sequence $m = m(k)$ such that

$$\mathbb{P}\left[|\xi_n| > \frac{a_n}{k}\right] \leq \frac{1}{k} \qquad \text{for all } n \geq m(k).$$

For $m(k) \leq n \leq m(k+1) - 1$, define

$$\xi_n^\star = \begin{cases} \xi_n & \text{if } |\xi_n| < k^{-1}a_n \\ 0 & \text{otherwise.} \end{cases}$$

It follows from the construction of $\xi_n^\star$ that for $n \in (m(k), m(k+1)-1)$,

$$\mathbb{P}[\xi_n \neq \xi_n^*] = \mathbb{P}[|\xi_n| > k^{-1}a_n] \leq \frac{1}{k},$$

which converges to zero as n goes to infinity. Finally set $k(n) = \sup\{k : m(k) \leq n\}$, we obtain

$$\mathbb{E}[|\xi_n^\star|] \leq \frac{a_n}{k(n)} = \mathrm{o}(a_n).$$

6.3 Proof of Theorem 3.1

Recall that $\widehat{m}_n(x) = \mu_n(x) + \sum_{j=1}^n \omega_{1j}(x)A_j(x) + \sum_{j=1}^n \omega_{1j}(x)B_j(x) + \mathrm{O}_P\left(\frac{1}{nh_0}\right)$. Focus on the conditional bias, we get

$$\mathbb{E}(\mu_n(x)|X_1,\dots,X_n) = \mu_n(x), \quad \mathbb{E}(A_j(x)|X_1,\dots,X_n) = 0$$

and

$$\mathbb{E}(B_j(x)|X_1,\dots,X_n) = \frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}\sigma^2\Big(\frac{\omega_{0j}(x)}{\bar{m}_n(x)} - \frac{\omega_{0j}(X_j)}{\bar{m}_n(X_j)}\Big).$$

Since

$$\left|\sum_{j=1}^n \omega_{1j}(x)\omega_{0j}(x)\right| \le \sqrt{\sum_{j=1}^n \omega_{1j}(x)^2}\sqrt{\sum_{j=1}^n \omega_{0j}(x)^2} = \mathrm{O}_p\left(\frac{1}{n\sqrt{h_0h_1}}\right),$$

we deduce that

$$\mathbb{E}\left(\sum_{j=1}^n \omega_{1j}(x)B_j(x)\Big|X_1,\dots,X_n\right) = \mathrm{O}_p\left(\frac{1}{n\sqrt{h_0h_1}}\right).$$

This proves the first part of the Theorem. For the conditional variance, we use the following expansion of the two stages estimator

$$\widehat{m}_n(x) = \sum_{j=1}^n \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}Y_j\,(1+[\Delta_n(x)-\Delta_n(X_j)]) + \mathrm{O}_p\left(\frac{1}{nh_0}\right).$$

Using the fact that the residuals have four finite moments and have a symmetric distribution around 0, a moment's thought shows that

$$\mathbb{V}(Y_j\,[\Delta_n(x)-\Delta_n(X_j)]\,|X_1,\dots,X_n) = \mathrm{O}_p\left(\frac{1}{nh_0}\right)$$

and

$$\mathrm{Cov}(Y_j,Y_j\,[\Delta_n(x)-\Delta_n(X_j)]\,|X_1,\dots,X_n) = \mathrm{O}_p\left(\frac{1}{nh_0}\right).$$

Hence

$$\mathbb{V}_\star(\widehat{m}_n(x)|X_1,\dots,X_n) = \mathbb{V}\left(\sum_{j=1}^n \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}Y_j\Big|X_1,\dots,X_n\right) + \mathrm{O}_p\left(\frac{1}{nh_0}\right).$$

Observe that the first term on the right hand side of this equality can be seen as the variance of the two stages estimator with a deterministic pilot estimator. It follows from [10] that

$$\mathbb{V}\left(\sum_{j=1}^n \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)}Y_j\Big|X_1,\dots,X_n\right) = \sigma^2\sum_{j=1}^n \omega_{1j}^2(x) + \mathrm{o}_p\left(\frac{1}{nh_1}\right),$$

which proves the theorem.

6.4 Proof of Theorem 3.2

Recall that

$$\mu_n(x) = \sum_{j\le n} \omega_{1j}(x)\frac{\bar{m}_n(x)}{\bar{m}_n(X_j)} m(X_j).$$

We consider the limited Taylor expansion of the ratio

$$\frac{m(X_j)}{\bar{m}_n(X_j)} = \frac{m(x)}{\bar{m}_n(x)} + (X_j - x)\left(\frac{m(x)}{\bar{m}_n(x)}\right)' + \frac{1}{2}(X_j - x)^2\left(\frac{m(x)}{\bar{m}_n(x)}\right)''(1+\mathrm{o}_p(1)),$$

then

$$\begin{aligned}\mu_n(x) = \bar{m}_n(x)\Bigg\{&\frac{m(x)}{\bar{m}_n(x)}\sum_{j=1}^n \omega_{1j}(x) + \left(\frac{m(x)}{\bar{m}_n(x)}\right)'\sum_{j=1}^n (X_j - x)\omega_{1j}(x)\\ &+\frac{1}{2}\left(\frac{m(x)}{\bar{m}_n(x)}\right)''\sum_{j=1}^n (X_j - x)^2\omega_{1j}(x)(1+\mathrm{o}_p(1))\Bigg\}.\end{aligned}$$

It is easy to verify that $\sum_{j=1}^n \omega_{1j}(x) = 1$, $\sum_{j=1}^n (X_j - x)\omega_{1j}(x) = 0$, and

$$\Sigma_2(x;h_1) = \sum_{j=1}^n (X_j - x)^2\omega_{1j}(x) = \frac{S_2^2(x;h_1) - S_3(x;h_1)S_1(x;h_1)}{S_2(x;h_1)S_0(x;h_1) - S_1^2(x;h_1)}.$$

For random designs, we can further approximate (see, e.g., [27])

$$S_k(x,h_1) = \begin{cases} h^k\sigma_K^k f(x) + \mathrm{o}_p(h^k) & \text{for } k \text{ even}\\ h^{k+1}\sigma_K^{k+1} f'(x) + \mathrm{o}_p(h^{k+1}) & \text{for } k \text{ odd,}\end{cases}$$

where $\sigma_K^k = \int u^k K(u)\,\mathrm{d}u$. Therefore

$$\begin{aligned}\Sigma_2(x;h_1) &= h_1^2\int u^2 K(u)\,\mathrm{d}u + \mathrm{o}_p(h_1^2)\\ &= \sigma_K^2 h_1^2 + \mathrm{o}_p(h_1^2),\end{aligned}$$

so that we can write $\mu_n(x)$ as

$$\begin{aligned}\mu_n(x) =&\bar{m}_n(x)\left\{\frac{m(x)}{\bar{m}_n(x)} + \frac{\sigma_K^2 h_1^2}{2}\left(\frac{m(x)}{\bar{m}_n(x)}\right)'' + \mathrm{o}_p(h_1^2)\right\}\\ =&m(x) + \frac{\sigma_K^2 h_1^2}{2}\bar{m}_n(x)\left(\frac{m(x)}{\bar{m}_n(x)}\right)'' + \mathrm{o}_p(h_1^2).\end{aligned}$$

Moreover

$$\left(\frac{m(x)}{\bar{m}_n(x)}\right)'' = \frac{\bar{m}_n^2(x)m''(x)}{\bar{m}_n^3(x)} - 2\frac{\bar{m}_n(x)\bar{m}_n'(x)m'(x)}{\bar{m}_n^3(x)} - \frac{m(x)\bar{m}_n(x)\bar{m}_n''(x)}{\bar{m}_n^3(x)} + 2\frac{m(x)(\bar{m}_n'(x))^2}{\bar{m}_n^3(x)}$$

and applying the usual approximations, we conclude that

$$\left(\frac{m(x)}{\bar{m}_n(x)}\right)'' = \mathrm{o}_p(1).$$

Putting all pieces together, we obtain

$$\mathbb{E}_\star(\widehat{m}_n(x)|X_1,\ldots,X_n) - m(x) = \mathrm{o}_p(h_1^2) + \mathrm{O}_p\left(\frac{1}{n\sqrt{h_0h_1}}\right) + \mathrm{O}_p\left(\frac{1}{nh_0}\right).$$

Since $nh_1^3 \longrightarrow \infty$ and $\frac{h_1}{h_0} \longrightarrow 0$, we conclude that the bias is of order $\mathrm{o}_p(h_1^2)$.

References

1. P.J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1:1071–1095, 1973.
2. T. Burr, N. Hengartner, E. Matzner-Løber, S. Myers, and L. Rouvière. Smoothing low resolution gamma spectra. *IEEE Transactions on Nuclear Science*, 57:2831–2840, 2010.
3. P-A. Cornillon, N. Hengartner, N. Jegou, and E. Matzner-Løber. Iterative bias reduction: a comparative study. *Statistics and Computing*, 23(6):777–791, 2013.
4. P-A. Cornillon, N. Hengartner, and E. Matzner-Løber. Recursive bias estimation for multivariate regression smoothers. *ESAIM*, 18:483–502, 2014.
5. L. Desmet and I. Gijbels. Local linear fitting and improved estimation near peaks. *The Canadian Journal of Statistics*, 37:473–475, 2009.
6. M. Di Marzio and C. Taylor. Multiple kernel regression smoothing by boosting. 2007.
7. J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Application, Theory and Methodologies*. Chapman et Hall, New York, 1996.
8. J. Fan, Y. Wu, and Y. Feng. Local quasi-likelihood with a parametric guide. *Annals of Statistics*, 37:4153–4183, 2009.
9. I. Glad. A note on unconditional properties of parametrically guided nadaraya-watson estimator. *Statistics and Probability Letters*, 37:101–108, 1998.
10. I. Glad. Parametrically guided non-parametric regression. *Scandinavian Journal of Statistics*, 25:649–668, 1998.
11. J. Gustafsson, M. Hagmann, J.P. Nielsen, and O. Scaillet. Local transformation kernel density estimation of loss distributions. *Journal of Business and Economic Statistics*, 27:161–175, 2009.
12. M. Hagmann and O. Scaillet. Local multiplicative bias correction for asymmetric kernel density estimators. *Journal of Econometrics*, 141:213–249, 2007.
13. N. Hengartner and E. Matzner-Løber. Asymptotic unbiased density estimators. *ESAIM: Probability and Statistics*, 13:1–14, 2009.
14. N. Hengartner and M. Wegkamp. Estimation and selection procedures in regression: an l_1 approach. *Canadian Journal of Statistics*, 29(4):621–632, 2001.
15. N. Hengartner, M. Wegkamp, and E. Matzner-Løber. Bandwidth selection for local linear regression smoothers. *JRSS B*, 64:1–14, 2002.

16. Hirukawa. A modified nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis*, 27:441–476, 2006.
17. M. Hirukawa. Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval. *Computational Statistics and Data Analysis*, 54:473–495, 2010.
18. C. Hurvich, G. Simonoff, and C. L. Tsai. Smoothing parameter selection in nonparametric regression using and improved akaike information criterion. *J. R. Statist. Soc. B*, 60:271–294, 1998.
19. M.C. Jones, O. Linton, and J.P. Nielsen. A simple and effective bias reduction method for kernel density estimation. *Biometrika*, 82:327–338, 1995.
20. O. Linton and J. P. Nielsen. A multiplicative bias reduction method for nonparametric regression. *Statistics and Probability Letters*, 19(181–187), 1994.
21. C. Martins-Filho, S. Mishra, and A. Ullah. A class of improved parametrically guided nonparametric regression estimators. *Econometric Reviews*, 27:542–573, 2008.
22. S. Mishra, L. Su, and A. Ullah. Semiparametric estimator of time series conditional variance. *Journal of Business and Economic Statistics*, 28:256–274, 2010.
23. J.P. Nielsen. Multiplicative bias correction in kernel hazard estimation. *Scandinavian Journal of Statistics*, 25:541–553, 1998.
24. J.P. Nielsen and C. Tanggaard. Boundary and bias correction in kernel hazard estimation. *Scandinavian Journal of Statistics*, 28:675–698, 2001.
25. D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New-York, 1992.
26. J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.
27. M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.
28. Z. Xiao and O. Linton. A nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis*, 23:215–250, 2002.