An alternative view of the deconvolution problem

Aurore Delaigle

Department of Mathematics, University of Bristol, BS8 1TW Bristol, UK and

Department of Mathematics and Statistics, University of Melbourne, Australia.

email: Aurore.Delaigle@bris.ac.uk, Phone: +44 1179289127, Fax: +44 1179287999

Abstract

The deconvolution kernel density estimator is a popular technique for solving the deconvolution problem, where the goal is to estimate a density from a sample of contaminated observations. Although this estimator is optimal, it suffers from two major drawbacks: it converges at very slow rates (inherent to the deconvolution problem) and can only be calculated when the density of the errors is completely known. These properties, however, follow from a classical asymptotic view of the problem which lets the sample size $n \to \infty$ but where the error variance σ^2 is supposed to be fixed. We argue that, in many situations, a more appropriate way to derive asymptotic properties for the deconvolution problem is to consider that both $\sigma^2 \to 0$ and $n \to \infty$. In this context, not only do the rates of convergence of the deconvolution kernel density estimator improve considerably, but it is also possible to consistently estimate the target density with only very little knowledge on the error density. In particular, the deconvolution kernel density estimator becomes robust against error misspecification and a low-order approximation developed in the literature becomes consistent. We propose a data-driven procedure for the low-order method and investigate the numerical performance of the various estimators on simulated and real data examples.

Key words and phrases: asymptotic results, bandwidth selection, classical errors, kernel method, measurement errors, smoothing.

Running title: Deconvolution problem.

1 Introduction

The conventional deconvolution problem for density estimation is one where a sample of independent and identically distributed (i.i.d.) variables $Y_1, \ldots, Y_n$ are observed with random measurement error. More precisely, the observations are generated by the model

$$Y_j = X_j + \varepsilon_j,\ X_j \sim f_X \text{ and } \varepsilon_j \sim f_\varepsilon, \tag{1.1}$$

where the density f_X of X_j is the unknown quantity to estimate, ε_j is the error variable, independent of X_j, and f_ε is a known and fixed density. This problem has received considerable attention in the literature and has numerous applications in different fields such as, for example, astronomy, public health or econometrics.

In this context, the most popular and extensively studied nonparametric estimator of f_X is the deconvolution kernel density estimator developed by Carroll and Hall (1988) and Stefanski and Carroll (1990). Let K be a kernel function integrating to 1, $h = h_n$ be a positive smoothing parameter (the bandwidth) and let ϕ_g denote the Fourier transform (resp. characteristic function) of a function (resp. random variable) g. Then if $\phi_\varepsilon(t) \neq 0$ $\forall t \in I\!R$ and $\widehat{\phi}_{Y,n}(t) = n^{-1}\sum_{j=1}^n e^{itY_j}$, the estimator is defined by

$$\tilde{f}_X(x;h) = \frac{1}{2\pi}\int e^{-itx}\widehat{\phi}_{Y,n}(t)\frac{\phi_K(ht)}{\phi_\varepsilon(t)}\,dt, \tag{1.2}$$

if we assume that the integral exists. See van Es and Uh (2005), Meister (2004,2006) or Hall and Qiu (2005) for recent contributions. See also Carroll et al. (2006) and Delaigle, Hall and Qiu (2006).

The rates of convergence to zero of the Mean Integrated Squared Error (MISE) of this estimator have been studied by Fan (1991a,b) in the class of functions $f_X \in \mathcal{F}_{\alpha,C} = \{\text{densities } f \in \mathcal{C}^\alpha \text{ s.t. } ||f^{(\alpha)}||_\infty < C \text{ and } \int\{f^{(\alpha)}\}^2 < C\}$, with $\mathcal{C}^\alpha$ the class of α times continuously differentiable functions. These rates depend on the behaviour of ϕ_ε in the tails: if ε is ordinary smooth of order β (see (2.1)), the optimal rates are $\sup_{f_X\in\mathcal{F}_{\alpha,C}} \text{MISE}\{\widehat{f}_X(\cdot;h)\} \sim n^{-2\alpha/(2\alpha+2\beta+1)}$, whereas if ε is supersmooth of order β (see (2.2)), the optimal rates are

$\sup_{f_X \in \mathcal{F}_{\alpha,C}} \text{MISE}\{\widehat{f}_X(\cdot; h)\} \sim (\ln n)^{-2\alpha/\beta}$. See also Masry (1993). Although these rates are optimal – no nonparametric estimator can improve them – they are particularly slow, especially when the error density is 'too regular'. In the very common case of Gaussian errors, for example, the logarithmic rates of convergence often make the deconvolution problem appear unpractical. Carroll and Hall (2004) argue that finding consistent estimators for the deconvolution problem is a goal that is often unattainable, and, in practice, one may obtain better practical results by constructing a less ambitious low-order approximation of f_X, and accurately estimate that approximation rather than the density f_X itself.

In practice, however, reasonable results can be obtained with the deconvolution kernel density estimator, even with moderate sample sizes. In such cases, the rates predicted by the classical theory appear too pessimistic and not flexible enough to capture some of the subtleties of the contamination problem. In standard asymptotic theory, the quality of an estimator is assessed through its behaviour when the quality of the sample improves, which, in the classical approach, amounts to studying the 'ideal situation' where the sample size n tends to infinity. However, when the observations contain measurement errors, the quality of a sample does not only depend on its size but also crucially on the magnitude of the error variance σ^2. Clearly, here, the 'ideal situation' does not reduce to having a sample of very large size, but also a very small error variance. Hence, its seems natural, when studying asymptotic properties of estimators for the deconvolution problem, to adopt the alternative approach where both $n \to \infty$ and $\sigma^2 \to 0$.

Of course, in practice, σ^2 is not necessarily small. However, in the classical approach, n is not especially large either and yet the interest of analyzing theoretical properties for the unrealistic situation where $n \to \infty$ is by now well understood. In particular, it allows to uncover some important properties of an estimator when n is not too small. Hence, just like any given sample size can be considered as a finite sample approximation of $n \to \infty$, any given σ^2 can be considered as a finite sample version of $\sigma^2 \to 0$ and we can expect the double asymptotics to be a helpful description of an estimator as long as σ^2 is not too large. This alternative approach can also be motivated by real data applications where the error variance is small compared with the variance of X, but we will see later (numerical section) that it is not necessary to have a small error variance for this theory to be appropriate.

From this discussion, it becomes natural to rewrite model (1.1) as

$$Y_j = X_j + \sigma Z_j,\ X_j \sim f_X,\ Z \sim f_Z,\ \mathrm{Var}(Z) = 1 \tag{1.3}$$

where, here and below, when we refer to this model, we imply that the asymptotic properties we consider are for $\sigma \to 0$ and $n \to \infty$; when we refer to model (1.1), we imply that the asymptotics are for $n \to \infty$ only. Hall and Simar (2002) study a related problem in the context of boundary estimation. Fan (1992) studies the behaviour of the deconvolution kernel density estimator in a subclass of model (1.3). One of the contributions of this paper is to fill the gaps between the classical theory and model (1.3). In section 2, we revisit the behaviour of the deconvolution kernel density estimator under the alternative model and show that its rates of convergence improve considerably compared to the classical theory. We apply our results to the interesting case of replicated observations.

Despite these theoretical improvements, the deconvolution kernel density estimator can only be calculated if the error density f_ε is known. In section 3, however, we show that, under model (1.3), consistent estimation of the density f_X can be achieved when only a few low-order moments of f_ε are known. We prove that, in this setting, the low-order approximation of Carroll and Hall (2004) is consistent; further, our results imply that the deconvolution kernel density estimator is robust against error misspecification. We derive simple data-driven procedures of bandwidth selection for the low-order estimator and prove that its convergence rates compare fairly with those of $\tilde{f}_X$ which, in some particular cases, loses its optimality properties.

We investigate the numerical performance of the estimators in section 4 via simulated and real data examples. We show that, as expected by the theory, the low-order estimator and the deconvolution kernel density estimator with misspecified error density work very well for moderately large error variances, but also that their quality (relative to the deconvolution kernel density estimator with known error) does not deteriorate very rapidly when the error variance increases. We conclude in section 5 and defer the proofs to the appendix.

2 Properties of the deconvolution kernel density estimator

Suppose we have a sample $Y_1, \dots, Y_n$ of i.i.d. observations generated by model (1.3) – remember that when we refer to this model, we imply that the asymptotics will be for $\sigma \to 0$ and $n \to \infty$. In this context, the asymptotic behaviour of the deconvolution kernel density estimator changes drastically and depends, in a crucial way, on the magnitude of the error variance. Define a kernel of order α by a function K for which $\mu_{K,0} = 1$, $\mu_{K,j} = 0$ for $1 \le j < \alpha$ and $\mu_{K,\alpha} = c$, where $\mu_{K,i} \equiv \int x^i K(x)\,dx$, $\alpha \ge 1$ is an integer and $c \ne 0$ is some finite constant. Theorems 2.1 and 2.2 describe the rates of the deconvolution kernel density estimator when $n \to \infty$ and $\sigma \to 0$ for two classes of errors usually considered in the literature: ordinary smooth errors ε of order $\beta > 0$, which are such that

$$d_1|t|^{-\beta} \le |\phi_\varepsilon(t)| \le d_2|t|^{-\beta} \text{ for all } |t| > M, \tag{2.1}$$

with M, d_1, d_2 some positive constants and supersmooth errors ε of order $\beta > 0$, which satisfy

$$d_1|t|^{\gamma_1}\exp(-d_3|t|^\beta) \le |\phi_\varepsilon(t)| \le d_2|t|^{\gamma_2}\exp(-d_3|t|^\beta) \text{ for all } |t| > M, \tag{2.2}$$

with M, d_1, d_2, d_3, γ_1 and γ_2 some positive constants. For two sequences of numbers a_n and b_n, we use the notation $a_n >> b_n$ (resp. $a_n << b_n$) to represent $b_n = o(a_n)$ (resp. $a_n = o(b_n)$). The proofs of the theorems are given in the appendix.

Theorem 2.1. *For model* (1.3)*, if Z is ordinary smooth of order β, K is of order α and $\int |t|^{2\beta}|\phi_K(t)|^2\,dt < \infty$,* *(i) if* $\sigma = O(n^{-1/(2\alpha+1)})$ *and* $h \sim n^{-1/(2\alpha+1)}$*, we have*

$$\sup_{f_X \in \mathcal{F}_{\alpha,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O(n^{-2\alpha/(2\alpha+1)});$$

(ii) if $\sigma >> n^{-1/(2\alpha+1)}$ *and* $h \sim \sigma^{2\beta/(2\alpha+2\beta+1)} n^{-1/(2\alpha+2\beta+1)}$*, we have*

$$\sup_{f_X \in \mathcal{F}_{\alpha,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O(\sigma^{4\alpha\beta/(2\alpha+2\beta+1)} n^{-2\alpha/(2\alpha+2\beta+1)}).$$

Theorem 2.2. *For model* (1.3)*, if* Z *is supersmooth of order* β*,* K *is of order* α*,* ϕ_K *is supported on* $[-1,1]$ *and* $\int [|t|^{-2\gamma_1} + |t|^{-2\gamma_2}] |\phi_K(t)|^2 \, dt < \infty$*,*
(i) if $\sigma = O(n^{-1/(2\alpha+1)})$ *and* $h \sim n^{-1/(2\alpha+1)}$*, we have*

$$\sup_{f_X \in \mathcal{F}_{\alpha,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O(n^{-2\alpha/(2\alpha+1)});$$

(ii) if $\sigma = n^{-1/(2\alpha+1)} a(n)$*, where* $1 << a(n) << n^{1/(2\alpha+1)}$ *and* $h = (2d_3/D)^{1/\beta} \sigma \{\ln a(n)\}^{-1/\beta}$*, with* $D < 2\alpha + 1$*, we have*

$$\sup_{f_X \in \mathcal{F}_{\alpha,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O\big(\sigma^{2\alpha} \{\ln a(n)\}^{-2\alpha/\beta}\big).$$

Note that no bandwidth can improve the rates provided above. These results generalize those of Fan (1992), who derives Theorem 2.2 (*i*). They show that, when $\sigma = O(n^{-1/(2\alpha+1)})$, the rates of the deconvolution kernel density estimator are the error-free rates $n^{-2\alpha/(2\alpha+1)}$. For larger error variances ($\sigma >> n^{-1/(2\alpha+1)}$), the rate of the MISE of the estimator to zero ranges from $n^{-2\alpha/(2\alpha+1)}$ to the classical deconvolution rates.

The sheer fact of knowing that the rates of the deconvolution kernel density estimator improve considerably under model (1.3) is already quite enlightening, but it also helps understanding the situation of growing interest where $r \geq 2$ replicated observations of the form $Y_{ij} = X_i + \varepsilon_{ij}$, $j = 1, \ldots, r$, are available for each individual. See Carroll et al. (2006), among others. There, it is rather common to use the averaged observations $\bar{Y}_{i.} = X_i + \bar{\varepsilon}_{i.}$ because these data have an error variance $r \geq 2$ times smaller than the original sample. However, (in the ordinary smooth case) the averaged errors become smoother and hence, in the classical theory, the rates of the deconvolution estimator worsen, suggesting that we should rather use the original non averaged observations. Nevertheless, in finite sample, the variance reduction induced by the averaging process can lead to significant improvement of performance of the estimator. In theory, this can be justified by our results, since, in model (1.3), averaging the data, i.e. reducing the error variance, leads to an improvement, rather than a deterioration, of the convergence rates of the estimator.

3 Consistency without knowledge of the error density

Despite the fast rates derived in the previous section, the deconvolution kernel density estimator suffers from a severe drawback: it can only be calculated when the error density f_ε is known, which is not always realistic. In the context of model (1.3) however, we show that it is not necessary to know more than just a few low-order moments of f_ε in order to obtain consistent estimators of f_X.

3.1 Low-order approximation

We start by studying the theoretical properties of the low-order approximation developed by Carroll and Hall (2004). In their approach, based on the ‘classical’ theoretical point of view, it is seen as a non consistent estimator of f_X whose properties remain relatively obscure. Below, we show that, in our context, their approximation is a consistent estimator of f_X. Suppose we have i.i.d. observations $Y_1, \ldots, Y_n$ generated by model (1.3). Then $f_Y(x) = \int f_X(x-\sigma z) f_Z(z)\, dz$ and, by recursive application of Taylor expansions of $f_X(x-\sigma z)$ and its derivatives, it is readily shown that, if f_Z has α finite absolute moments and f_X has α continuous bounded derivatives,

$$f_X(x) = f_Y(x) + \sum_{m=1}^{\alpha} (-1)^m S_m \sigma^m f_Y^{(m)}(x) + o(\sigma^\alpha), \tag{3.1}$$

if we define $S_m = \sum_{r=1}^m \sum_{\substack{i_1,\ldots,i_r \geq 1 \\ i_1+\ldots+i_r=m}} (-1)^r \prod_{j\in\{i_1,\ldots,i_r\}} \mu_{Z,j}/(j!)$, with $\mu_{Z,j} = \int z^j f_Z(z)\, dz$.

Based on this equality, an estimator of f_X can be defined by

$$\widehat{f}_X(x;h) = \widehat{f}_Y(x;h) + \sum_{m=1}^{\alpha}(-1)^m S_m \sigma^m \widehat{f}_Y^{(m)}(x;h), \tag{3.2}$$

where $\widehat{f}_Y^{(m)}(.;h)$ is the error-free kernel density estimator of $f_Y^{(m)}$, defined by

$$\widehat{f}_Y^{(m)}(x;h) = \frac{1}{nh^{m+1}} \sum_{j=1}^{n} K^{(m)}\Big(\frac{x - Y_i}{h}\Big), \tag{3.3}$$

with K and h as at page 2. It is straightforward to check that this estimator is equal to the low-order approximation of Carroll and Hall (2004). Here and below, we refer to an 'error-free' estimator of a density f_T or its derivatives, as an estimator obtained from an error-free sample, i.e. from a sample $T_1, \dots T_n$ where $T_i \sim f_T$, $1 \leq i \leq n$. Similarly, we refer to the 'error-free' case as the case where the observations are not contaminated by a measurement error. Note that the condition on f_X is commonly used in kernel density estimation, where it is usually assumed that $\alpha = 2$.

One of the interests of the estimator (3.2) lies in the fact that, contrary to the deconvolution kernel density estimator $\tilde{f}_X$, it requires very little information about the error density, since only σ and low-order moments $\mu_{Z,j}$, $j \leq \alpha$, are needed; if these are unknown, they can be easily estimated, either via the empirical variance of the difference of replicated observations or, as proposed by Dunn (2004), by the method of moments via instrumental variables; see our real data example in section 4. From a practical point of view, it is also very easy to calculate; for example, under the usual assumption that $\alpha = 2$ and the error density is symmetric, the estimator (3.2) simplifies into $\widehat{f}_X(x;h) = \widehat{f}_Y(x;h) - \sigma^2 \widehat{f}_Y^{(2)}(x;h)/2$. Finally, unlike the estimator $\tilde{f}_X$, it does not restrict to the cases where the characteristic function of the error does not vanish.

Our alternative derivation of the estimator allows a simple understanding of its asymptotic behaviour, which depends on h and σ and on the relative magnitude of both. In the case where σ is sufficiently small, the $o(\sigma^\alpha)$ error of the approximation of f_X by the main terms of (3.1) is negligible and the main source of error for the estimator comes from the kernel estimation of f_Y and its derivatives. For σ larger, the error comes from both the approximation and the kernel estimators. In this case, the exact behaviour of the approximation error, of order $o(\sigma^\alpha)$, can only be established under an additional condition on f_X involving the smallest integer $k \geq \alpha + 1$ such that $S_k \neq 0$. We note that this condition is not needed for constructing the estimator but to establish the main term of the bias when σ is 'large'. In most practical situations, the kernel K is symmetric (α is even) and the error is symmetric. There, $S_m = 0$ for odd values of m and, typically, $k = \alpha + 2$. Let $\mathcal{L}_2$ denote the class of square integrable functions. The following conditions will be useful.

Condition A

(A1) f_X has α continuous and uniformly bounded derivatives and $f_X^{(\alpha)} \in \mathcal{L}_2$;

(A2) f_Z has α finite absolute moments;

(A3) K is of order α and has α continuous, bounded and absolutely integrable derivatives.

For a function $g \in \mathcal{L}_2$, we denote $R(g) = \int g^2$. We refer to the bandwidth that minimizes the MISE of the estimator as the optimal bandwidth and we denote it by h_{MISE}. In the theorem, k is as defined above.

Theorem 3.1. *Under Condition A,* $\text{MISE}\{\widehat{f}_X(.;h)\} = \text{AMISE}\{\widehat{f}_X(.;h)\} \times (1 + o(1))$*, where*

(i) if $\sigma = o(n^{-1/(2\alpha+1)})$*, we have, for* $h = h_{\text{MISE}} \sim n^{-1/(2\alpha+1)}$

$$\text{AMISE}\{\widehat{f}_X(.;h)\} = R(f_Y^{(\alpha)})(\alpha!)^{-2}\mu_{K,\alpha}^2 h^{2\alpha} + (nh)^{-1}R(K).$$

(ii) if $n^{-1/(2\alpha+1)} << \sigma << n^{-\alpha/(4\alpha k + k - 2\alpha^2)}$*,* f_Y *has* 2α *continuous and uniformly bounded derivatives,* f_X *has* k *continuous, uniformly bounded derivatives,* $f_X^{(k)} \in \mathcal{L}_2$ *and* $|\mu_{Z,k}| < \infty$*, we have, for* $h = h_{\text{MISE}} \sim \sigma^{2\alpha/(4\alpha+1)} n^{-1/(4\alpha+1)}$

$$\text{AMISE}\{\widehat{f}_X(.;h)\} = R(f_Y^{(\alpha)})(\alpha!)^{-2}\mu_{K,\alpha}^2 h^{2\alpha} + \sigma^{2\alpha}(nh^{2\alpha+1})^{-1}S_\alpha^2 R(K^{(\alpha)}).$$

(iii) if $\sigma >> n^{-\alpha/(4\alpha k + k - 2\alpha^2)}$*, under the same conditions on* f_X*,* f_Y *and* f_Z *as in (ii), we have, for* $h = h_{\text{MISE}} \sim \sigma^{(2\alpha-k)/(3\alpha+1)} n^{-1/(3\alpha+1)}$

$$\text{AMISE}\{\widehat{f}_X(.;h)\} = \sigma^{2k} R(f_Y^{(k)}) S_k^2 + \sigma^{2\alpha}(nh^{2\alpha+1})^{-1}S_\alpha^2 R(K^{(\alpha)}).$$

Note that since $k \geq \alpha+1$, we always have that $n^{-1/(2\alpha+1)} << n^{-\alpha/(4\alpha k+k-2\alpha^2)}$. It is clear that, as for the deconvolution kernel density estimator, the rates of convergence strongly depend on the magnitude of the error variance. A discussion on these rates will be provided later but we already note that, for error variances of order $O(n^{-1/(2\alpha+1)})$, they are the same error-free rates as for the deconvolution kernel density estimator. In practice, this means that, when the error variance is small, we can expect both estimators to perform very well. In the simulation section, we will see that, in fact, the error variance does not need to be extremely small for the estimator $\widehat{f}_X$ to work well. In the theorem, for simplicity, we disregarded the case $\sigma \sim n^{-1/(2\alpha+1)}$, for which the optimal bandwidth and the corresponding MISE are both of the same order as those described in $(i)-(ii)$ but with a more complicated expression. A similar remark applies to the case $\sigma \sim n^{-\alpha/(4\alpha k+k-2\alpha^2)}$, which behaves like $(ii)-(iii)$. These expressions, as well as the proof of the theorem, are readily obtained from Theorems A.1 and A.2 of the appendix.

Bandwidth selectors. We obtain analytic expressions for the asymptotic optimal bandwidth, h_{AMISE}, by minimizing the AMISE given in the three cases of the theorem. In each case, in order to come up with a practical bandwidth, we estimate the unknown quantity $R(f_Y^{(\ell)})$ by a plug-in estimator. See for example Silverman (1986). We examine the performance of these bandwidths in the simulation section and see that they work well in practice. We have:

(i) If $\sigma = o(n^{-1/(2\alpha+1)})$, then for $C_1 = (\alpha!)^2 R(K)/\{2\alpha\mu_{K,\alpha}^2 R(f_Y^{(\alpha)})\}$,

$$h_{\text{AMISE}} = C_1^{1/(2\alpha+1)} n^{-1/(2\alpha+1)}, \tag{3.4}$$

which is the same bandwidth as for the usual (error-free) kernel density estimator of f_Y.

(ii) If $n^{-1/(2\alpha+1)} << \sigma << n^{-\alpha/(4\alpha k+k-2\alpha^2)}$, then for $C_2 = C_1(2\alpha+1)S_\alpha^2 R(K^{(\alpha)})/R(K)$,

$$h_{\text{AMISE}} = C_2^{1/(4\alpha+1)} \sigma^{2\alpha/(4\alpha+1)} n^{-1/(4\alpha+1)}. \tag{3.5}$$

(iii) If $\sigma >> n^{-\alpha/(4\alpha k+k-2\alpha^2)}$, h_{AMISE} can only be found by reintroducing second order terms

in the AMISE expression (see appendix). For $C_3 = (-1)^{\alpha+k} R(K^{(\alpha)})\alpha!\, S_\alpha^2 / \big(2S_k \mu_{K,\alpha} \int f_Y^{(\alpha)} f_Y^{(k)}\big)$ and $C_4 = -C_3(2\alpha+1)/\alpha$, this gives

$$h_{\text{AMISE}} = \max(C_3, C_4)^{1/(3\alpha+1)} \sigma^{(2\alpha-k)/(3\alpha+1)} n^{-1/(3\alpha+1)}. \tag{3.6}$$

In particular, when $\alpha = 2$ and the error density is symmetric, we have $k = 4$, $S_2 = -1/2$, $S_4 = 1/4 - \mu_{Z,4}/(4!)$ and $\int f_Y^{(\alpha)} f_Y^{(k)} = -R(f_Y^{(3)})$.

Exact expression. In some cases, (3.1) is an exact expression for f_X, rather than just an approximation. This is for example the case for errors whose Fourier transform can be written as $\phi_Z(t) = (1 + \sum_{j=1}^{\beta} a_j t^j)^{-1}$ for all t, as shown in the appendix (page 24). For example, the Laplace error satisfies this condition for $\beta = 2$. Then, if $\alpha \geq \beta$, the formula (3.1) is exact (and the terms of order higher than β vanish). Moreover, in this case, the estimator (3.2) is equal to the deconvolution kernel density estimator $\tilde{f}_X(x; h)$. Here, although both estimators are equal, the deconvolution kernel density estimator can only be calculated if the error density f_Z is known, whereas the estimator (3.2) only requires the first few moments of f_Z. In the case where $\alpha < \beta$, our simulation results indicate that the estimator (3.2) remains a good alternative to the deconvolution kernel density estimator. Further, in this case, we show at page 12 that, in some occasions, the estimator (3.2) has better rates of convergence than the deconvolution kernel density estimator, which loses its optimality properties. In such cases, σ is small and the approximation error (of order $o(\sigma^\alpha)$) in (3.1) is negligible compared with the variance increase produced by the additional $\beta - \alpha$ kernel estimates $\widehat{f}_Y^{(j)}$, $j = \alpha+1, \ldots, \beta$, used by the deconvolution kernel density estimator.

Estimation of a cumulative distribution function. The same idea can be used to develop an estimator of the cumulative distribution function of X. Namely, from $F_X(x) = F_Y(x) + \sum_{m=1}^{\alpha} (-1)^m S_m \sigma^m f_Y^{(m-1)}(x) + o(\sigma^\alpha)$, we define the following estimator of F_X: $\widehat{F}_X(x) = \widehat{F}_Y(x; h) + \sum_{m=1}^{\alpha} (-1)^m S_m \sigma^m \widehat{f}_Y^{(m-1)}(x; h)$, where $\widehat{F}_Y(x; h) = n^{-1} \sum_{j=1}^{n} \kappa\{(x - Y_j)/h\}$, with $\kappa(x) = \int_{-\infty}^{x} K(u)\, du$, is the kernel estimator of F_Y. The MISE of this estimator is obtained by calculations similar to the density case. In particular, the MISE is of order n^{-1} whenever $\sigma = O(n^{-1/(2\alpha)})$.

Comparison with the deconvolution kernel density estimator. Before we compare the estimators $\widehat{f}_X$ and $\tilde{f}_X$, it is important to realize that cases (*ii*) and (*iii*) of Theorem 3.1 were obtained under the additional condition that $f_X \in \mathcal{F}_{k,C}$, with $k \geq \alpha + 1$ (without such an assumption, it is impossible to determine the order of the bias of the estimator $\widehat{f}_X$, which depends on a $o(\sigma^\alpha)$ term (see Theorem A.1), and hence it is impossible to compare the two estimators). If we change the conditions on the kernel and use a kernel of order k instead of a kernel of order α, then the rates of the deconvolution kernel density estimator can be improved by replacing α by k in Theorems 2.1 and 2.2. Nevertheless, it is well known that, in practice, increasing the order of the kernel introduces extra variability of estimators and generally does not improve their quality (see, for example, Marron and Wand (1992)). Similarly, there exist infinite order kernels, which have the property that the behaviour of the bias of kernel estimators depends only on the smoothness of the target density f_X (and these estimators have optimal rates of convergence), but, in practice, they usually suffer from some drawbacks which make their use quite unpopular. For example, the resulting estimators are often too wiggly and the good standard bandwidth selectors usually do not apply (the cross-validation method can be used, but this procedure is usually not very satisfactory, see for example Delaigle and Gijbels (2004)).

Therefore, and since the exact smoothness properties of the density f_X are usually unknown, the most commonly used kernels are of order 2 or 4. In view of these facts, it is thus legitimate to compare the rates of the deconvolution kernel density estimator with these of the alternative estimator, in the case where the kernel is of order $\alpha < k$ and $f_X \in \mathcal{F}_{k,C}$. It is in this most interesting case that the alternative estimator sometimes enjoys better theoretical properties than the deconvolution kernel density estimator, because its rates of convergence improve with the smoothness of f_X whether or not we increase the order of the kernel.

Suppose $f_X \in \mathcal{F}_{k,C}$ and K is of order α, with $k \geq \alpha+1$. From Theorems 2.1 and 2.2, we have $\sup_{f_X \in \mathcal{F}_{k,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O(\sigma^{4\alpha\beta/(2\alpha+2\beta+1)} n^{-2\alpha/(2\alpha+2\beta+1)})$ in the ordinary smooth case and $\sup_{f_X \in \mathcal{F}_{k,C}} \mathrm{MISE}\{\tilde{f}_X(.;h)\} = O(\sigma^{2\alpha}\{\ln a(n)\}^{-2\alpha/\beta})$ in the supersmooth case. In case (ii) of Theorem 3.1, we have $\sup_{f_X \in \mathcal{F}_{k,C}} \mathrm{MISE}\{\widehat{f}_X(.;h)\} \sim \sigma^{4\alpha^2/(4\alpha+1)} n^{-2\alpha/(4\alpha+1)}$. It follows that when the error is ordinary smooth, the MISE of the estimator (3.2) is of lower order than the MISE of the deconvolution kernel density estimator if and only if $\alpha < \beta$; they have the same rate when $\alpha = \beta$. In the supersmooth error case, the estimator (3.2) beats the deconvolution kernel density estimator whatever the value of α and β. In case (iii) of Theorem 3.1, we have $\sup_{f_X \in \mathcal{F}_{k,C}} \mathrm{MISE}\{\widehat{f}_X(.;h)\} \sim \sigma^{2k}$. It follows that, when the error is ordinary smooth, the estimator (3.2) beats the deconvolution kernel density estimator if and only if $\sigma = o(n^{-\alpha/(2\alpha k+2\beta k-2\alpha\beta+k)})$, which is only possible when $\alpha > \beta$; they have the same rate when $\alpha = \beta$. In the supersmooth error case, the estimator (3.2) has better rates of convergence than the deconvolution estimator if and only if $\sigma^{\beta(\alpha-k)/\alpha} >> \ln\left(\sigma n^{1/(2\alpha+1)}\right)$, which is satisfied unless the error variance tends very slowly to zero, i.e. is rather large.

3.2 Deconvolution kernel density estimator

It follows from the discussion at page 11 that, under model (1.3), the deconvolution kernel density estimator $\tilde{f}_X$ is robust against certain error misspecifications since, as long as the first α moments of f_ε are correctly specified, the estimator $\widehat{f}_X$ is consistent and equal to the deconvolution kernel density estimator which pretends that the error density f_ε is such that $\phi_\varepsilon(t) = (1 + \sum_{j=1}^{\alpha} a_j \sigma^j t^j)^{-1}$. More generally, the misspecified error density, say f_η, does not need to be of the form above but can be taken from any parametric family large enough to contain densities that match the first α moments of ε. It is not hard to prove, using the findings of the previous sections, that, as long as the first α moments of f_η equal those of f_ε, the deconvolution kernel density estimator is consistent: its bias is of order

$O(h^\alpha) + o(\sigma^\alpha)$, like the estimator $\widehat{f}_X$, and its variance is of the same order as the variance of the deconvolution kernel density estimator $\tilde{f}_X$ for the situation where the errors genuinely come from f_η. Since this variance is larger with supersmooth errors, this indicates that we should preferably select f_η in the ordinary smooth class.

In our simulations, we found that, when the error variance was not huge, the finite sample performance of the deconvolution kernel density estimator with misspecified error was often similar to that of the known error case, even when the wrong error η was normal. In their real data example, Delaigle and Gijbels (2004), page 18, already noted that the deconvolution estimators which assume Laplace or normal error densities with a same variance do not differ much unless the error variance is very large. For large error variance, their estimator becomes more erratic when they assume normal errors, which supports our preference for ordinary smooth errors. See also Delaigle (2007) for simulated examples on the robustness in problems of measurement errors. Note that, in the classical theory, the estimator is generally not robust against error misspecification (see Meister (2004)) and, once again, the alternative asymptotic approach we adopted in this paper allows to account for a behaviour of the estimator often encountered in practice and yet invalidated by the classical theory.

4 Numerical properties

We examine and compare the numerical properties of the two methods of estimation of f_X and of the kernel density estimator of f_Y, i.e. the estimator (3.3) with $m = 0$ *that ignores the error present in the data*, for kernels of order $\alpha = 2$. For the deconvolution kernel density estimator (DKDE), we use the plug-in bandwidth of Delaigle and Gijbels (2002,2004), and for the kernel density estimator (KDE), we use the plug-in bandwidth described in Silverman (1986). For the estimator (3.2), which we denote by LOE, we use bandwidths $h_1 =$ (3.4), $h_2 =$ (3.5) or $h_3 =$ (3.6), where $R(f_Y'')$ and $R(f_Y^{(3)})$ are estimated by the plug-in method described in Silverman (1986). We then write LOE$_i$ when we refer to (3.2) with bandwidth h_i. We do not report the results for bandwidth (3.6), as it was systematically outperformed by the other bandwidths. We used the standard normal kernel in the case of Laplace errors and their convolutions, and we used the kernel with characteristic function $\phi_K(t) = (1 - t^2)^3 \cdot 1_{[-1,1]}(t)$ for Gaussian errors (to ensure existence of the DKDE).

4.1 Simulated examples

We consider four target densities f_X corresponding to: (i) $X \sim 0.5\, N(-2; 1) + 0.5\, N(2; 1.5^2)$, (ii) $X \sim 0.5\, N(-3; 1) + 0.5\, N(2; 1)$, (iii) $X \sim 1/3\, N(0; 1.2^2) + 1/3\, N(1; 4) + 1/3\, N(2; 4)$, (iv) $X \sim \sum_{\ell=0}^{5}(2^{5-\ell}/63)\, N(65 - 96(1/2)^{\ell}/21; (32/63)^2/2^{2\ell})$ – smooth comb density from Marron and Wand (1992). Note that, even in the error-free case, these densities are hard to estimate.

In each case, we generate 500 samples of sizes $n = 50$, 100 or 250 from f_X and add some random noise $\varepsilon \sim f_\varepsilon$, where f_ε is either a normal, a Laplace, a 2- or 8-fold Laplace, where a p-fold Laplace is a Laplace convolved $p-1$ times with itself; the noise-to-signal ratio, defined by $\text{NSR} = \sigma^2/\operatorname{Var}(X)$, ranges from 5% to 30%. To evaluate performance, we calculate the 500 values of the Integrated Squared Error (ISE), defined by $\text{ISE}_{\widehat{f}} = \int(\widehat{f} - f_X)^2$, where $\widehat{f}$ is a calculated estimator. We show boxplots of these calculated ISE's or of the quantity $\log(\text{ISE}_m / \text{ISE}_{\text{DKDE}})$, where m is a method we compare with the DKDE. We also show, for one sample, the estimators found by each method. We use the same sample for each method; it is the sample giving the 249th or 250th smallest calculated ISE for the method LOE_2. We denote these samples by $\mathcal{S}_{249}$ and $\mathcal{S}_{250}$, respectively. We only present a portion of the results; the conclusions are also supported by the simulations not presented here.

Figure 1 shows the results for the estimation of density (ii) when $\text{NSR} = 5\%$. Since the error variance is small, we want to see if we could ignore the error in the analysis, i.e. use the KDE of f_Y to estimate f_X. For $\varepsilon \sim N(0, \sigma^2)$, we present boxplots of $\log(\text{ISE}_m / \text{ISE}_{\text{DKDE}})$ where m denotes the LOE_1, LOE_2 or the KDE of f_Y. In this case, the LOE (with any of the bandwidths (3.4) or (3.5)) outperforms the DKDE. These three estimators strongly outper-

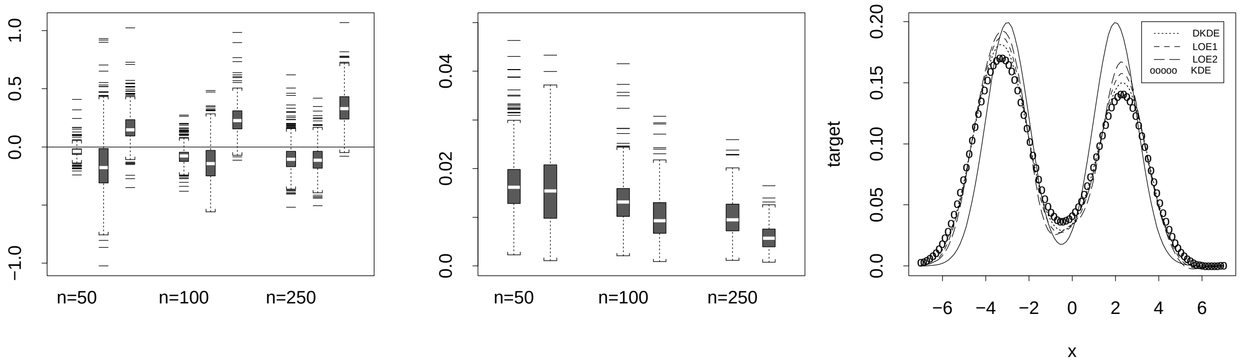

Figure 1: Estimation of density (ii) when NSR=5%: left panel: boxplots of $\log(\text{ISE}_m \,/\, \text{ISE}_{\text{DKDE}})$ for $\varepsilon \sim N$ and $n = 50$, 100 or 250; in each group of boxplots, m is, from left to right, LOE_1, LOE_2 or the KDE of f_Y; centre panel: boxplots of ISE_{DKDE} for $n = 50$, 100 or 250; in each group of boxplots, the 1st is for ε ∼Laplace and the 2nd for $\varepsilon \sim N$; right panel: estimated curves by the four methods when $\varepsilon \sim N$, $n = 100$ and using the sample $\mathcal{S}_{250}$.

form the KDE, which oversmoothes the data; this illustrates the non negligible improvement one can get by taking the error into account, even if this error is small. We also compare boxplots of the 500 calculated values of the ISE of the DKDE when the error is Laplace or Gaussian. Here, from the classical deconvolution theory, we expect the estimator to perform considerably better for Laplace than for Gaussian error, but we see that the both estimators are comparable (here the Gaussian error even works better). For such small error variances, the less conventional theory for model (1.3) seems more appropriate. On the right panel, we show, for one sample, the estimated curve for each method when $\varepsilon \sim N$ and $n = 100$.

In Figure 2, we check further the appropriateness of the LOE. The target is density (iv), the sample size is $n = 250$ and we consider Laplace, 2-fold Laplace and normal errors with NSR = 10% or 25%. We present boxplots of $\log(\text{ISE}_m \,/\, \text{ISE}_{\text{DKDE}})$ for m as in Figure 1 and we compare, for two samples, the curves found by each method. Without any surprise, all methods strongly outperform the KDE of f_Y which oversmoothes the data. The LOE still compares very fairly with the DKDE: here, although the error variance is not very small, LOE_1 even beats the DKDE but LOE_2 is less good when NSR = 25%. We note that the

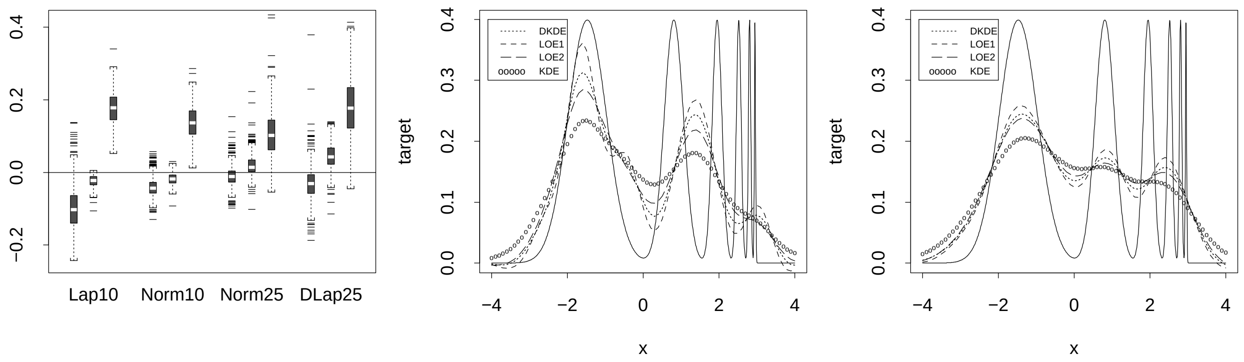

Figure 2: Estimation of density (iv): boxplots of $\log(\mathrm{ISE}_m\,/\,\mathrm{ISE}_{\mathrm{DKDE}})$ for $n = 250$, when $\varepsilon \sim$ Laplace and NSR = 10%, $\varepsilon \sim N$ and NSR = 10% or 25%, or $\varepsilon \sim$ 2-fold Laplace (DLap) and NSR = 25%; in each group of boxplots, m is, from left to right, LOE_1, LOE_2 or the KDE of f_Y (left panel). Estimated curves by the four methods when ε is DLap with NSR = 25% and using the sample $\mathcal{S}_{249}$ (centre panel) or $\mathcal{S}_{250}$ (right panel).

target density is particularly hard to estimate and, like in the error-free case, only the first mode is well estimated. In the case where $\varepsilon \sim$Laplace and NSR = 10%, $\widehat{f}_X$ and $\tilde{f}_X$ are equal except for the value of the bandwidth and we see the amount of improvement one can get by using bandwidth (3.4) when σ^2 is not too large.

In Figure 3, we compare the procedures for estimating density (i) with an 8-fold Laplace error. The bimodal and asymmetric shape of this density is similar to that of the target density of our real data example, and we choose the same error variance as in that example, i.e. NSR = 30%. As for the previous figures, we show boxplots of $\log(\mathrm{ISE}_m\,/\,\mathrm{ISE}_{\mathrm{DKDE}})$ and compare the estimated curves by each method for two samples. Here, the error variance is moderately large and the DKDE and LOE_2 give similar results. Once again, the KDE of f_Y systematically undersmoothes the data much more than the other methods. We obtained similar results when estimating the simpler unimodal asymmetric density (iii).

Finally, in Figure 4, we illustrate further the robustness of the DKDE by comparing, for samples of size $n = 50, 100$ or 250, boxplots for the estimation of densities (iv), (ii) and (i), for the DKDE with known f_ε, DKDE assuming normal error, LOE_1, LOE_2 and the KDE

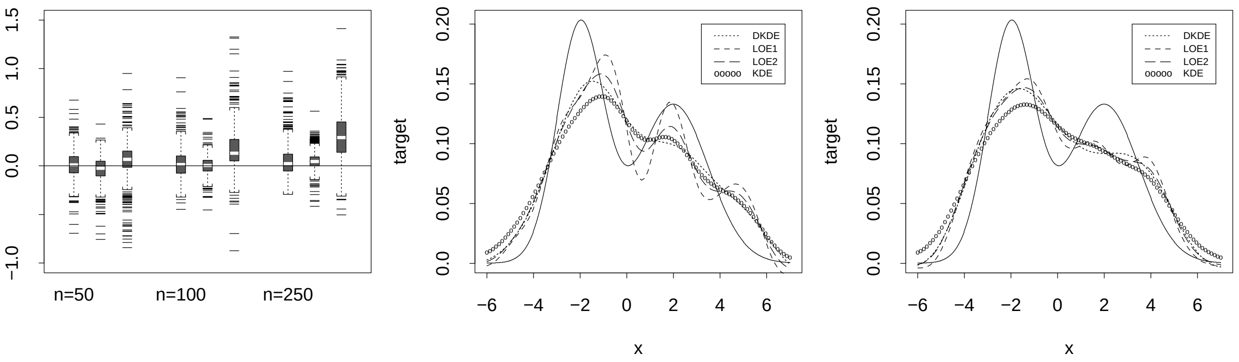

Figure 3: Estimation of density (i) when ε is 8-fold Laplace with variance NSR = 30%: boxplots of $\log(\text{ISE}_m \,/\, \text{ISE}_{\text{DKDE}})$ for $n = 50$, 100 or 250; in each group of boxplots, m is, from left to right, LOE$_1$, LOE$_2$ or the KDE of f_Y (left panel). Estimated curves by the four methods when $n = 250$ and using the sample $\mathcal{S}_{249}$ (centre panel) or $\mathcal{S}_{250}$ (right panel).

of f_Y. We see that the DKDE is robust against error misspecification and even assuming normal error gives reasonable results, although in the first panel, for $n = 50$, it just slightly outperforms the KDE that ignores the error. Overall, the DKDE with correct or wrong error density and the LOE gave quite similar results and strongly outperformed the KDE. We obtained similar results for other simulations we carried out, but for very large error variances, assuming normal error sometimes resulted in a bigger loss of performance.

In most of our simulation results, the best bandwidth for the LOE was (3.4), whereas bandwidths (3.5) and (3.6) tended to be slightly too large. We also tried larger sample sizes ($n \geq 1000$) and NSR ($> 30\%$) and there, (3.4) tended to be too small whereas the smallest of (3.5) and (3.6) gave better results, usually close but sometimes slightly less good than the DKDE. A 'conservative' approach, for large error variances, thus seems to be to select the smallest of the bandwidths (3.5) and (3.6). Our results for the Laplace case, where the DKDE and the LOE are equal except for the value of the bandwidth, raise the question of whether (and when) it could suffice or be preferable to use f_Y-related bandwidths, such as (3.4) to (3.6), which are much easier to calculate than the usual bandwidths.

We have seen that, for moderate sample size and error variances, the DKDE with mis-

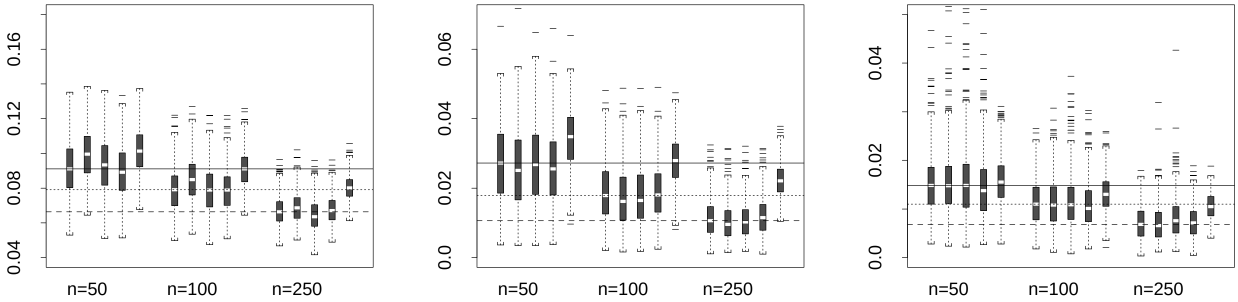

Figure 4: Boxplots of ISE_m for $n = 50, 100$ or 250. In each group of boxplots, m is, from left to right, DKDE with known f_ε, DKDE assuming $\varepsilon \sim N$, LOE_1, LOE_2 and the KDE of f_Y. Left panel: target is density (*iv*), $\varepsilon \sim$ 2-fold Laplace and NSR = 10%; centre panel: target is density (*ii*), $\varepsilon \sim$ Laplace and NSR = 25%; right panel: target is density (*i*), $\varepsilon \sim$ Laplace and NSR = 30%. The horizontal lines show the median ISE of the 1st boxplot of each group.

specified error (preferably ordinary smooth) and the LOE (which can be seen as a DKDE which uses a different bandwidth) can be used quite confidently as substitutes to the DKDE with known error. It is clear however that, for huge error variances, these estimators are less appropriate since the approximation error in (3.1) can sometimes get quite large. One might argue that, in that case, no estimator will give good results, but, if the error density is known, we should preferably use the DKDE.

4.2 Real data example: the sucrase data

The data concern the measurement of the enzyme sucrase in intestinal tissues of 24 patients. In this example, the sucrase (X) was measured by two different methods to which we refer as the pellet (Y) and the homogenate (T) methods, see Carter (1981) for a complete description. Our goal is to estimate the density of the actual content of sucrase X in the intestinal tissues from one of the two measurements (in this case we use Y). The error density is unknown but a third (instrumental) variable U, the alkaline phosphate, was also measured for each patient. In this example, the variables can be modelled as $Y = X + \varepsilon$, $T = \alpha + \beta X + \delta$ and $U = \gamma + \lambda X + \nu$, where α, β, γ and λ are unknown constants and ε, δ and ν are uncorrelated

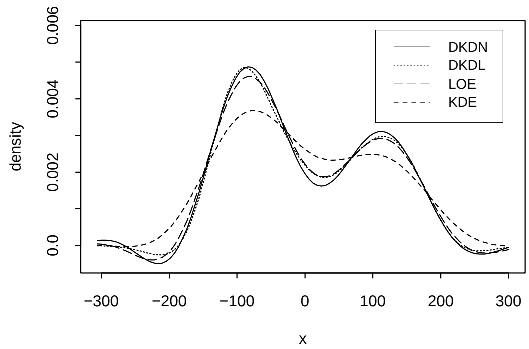

Figure 5: Estimation of the density of sucrase, using the LOE, the KDE which ignores the error or the DKDE when assuming a normal error (DKDN) or a Laplace error (DKDL).

error variables of zero mean, see Dunn (2004). From this relation, the variance of ε can then be estimated by the method of moments through the 24 observations on the three variables, which yields approximately $\sigma^2 = (1/3)\,\mathrm{Var}(X)$. See Dunn (2004) for detailed calculations.

Here the error density is unknown and we calculate the DKDE assuming Gaussian or Laplace error with a variance $\sigma^2 = (1/3)\,\mathrm{Var}(X)$. From section 3.2, both estimators should be quite similar. We compare the results with the KDE of f_Y (i.e. the estimator obtained when ignoring the error in the data) and the LOE for $\alpha = 2$, $\mu_{Z,1} = 0$ and $\sigma^2 = (1/3)\,\mathrm{Var}(X)$. The results are depicted in Figure 5, where we present the estimated densities of the centered sucrase. Here, LOE$_1$ and LOE$_2$ gave the same curve, which we denoted by LOE. The LOE and the DKDE with normal or Laplace error are very close and, as was the case in our simulations, the KDE seems to strongly oversmooth the data. The estimated density is bimodal, suggesting two groups of patients for which the sucrase concentration differs significantly.

5 Conclusion

We have studied the deconvolution problem in the asymptotic context where $\sigma^2 \to 0$ and $n \to \infty$. This alternative approach of describing the asymptotics has allowed us to theoretically account for several results that are encountered in practice but which are yet invalidated by the classical theory. In particular, we have seen why the deconvolution kernel density estimator does not especially work as bad as expected, we have proved and illustrated its robustness to error misspecification, we have justified the procedure of averaging replicated observations and we have proved, both in theory and in practice, that, even when the error is small, the improvement one can get by taking it into account is usually non negligible.

We have been able to clarify the properties of a low-order approximation proposed in Carroll and Hall (2004) as a substitute to the (seemingly too hard) deconvolution problem. We have shown that it is a consistent estimator and is indeed a good alternative of the deconvolution kernel density estimator, especially when little information is available about the error density. While our results show that, if the error variance is not too large, the low-order method can occasionally beat the deconvolution kernel density estimator when the deconvolution problem is very hard, they also imply that, when the error variance is large, this alternative estimator can not be expected to work better than the deconvolution kernel density estimator, even in cases where the latter has very slow convergence rates.

A Proofs of the main results

Rates for the deconvolution kernel density estimator. Similarly as for model (1.1), it is easy to prove that the integrated squared bias satisfies

$$\int [\mathrm{Bias}\{\tilde{f}_X(x;h)\}]^2\,dx = \frac{h^{2\alpha}\mu_{K,\alpha}^2}{(\alpha!)^2}\int (f_X^{(\alpha)})^2 + o(h^{2\alpha}), \tag{A.1}$$

whereas the integrated variance can be written as

$$\int \mathrm{Var}\{\tilde{f}_X(x;h)\}\,dx = \frac{1}{2\pi nh}\int |\phi_K(t)|^2|\phi_Z(\sigma t/h)|^{-2}\,dt + O(n^{-1}). \tag{A.2}$$

The next two proofs follow from this result.

Proof of Theorem 2.1. From (A.1) and (A.2), we can write AMISE $= c_1h^{2\alpha} + I$, where $I = (2\pi nh)^{-1}\int |\phi_K(t)|^2|\phi_Z(\sigma t/h)|^{-2}\,dt$ and c_1 is a positive constant. From (2.1), we find the following upper bound for I

$$I \leq \frac{c}{2\pi nh} \int_{|t| \leq Mh/\sigma} |\phi_K(t)|^2 \, dt + \frac{d_1^{-2} \sigma^{2\beta}}{2\pi nh^{2\beta+1}} \int_{|t| > Mh/\sigma} |\phi_K(t)|^2 |t|^{2\beta} \, dt, \tag{A.3}$$

where $c = (\inf_{|u| \leq M} |\phi_Z(u)|^2)^{-1} < \infty$. The behaviour of (A.3) and a lower bound for I depend on the behaviour of σ/h.

(a) If $\sigma = O(h)$, (A.3) $\leq c_2/(nh)$, with c_2 a positive constant, and, for n large enough, $I \geq c/(2\pi nh) \int_{|t| \leq 1/2} |\phi_K(t)|^2 \, dt = c_3/(nh)$, with c_3 a positive constant. It follows that the optimal bandwidth satisfies $h \sim n^{-1/(2\alpha+1)}$ and $\sigma = O(n^{-1/(2\alpha+1)})$.

(b) If $\sigma >> h$, we have (A.3) $\leq c_2 \sigma^{2\beta}/(nh^{2\beta+1})$, with c_2 a positive constant, and, for n large enough, $I \geq d_2^{-2} \sigma^{2\beta}/(2\pi nh^{2\beta+1}) \int_{|t| > 1/2} |\phi_K(t)|^2 |t|^{2\beta} \, dt = c_3 \sigma^{2\beta}/(nh^{2\beta+1})$, with c_3 a positive constant. It follows that the optimal bandwidth satisfies $h \sim \sigma^{2\beta/(2\alpha+2\beta+1)} n^{-1/(2\alpha+2\beta+1)}$ and $\sigma >> n^{-1/(2\alpha+1)}$. □

Proof of Theorem 2.2. Similarly to the proof of Theorem 2.1, we need to study the behaviour of I, which depends on the behaviour of σ/h. The case $\sigma = O(h)$ is similar to Theorem 2.1. For $\sigma >> h$, from (2.2) and the fact that ϕ_K is supported on $[-1, 1]$, we have for n large enough, $c_1(\sigma/h)^{-2\gamma_2} \exp(2d_3|\sigma/(2h)|^\beta)/(nh) \leq I \leq c_2(\sigma/h)^{-2\gamma_1} \exp(2d_3|\sigma/h|^\beta)/(nh)$, with c_1 and c_2 two positive and finite constants.

Take $h = (2d_3/D)^{1/\beta} \sigma \{\ln a(n)\}^{-1/\beta}$, with $0 < D < 2\alpha + 1$ a constant. We get

$$\frac{(\sigma/h)^{-2\gamma_1}}{nh} \exp(2d_3|\sigma/h|^\beta) \sim \frac{\{\ln a(n)\}^{(-2\gamma_1+1)/\beta}}{n^{2\alpha/(2\alpha+1)}} a(n)^{D-1},$$

and $h^{2\alpha}$, the squared bias term, behaves like $n^{-2\alpha/(2\alpha+1)} a(n)^{2\alpha} \{\ln a(n)\}^{-2\alpha/\beta}$ which dominates the upper bound of the variance term. Hence, for that bandwidth, we have MISE $\sim n^{-2\alpha/(2\alpha+1)} a(n)^{2\alpha} \{\ln a(n)\}^{-2\alpha/\beta}$. Clearly, a bandwidth of larger order would increase the squared bias term, and hence would increase this rate. It is not difficult to see that, for a bandwidth of smaller order, the lower bound of the variance is of an order larger than this rate. □

Proof of Theorem 3.1. The proof follows from the next two theorems describing the behaviour of the bias and variance of the estimator.

Theorem A.1. *Under Condition A, we have*

(i) if $\sigma = O(h)$*,* $\text{Bias}\{\widehat{f}_X(x;h)\} = (-1)^\alpha f_Y^{(\alpha)}(x)h^\alpha(\alpha!)^{-1}\mu_{K,\alpha} + o(h^\alpha)$*, where the remainder terms are uniform in* x*.*

(ii) if $\sigma >> h$*, then if* f_Y *has* 2α *continuous and uniformly bounded derivatives,* f_X *has* k *continuous and uniformly bounded derivatives and* $|\mu_{Z,k}| < \infty$*, we have*

$$\text{Bias}\{\widehat{f}_X(x;h)\} = (-1)^\alpha f_Y^{(\alpha)}(x)\frac{h^\alpha}{\alpha!}\mu_{K,\alpha} + (-1)^{k+1}\sigma^k f_Y^{(k)}(x)S_k + o(h^\alpha) + o(\sigma^k), \tag{A.4}$$

where the remainder terms are uniform in x*.*

Proof of Theorem A.1. We prove the two cases separately.

(i) Under (A3), we have, if we set $\nu_m = \alpha - m$, $m \leq \alpha - 1$,

$$\text{E}\{\widehat{f}_Y^{(m)}(x;h)\} = f_Y^{(m)}(x) + \sum_{j=1}^{\nu_m}(-1)^j f_Y^{(m+j)}(x)\frac{h^j}{j!}\mu_{K,j} + o(h^{\nu_m}) = f_Y^{(m)}(x) + o(h^{\nu_m}), \tag{A.5}$$

since $\mu_{K,\nu_m} = 0$ for $\nu_m = 1, \ldots, \alpha - 1$ and where the last term is uniform in x. From (3.1), (3.2) and (A.5), we deduce

$$\begin{aligned}\text{E}\{\widehat{f}_X(x;h)\} =& f_Y(x) + (-1)^\alpha f_Y^{(\alpha)}(x)\frac{h^\alpha}{\alpha!}\mu_{K,\alpha} + \sum_{m=1}^{\alpha}(-1)^m S_m\sigma^m\{f_Y^{(m)}(x) + o(h^{\nu_m})\} + o(h^\alpha)\\ =& f_X(x) + (-1)^\alpha f_Y^{(\alpha)}(x)\frac{h^\alpha}{\alpha!}\mu_{K,\alpha} + o(h^\alpha) + o(\sigma^\alpha),\end{aligned}$$

where we used the fact that $\sigma^m h^{\nu_m} = O(h^\alpha)$. The conclusion follows from $\sigma = O(h)$.

(ii) Under the additional conditions, the term $o(\sigma^\alpha)$ equals minus the first non zero higher order term in the Taylor expansion of (3.1), giving $(-1)^{k+1}S_k\sigma^k f_Y^{(k)}(x) + o(\sigma^k)$, while the $o(h^{\nu_m})$ term of (A.5) is replaced by a $O(h^\alpha)$ term. □

Theorem A.2. *Under Condition A, we have*

$$\text{Var}\{\widehat{f}_X(x;h)\} = \frac{f_Y(x)}{nh}\Big(\int K^2 + T_{\alpha,1} + 2\,T_{\alpha,2}\Big) + o\big\{(nh)^{-1}\big\} + o\big\{(\sigma/h)^{2\alpha}(nh)^{-1}\big\}, \tag{A.6}$$

where we introduced the notations $T_{\alpha,1} = \sum_{m,l=1}^{\alpha}(-1)^{m+l}S_mS_l(\sigma/h)^{m+l}\int K^{(m)}(u)K^{(l)}(u)\,du$ *and* $T_{\alpha,2} = \sum_{m=1}^{\alpha}(-1)^m S_m(\sigma/h)^m\int K(u)K^{(m)}(u)\,du$ *and where the remainder terms are uniform in* x*.*

Proof of Theorem A.2. We have

$$\begin{aligned}\mathrm{Var}\{\widehat{f}_X(x;h)\} =& \mathrm{Var}\{\widehat{f}_Y(x;h)\} + \sum_{m,l=1}^{\alpha}(-1)^{m+l}S_mS_l\sigma^{m+l}\,\mathrm{cov}\{\widehat{f}_Y^{(m)}(x;h), \widehat{f}_Y^{(l)}(x;h)\}\\ &+ 2\sum_{m=1}^{\alpha}(-1)^m S_m\sigma^m\,\mathrm{cov}\{\widehat{f}_Y(x;h), \widehat{f}_Y^{(m)}(x;h)\},\end{aligned}$$

where, for any two positive integers $r, s \leq \alpha$, it is easy to check that

$$\mathrm{cov}\{\widehat{f}_Y^{(r)}(x;h), \widehat{f}_Y^{(s)}(x;h)\} = \frac{f_Y(x)}{nh^{s+r+1}}\int K^{(r)}(u)K^{(s)}(u)\,du + o\Big(\frac{1}{nh^{s+r+1}}\Big),$$

where the lower order terms are negligible uniformly in x. □

Derivation of bandwidth (3.6)**.** Here the bandwidth can not be found via the AMISE expression of case (*iii*) of Theorem 3.1, but can be found by reintroducing second order (bias) terms in this AMISE expression. Proceeding that way, we find

$$\mathrm{AMISE} = R(f_Y^{(k)})\sigma^{2k}S_k^2 + 2(-1)^{\alpha+k+1}\sigma^k S_k\frac{h^\alpha}{\alpha!}\mu_{K,\alpha}\int f_Y^{(\alpha)}f_Y^{(k)} + \sigma^{2\alpha}S_\alpha^2\frac{R(K^{(\alpha)})}{nh^{2\alpha+1}}. \qquad \text{(A.7)}$$

If $(-1)^{\alpha+k+1}S_k\mu_{K,\alpha}\int f_Y^{(\alpha)}f_Y^{(k)} > 0$, the optimal bandwidth is found by differentiating the AMISE, which gives $h = C_4^{1/(3\alpha+1)}\sigma^{(2\alpha-k)/(3\alpha+1)}n^{-1/(3\alpha+1)}$. Otherwise, the optimal bandwidth cancels the sum of the last two terms of (A.7), which gives $h = C_3^{1/(3\alpha+1)}\sigma^{(2\alpha-k)/(3\alpha+1)}$ $n^{-1/(3\alpha+1)}$.

Exact expression at page 11. Consider ordinary smooth errors whose Fourier transform can be written as $\phi_Z(t) = (1 + \sum_{j=1}^{\beta} a_jt^j)^{-1}$ for all t. We note that, for $j = 0, \ldots, \beta$, we have $\phi_Z^{(j)}(0) = i^j\mu_{Z,j}$ and $a_j = i^jS_j$. By the Fourier inversion theorem, we have $f_X(x) = (2\pi)^{-1}\int e^{-itx}\phi_Y(t)\phi_Z^{-1}(\sigma t)\,dt$, and we deduce

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\phi_Y(t)\,dt + \sum_{j=1}^{\beta}a_j\sigma^j\frac{1}{2\pi}\int e^{-itx}t^j\phi_Y(t)\,dt = f_Y(x) + \sum_{j=1}^{\beta}(-1)^jS_j\sigma^jf_Y^{(j)}(x).$$

It follows that if $\alpha \geq \beta$, the formula (3.1) is exact (the terms of order higher than β vanish).

Acknowledgement

Research supported by a Maurice Belz Fellowship and a Hellman Fellowship.

References

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. and Hall, P. (2004). Low-order approximations in deconvolution and regression with errors in variables. *J. Roy. Statist. Soc. Ser.B*, **66**, 31 – 46.

Carroll, R.J., Ruppert, D., Stefanski, L. and Crainiceanu, C.(2006). *Measurement error in nonlinear models, Second Edition.* Chapman and Hall/CRC, Boca Raton.

Carter, R.L. (1981). Restricted maximum likelihood estimation of bias and reliability in the comparison of several measuring methods. *Biometrics*, **37**, 733 – 741.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of Berkson and classical errors. *Canadian J. Statist.*, **35**, 1 – 16.

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. (2004). Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Comp. Statist. Data Anal.*, **45**, 249 – 267.

Delaigle, A. and Hall, P. and Qiu, P (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *J. Roy. Statist. Soc. Ser.B*, **68**, 201 – 220.

Dunn, G. (2004). *Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies, Second Edition.* Oxford U. P., New York.

Es, van, A.J. and Uh, H.-W. (2005). Asymptotic normality of kernel type deconvolution estimators. *Scand. J. Statist.*, **32**, 467 – 483.

Fan, J. (1991a). Global behaviour of deconvolution kernel estimates. *Statist. Sinica*, **1**, 541 – 551.

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

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian J. Statist.*, **20**, 155 – 169.

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

Hall, P. and Simar, L. (2002). Estimating a changepoint, boundary or frontier in the presence of observation error. *J. Amer. Statist. Assoc.*, **97**, 523 – 534.

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

Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Processes and their Applications*, **47**, 53 – 74.

Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canadian J. Statist.*, **32**, 439 – 449.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statist. Sinica*, **16**, 195 – 211.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **2**, 169 – 184.