Local bandwidth selectors for deconvolution kernel density estimation

Achilleas Achilleos [1] and Aurore Delaigle[2]

[1] Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

[2] Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia.

Abstract

We consider kernel density estimation when the observations are contaminated by measurement errors. It is well known that the success of kernel estimators depends heavily on the choice of a smoothing parameter called the bandwidth. A number of data-driven bandwidth selectors exist in the literature, but they are all global. Such techniques are appropriate when the density is relatively simple, but local bandwidth selectors can be more attractive in more complex settings. We suggest several data-driven local bandwidth selectors and illustrate via simulations the significant improvement they can bring over a global bandwidth.

Keywords: contaminated data; data-driven bandwidth; EBBS; errors-in-variables; kernel smoothing; measurement errors; plug-in; smoothing parameter.

1 Introduction

Nonparametric estimation of a density from data observed with additive measurement errors has received a lot of attention in the last two decades. In this problem the goal is to estimate the density f_X of a random variable X from observations on a random variable W contaminated by additive measurement errors. One of the most popular methods for estimating a density from such data is the deconvolution kernel method of Carroll and Hall (1988) and Stefanski and Carroll (1990). See also Fan (1991a,b).

It is well known that the performance of this estimator depends crucially on the choice of a smoothing parameter h called the bandwidth. Several procedures of bandwidth selection for the deconvolution estimator have been developed in the literature, but these methods are global, which means that a single bandwidth is used to construct the estimator of $f_X(x)$ at every point x.

In the error-free context (i.e. when the data are observed without measurement errors), a global bandwidth is appropriate for estimating simple densities, but when the target density f_X has various local features, the results can be improved by using a local bandwidth $h(x)$ which depends on x. Since the deconvolution problem is notoriously very hard, this sort of improvement would be particularly attractive in the measurement error context. To our knowledge, this problem has not been studied before in the deconvolution context. Techniques for choosing a local bandwidth in the standard error-free kernel density estimation problem have been suggested by various authors, including Sheather (1983, 1986), Brockmann et al. (1993), Schucany (1995), Hazelton (1996), Fan et al. (1996) and Farmen and Marron (1999). However, these methods are difficult to adapt to the error case as either they are based on an explicit form of the optimal bandwidth, which is not available in the error context, or they require estimation of pointwise quantities, which is considerably harder in the error setting. The EBBS technique developed by Ruppert (1997) in the context of error-free regression is less standard, but it seems more attractive to adapt to the error case, since it does not suffer from these drawbacks. We develop an EBBS local bandwidth selector for the deconvolution problem and suggest two other local bandwidth selectors: an integrated plug-in approach and a local SIMEX (SIMulation EXtrapolation) procedure.

This paper is organized as follows. In section 2 we define the model and estimators. We develop data-driven local bandwidth selectors in section 3. In section 4 we describe the implementation of the procedures and define variants of the methods developed in section 3. We present results of our extensive comparative simulation study in section 5. Finally, section 6 summarizes the important conclusions from our simulation study. All proofs are deferred to the appendix, and additional simulation results are provided in Achilleos and Delaigle (2011).

2 Model and estimators

Suppose we want to estimate the density f_X of a random variable X but only observe an i.i.d. sample $W_1, \ldots, W_n$ of contaminated observations generated by the model

$$W = X + U \quad \text{and} \quad X \perp\!\!\!\perp U, \tag{2.1}$$

where $U \sim f_U$ represents a measurement error and is such that $E(U) = 0$ and $\text{var}(U) = \sigma_U^2 < \infty$. Throughout this work we assume, as is commonly done in the literature, that the error density f_U is known. When this density is unknown, it can be easily estimated from replicated data without affecting first order asymptotic properties of estimators. See for example Delaigle et al. (2008). The literature on deconvolution is large; see for example Butucea and Matias (2005), Hall and Qiu (2005), Carroll et al. (2006), Holzmann et al. (2007), Staudenmayer et al. (2008), Hazelton and Turlach (2009), Meister (2009) and Wang et al. (2010) for recent related work. For kernel deconvolution in the presence of boundary points, which we do not consider in this work, see Delaigle and Gijbels (2006) and Zhang and Karunamuni (2009).

Let $h > 0$ denote a smoothing parameter called the bandwidth, K a smooth and symmetric function called the kernel, and let $\phi_K(t) = \int e^{itx} K(x)\, dx$ and $\phi_U(t) = E(e^{iUt})$. The deconvolution kernel density estimator of f_X, developed by Carroll and Hall (1988) and Stefanski and Carroll (1990), is defined by

$$\hat{f}_X(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K_U\Big(\frac{x - W_i}{h}\Big), \tag{2.2}$$

where

$$K_U(v) = \frac{1}{2\pi} \int e^{-itv} \frac{\phi_K(t)}{\phi_U(t/h)}\, dt.$$

Our goal is to develop data-driven methods for selecting the bandwidth h in practice. Techniques already exist in the literature. They include the cross-validation method of Stefanski and Carroll (1990) (see also Hesse, 1999), the plug-in approach of Delaigle and Gijbels (2002, 2004a), and the bootstrap approach of Delaigle and Gijbels (2004b). However, these methods construct a global bandwidth, that is, they focus on the case where the same bandwidth h is used at all points x. We wish to develop local bandwidth selectors, where the bandwidth $h = h(x)$ depends on x.

A theoretical optimal local bandwidth at x is a bandwidth that minimizes a distance between $f_X(x)$ and its estimator $\hat{f}_X(x;h)$. In this paper, we use the Mean Squared Error (MSE) distance, defined by

$$\text{MSE}(x;h) = \text{E}\left[\left\{\hat{f}_X(x;h) - f_X(x)\right\}^2\right],$$

and define the theoretical optimal local bandwidth by

$$h_{\text{MSE}}(x) = \text{argmin}_h \, \text{MSE}(x;h).$$

Of course, the MSE depends on the unknown f_X and in practice, $h_{\text{MSE}}(x)$ has to be estimated. We construct such an estimator in section 3.3. In the global setting, bandwidths based on asymptotic expressions of the integrated MSE, rather than integrated MSE itself, are often very competitive. Since, a priori, the same can be expected in the local case, in section 3 we also construct bandwidth selectors based on the asymptotically dominating part of the MSE, the Asymptotic Mean Squared Error (AMSE). Let $\mu_{K,j} = \int u^j K(u)\, du$ and recall that a kernel K is of order k if $\mu_{K,0} = 1$, $\mu_{K,j} = 0$ for $j = 1, \ldots, k-1$, and $\mu_{K,k} = c$, where $c \neq 0$ is a finite constant. We make the following assumptions.

Condition A:

(A1) $\phi_U(t) \neq 0$ for all t;

(A2) $h \to 0$ and $nh \to \infty$ as $n \to \infty$;

(A3) K is of order k and is such that $\int |u|^{k+1} K(u)\, du < \infty$, $\sup_t |\phi_K(t)/\phi_U(t/h)| < \infty$, $\int |\phi_K(t)/\phi_U(t/h)|\, dt < \infty$;

(A4) $\int |\phi_X| < \infty$ and f_X is $k+1$ times differentiable, and $\|f_X^{(j)}\|_\infty < \infty$ for $j = 0, \ldots, k+1$.

Under these conditions, the AMSE of $\hat{f}_X$ at x is given by

$$\begin{aligned} \text{AMSE}(x;h) &= \text{ABias}^2(x;h) + \text{AVar}(x;h) \\ &= \frac{h^{2k}}{(k!)^2} \mu_{K,k}^2 \{f_X^{(k)}(x)\}^2 + \frac{1}{nh^2} \int \left\{K_U\left(\frac{x-w}{h}\right)\right\}^2 f_W(w)\, dw, \qquad (2.3) \end{aligned}$$

where ABias and AVar denote, respectively, the asymptotic bias and variance of $\hat{f}_X$, and ABias^2 and AVar are equal to, respectively, the first and second terms of

(2.3). See for example Stefanski and Carroll (1990). The asymptotically optimal local bandwidth is defined by

$$h_{\text{AMSE}}(x) = \text{argmin}_h \, \text{AMSE}(x; h). \tag{2.4}$$

In the sequel, to simplify notation, we denote $h(x)$ by h (i.e. we do not indicate the dependence on x) unless the context is unclear.

Remark 1. *In the error-free case, it has been proved that, for many densities, there are some points x at which there is a nonzero bandwidth $h(x)$ for which the bias at x vanishes. See for example Hazelton (1998) and Sain (2003). In such cases, it can happen that, at those particular points x, $h_{\text{AMSE}}(x)/h_{\text{MSE}}(x) \not\to 1$. Since the bias of the deconvolution kernel density estimator is exactly equal to the bias of the error-free kernel density estimator, then such bias annihilating bandwidths exist in the deconvolution setting too. However, as in the error-free case, as far as practice is concerned, the existence of such bandwidths is not very important; see Hazelton (1998) and Sain (2003).*

3 Local bandwidth selection

In this section we develop three local bandwidth selectors based on the AMSE (sections 3.1 and 3.2) or on the MSE (section 3.3).

3.1 Empirical Bias Bandwidth selection

To estimate $h_{\text{AMSE}}(x)$ at (2.4), we need to estimate ABias and AVar and deduce an estimator $\widehat{\text{AMSE}}$ of the AMSE at (2.3). Then we can estimate $h_{\text{AMSE}}(x)$ by

$$\widehat{h}_{\text{AMSE}}(x) = \text{argmin}_h \widehat{\text{AMSE}}(x; h).$$

Estimating the variance part is not difficult. As in Fan (1991a), we can take

$$\widehat{\text{AVar}}(x; h) = (n^2 h^2)^{-1} \sum_{i=1}^{n} \Big\{ K_U\Big(\frac{x - W_i}{h}\Big) \Big\}^2, \tag{3.5}$$

which is unbiased. We will see later that this simple variance estimator works well in practice. The bias part is considerably more complex to estimate and standard approaches which work well in the global case (e.g. the plug-in method) do not perform well in the local context (see the discussion in section 3.2). Therefore, alternative less standard procedures have to be developed.

To estimate ABias, we first consider adapting to our density deconvolution context, the Empirical Bias Bandwidth Selector (EBBS) approach developed by Ruppert (1997) in the error-free regression context. Let $m(x)$ be a regression curve, $\hat{m}(x;h)$ be a local polynomial estimator of m, and let $\mu_m(x)$ be the mean of the asymptotic distribution of $\hat{m}(x;h)$. Then it is possible to show that $\mu_m(x) = m(x) + h^\ell B(x) + o(h^\ell)$ for some $\ell > 0$, where B is a function which depends on m but not on h. Clearly, this relation resembles a regression model, where the dependent variable is $\hat{m}(x;h)$ and the independent variable is h^ℓ. For a range of bandwidths h_0, Ruppert (1997) suggests estimating the bias $B(x)$ of $\hat{m}(x;h_0)$ by the least-squares estimator of the slope of a linear regression model based on the data $(h_1^\ell, \hat{m}(x;h_1)), \ldots, (h_T^\ell, \hat{m}(x;h_T))$, where $h_1, \ldots, h_T$ is a grid of bandwidths around h_0 (a more sophisticated polynomial approach is considered in Ruppert (1997), but for simplicity we only discuss this linear version).

To adapt the EBBS procedure to the deconvolution density context, note that under Condition A, we have $E\{\hat{f}_X(x;h)\} = f_X(x) + h^k B(x) + o(h^k)$, where $B(x) = f_X^{(k)}(x)\mu_{K,k}/(k!)$. That is,

$$\hat{f}_X(x;h) = f_X(x) + h^k B(x) + o(h^k) + \epsilon, \tag{3.6}$$

where $\epsilon = \hat{f}_X(x;h) - E\{\hat{f}_X(x;h)\}$ and $E(\epsilon) = 0$. Therefore, for small h, the model at (3.6) is approximately a linear regression model where the dependent variable is $\hat{f}_X(x;h)$ and the fixed covariate is h^k. As in the regression case of Ruppert (1997), this suggests estimating $B(x)$ by the least-squares estimator of the slope of this linear model. Replacing sums in the least-squares estimator by integrals, which is simpler to write and to analyse theoretically, we suggest estimating the bias of the deconvolution kernel density estimator $\hat{f}_X(x; h_0)$ by $h_0^k \hat{B}(x)$, where

$$\hat{B}(x) = \Big(I_0 \int_{\mathcal{I}} h^k \hat{f}_X(x; h)\, dh - I_k \int_{\mathcal{I}} \hat{f}_X(x; h)\, dh\Big) / \big(I_{2k} I_0 - I_k^2\big)\,, \tag{3.7}$$

where $I_j = \int_{\mathcal{I}} h^j\, dh$ and $\mathcal{I}$ is a compact interval of the form $[ah_*, bh_*]$, with a and b some positive finite constants and $h_* > 0$. To the best of our knowledge, this approach is new in the error-free context too. The following lemmas, which describe the asymptotic bias and variance of the estimator $\hat{B}(x)$, will help us choose a, b and h_*.

Lemma 1. *Assume condition A and suppose that f_X has $k+3$ continuous and bounded derivatives and $|K|$ has $k+3$ finite absolute moments. Then, if we take $\mathcal{I} = [ah_*, bh_*]$, where $a < 1$ and $b > 1$ are some finite constants and $h_* \to 0$ as $n \to \infty$,*

$$E\{\hat{B}(x)\} = B(x) + h_*^2\, C_1\, f_X^{(k+2)}(x) \frac{\mu_{K,k+2}}{(k+2)!} + o(h_*^2),$$

where C_1 is a finite constant defined in the proof of the lemma.

Lemma 2. *Assume condition A and suppose that $||\phi_K'||_\infty < \infty$, $||\phi_U'||_\infty < \infty$ and $\int \big(|t|^\beta + |t|^{\beta-1}\big)\big\{|\phi_K'(t)| + |\phi_K(t)|\big\}\, dt < \infty$. Assume too that, for some constants $c > 0$ and $\beta > 1$,*

$$\lim_{t \to +\infty} t^\beta \phi_U(t) = c \text{ and } \lim_{t \to +\infty} t^{\beta+1} \phi_U'(t) = -c\beta. \tag{3.8}$$

Then, if we take $\mathcal{I} = [ah_*, bh_*]$, where $a < 1$ and $b > 1$ are some finite constants, $h_* \to 0$ and $nh_*^{2\beta+2k+1} \to \infty$ as $n \to \infty$, we have

$$\text{var}\{\hat{B}(x)\} = \frac{C_2}{nh_*^{2\beta+2k+1}} \frac{f_W(x)}{c^2} + o(n^{-1} h_*^{-2\beta-2k-1}),$$

where C_2 is a finite constant defined in the proof of the lemma.

From these two lemmas we see that in order to get the fastest convergence rates for $\hat{B}(x)$, we should take $h_* \asymp n^{-1/(2\beta+2k+5)}$. When $k = 2$, this choice corresponds to the optimal order of the bandwidth for estimating $f_X^{(k)}$ by a deconvolution kernel estimator (of course, this is not surprising since estimating B is equivalent to estimating $f_X^{(k)}$). On the other hand, let h_0 be a bandwidth which is candidate for being the local bandwidth $h(x)$ to estimate $f_X(x)$. In theory, a good candidate should satisfy

$h_0 \asymp n^{-1/(2\beta+2k+1)}$, since this is the order of the optimal bandwidth for estimating $f_X(x)$. Hence, asymptotically, the bandwidths in $\mathcal{I}$ should be an order of magnitude larger than h_0. However, Ruppert (1997) noted that, in practice, if the target curve changes rapidly, which is where a local bandwidth can be useful, then better results can usually be obtained if $\mathcal{I}$ contains h_0. Our extensive simulation results indicated that, in the deconvolution density context too, for finite samples it was preferable to take a small enough so that $ah_* < h_0$. This does not necessarily contradict the asymptotic theory since in finite sample, where n is fixed, h_* and h_0 are both constants. Rather, the asymptotic results can be used as a guide to help us choose the interval $\mathcal{I}$ appropriately. In particular, since the practice leads us to take $ah_* < h_0$, and the theory indicates that the interval $\mathcal{I}$ should also contain larger bandwidths, this suggests to take the interval $\mathcal{I}$ wider for small h_0 than for large h_0. These considerations motivate the practical implementation of the method described in section 4.

Remark 2. *Since our goal with the theory is only to provide some insight into the method, Lemma 2 only studies the variance in the case where the errors U_i are ordinary smooth, that is, where the characteristic function $\phi_U(t)$ behaves like $|t|^{-\beta}$ in the tails. It would be possible to also study the behaviour of the variance in the supersmooth case, where, for t large, $\phi_U(t)$ behaves like a $\exp(-|t|^\beta)$ for some $\beta > 0$. However this would not be very useful for practice since, as usual in deconvolution problems, the message would be that the bias is negligible compared to the variance, and that the latter tends to zero at a logarithmic rate. We consider supersmooth errors in our numerical investigation in section 5.*

Remark 3. *In section 4.3 we will discuss a variant of this approach and of the approach of section 3.2, where the so-called diagonal terms are left out of the squared bias.*

3.2 Integrated plug-in approach

Since one of the most effective methods of global bandwidth selection is the plug-in procedure, it seems natural to consider a plug-in approach in the local context too.

Assume that

$$\sup_t |t^k \phi_K(t)/\phi_U(t/h)| < \infty \quad \text{and} \quad \int |t^k \phi_K(t)/\phi_U(t/h)| \, dt < \infty. \tag{3.9}$$

Then a plug-in estimator of the asymptotic squared bias of $\hat{f}_X$ could be defined by replacing $f_X^{(k)}(x)$ in (2.3) by $\hat{f}_X^{(k)}(x; g)$, the kth derivative of $\hat{f}_X(x; g)$ with the bandwidth $g = g(x)$ chosen to minimise the MSE of $\hat{f}_X^{(k)}(x; g)$. However, unlike in the global setting, this MSE depends in a complex way on several unknown pointwise f_X-related quantities, which makes this approach unattractive. In particular, choosing for each x a bandwidth $g(x)$ can turn out to work rather poorly in practice. See Hall (1993) for a more thorough description of the problems associated with a local plug-in procedure in the simpler error-free case.

Inspired by the EBBS procedure, in order to reduce the effect of a particular local bandwidth $g(x)$, we suggest taking an average over a range of bandwidth values. More precisely, we suggest estimating $f_X^{(k)}(x)$ by

$$\widetilde{f}_X^{(k)}(x) = I_0^{-1} \int_{\mathcal{I}} \hat{f}_X^{(k)}(x; h) \, dh ,$$

where I_j and $\mathcal{I}$ are defined in section 3.1. We call this method "integrated plug-in" (IPI).

The following lemmas describe the asymptotic bias and variance of the estimator $\widetilde{f}_X^{(k)}(x)$.

Lemma 3. *Under the conditions of Lemma 1, if* (3.9) *holds, then for* $\mathcal{I} = [ah_*, bh_*]$, *where* $a < 1$ *and* $b > 1$ *are some finite constants and* $h_* \to 0$ *as* $n \to \infty$, *we have* $E\{\widetilde{f}_X^{(k)}(x)\} = f_X^{(k)}(x) + O(h_*^2)$. *Moreover, if* f_X *has* $2k + 1$ *bounded derivatives, then*

$$E\{\widetilde{f}_X^{(k)}(x)\} = f_X^{(k)}(x) + C_3 h_*^k f_X^{(2k)}(x) \frac{\mu_{K,k}}{k!} + o(h_*^k),$$

where C_3 *is a finite constant defined in the proof of the lemma.*

Lemma 4. *Assume condition A and suppose that* $||\phi_K'||_\infty < \infty$, $||\phi_U'||_\infty < \infty$ *and* $\int \left(|t|^\beta + |t|^{\beta-1}\right)\left\{|t^k \phi_K'(t)| + |t^{k-1}\phi_K(t)| + |t^k \phi_K(t)|\right\} dt < \infty$. *Then if* (3.8) *and* (3.9) *hold, we have*

$$\text{var}\{\widetilde{f}_X^{(k)}(x)\} = C_4 I_0^{-2} \frac{f_W(x)}{n h_*^{2k+2\beta+1}},$$

where C_4 *is a finite constant defined in the proof of the lemma.*

Comparing these two lemmas with Lemmas 1 and 2, we see that the conditions on f_X and the rates coincide if we use a kernel of order $k = 2$. For higher order kernels, the bias of the IPI method is of smaller order if the density has enough derivatives, but in this case the same rate could be obtained by EBBS by using a higher order kernel. In practice anyway, it is often recommended to work with second order kernels to avoid too high variability in the estimator and to avoid assuming too many unverifiable conditions on the unknown density. As in the EBBS case, the lemmas suggest that the bandwidths in $\mathcal{I}$ should be an order of magnitude larger than candidate local bandwidths h_0, but in finite samples better results are obtained if $h_0 \in \mathcal{I}$. As in the EBBS case, the theory also suggests that wider intervals be employed for smaller values of h_0. The detailed implementation of our procedure is described in section 4.

3.3 Local SIMEX bandwidth selection

Instead of constructing estimators of the asymptotic optimal bandwidth, we could also try to estimate the bandwidth $h_{\mathrm{MSE}}(x)$. However, this requires to estimate the MSE, which is not easy to do. One possibility for doing this would be to use the SIMEX ideas developed recently by Delaigle and Hall (2008) in the errors-in-variable regression context. SIMEX is a general technique originally suggested by Cook and Stefanski (1994) and Stefanski and Cook (1995) for parametric curve estimation in errors-in-variables contexts. The idea is to learn the mechanism of contamination by adding extra errors to the data, examining the effect on the quantities of interest (in our case the local bandwidth), and extrapolating back to the original data.

These ideas can be applied as follows. Let $D(x;h) = \big\{f_X(x) - \hat{f}_X(x;h)\big\}^2$ denote the squared error at x when using the bandwidth h. If we knew f_X, we would choose the local bandwidth $h(x)$ at x to minimize $D(x;h)$. Since f_X is unknown, we consider versions of $D(x;h)$ that we can calculate, using data contaminated by more errors. As in any SIMEX procedure our first step is to generate data which contain more errors than the original data. We take them as follows:

1. Generate $U_1^*, \dots, U_n^*$ and $U_1^{**}, \dots, U_n^{**}$ independent of the data $W_1, \dots W_n$ and

having the distribution of U;

2. Put $W_i^* = W_i + U_i^*$ and $W_i^{**} = W_i^* + U_i^{**}$, for $1 \le i \le n$;

Next, let f_W (resp. f_{W^*}) denote the density of W (resp. W_*), and let $\hat{f}_W(x;h)$ (resp. $\hat{f}_{W^*}(x;h)$) denote the deconvolution kernel density estimator of f_W (resp. f_{W^*}) constructed from the contaminated data $W_1^*, \ldots, W_n^*$ (resp. $W_1^{**}, \ldots, W_n^{**}$). Define

$$D^*(x;h) = \big\{f_W(x) - \hat{f}_W(x;h)\big\}^2 \text{ and } D^{**}(x;h) = \big\{f_{W^*}(x) - \hat{f}_{W^*}(x;h)\big\}^2.$$

Let $h^*(x)$ and $h^{**}(x)$ be the bandwidths which minimise $D^*(x;h)$ and $D^{**}(x;h)$ with respect to h. Since W^{**} measures W^* in the same way that W^* measures W, and W measures X, we can expect the relationship between $h(x)$ and $h^*(x)$ to be similar to that between $h^*(x)$ and $h^{**}(x)$. As in Delaigle and Hall (2008), this suggests approximating $h(x)$ by $\big\{h^*(x)\big\}^2/h^{**}(x)$. Note that since $h^*(x)$ and $h^{**}(x)$ converge to zero at the same rate as $h(x)$, the ratio $\big\{h^*(x)\big\}^2/h^{**}(x)$ converges to zero at the same rate as $h(x)$.

Although we can not calculate $D^*(x;h)$ and $D^{**}(x;h)$, and thus $h^*(x)$ and $h^{**}(x)$, exactly, we can construct very good estimators of these quantities. Indeed, since we have access to the data $W_1, \ldots, W_n$ and $W_1^*, \ldots, W_n^*$, we can construct standard "error-free" kernel estimators of f_W and f_{W^*}, defined by

$$\hat{f}_{W,\mathrm{EF}}(x) = \frac{1}{ng_1}\sum_{j=1}^n K\Big(\frac{x - W_j}{g_1}\Big) \text{ and } \hat{f}_{W^*,\mathrm{EF}}(x) = \frac{1}{ng_2}\sum_{j=1}^n K\Big(\frac{x - W_j^*}{g_2}\Big),$$

where K is a kernel and g_1 and g_2 are bandwidths (calculated using, e.g., the plug-in method of Sheather and Jones, 1991). Since the estimation error of these estimators is negligible compared to that of deconvolution estimators, we can estimate $D^*(x;h)$ and $D^{**}(x;h)$ by $\hat{D}^*(x;h) = \big\{\hat{f}_{W,\mathrm{EF}}(x) - \hat{f}_W(x;h)\big\}^2$ and $\hat{D}^{**}(x;h) = \big\{\hat{f}_{W^*,\mathrm{EF}}(x) - \hat{f}_{W^*}(x;h)\big\}^2$. Let $\hat{h}^*(x) = \mathrm{argmin}_h \hat{D}^*(x;h)$ and $\hat{h}^{**}(x) = \mathrm{argmin}_h \hat{D}^{**}(x;h)$. Our local SIMEX estimator of the local bandwidth $h(x)$ is defined by $\hat{h}(x) = \{\hat{h}^*(x)\}^2/\hat{h}^{**}(x)$. To avoid too strong dependence on the SIMEX samples $W_1^*, \ldots, W_n^*$ and $W_1^{**}, \ldots, W_n^{**}$, we proceed as in Delaigle and Hall (2008) and generate B such samples $W_{b_1}^*, \ldots, W_{b,n}^*$ and $W_{b,1}^{**}, \ldots, W_{b,n}^{**}$, for $b = 1, \ldots, B$. Then we take the average of the B corresponding estimators of D^* and D^*. That is, in the procedure described above, we replace

$\hat{D}^*(x;h)$ and $\hat{D}^{**}(x;h)$ by

$$\bar{D}^*(x;h) = \frac{1}{B}\sum_{b=1}^{B} \hat{D}_b^*(x;h) \text{ and } \bar{D}^{**}(x;h) = \frac{1}{B}\sum_{b=1}^{B} \hat{D}_b^{**}(x;h),$$

where $\hat{D}_b^*(x;h)$ and $\hat{D}_b^{**}(x;h)$ denote the version of $\hat{D}^*(x;h)$ and $\hat{D}^{**}(x;h)$ calculated from the bth sample.

4 Details of implementation

In this section we give details of our implementation of the various local bandwidth selectors. First, note that in practice it is reasonable to expect that for most points x, the optimal local bandwidth at x will lie in the interval $[h_{\mathrm{PI}}/4, 2h_{\mathrm{PI}}]$, where h_{PI} denotes the global plug-in bandwidth selector of Delaigle and Gijbels (2004a). Therefore, when searching for a local bandwidth, we restrict our attention to that interval. More precisely, we search for our local bandwidth on a discrete grid $H = \{h_1, \ldots, h_M\}$, where $h_{j+1} = C^j h_{\mathrm{PI}}/4$, with $C = 1.1$ and where M is such that $h_{M-1} < 2h_{\mathrm{PI}} \leq h_M$. Such geometric grids of points are frequently used in practice (see e.g. Fan and Gijbels, 1996, page 122). Similarly, Ruppert (1997) takes an equally spaced grid of bandwidths on the log scale.

This is the default construction of H that we suggest using. In general, replacing $h_{\mathrm{PI}}/4$ by a smaller value such as $h_{\mathrm{PI}}/10$ significantly worsens the results and produces too wiggly estimators. On the other hand, we can replace $2h_{\mathrm{PI}}$ by a value as large as $10h_{\mathrm{PI}}$ without affecting the results significantly, but this increases the computational time. Of course, we can't exclude the possibility that, in some cases, for some points x, our method could work better with wider intervals. Nevertheless, since the interval $[h_{\mathrm{PI}}/4, 2h_{\mathrm{PI}}]$ contains the global PI bandwidth, we can expect to get reasonable results in the majority of cases (remember that h_{PI} estimates the global optimal bandwidth).

4.1 Main methods

For the local SIMEX bandwidth selector, the only quantity we need to choose is the number B of SIMEX samples. To keep the computational time reasonable, we took

$B = 15$ as in Delaigle and Hall (2008). In general, taking B larger does not affect the results much, but it can significantly increase the computational time.

We implemented the EBBS and the IPI methods as follows: for all $h_0 \in H$:

1. Estimate the variance of $\hat{f}_X(x; h_0)$ by $\widehat{\mathrm{AVar}}(x; h_0)$ defined at (3.5);
2. Construct the interval $\mathcal{I}$ as follows:

$$\mathcal{I} = \begin{cases} \left[h_0\, C^{-\alpha_1}, h_0\, C^{\alpha_2}\right] & \text{if } h_0 \in [h_{\mathrm{PI}}/4, 2h_{\mathrm{PI}}/3) \\ \left[h_0\, C^{-\beta_1}, h_0\, C^{\beta_2}\right] & \text{if } h_0 \in [2h_{\mathrm{PI}}/3, 2h_{\mathrm{PI}}]; \end{cases} \tag{4.10}$$

3. (a) For the EBBS method, calculate $\hat{B}(x)$ with $\mathcal{I}$ defined in step 2;
   (b) For the IPI method, calculate $\widetilde{f}_X^{(k)}(x)$ with $\mathcal{I}$ defined in step 2;
4. (a) For the EBBS method, estimate $\mathrm{AMSE}(x; h_0)$ by

$$\widehat{\mathrm{AMSE}}(x; h_0) = \widehat{\mathrm{AVar}}(x; h_0) + h_0^{2k} \hat{B}^2(x);$$

   (b) For the IPI method, estimate $\mathrm{AMSE}(x; h_0)$ by

$$\widehat{\mathrm{AMSE}}(x; h_0) = \widehat{\mathrm{AVar}}(x; h_0) + h_0^{2k} \frac{\mu_{K,k}^2}{(k!)^2} \{\widetilde{f}_X^{(k)}(x)\}^2;$$

Then, estimate $h_{\mathrm{AMSE}}(x)$ by $\hat{h}_{\mathrm{AMSE}}(x) =$ first local minimiser of $\widehat{\mathrm{AMSE}}(x; h_0)$ where the minimum is searched over $h_0 \in H$. The motivation for taking the first local minimum of $\widehat{\mathrm{AMSE}}(x; h_0)$ instead of the global minimum is the same as in Ruppert (1997) and comes from the fact that as h_0 increases, $\hat{B}(x)$ becomes more and more biased, thereby producing less meaningful results. Guided by our theoretical results, an extensive simulation study led us to set the parameters used in the definition of the interval $\mathcal{I}$, at (4.10), equal to $\alpha_1 = 6$, $\alpha_2 = 3$, $\beta_1 = 3$ and $\beta_2 = 1$. We found this choice to work well for the various densities we considered, even though these had very different scales and features. This is the default choice we recommend to use in practice. It is clear from the definition of the α_j's and the β_j's that these parameters are closely connected to the value of C used to construct the geometric grid H. The values we suggest here are appropriate for $C = 1.1$, our default choice. If the user changes the value of C, then the values of α_j and β_j have to be modified accordingly.

In general, for all three methods, the final local bandwidth $\hat{h}(x)$ tends to be somewhat rough as a function of x. Hence, as in Ruppert (1997), we smooth the local bandwidth, using a triangular function. At a fixed point x, the smoothed version $\bar{h}(x)$ is constructed by taking a weighted average of the bandwidths $\hat{h}(y)$ for all y's in a neighborhood of x. We use the smoothing scheme of Ruppert (1997), page 1054, where we take `span` equal to 0.05 times the length of the grid G of x-values defined below (at page 19). As in Ruppert (1997), we can not guarantee that this is always the best choice, but we found that it worked well in all cases that we tried. See Ruppert (1997) for more details and recommendations.

4.2 Discrete version

In practice, when a grid of equispaced bandwidths is used, our integral version of EBBS is essentially equivalent to the discrete least-squares version employed by Ruppert (1997). However, the two methods differ a little when using a geometric grid. For comparison, in our simulation study, we also implemented discrete versions of the EBBS and the IPI procedures in a way similar to Ruppert (1997). More precisely, in his implementation, to estimate the bias corresponding to a bandwidth $h_0 = h_j \in H$, Ruppert (1997) uses a geometric grid $\{h_{j+k}\}_{k=-J_1}^{J_2}$ of neighboring points of h_j, where J_1 and J_2 are two fixed numbers (of course, this is only done for bandwidths h_j that have J_1 neighbours to the left and J_2 neighbours to the right). Instead of our integrals I_i, Ruppert (1997) uses sums $L_i = \sum_{\ell=-J_1}^{J_2} h_{j+\ell}^i$. We implemented a similar discrete version of our procedures as follows. For $j = 1 + J_1, \ldots, M - J_2$:

1. Same as step 1 of section 4.1, replacing there h_0 by h_j;
2. (a) For EBBS, estimate $B(x)$ by $\hat{B}(x; h_j) = \big\{L_0 \sum_{\ell=-J_1}^{J_2} h_{j+\ell}^k \hat{f}_X(x; h_{j+\ell}) - L_k \sum_{\ell=-J_1}^{J_2} \hat{f}_X(x; h_{j+\ell})\big\}/(L_{2k}L_0 - L_k^2)\cdot$
   (b) For IPI, estimate by $f_X^{(k)}(x; h_j)$ by $\widetilde{f}_X^{(k)}(x; h_j) = L_0^{-1} \sum_{\ell=-J_1}^{J_2} \hat{f}_X^{(k)}(x; h_{j+\ell})$;
3. Same as step 4 of section 4.1, replacing there h_0 by h_j;

The local bandwidth $\hat{h}(x)$ for estimating $f_X(x)$ is defined by the first local minimum of $\widehat{\text{AMSE}}(h)$ among all $h \in \{h_{1+J_1}, \ldots, h_{M-J_2}\}$. Then, we smooth the local bandwidth as described at the end of section 4.1. We took $J_1 = 6$ and $J_2 = 3$ for

$h_0 \in [h_{\text{PI}}/4, 2h_{\text{PI}}/3)$, and $J_1 = 3$ and $J_2 = 1$ when $h_0 \in [2h_{\text{PI}}/3, 2h_{\text{PI}}]$, which corresponds to the method employed in section 4.1. Like there, by definition, the values of J_1 and J_2 are strongly connected to the value of C and the values we suggest here are appropriate for our default choice $C = 1.1$, and if another value of C is used then J_1 and J_2 have to be modified accordingly.

4.3 Leaving the diagonal terms out

When taking the square of nonparametric estimators, it often happens that some terms essentially contribute to bias. These terms are often referred to as diagonal terms, and diagonals-out versions of estimators are sometimes considered as more attractive. See for example Hall and Marron (1987). As we shall see below, a similar phenomenon occurs with the IPI and the EBBS approaches when estimating $\{f_X^{(k)}(x)\}^2$, which is required to estimate the squared bias.

For the method of section 3.2, $\{f_X^{(k)}(x)\}^2$ is estimated by $\{\widetilde{f}_X^{(k)}(x)\}^2 = \sum_{j,\ell=1}^n T_{j,\ell} = D + \sum_{j\neq\ell} T_{j,\ell}$, where

$$T_{j,\ell} = I_0^{-2} n^{-2} \int_{\mathcal{I}} \int_{\mathcal{I}} (hg)^{-k-1} K_U^{(k)}\Big(\frac{x - W_j}{h}\Big) K_U^{(k)}\Big(\frac{x - W_\ell}{g}\Big)\, dh\, dg,$$

$D = \sum_{j=1}^n T_{j,j}$, and the $T_{j,j}$'s are the diagonal terms. Using techniques similar to those employed in Lemmas 3 and 4, it can be proved that $\sum_{j\neq\ell} E(T_{j,\ell}) = \{f_X^{(k)}(x)\}^2 + R_n$, where the bias term R_n is of order $o(1)$ and depends on derivatives of f_X. On the other hand, $E[D(x)] \sim \text{const.} n^{-1} h_*^{-2k-2\beta-1} f_W(x)$, where $0 < \text{const.} < \infty$ is a constant, so that this term has no opportunity to compensate for the dominant part of R_n. As in Hall and Marron (1987), this motivates considering a diagonals-out variant of the estimator, defined by

$$I_0^{-2} n^{-1} (n-1)^{-1} \sum_{j=1}^n \sum_{\ell\neq j} \int_{\mathcal{I}} \int_{\mathcal{I}} (hg)^{-k-1}\ K_U^{(k)}\Big(\frac{x - W_j}{h}\Big) K_U^{(k)}\Big(\frac{x - W_\ell}{g}\Big)\, dh\, dg.$$

Similarly, a variant of the EBBS method can be obtained by leaving the diagonal terms out of $\hat{B}^2(x)$. Here we replace $\hat{B}^2(x)$ by $\tilde{B}^2(x) = N/D^2$ where $D = I_{2k} I_0 - I_k^2$,

$$N = I_0^2 \int_{\mathcal{I}} \int_{\mathcal{I}} h^k g^k \tilde{f}_{Xhg}(x)\, dh\, dg + I_k^2 \int_{\mathcal{I}} \int_{\mathcal{I}} \tilde{f}_{Xhg}(x)\, dh\, dg - 2 I_0 I_k \int_{\mathcal{I}} \int_{\mathcal{I}} h^k \tilde{f}_{Xhg}(x)\, dh\, dg,$$

and

$$\tilde{f}_{Xhg}(x) = \frac{1}{n(n-1)hg} \sum_{j=1}^{n} \sum_{k \neq j} K_U\Big(\frac{x - W_j}{h}\Big) K_U\Big(\frac{x - W_k}{g}\Big).$$

These diagonals-out estimators have the same convergence rates as the estimators of the previous sections, and we study their numerical merit in section 5.

5 Simulations

5.1 Methods compared

We implemented the following local bandwidth selectors: integral versions of EBBS (EBBSI) and of IPI (IPII) as in section 4.1, discrete versions of EBBS (EBBSD) and of IPI (IPID) as in section 4.2, diagonals-out versions of the discrete selectors (EBBSD1 and IPID1) and integral selectors (EBBSI1 and IPII1), and local SIMEX procedure (LSIMEX). In most cases, the diagonals-out versions worked better than the diagonals-in, and thus we do not show the results of the latter. We compared the performance of our local bandwidth selectors with the global plug-in bandwidth (GPI) of Delaigle and Gijbels (2002, 2004a), which, as in the error-free case, is a very good global bandwidth. For illustration, we also implemented a global version of our SIMEX bandwidth selector, which we denote by GSIMEX. The procedure is exactly the same as in section 3.3 except that we replace the local distances $\hat{D}^*(x;h)$ and $\hat{D}^{**}(x;h)$ by estimators of the Asymptotic Mean Integrated Squared Errors of the deconvolution estimators $\hat{f}_W$ and $\hat{f}_{W^*}$. That is, we replace $\hat{D}^*(x;h)$ and $\hat{D}^{**}(x;h)$ by

$$\widehat{\text{AMISE}}(\hat{f}_W) = (2\pi nh)^{-1} \int |\phi_K(t)|^2 |\phi_U(t/h)|^{-2}\, dt + \frac{h^{2k}}{(k!)^2}\, \mu_{K,k}^2 \hat{R}(f_W^{(2)}) \qquad (5.11)$$

and

$$\widehat{\text{AMISE}}(\hat{f}_{W^*}) = (2\pi nh)^{-1} \int |\phi_K(t)|^2 |\phi_U(t/h)|^{-2}\, dt + \frac{h^{2k}}{(k!)^2}\, \mu_{K,k}^2 \hat{R}(f_{W^*}^{(2)}), \qquad (5.12)$$

where, for any square integrable function p, we used the notation $R(p) = \int p^2(v)\, dv$, and where $\hat{R}(f_W^{(2)})$ and $\hat{R}(f_{W^*}^{(2)})$ denote the estimators of Jones and Sheather (1991) constructed from the data W_i and W_i^*, respectively. For a derivation of the AMISE of the deconvolution kernel estimator, see for example Stefanski and Carroll (1990). Of course, in the GSIMEX approach the same bandwidth was used at all points x.

5.2 Simulation settings

We considered the following eight densities:

1. Comb density: $\sum_{l=0}^{5}(2^{5-l}/63)N\left[\{65-96(0.5)^l\}/21,(\frac{32}{63})^2/2^{2l}\right]$;
2. Gammas and a Normal Mixture: $0.3G(10)+0.4N(40,25)+0.3G(70)$;
3. Mixture of three Normals: $0.5N(0,0.25)+0.3N(1.5,0.01)+0.2N(-1,0.025)$;
4. Beta and Normal mixture density: $0.35B(2,2)+0.3N(4,0.25^2)+0.35N(-3,1)$;
5. Gamma Mixture: $0.4G(5)+0.6G(13)$;
6. Kurtotic density: $(2/3)N(0,1)+(1/3)N(0,0.1^2)$;
7. Bimodal Normal mixture: $0.5N(-3,1)+0.5N(2,1)$;
8. Standard Normal: $N(0,1)$.

Densities 1, 6, 7 and 8 are taken form Marron and Wand (1992) and density 5 is taken from Delaigle and Gijbels (2004a). Some of these densities involve various complex features, such as several modes of various sizes and shapes. For these densities, we can expect that our local bandwidths will be beneficial. Other densities are much simpler and for them we do not expect any gain by using a local bandwidth, but we want to check that using a local bandwidth is not going to deteriorate the results too much.

We used the second order ($k=2$) kernel K with characteristic function $\phi_K=(1-t^2)^3 1_{[-1,1]}$, which is often used in deconvolution methods. See for example Delaigle and Hall (2006). We considered Laplace errors (which are ordinary smooth) and normal errors (which are supersmooth). The error variance was controlled by the noise to signal ratio, denoted by NSR, and defined by NSR $=\operatorname{var}(U)/\operatorname{var}(X)$. We estimated each density from each of 100 randomly generated samples of size 100, 250 or 500 contaminated by either Laplace or Normal errors such that the NSR was either 10% or 25%. To calculate K_U, note that in the normal case, there is no analytic expression and we need to use numerical integration especially adapted to the calculation of Fourier transforms. In the Laplace case, there is an analytic expression for K_U, but it can not be used near zero. For details about these issues and how to address them in practice, see Delaigle and Gijbels (2007).

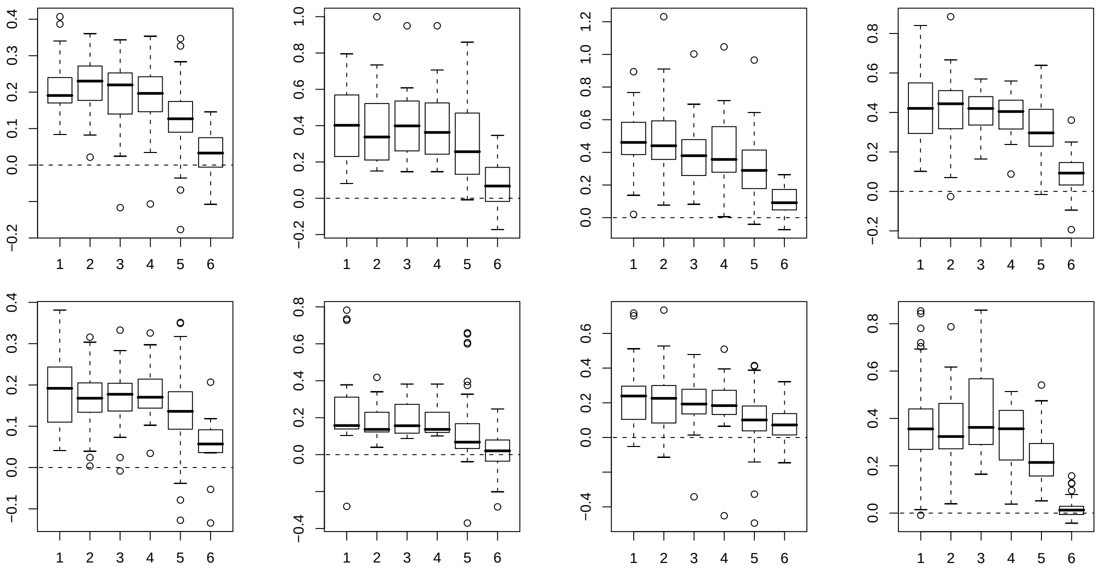

Figure 1: Boxplots of $\log(\mathrm{ASE}(h_{GPI})/\mathrm{ASE}(\hat{h}))$ indicating the efficiency gain over the GPI method calculated from the 100 replications for, from left to right, densities 1 to 4. The boxes show 1: EBBSD1, 2: EBBSI1, 3: IPID1, 4: IPII1, 5: LSIMEX, 6: GSIMEX. Row 1: $n = 100$ and Laplace with NSR 10%; row 2: $n = 250$ and Laplace errors with NSR 25%.

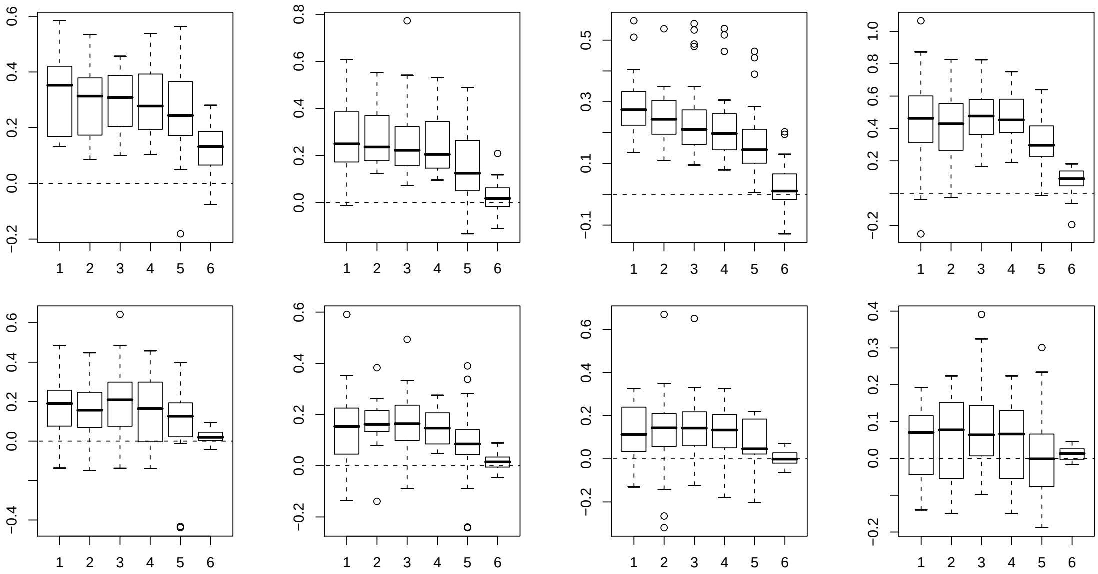

Figure 2: Boxplots of $\log\ \mathrm{ASE}(h_{GPI})/\mathrm{ASE}(\hat{h})$ over the 100 replications for, from left to right, densities 1 to 4. The boxes show 1: EBBSD1, 2: EBBSI1, 3: IPID1, 4: IPII1, 5: LSIMEX, 6: GSIMEX. Row 1: $n = 100$ and normal errors with NSR 10%; row 2: $n = 250$ and normal errors with NSR 25%.

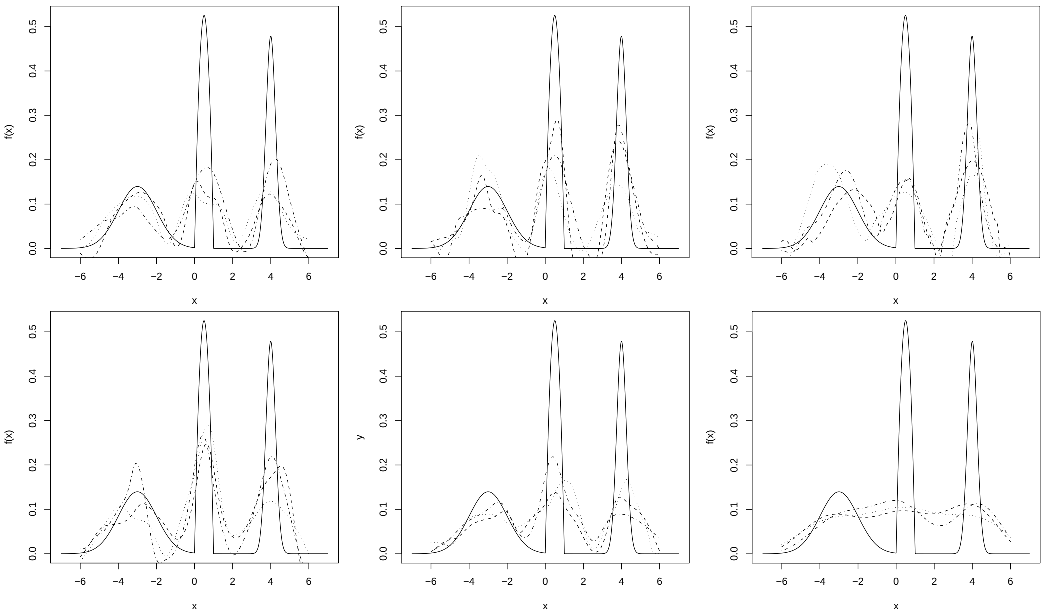

Figure 3: Quartile estimators for density 4 when $n = 100$ and measurement errors are Laplace with 10% NSR; Dotted line: 1st quartile estimator, Dashed line: Median estimator, Dotted-dashed line: 3rd quartile estimator. From left to right, row 1: EBBSD1, EBBSI1, IPID1;row 2: IPII1, LSIMEX, GPI.

We estimated each density on a grid G of $n_G = 80$ equally spaced points (different for each density). We assessed the quality of each density estimator by calculating the Averaged Squared Error, defined by:

$$\text{ASE}(\hat{f}_{X,d}) = \frac{1}{n_G} \sum_{i=1}^{n_G} \Big[f_X(x_i) - \hat{f}_X\big\{x_i; h(x_i)\big\} \Big]^2.$$

In each graph of figures 3 to 5 below, we show the target curve (in solid line), and three estimators. These three estimators, which we will refer to as "the quartile estimators", are those which correspond to the first, second and third quartiles of the 100 values of ASE (remember that we generate 100 samples each time). We show only a summary of the results, but the other results were similar. They are available in Achilleos and Delaigle (2011). All codes were written in R.

In figures 1 and 2, we show, for densities 1 to 4, boxplots of the log of the ratio $\text{ASE}(h_{GPI})/\text{ASE}(\hat{h})$ over the 100 replications, where h_{GPI} denotes the GPI bandwidth and $\hat{h}$ is one of our new procedures. A value larger than 0 means that the local bandwidth gave a smaller ASE than h_{GPI} (and thus worked better). Figure 1 is for Laplace errors with NSR = 10% and $n = 100$, and with NSR = 25% and $n = 250$.

Figure 2 shows the results for normal errors with either NSR = 10% and $n = 100$ or NSR = 25% with $n = 250$. The other combinations of n and NSR are shown in Achilleos and Delaigle (2011). Clearly, for these four densities all local bandwidth selectors significantly outperformed the two global approaches, but the LSIMEX did not compete with the other four local procedures. The GSIMEX method did not seem to be particularly useful compared to the GPI. The EBBSD1 and IPID1 bandwidths performed slightly better than the EBBSI1 and IPII1 selectors.

Next we show the quartile estimators for the EBBSD1, EBBSI1, IPID1, IPII1, LSIMEX and GPI methods, for densities 2, 3 and 4, various sample sizes, and for Laplace or Normal errors. Figure 3 shows that GPI failed to capture the refined features of density 4. The local bandwidth selectors all performed much better, but as already indicated by the boxplots in the previous figures, the LSIMEX method did not work nearly as well as the other four methods. The same conclusions can be drawn for figures 4 and 5. We can not show the results for all the cases that we tried, but of course, as usual, the results improved as the sample size increased or as the NSR decreased, and normal errors were harder to deal with than Laplace errors.

Boxplots of the log of the ratio $\text{ASE}(h_{GPI})/\text{ASE}(\hat{h})$ over 100 replications for densities 5 to 8 are shown in figure 6 for Laplace errors with NSR = 10% and $n = 100$, and with NSR = 25% and $n = 250$. Here, the local bandwidth selectors performed comparably with, but did not improve, the global methods for the simpler densities 5, 7 and 8, but they clearly improved results for density 6. Unsurprisingly, our other simulation results showed that, for simpler densities like densities 7 and 8, the local bandwidth selectors gave more variable estimators than the global approaches, and that the latter worked slightly better. Nevertheless, the small loss incurred by local bandwidths in simple cases such as these is negligible compared to the significant gain in more complex cases.

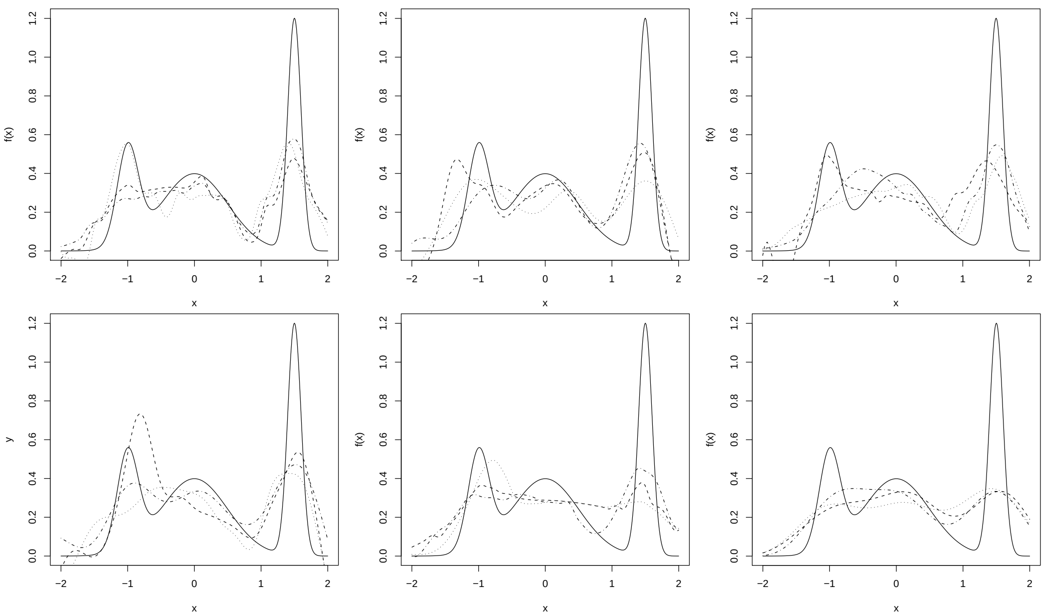

Figure 4: Quartile estimators for density 3 when $n = 250$ and measurement errors are Laplace with 25% NSR; Dotted line: 1st quartile estimator, Dashed line: Median estimator, Dotted-dashed line: 3rd quartile estimator. From left to right, row 1: EBBSD1, EBBSI1, IPID1;row 2: IPII1, LSIMEX, GPI.

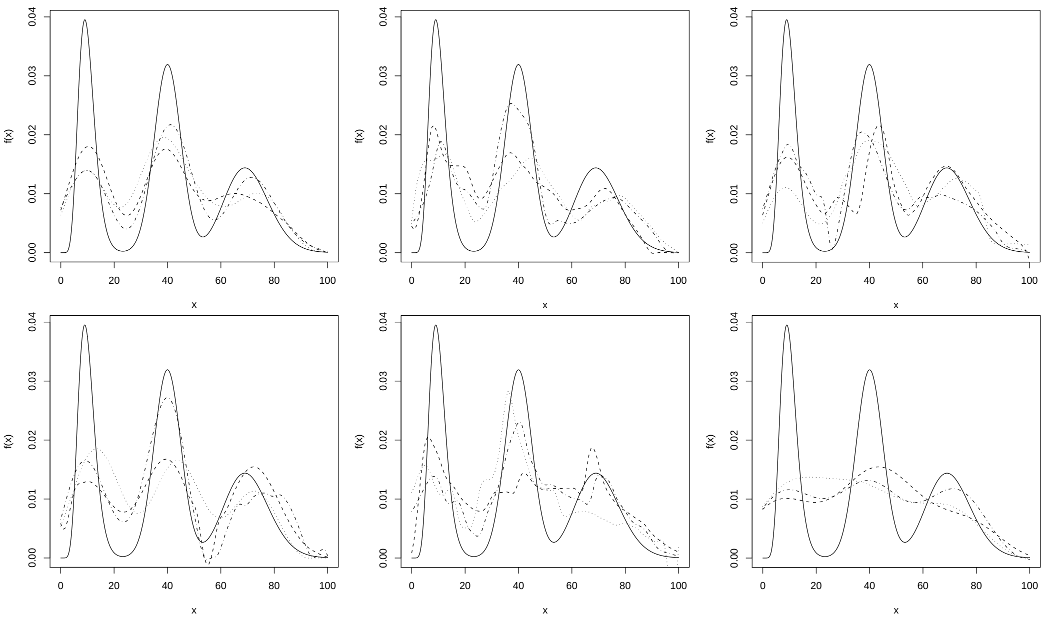

Figure 5: Quartile estimators for density 2 when $n = 100$ and measurement errors are Normal with 10% NSR; Dotted line: 1st quartile estimator, Dashed line: Median estimator, Dotted-dashed line: 3rd quartile estimator. From left to right, row 1: EBBSD1, EBBSI1, IPID1;row 2: IPII1, LSIMEX, GPI.

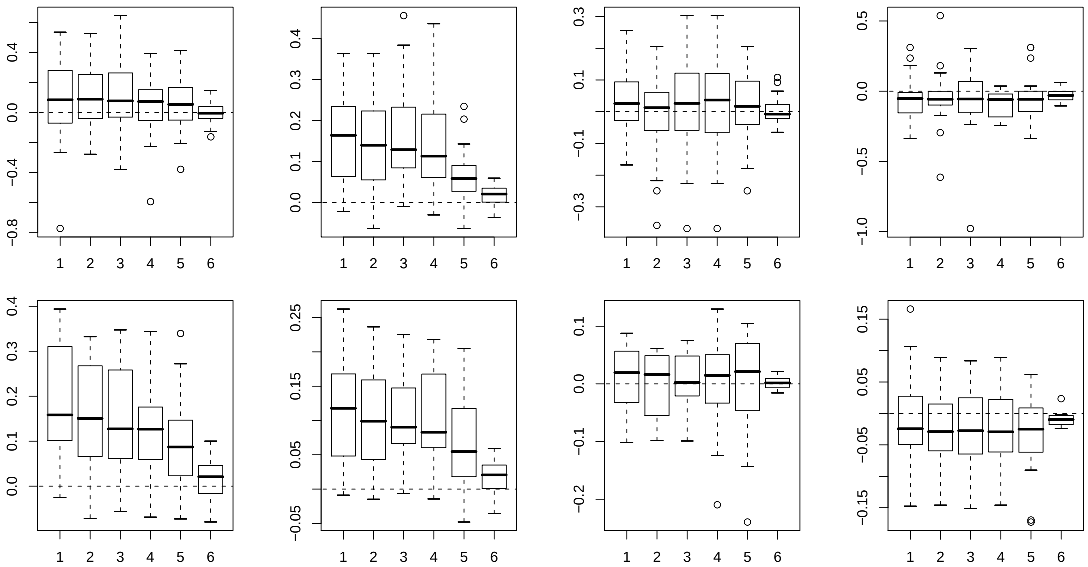

Figure 6: Boxplots of $\log(\text{ASE}(h_{GPI})/\text{ASE}(\hat{h}))$ indicating the efficiency gain over the GPI method calculated from the 100 replications for, from left to right, densities 5 to 8. The boxes show 1: EBBSD1, 2: EBBSI1, 3: IPID1, 4: IPII1, 5: LSIMEX, 6: GSIMEX. Row 1: $n = 100$ and Laplace with NSR 10%; row 2: $n = 250$ and Laplace errors with NSR 25%.

6 Conclusions

We have proposed several local bandwidth selectors for the errors-in-variables density estimation problem. We have illustrated the significant improvement they can bring over global bandwidths when the density to estimate has various local features, without significantly worsening performance in other cases. Of all the local bandwidth selectors we suggested, we found that the EBBSD1, EBBSI1, IPID1 and IPII1 methods performed the best, with, depending on the cases, one working better than the other.

Acknowledgement

Delaigle's research was supported by grants from the Australian Research Council and by a QEII Fellowship. We thank two referees for valuable comments that helped improve the manuscript.

References

Achilleos, A. and Delaigle, A. (2011). Supplementary file for the paper "Local bandwidth selectors for deconvolution kernel density estimation". Available at `http://www.ms.unimelb.edu.au/~aurored`.

Brockmann, M., Gasser, T. and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.*, **88**, 1302–1309.

Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, **11**, 309–340.

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, **83**, 1184–1186.

Carroll, R.J., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A modern prospective.* Second edition. Chapman and Hall CRC Press.

Cook, J.R. and Stefanski, L. (1994). Simulation-Extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.*, **89**, 1314–1328

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc., Ser. B*, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2004a). Practical bandwidth selection in deconvolution kernel density estimation. *Comp. Statist. Data Anal.*, **45**, 249–267.

Delaigle, A. and Gijbels, I. (2004b). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, **56**, 19–47.

Delaigle, A. and Gijbels, I. (2006). Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, **16**, 773–788.

Delaigle, A. and Gijbels, I. (2007). Frequent problems in calculating integrals and optimizing objective functions: a case study in density deconvolution. *Statist. Comput.*, **17**, 349–355.

Delaigle, A. and Hall, P. (2006). On the optimal kernel choice for deconvolution. *Statist. Probab. Lett.*, **76**, 1594–1602.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.*, **103**, 280–287.

Delaigle, A., Hall, P. and Meister, A. (2008). On Deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665-685.

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhya A*, **53**, 97–110.

Fan, J. (1991b). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.

Fan, J., Hall, P., Martin, M. and Patil, P. (1996). On local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc.*, **91**, 258–266.

Farmen, M., and Marron, J.S. (1999). An assessment of finite sample performance of adaptative methods in density estimation. *Comp. Statist. Data Anal.*, **30**, 143–168.

Hall, P. (1993). On plug-in rules for local smoothing of density estimators. *Ann. Statist.*, **21**, 694–710.

Hall, P. and Marron, J.S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, **6**, 109-115.

Hall, P. and Qiu, P.H. (2005). Discrete-transform approach to deconvolution problems. *Biometrika*, **92**, 135–148.

Hazelton, M.L. (1996). Bandwidth selection for local density estimators. *Scand. J. Statist.*, **23**, 221–232.

Hazelton, M.L. (1998). Bias annihilating bandwidths for kernel density estimation at a point. *Statist. Probab. Lett.*, **38**, 305–309.

Hazelton, M.L. and Turlach, B.A. (2009). Nonparametric density deconvolution by weighted kernel estimators. *Statist. Computing*, **19**, 217–228.

Hesse, C.H. (1999). Data-driven deconvolution. *J. Nonparam. Statist.*, **10**, 343–373.

Holzmann, H., Bissantz, N., Munk, A. (2007). Density testing in a contaminated sample. *J. Multivar. Anal.*, **98**, 57–75.

Jones, M. C. and Sheather, S. J. (1991). Using non stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives *Statist. Probab. Lett.*, **6**, 511–514.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.

Meister, A. (2009). On testing for local monotonicity in deconvolution problems. *Statist. Probab. Lett.*, **79**, 312–319.

Ruppert, D. (1997). Empirical Bias Bandwidths for Local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.*, **92**, 1049–1062.

Sain, S. R. (2003). A new characterization and estimation of the zero-bias bandwidth. *Aust. N. Z. J Statist.*, **45**, 29–42

Schucany, W.R. (1995). Adaptive bandwidth selection for kernel regression. *J. Amer. Statist. Assoc.*, **90**, 535–540.

Sheather, S. (1983). A data-based algorithm for choosing the window width when estimating the density at a point. *Comp. Statist. Data Anal.*, **1**, 229–238.

Sheather, S.J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comp. Statist. Data Anal.*, **4**, 61–65.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc., Ser. B*, **53**, 683–690.

Staudenmayer J., Ruppert, D., Buonaccorsi, J.R. (2008). Density estimation in the presence of heteroscedastic measurement error. *J. Amer. Statist. Assoc.*, **103**, 726–736.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.

Stefanski, L. and Cook, J.R. (1995). Simulation-extrapolation: the measurement error jackknife. *J. Amer. Statist. Assoc.*, **90**, 1247–1256.

Wang, X.F., Fan, Z.Z., Wang, B. (2010). Estimating smooth distribution function in the presence of heteroscedastic measurement errors. *Comp. Statist. Data Anal.*, **54**, 25–36.

Zhang, S.P. and Karunamuni, R.J. (2009). Deconvolution boundary kernel method in nonparametric density estimation. *J. Statist. Plan. Infer.*, **139**, 2269–2283.

A Proofs

Proof of Lemma 1. Let $A_k = (b^k - a^k)/k$. For $\ell = 0, 2$, we have

$$\begin{aligned}
E\Big\{\int_{\mathcal{I}} h^{\ell} \hat{f}_X(x;h)\,dh\Big\} = \int_{\mathcal{I}} h^{\ell} E\{\hat{f}_X(x;h)\}\,dh \\
=& f_X(x)\int_{\mathcal{I}} h^{\ell} + f_X^{(k)}(x)\frac{\mu_{K,k}}{k!}\int_{\mathcal{I}} h^{\ell+k}\,dh \\
&+ f_X^{(k+2)}(x)\frac{\mu_{K,k+2}}{(k+2)!}\int_{\mathcal{I}} h^{\ell+k+2}\,dh + o\Big(\int_{\mathcal{I}} h^{\ell+k+2}\,dh\Big) \\
=& f_X(x)h_*^{\ell+1}A_{\ell+1} + f_X^{(k)}(x)\frac{\mu_{K,k}}{k!}h_*^{\ell+k+1}A_{\ell+k+1} \\
&+ f_X^{(k+2)}(x)\frac{\mu_{K,k+2}}{(k+2)!}h_*^{\ell+k+3}A_{\ell+k+3} + o(h_*^{\ell+k+3}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E\Big\{\int_{\mathcal{I}} h^k \hat{f}_X(x;h)\,dh \int_{\mathcal{I}} dh - \int_{\mathcal{I}} \hat{f}_X(x;h)\,dh \quad \int_{\mathcal{I}} h^k\,dh\Big\} \\
&= f_X(x)h_*^{k+2}A_{k+1}A_1 + f_X^{(k)}(x)\frac{\mu_{K,k}}{k!}h_*^{2k+2}A_{2k+1}A_1 \\
&\quad + f_X^{(k+2)}(x)\frac{\mu_{K,k+2}}{(k+2)!}h_*^{2k+4}A_{2k+3}A_1 + o(h_*^{2k+4}) - f_X(x)h_*^{k+2}A_1A_{k+1} \\
&\quad - f_X^{(k)}(x)\frac{\mu_{K,k}}{k!}h_*^{2k+2}A_{k+1}^2 - f_X^{(k+2)}(x)\frac{\mu_{K,k+2}}{(k+2)!}h_*^{2k+4}A_{k+3}A_{k+1} + o(h_*^{2k+4}) \\
&= f_X^{(k)}(x)\frac{\mu_{K,k}}{k!}h_*^{2k+2}\big(A_{2k+1}A_1 - A_{k+1}^2\big) \\
&\quad + f_X^{(k+2)}(x)\frac{\mu_{K,k+2}}{(k+2)!}h_*^{2k+4}\big(A_{2k+3}A_1 - A_{k+3}A_{k+1}\big) + o(h_*^{2k+4}),
\end{aligned}$$

and the result follows from the definition of $\hat{B}(x)$, if we take $C_1 = (A_{2k+3}A_1 - A_{k+3}A_{k+1})/(A_{2k+1}A_1 - A_{k+1}^2)$. □

Proof of Lemma 2. Let $A_k = (b^k - a^k)/k$. We have

$$\mathrm{var}\Big\{ \int_{\mathcal{I}} h^k \hat{f}_X(x;h)\,dh \int_{\mathcal{I}} dh - \int_{\mathcal{I}} \hat{f}_X(x;h)\,dh \quad \int_{\mathcal{I}} h^k\,dh \Big\} = \mathrm{var}\Big(\frac{1}{n}\sum_{j=1}^{n} T_j\Big) = \frac{1}{n}\,\mathrm{var}(T_1),$$

where

$$T_j = \int_{\mathcal{I}} h^{k-1} K_U\Big(\frac{x-W_j}{h}\Big)\,dh \int_{\mathcal{I}} dh - \int_{\mathcal{I}} h^{-1} K_U\Big(\frac{x-W_j}{h}\Big)\,dh \quad \int_{\mathcal{I}} h^k\,dh.$$

Let

$$E_1(h,g) = \mathrm{E}\Big\{K_U\Big(\frac{x-W}{h}\Big) K_U\Big(\frac{x-W}{g}\Big)\Big\}.$$

We have

$$\begin{aligned}\mathrm{var}(T_1) = &\iint_{\mathcal{I}} h^{k-1} g^{k-1} E_1(h,g)\,dh\,dg\Big(\int_{\mathcal{I}} dh\Big)^2 + \iint_{\mathcal{I}} h^{-1} g^{-1} E_1(h,g)\,dh\,dg\Big(\int_{\mathcal{I}} h^k\,dh\Big)^2 \\ &- 2\iint_{\mathcal{I}} h^{k-1} g^{-1} E_1(h,g)\,dh\,dg \int_{\mathcal{I}} dh \int_{\mathcal{I}} h^k\,dh + O(h_*^{2k+2}).\end{aligned}$$

Now for any integers ℓ_1 and ℓ_2, we have

$$\begin{aligned}&\int_{\mathcal{I}}\int_{\mathcal{I}} h^{\ell_1} g^{\ell_2} E_1(h,g)\,dh\,dg \\ &\quad= \int_{\mathcal{I}}\int_{\mathcal{I}}\int h^{\ell_1} g^{\ell_2} K_U\Big(\frac{x-w}{h}\Big) K_U\Big(\frac{x-w}{g}\Big) f_W(w)\,dw\,dh\,dg \\ &\quad= h_*^{\ell_1+\ell_2-2\beta+3}\int_a^b\int_a^b u^{\ell_1} v^{\ell_2}\int h_*^{2\beta-1} K_U\Big(\frac{x-w}{uh_*}\Big) K_U\Big(\frac{x-w}{vh_*}\Big) f_W(w)\,dw\,du\,dv \\ &\quad= h_*^{\ell_1+\ell_2-2\beta+3}\int_a^b\int_a^b u^{\ell_1} v^{\ell_2}\int h_*^{2\beta} K_U(y/u) K_U(y/v) f_W(x-h_*y)\,dy\,du\,dv \\ &\quad\sim h_*^{\ell_1+\ell_2-2\beta+3} f_W(x) c^{-2} A_{\ell_1,\ell_2},\end{aligned}$$

if we define

$$A_{j_1,j_2} = \frac{1}{2\pi}\int_a^b\int_a^b\int u^{j_1+\beta+1} v^{j_2+\beta+1} |t|^{2\beta} \phi_K(tu)\phi_K(tv)\,dt\,du\,dv,$$

and where we used Lemma 7. Therefore,

$$\mathrm{var}(T_1) \sim h_*^{2k-2\beta+3} f_W(x) c^{-2}\big(A_{k-1,k-1}A_1^2 + A_{-1,-1}A_{k+1}^2 - 2A_{k-1,-1}A_1A_{k+1}\big).$$

The result follows by taking $C_2 = \big(A_{k-1,k-1}A_1^2 + A_{-1,-1}A_{k+1}^2 - 2A_{k-1,-1}A_1A_{k+1}\big)/(A_{2k+1}A_1 - A_{k+1}^2)^2$. □

Proof of Lemma 3. We have

$$\begin{aligned}
E\Big\{\int_{\mathcal{I}} \hat f_X^{(k)}(x;h)\,dh\Big\} =& \int_{\mathcal{I}} E\{\hat f_X^{(k)}(x;h)\}\,dh\\
=& \int_{\mathcal{I}} h^{-k-1} E\Big\{K_U^{(k)}\Big(\frac{x-W}{h}\Big)\Big\}\,dh\\
=& \int_{\mathcal{I}} h^{-k-1} E\Big\{K^{(k)}\Big(\frac{x-X}{h}\Big)\Big\}\,dh\\
=& \int_{\mathcal{I}} h^{-k} K^{(k)}(u) f_X(x-hu)\,du\,dh\\
=& \int_{\mathcal{I}} K(u) f_X^{(k)}(x-hu)\,du\,dh\\
=& f_X^{(k)}(x)\int_{\mathcal{I}} dh + O(h_*^2),
\end{aligned}$$

where we used a first order Taylor expansion. If f_X has $2k+1$ continuous and bounded derivatives then we can go further in the Taylor expansion and replace the last term of the expression above by $f_X^{(k)}(x)\int_{\mathcal{I}} dh + f_X^{(2k)}(x)\mu_{K,k}/(k!)\int_{\mathcal{I}} h^k\,dh + o(h_*^{k+1})$. This proves the result if we define $C_3 = A_{k+1}A_1^{-1}$ with $A_k = (b^k - a^k)/k$. $\square$

Proof of Lemma 4. We have

$$\mathrm{var}\Big\{\int_{\mathcal{I}} \hat f_X^{(k)}(x;h)\,dh\Big\} = \int_{\mathcal{I}}\int_{\mathcal{I}} E\{\hat f_X^{(k)}(x;h)\hat f_X^{(k)}(x;g)\}\,dh\,dg - \Big[\int_{\mathcal{I}} E\{\hat f_X^{(k)}(x;h)\}\,dh\Big]^2.$$

Let

$$E_k(h,g) = \mathrm{E}\Big\{K_U^{(k)}\Big(\frac{x-W}{h}\Big)K_U^{(k)}\Big(\frac{x-W}{g}\Big)\Big\}.$$

We have

$$\begin{aligned}
&\int_{\mathcal{I}}\int_{\mathcal{I}} E\Big\{\hat f_X^{(k)}(x;h)\hat f_X^{(k)}(x;g)\Big\}\,dh\,dg\\
&\quad= \sum_{j_1,j_2=1}^{n}\int_{\mathcal{I}}\int_{\mathcal{I}} \frac{1}{n^2h^{k+1}g^{k+1}} E\Big\{K_U^{(k)}\Big(\frac{x-W_{j_1}}{h}\Big)K_U^{(k)}\Big(\frac{x-W_{j_2}}{g}\Big)\Big\}\,dh\,dg\\
&\quad= \frac{1}{n}\int_{\mathcal{I}}\int_{\mathcal{I}} h^{-k-1}g^{-k-1}E_k(h,g)\,dh\,dg + \frac{n-1}{n}\Big[\int_{\mathcal{I}} h^{-k-1}E\Big\{K^{(k)}\Big(\frac{x-X}{h}\Big)\Big\}\,dh\Big]^2\\
&\quad= \frac{1}{n}\int_{\mathcal{I}}\int_{\mathcal{I}} h^{-k-1}g^{-k-1}E_k(h,g)\,dh\,dg + \Big[\int_{\mathcal{I}} E\{\hat f_X^{(k)}(x;h)\}\,dh\Big]^2 + O(1/n),
\end{aligned}$$

so that

$$\begin{aligned}
&\mathrm{var}\Big\{\int_{\mathcal{I}} \hat f_X^{(k)}(x;h)\,dh\Big\}\\
&\quad\sim n^{-1}\int_{\mathcal{I}}\int_{\mathcal{I}} h^{-k-1}g^{-k-1}E_k(h,g)\,dh\,dg
\end{aligned}$$

$$
\begin{aligned}
=&n^{-1}\int_{\mathcal{I}}\int_{\mathcal{I}}\int h^{-k-1}g^{-k-1}K_U^{(k)}\Big(\frac{x-w}{h}\Big)K_U^{(k)}\Big(\frac{x-w}{g}\Big)f_W(w)\,dw\,dh\,dg\\
=&n^{-1}h_*^{-2k-2\beta+1}\int_a^b\int_a^b(uv)^{-k-1}\int h_*^{2\beta-1}K_U^{(k)}\Big(\frac{x-w}{uh_*}\Big)K_U^{(k)}\Big(\frac{x-w}{vh_*}\Big)f_W(w)\,dw\,du\,dv\\
=&n^{-1}h_*^{-2k-2\beta+1}\int_a^b\int_a^b(uv)^{-k-1}\int h_*^{2\beta}K_U^{(k)}(y/u)K_U^{(k)}(y/v)f_W(x-h_*y)\,dy\,du\,dv\\
\sim&n^{-1}h_*^{-2k-2\beta+1}f_W(x)c^{-2}\frac{1}{2\pi}\int_a^b\int_a^b\int(uv)^{\beta}|t|^{2\beta+2k}\phi_K(tu)\phi_K(tv)\,dt\,du\,dv,
\end{aligned}
$$

where we used Lemma 7. The result follows from the definition of $\widetilde{f}_X^{(k)}(x)$ if we take $C_4=(2\pi)^{-1}\int_a^b\int_a^b\int(uv)^{\beta}|t|^{2\beta+2k}\phi_K(tu)\phi_K(tv)\,dt\,du\,dv$. □

Lemma 5. *Suppose that K is symmetric and is such that $||\phi_K||_\infty<\infty$, $||\phi_K'||_\infty<\infty$, $\phi_U(t)\neq 0$ for all t, $||\phi_U'||_\infty<\infty$ and $\int\big[|t|^\beta+|t|^{\beta-1}\big]\big[|t^k\phi_K'(t)|+|t^{k-1}\phi_K(t)\cdot 1\{k\geq 1\}|+|t^k\phi_K(t)|\big]\,dt<\infty$. Then, if (3.8) holds, we have*

$$h^{2\beta}|K_U^{(k)}(x)|^2\leq c\min(1,|x|^{-2}).$$

Proof of Lemma 5. See Lemma 3.1 of Fan (1991a). □

Lemma 6. *Suppose that K is symmetric and satisfies $||\phi_K||_\infty<\infty$, $\phi_U(t)\neq 0$ for all t, $\int|t|^{\beta+k}|\phi_K(t)|\,dt<\infty$. Then, if (3.8) holds, we have, for k even,*

$$\lim_{n\to\infty}h^\beta K_U^{(k)}(x)=(-i)^k\frac{1}{2\pi c}\int e^{-itx}|t|^{\beta+k}\phi_K(t)\,dt,$$

Proof of Lemma 6. See Fan (1991a), page 105, but replacing there t^β by $|t|^\beta$. □

Lemma 7. *Let u and v be two finite and strictly positive constants, and assume that the conditions of the above two lemmas are satisfied for k even and let $K_n(x)=h^{2\beta}K_U^{(k)}(x/u)K_U^{(k)}(x/v)$ and g a bounded function. Then we have at any continuity point x of g*

$$\lim_{n\to\infty}\int K_n(y)g(x-hy)\,dy=\frac{g(x)}{c^2}\frac{(uv)^{\beta+k+1}}{2\pi}\int|t|^{2\beta+2k}\phi_K(tu)\phi_K(tv)\,dt \qquad (1.13)$$

with c as in (3.8).

Proof of Lemma 7. It follows from Lemma 6 that

$$\lim_{n\to\infty}K_n(x)=\frac{1}{4\pi^2c^2}\int e^{-itx/u}|t|^{\beta+k}\phi_K(t)\,dt\int e^{-itx/v}|t|^{\beta+k}\phi_K(t)\,dt.$$

We also have, from Lemma 5 that

$$\sup_n |K_n(x)| \le K^*(x)$$

where $K^*(x) = C\min(1, |x|^{-2})$ is such that $\int K^*(x)\,dx < \infty$ and $\lim_{x\to\infty} |xK^*(x)| = 0$, and C denotes a generic finite constant. Therefore, from Lemma 2.1 of Fan (1991a), we have, at any point x of continuity of g

$$\begin{aligned}
&\lim_{n\to\infty} \int h^{2\beta-1} K_U^{(k)}\Big(\frac{x-y}{hu}\Big) K_U^{(k)}\Big(\frac{x-y}{hv}\Big) g(y)\,dy \\
&\quad= \lim_{n\to\infty} \int K_n(z) g(x-hz)\,dz \\
&\quad= \frac{g(x)}{c^2} \int \Big(\frac{1}{2\pi}\int e^{-itx/u}|t|^{\beta+k}\phi_K(t)\,dt\Big)\Big(\frac{1}{2\pi}\int e^{-isx/v}|s|^{\beta+k}\phi_K(s)\,ds\Big)\,dx \\
&\quad= \frac{g(x)}{c^2}(uv)^{\beta+k+1} \int \Big(\frac{1}{2\pi}\int e^{-itx}|t|^{\beta+k}\phi_K(tu)\,dt\Big)\Big(\frac{1}{2\pi}\int e^{-isx}|s|^{\beta+k}\phi_K(sv)\,ds\Big)\,dx \\
&\quad= \frac{g(x)}{c^2}\frac{(uv)^{\beta+k+1}}{2\pi} \int |t|^{2\beta+2k}\phi_K(tu)\phi_K(tv)\,dt.
\end{aligned}$$

$\square$