Nonparametric function estimation under Fourier-oscillating noise

Aurore Delaigle
Department of Mathematics and Statistics,
University of Melbourne,
VIC, 3010, Australia
email: A.delaigle@ms.unimelb.edu.au

Alexander Meister
Institut für Mathematik,
Universität Rostock,
D-18051 Rostock, Germany
email: alexander.meister@uni-rostock.de

Abstract: In the popular deconvolution problem, the goal is to estimate a curve f from data that only allow direct estimation of another curve g, the convolution of f and a so-called error density. Unlike the standard assumption in deconvolution, we consider a more general setting where the characteristic function of the error density can have zeros. This problem is important as the characteristic functions of uniform distributions, and more generally of many compactly supported distributions, have some zeros. We propose a new nonparametric deconvolution estimator, prove that its convergence rates are not affected by the zeros if f has a finite left endpoint, and we show rate-adaptivity. We suggest data-driven bandwidth selectors and examine their finite sample behaviour via simulated examples.

Keywords: contaminated data; deconvolution; density estimation; errors-in-variables regression; kernel smoothing; measurement errors.

AMS 2000 subject classification: 62G07, 62G08.

1 Introduction

We consider curve estimation from a sample of data observed with additive measurement errors. This problem, often referred to as a deconvolution problem, or an errors-in-variables problem, has received a lot of attention. Fourier methods are very popular in the deconvolution literature because they turn convolution equations into simple products. However, their drawback is that they can only be used when the characteristic function of the measurement error has no zeros, because they involve dividing by it. Several authors have proposed ways to modify deconvolution estimators so that they can be used also when the characteristic function of the error has some zeros. However, the techniques developed so far are either cumbersome (they require the introduction of additional smoothing parameters) or they suffer from very poor rates of convergence.

In this paper, we show that, if the target curve has a finite left endpoint then, without introducing any additional smoothing parameter, it is possible to construct a simple deconvolution kernel estimator that is consistent even if the characteristic function of the error has some zeros. Our technique can be applied to general deconvolution problems, and we describe it in detail for the density case in Section 2, then show in Section 3 how to use it in the errors-in-variables regression problem. Other possible applications include image reconstruction. In Section 4, we prove consistency of the estimator in the general deconvolution problem and show that its rates are not affected by the zeros of the characteristic function of the error. We apply these results to derive asymptotic properties of the density and regression estimators. Like other kernel techniques, our estimator requires the selection of a smoothing parameter called the bandwidth. We suggest two data-driven ways to select it in practice. In Section 5.1, we suggest a general cross-validation procedure, study its theoretical properties, and show that it is rate-adaptive. In Section 5.2, we propose an alternative SIMEX bandwidth selection procedure. Finite sample performance of the method is illustrated in Section 6 via simulations and on a data example.

2 Density deconvolution

One of the most popular deconvolution problems is density estimation from data contaminated by additive measurement error, often referred to as density deconvolution. The goal is to estimate the density f_X of a random variable X when the only available data are an i.i.d. sample $W_1, \ldots, W_n$, with

$$W_j = X_j + \delta_j \,,\; j = 1, \ldots, n \,,$$

$X_j \sim f_X$, $\delta_j \sim f_\delta$ with f_δ known, and all the X_j's and δ_j's are independent. Thus, each observation $W_j \sim f_W$ is a contaminated version of X_j. Throughout, we follow the classical approach to deconvolution in which the error density f_δ is known so that one can focus on the deconvolution techniques themselves. In practice, the error density might be unknown but empirically accessible by some additional data, see e.g. Efromovich (1997) and Neumann (1997).

2.1 Classical deconvolution

Let ℓ^{ft} denote the Fourier transform of a function ℓ. Then $f_W = f_X * f_\delta$ is equivalent to $f_W^{\text{ft}} = f_X^{\text{ft}} f_\delta^{\text{ft}}$, so that if $f_\delta^{\text{ft}}(t) \neq 0 \; \forall t$, we can write $f_X^{\text{ft}} = f_W^{\text{ft}} / f_\delta^{\text{ft}}$. Motivated by these considerations, if $f_\delta^{\text{ft}}(t) \neq 0 \; \forall t$, the deconvolution kernel density estimator of Carroll and Hall (1988) and Stefanski and Carroll (1990) is

$$\widehat{f}_X(x) = (2\pi)^{-1} \int e^{-itx} K^{\text{ft}}(ht) \widehat{f}_W^{\text{ft}}(t) / f_\delta^{\text{ft}}(t) \, dt, \tag{2.1}$$

where $\widehat{f}_W^{\text{ft}}(t) = n^{-1} \sum_{j=1}^n e^{itW_j}$, $h > 0$ is a smoothing parameter called the bandwidth, and K is a symmetric function called the kernel. Thus, $K^{\text{ft}}(ht) \widehat{f}_W^{\text{ft}}(t)$ is an estimator of f_W^{ft}, kernel-regularized so that the integral in (2.1) exists.

In the classical deconvolution literature, it is common to consider errors of two types, ordinary smooth and supersmooth. Ordinary smooth errors of order $\alpha > 0$ satisfy $d_1(1 + |t|)^{-\alpha} \leq |f_\delta^{\text{ft}}(t)| \leq d_2(1 + |t|)^{-\alpha}$ for all $t \in \mathbb{R}$, with $0 < d_1 \leq d_2$ some constants. Supersmooth errors of order $\alpha > 0$ satisfy $d_1 |t|^\gamma \exp(-d_3 |t|^\alpha) \leq |f_\delta^{\text{ft}}(t)| \leq d_2 |t|^\gamma \exp(-d_3 |t|^\alpha)$ for all $t \in \mathbb{R}$, with $d_1 > 0$, $d_2 > 0$, $d_3 > 0$, and $\gamma \geq 0$, some

constants. In particular, both classes of errors have a characteristic function that never vanishes and decays approximately monotonically, i.e. the absolute value of the characteristic function is bounded from above and below by two positive and decreasing functions that coincide up to some different constants.

2.2 Fourier-oscillating errors

Although the estimator (2.1) is only defined when $f_\delta^{\text{ft}}(t) \neq 0 \ \forall t$, there exist many error densities f_δ whose characteristic functions have zeros. For example, uniform densities, self-convolved uniform densities, or the convolution of uniform densities and other densities all have a characteristic function f_δ^{ft} that has isolated zeros. In this context, Hall and Meister (2007) generalise the ordinary smooth and supersmooth errors into, respectively

$$|f_\delta^{\text{ft}}(t)| \ \geq \ c_1 \big|\sin(\pi t/\lambda)\big|^\nu (1+|t|)^{-\alpha}\,, \tag{2.2}$$

for all $t \in \mathbb{R}$ with $c_1 > 0$ a constant, and

$$|f_\delta^{\text{ft}}(t)| \ \geq \ c_1 \big|\sin(\pi t/\lambda)\big|^\nu |t|^\gamma \exp\{-d|t|^\alpha\}\,, \tag{2.3}$$

for all $t \in \mathbb{R}$, with $c_1 > 0$, $d > 0$, and $\gamma \geq 0$ some constants, and where $f_\delta \in L_2(\mathbb{R})$. See also Meister (2008). In both (2.2) and (2.3) we have $\nu \in \mathbb{Z}$ and $\nu, \lambda > 0$. Clearly, for errors in this class, f_δ^{ft} can have isolated zeros at non-zero integer multiples of λ, and the standard method at (2.1) cannot be used.

This type of error includes simple errors like uniforms, which arise frequently when a device is used to measure a physical quantity. For example, Sun et al. (2002) describe an experiment where data on the velocity of halo stars in the Milky Way are collected, and where the measurement errors due to effects such as the mechanical stiffness of the spectrograph are assumed to be uniformly distributed. Other examples include resolution limits of chronometers in problems where the variable of interest is the time taken for a task to be completed. If the timing resolution limit is L μs, assigning detection times to the center of the interval where the measured time falls

gives rise to a uniform error $U[-L/2, L/2]$. See for example Dosso et al. (1998). Similar problems also arise when an analog signal is digitized.

Our error model can also be employed to model more complex situations where the measurement process is affected by a superposition of several sources of errors (operator errors, machine imperfections, etc). For example, quantization devices are often assumed to produce uniformly distributed errors, on which instrumentation noise and the effect of external random factors is superimposed. See Knyupfer (1966) for an early consideration of this problem. Similarly, the error due to limited resolution of a machine can be superimposed to other sources of errors, such as those made by the experimenter. We will also see another interesting application of our model in the data section.

Several authors have already studied the problem of a vanishing f_δ^{ft}. A first method was developed by Devroye (1989), who modified the deconvolution kernel estimator so as to exclude neighbourhoods of the zeros of f_δ^{ft} from the integration domain. His method requires the choice of three parameters (one of which is used to determine the size of these neighbourhoods), which seems quite unattractive from a practical viewpoint. Moreover, his work is restricted to consistency with convergence rates not investigated.

In Hall and Meister (2007), the main idea is to replace $f_\delta^{\text{ft}}(t)$ by a threshold function $\rho(t)$, called a ridge parameter function, when it takes values too close to zero. Their method requires the selection of a functional smoothing parameter, and can be reduced to the selection of two smoothing parameters. Although the authors suggest a way to choose them, the estimator cannot achieve better convergence rates than $n^{-1/(2\nu)}$, whether f_X has a finite or infinite left endpoint. In the next section we propose a new estimator which attains the "classical" deconvolution rate $n^{-2\beta/(2\beta+2\alpha+1)}$, regardless of the parameter ν. Here, β denotes the smoothness degree of the target function f (see the definition at (4.26)). Thus, for large β, our estimator significantly improves the rates achieved by the estimator of Hall and Meister (2007).

The method of Meister (2008) requires the construction of local neighbourhoods around the zeros of f_δ^{ft}. The estimator of f_X^{ft} in those neighbourhoods is defined

by local polynomial continuation of an estimated function, which is defined on a domain outside the neighbourhoods. Meister (2008) shows that the convergence rates can be improved compared to Hall and Meister (2007) by imposing some additional local smoothness constraints on f_X^{ft}. Still, the rates are slower than those derived for ordinary smooth error densities in Fan (1991). This approach requires three smoothing parameters: the bandwidth plus two smoothing parameters that determine the length of the small neighbourhoods and the approximation region. Choosing them is not easy. In addition, when f_X is compactly supported (this is included in our context of finite left endpoint), Meister (2008) shows that his estimator attains convergence rates of order $a(n) \cdot n^{-2\beta/(2\beta+2\alpha+1)}$, where $a(n)$, with $1/a(n) = o(1)$, is a logarithmic loss factor. As already indicated above, our estimator does not suffer from any logarithmic loss.

2.3 Our proposed estimation procedure when the errors are Fourier-oscillating

In the existing approaches discussed in Section 2.2, the idea is to replace f_δ^{ft} by an approximation of f_δ^{ft} that does not have any zeros, depends on one or more smoothing parameters, and becomes arbitrarily close to f_δ^{ft} as n increases. However, in the particular case where the error density is uniform U[0,1], Groeneboom and Jongbloed (2003) show that, if f_X has a finite left endpoint, it is possible to construct a kernel density estimator of f_X for which the only smoothing parameter required is a bandwidth.

Motivated by their finding, our goal was to develop a kernel deconvolution estimator for the general errors defined at (2.2) and (2.3), that does not require any other smoothing parameter than the bandwidth. In particular we wanted to see if it is possible, without using any approximation, to express f_X in a form that could be estimated from the data without having to deal with division by zero.

For errors of the type defined at (2.2) and (2.3), f_δ^{ft} can vanish at $t = k\lambda$, $k \in \mathbb{Z}$. To deal with these zeros, we propose a method with two main steps. In the first, we

avoid division by zero by estimating a function p which is not equal to f_X (not even approximately); In the second, we reconstruct f_X from the function p.

Let supp denote support. In order to involve our method, we assume that

$$f_X \in L_2(\mathbb{R}) \text{ and } \mathcal{I} \,=\, \text{supp}\, f_X \,\subseteq\, [a, \infty), \tag{2.4}$$

where a is known (the only restriction on a is that it be finite, but the infimum of $\mathcal{I}$ can be much larger than a). In practice, our assumption that the support of f_X is finite can be justified by the fact that many physical quantities cannot take arbitrarily small values. For example, lots of econometric, medical or astronomic quantities (e.g. the salary of an individual, the systolic blood pressure of a patient, the velocity of a star) cannot take negative values. However, we should mention that W may also be limited by such physical constraints. For example, the measured systolic blood pressure of an individual is always positive. In such cases, our conditions cover only the cases where the left endpoint of f_X is strictly larger than that of f_W, and $|\delta|$ does not take values larger than the difference of the two endpoints.

As a first step, instead of trying to estimate f_X directly, we estimate the function $p(x) = (2\pi)^{-1} \int e^{-itx} p^{\text{ft}}(t)\, dt$, where

$$p^{\text{ft}}(t) = \big(\exp(2it\pi/\lambda) - 1\big)^{\nu} f_X^{\text{ft}}(t). \tag{2.5}$$

Thus, compared to f_X, in the definition of p we replace $f_X^{\text{ft}}(t)$ by the *smoothing parameter-free* function $\big(\exp(2it\pi/\lambda) - 1\big)^{\nu} f_X^{\text{ft}}(t)$ (recall that ν and λ are known since f_δ is known). To estimate $p(x)$, we take

$$\widehat{p}(x) = \frac{1}{2\pi} \int e^{-itx} K^{\text{ft}}(th) \widehat{\Phi}_p(t)\, dt, \tag{2.6}$$

where $h > 0$ is a bandwidth, $K \in L_2(\mathbb{R})$ is a kernel function such that $\|K^{\text{ft}}\|_\infty < \infty$, K^{ft} is compactly supported, and where, for $\widehat{f}_W^{\text{ft}}$ as in (2.1),

$$\widehat{\Phi}_p(t) \,=\, \big(\exp(2it\pi/\lambda) - 1\big)^{\nu} \widehat{f}_W^{\text{ft}}(t) / f_\delta^{\text{ft}}(t) \tag{2.7}$$

for all $t \in \mathbb{R} \backslash \{k\lambda : k \in \mathbb{Z}, k \neq 0\}$. For $t = k\lambda$, $k \in \mathbb{Z} \backslash \{0\}$, we can conventionally put $\widehat{\Phi}_p(k\lambda) = 0$ or $\widehat{\Phi}_p(k\lambda) = \lim_{t \to k\lambda} \widehat{\Phi}_p(t)$ if it exists. As the function $\widehat{\Phi}_p$ arises only in

an integral, those values of $\widehat{\Phi}_p$ do not matter. Furthermore, we have

$$\limsup_{t\to k\lambda} \left|\widehat{\Phi}_p(t)\right| \;\le\; 2^{\nu+1}\left|\widehat{\Phi}_g(k\lambda)\right| c_1^{-1}(1+|k\lambda|)^{\alpha}\,, \quad \text{a.s.}\,,$$

and thus $\widehat{\Phi}_p$ is well defined for all t.

Clearly, for $\nu > 0$ (i.e. when $f_\delta^{\rm ft}$ has zeros), $p(x)$ is not equal (even approximately) to $f_X(x)$. By using $\widehat{p}$ instead of $\widehat{f}_X$ at (2.1) we avoid the problem of dividing by zero, though $\widehat{p}$ is not an estimator of f_X. However, if f_X has a finite left endpoint, then it is possible to reconstruct f_X from p, and thus to derive an estimator of f_X from $\widehat{p}$. To see this, note that, by the binomial expansion, we can write

$$p^{\rm ft}(t) = \sum_{k=0}^{\nu} \binom{\nu}{k} (-1)^{\nu-k} e^{2itk\pi/\lambda} f_X^{\rm ft}(t), \tag{2.8}$$

so that

$$p(x) \;=\; \sum_{k=0}^{\nu} \binom{\nu}{k} (-1)^{\nu-k} f_X(x - 2k\pi/\lambda). \tag{2.9}$$

Next we show how (2.9) can be inverted to write f_X as a function of p. Since the support $\mathcal{I}$ of f_X is in $[a,\infty)$, we have that $f_X(u) = 0$ for all $u < a$, and, in particular, for $u = x - 2k\pi/\lambda$, when $k > (x-a)\lambda/(2\pi)$. Therefore, as long as $J+1 > (x-a)\lambda/(2\pi)$ (for example, $J = \lceil \lambda(x-a)/(2\pi)\rceil$), we have, for any given $x \in \mathbb{R}$, $\mathbf{P}(x) \;=\; \mathbf{\Gamma}\,\mathbf{F}(x)$, where $\mathbf{P}(x) \;=\; \big(p(x),\ldots,p(x-2J\pi/\lambda)\big)^T$, $\mathbf{F}(x) \;=\; \big(f_X(x),\ldots,f_X(x-2J\pi/\lambda)\big)^T$, and $\mathbf{\Gamma} = \big(\Gamma_{j,k}\big)_{1\le j,k\le J+1}$, with $\Gamma_{j,k} = \gamma_{\nu,k-j}$, and

$$\gamma_{\nu,k} \;=\; \begin{cases} \binom{\nu}{k}(-1)^{\nu-k} & \text{if } k \in \{0,\ldots,\nu\}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Here we have used the fact that, when $k > J$, all the terms of (2.9), with x replaced by $x - 2k\pi/\lambda$, are zero. Note that $\mathbf{\Gamma}$ is an upper triangular matrix with all its diagonal components equal to $(-1)^\nu$. Hence, it is invertible and we can write $\mathbf{F}(x) = \mathbf{\Gamma}^{-1}\,\mathbf{P}(x)$. Therefore, for all $x \in \mathbb{R}$ we have

$$f_X(x) = (1,0,\ldots,0)\mathbf{F}(x) = \sum_{k=0}^{J} \eta_k p(x - 2k\pi/\lambda)\,, \tag{2.10}$$

where $(\eta_0, \ldots, \eta_J) = \big(1, 0, \ldots, 0\big)\mathbf{\Gamma}^{-1}$.

Finally, since the η_k's are known constants, and combining (2.10), (2.6) and the fact that we know that the support of f_X is included in $[a, \infty)$, we take

$$\begin{aligned}\widehat{f}_X(x) =& \sum_{k=0}^{J} \eta_k \widehat{p}(x - 2k\pi/\lambda) \cdot 1_{[a,\infty)}(x) \\ =& \frac{1}{2\pi} \int e^{-itx} K^{\mathrm{ft}}(th) \Big(\sum_{k=0}^{J} \eta_k \exp(2itk\pi/\lambda) \Big) \widehat{\Phi}_p(t)\, dt \cdot 1_{[a,\infty)}(x)\,. \end{aligned} \tag{2.11}$$

Our estimator can be written in the usual kernel form as

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_\delta\Big(\frac{x - W_j}{h}\Big) \cdot 1_{[a,\infty)}(x), \tag{2.12}$$

where

$$K_\delta(x) = \frac{1}{2\pi} \int e^{-itx} \frac{K^{\mathrm{ft}}(t)}{f_\delta^{\mathrm{ft}}(t/h)} \Big(\sum_{k=0}^{J} \eta_k e^{2itk\pi/(\lambda h)} \Big) \big(e^{2it\pi/(\lambda h)} - 1\big)^\nu\, dt. \tag{2.13}$$

Remark 1. Note that J depends on x, as already noted above. However, if we restrict to $x \le b$, where b is finite, then it is not necessary to use a different J for each x. As a matter of fact, in that case we can take $J = \lceil \lambda(b-a)/(2\pi) \rceil$ for all x.

Remark 2. If the target density does not have a finite left endpoint a but decays rather fast, our procedure still works well in practice, as we will see in Section 6. In theory, one could try to find a specific guideline for selecting $J = J(x)$, with $J(x) \to \infty$ and $a \to -\infty$ as $n \to \infty$. However, the results of Hall and Meister (2007) and Meister (2008) have it that no estimator (including ours) can have the classical deconvolution rates $n^{-2\beta/(2\beta+2\alpha+1)}$, and our estimator reaches these rates when a is finite (see Section 4). Moreover in practice, if we let $J \to \infty$, the estimator of f_X suffers from numerical instability. In particular, the size of the matrix $\mathbf{\Gamma}$ increases (thus the calculation of the inverse matrix $\mathbf{\Gamma}^{-1}$ becomes computationally more difficult).

Example 1. When $f_\delta = f_U^{*m} * f_V$, where f_U is a $U[-a_\delta, a_\delta]$ density, f_U^{*m} denotes f_U convolved m times with itself, and f_V is a density whose Fourier transform does

not vanish, we have $\nu = m$, $\lambda = \pi/a_\delta$ and $f_\delta^{\rm ft}(t) = (\sin a_\delta t)^m/(a_\delta t)^m\, f_V^{\rm ft}(t)$. Simple calculations show that $e^{2it\pi/(\lambda h)} - 1 = 2i\sin\{t\pi/(\lambda h)\}e^{it\pi/(\lambda h)}$, so that

$$K_\delta(x) = \frac{(2i)^m(a_\delta/h)^m}{2\pi}\sum_{k=0}^{J}\eta_k\int e^{-it\{x-a_\delta(2k+m)/h\}}K^{\rm ft}(t)t^m\,/f_V^{\rm ft}(t/h)\,dt.$$

When m is even, the sine part of the exponential above vanishes, and when m is odd its cosine part vanishes. To show an explicit example of calculation of the η_k's, consider the case where the error is a mixture of a uniform and another variable V, so $m = 1$. We have $\gamma_{\nu,k} = \gamma_{1,k} = \binom{1}{k}(-1)^{1-k}$ for $k = 0, 1$ and $\gamma_{1,k} = 0$ otherwise. Thus $\gamma_{1,0} = -1$, and $\gamma_{1,1} = 1$, so that if $J = 0$, then $\mathbf{\Gamma} = -1$ and $\eta_0 = \mathbf{\Gamma}^{-1} = -1$. If $J = 1$, then

$$\mathbf{\Gamma} = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix},\ \mathbf{\Gamma}^{-1} = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix}$$

and $(\eta_0, \eta_1) = (-1, -1)$, which is the first row of $\mathbf{\Gamma}^{-1}$. if $J = 2$, then

$$\mathbf{\Gamma} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix},\ \mathbf{\Gamma}^{-1} = \begin{pmatrix} -1 & -1 & -1 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

and $(\eta_0, \eta_1, \eta_3) = (-1, -1, -1)$, which corresponds to the first row of $\mathbf{\Gamma}^{-1}$. Similar calculations can be made for $J > 2$ and other values of m.

3 Errors-in-variables regression

Our procedure can be applied to the errors-in-variables regression problem. There the goal is to estimate a regression function g from i.i.d. data $(W_1, Y_1), \ldots, (W_n, Y_n)$ generated by the model

$$Y_j \,=\, g(X_j) + \varepsilon_j\,,\quad W_j = X_j + \delta_j$$

where $X_j \sim f_X$ is independent of $\delta_j \sim f_\delta$, the ε_j's are independent, $E\varepsilon_j = 0$, and $\sup_j E\varepsilon_j^2 < \infty$. See e.g. Fan and Truong (1993) and Carroll et al. (2006).

3.1 Nadaraya-Watson estimator

To construct a nonparametric estimator of g, a standard approach is to write $g = m/f_X$ where $m = g \cdot f_X$, construct estimators of m and f_X and take their ratio. In the error-free case, when kernel techniques are used, the resulting estimator of g is called the Nadaraya-Watson estimator. See Fan and Truong (1993) for an extension of that method to the errors-in-variables setting when f_δ^{ft} has no zeros. Suppose that g and f_X are supported on $\mathcal{I} \subset [a, b]$ where a and b are finite constants. By assuming that the right endpoint b is finite, integrability of the function $g^j f_X$ for $j = 1, 2$ is guaranteed. Take

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_\delta\Big(\frac{x - W_j}{h}\Big) \cdot 1_{[a,b]}(x), \tag{3.14}$$

$$\widehat{m}(x) = \frac{1}{nh} \sum_{j=1}^{n} Y_j K_\delta\Big(\frac{x - W_j}{h}\Big) \cdot 1_{[a,b]}(x), \tag{3.15}$$

with K_δ as in (2.13). We define a Nadaraya-Watson type estimator of g by

$$\widehat{g}(x) = \widehat{m}(x) / \max\{\widehat{f}_X(x), \rho\}, \tag{3.16}$$

where the truncation parameter $\rho > 0$ is used to avoid problems caused at points where $\widehat{f}_X(x)$ is too small. Note that, although it is not always mentioned in papers treating nonparametric regression, the problem of avoiding division by a too small number is quite standard.

3.2 Local polynomial estimator

One of the advantages of the Nadaraya-Watson estimator is that it is simple to understand and easy to implement in practice. A theoretical advantage is that it can be used with the sinc kernel, an infinite order kernel having the property that it adapts automatically to the smoothness of the curves. On the negative side, the convergence rate of the estimator corresponds to the lowest of the smoothness degrees of f_X and g. In particular, if g is smoother than f_X, then the estimator of g converges at a rate dictated by f_X and not g.

One way to avoid smoothness conditions on f_X when deriving convergence rates of estimators of g is to use local polynomial methods; see Fan and Gijbels (1996) for a thorough investigation of their properties in the error-free setting. An extension of those methods to errors-in-variables problems has recently been given by Delaigle et al. (2009). These estimators have the advantage of adapting automatically to the boundary of the design density and provide a natural and elegant solution to the problem of derivative estimation. Nevertheless, the construction of the local polynomial estimator requires inversion of a matrix depending on the kernel function; in order to ensure its invertibility, more specific kernels than the sinc kernel have to be used.

In the sequel, we show how to extend the concept of local polynomial regression estimators to the case where f_δ^{ft} has some zeros. Our local polynomial estimator of order p of $g^{(\ell)}$, with $p \geq \ell$, is

$$\widetilde{g}_p^{(\ell)}(x) = \ell! h^{-\ell} \boldsymbol{e}_{\ell+1}^T \widehat{\boldsymbol{S}}_n^{-1} \widehat{\boldsymbol{T}}_n \cdot 1_{[a,b]}(x), \tag{3.17}$$

where $\boldsymbol{e}_{\ell+1}^T = (0, , \ldots, 0, 1, 0, \ldots, 0)$ with 1 on the $(\ell+1)$th position, $\widehat{\boldsymbol{S}}_n = \{\widehat{S}_{n,j+k}(x)\}_{0 \leq j,k \leq p}$, and $\widehat{\boldsymbol{T}}_n = \{\widehat{T}_{n,0}(x), \ldots, \widehat{T}_{n,p}(x)\}^T$ with

$$\widehat{S}_{n,k}(x) = \frac{1}{nh} \sum_{j=1}^{n} K_{\delta,k}\Big(\frac{x - W_j}{h}\Big) \cdot 1_{[a,b]}(x), \tag{3.18}$$

$$\widehat{T}_{n,k}(x) = \frac{1}{nh} \sum_{j=1}^{n} Y_j K_{\delta,k}\Big(\frac{x - W_j}{h}\Big) \cdot 1_{[a,b]}(x), \tag{3.19}$$

where

$$K_{\delta,k}(x) = \frac{i^{-k}}{2\pi} \int e^{-itx} \frac{(K^{\mathrm{ft}})^{(k)}(t)}{f_\delta^{\mathrm{ft}}(-t/h)} \Big(\sum_{k=0}^{J} \eta_k e^{2itk\pi/(\lambda h)} \Big) \big(e^{2it\pi/(\lambda h)} - 1\big)^\nu \, dt. \tag{3.20}$$

The motivation for the construction of our estimator is essentially the same as in Delaigle et al. (2009), and we refer to that paper for details. Note that by Cramer's rule we can rewrite the estimator as

$$\tilde{g}_p^{(\ell)}(x) \;=\; \ell! h^{-\ell} \det(\widehat{\mathbf{A}}_{n,l}) / \det(\widehat{\mathbf{S}}_n) \cdot 1_{[a,b]}(x)\,,$$

where the matrix $\widehat{\mathbf{A}}_{n,l}$ is constructed by replacing the $\ell+1$st column of $\widehat{\mathbf{S}}_n$ by the vector $\widehat{\mathbf{T}}_n$. As in the Nadaraya-Watson case, to avoid division by zero or a very small number, we define a modified local polynomial estimator of $g^{(\ell)}$ by

$$\widehat{g}_p^{(\ell)}(x) = \frac{\ell! h^{-\ell} \det(\widehat{\mathbf{A}}_{n,l})}{\max\{\det(\widehat{\mathbf{S}}_n), \rho\}} \cdot 1_{[a,b]}(x), \tag{3.21}$$

where $\rho > 0$ is a truncation parameter. As usual, it is not hard to check that the Nadaraya-Watson estimator is equal to the local constant estimator, that is, to the local polynomial estimator of order $p = 0$.

4 Generalisation and asymptotic properties

In this section we prove that our procedure, applied to density, regression, or to estimate more general curves f, attains the same convergence rates whether $\nu = 0$ (i.e. f_δ^{ft} has no zeros) or $\nu > 0$ (i.e. f_δ^{ft} has zeros). Our results improve those of Hall and Meister (2007) and Meister (2008), who incur a logarithmic penalty when f_δ^{ft} has isolated zeros under slightly different smoothness constraints. To obtain these results, the condition that the support of f is included in $[a, \infty)$ is essential, as can be deduced from the lower bound proofs in Hall and Meister (2007) and Meister (2008). However, even under this condition, their estimators attain slower convergence rates than ours (see also Section 2.2).

4.1 Deconvolution in general

Our estimator can be applied in more general deconvolution problems than density and regression, and in this section we derive asymptotic theory for the general context. Suppose we wish to estimate the function f, but we only have data allowing direct empirical estimation of ξ, where $\xi = f * f_\delta$, for f_δ a known density satisfying (2.2) or (2.3). More specifically, we assume that the data are such that we can construct an estimator $\widehat{\Phi}_\xi(t)$ of ξ^{ft} for which

$$\sup_{t \in \mathbb{R}} E\big|\widehat{\Phi}_\xi(t) - \xi^{\text{ft}}(t)\big|^2 = O(1/n)\,, \tag{4.22}$$

where the constant contained in $O(\cdots)$ is uniform for all admitted functions f. For example, in the density deconvolution setting introduced in Section 2, we have $f = f_X$, $\xi = f_W$ and $\widehat{\Phi}_\xi = \widehat{f}_W^{\rm ft}$, the empirical characteristic function of the observed data.

We also assume that

$$f \in L_2(\mathbb{R}) \text{ and } \mathcal{I} \,=\, \text{supp}\, f \,\subseteq\, [a, \infty) \tag{4.23}$$

where a is known.

Based on arguments we developed in the density case, our general estimator of f is

$$\widehat{f}(x) = \frac{1}{2\pi} \int e^{-itx} K^{\rm ft}(th) \Big(\sum_{k=0}^{J} \eta_k \exp(2itk\pi/\lambda) \Big) \widehat{\Phi}_p(t)\, dt \cdot 1_{[a,\infty)}(x)\,. \tag{4.24}$$

where

$$\widehat{\Phi}_p(t) \,=\, \big(\exp(2it\pi/\lambda) - 1\big)^\nu \widehat{\Phi}_\xi(t)/f_\delta^{\rm ft}(t)\,. \tag{4.25}$$

Our result concerns the asymptotic behaviour of the mean integrated squared error (MISE) of $\widehat{f}$, where the integral is taken over a compact interval $[a, b]$ with an arbitrary but fixed constant $b > a$. In this case, we can take $J = \lceil \lambda(b-a)/(2\pi) \rceil$, a constant that does not depend on x in the construction of estimator (4.24). As noted earlier, this is sufficient for most practical applications, as one would rarely be interested in estimating a curve f outside a compact interval. In order to attain the convergence rates, we assume that f belongs to the Sobolev class

$$\mathcal{F}_{C,\beta}^S = \{ f \text{ supported on } [a,\infty) \text{ s.t. } \int \big|f^{\rm ft}(t)\big|^2 (1 + |t|^{2\beta}) dt \,\le\, C \}. \tag{4.26}$$

Thus $\beta > 0$ represents the smoothness degree of f.

Theorem 4.1 (Generalized ordinary smooth case)**.** *In the general deconvolution problem, suppose that f_δ satisfies* (2.2)*, that $K \in L_2(\mathbb{R})$, $\|K^{\rm ft}\|_\infty \le 1$, supp $K^{\rm ft} \subseteq [-1, 1]$, and $|K^{\rm ft}(t) - 1| \,=\, o(|t|^\beta)$, and assume that* (4.22) *holds. Then, if $h \asymp n^{-1/(2\beta+2\alpha+1)}$ and b denotes some arbitrary but fixed constant, for $\widehat{f}$ at* (4.24)*,*

$$\sup_{f \in \mathcal{F}_{C,\beta}^S} E \int_a^b \big|\widehat{f}(x) - f(x)\big|^2 dx \,=\, O\big(n^{-2\beta/(2\beta+2\alpha+1)}\big)\,.$$

These rates correspond to those derived by Fan (1993) for ordinary smooth f_δ in the case $\nu = 0$ (i.e. without zeros) and do not suffer from the isolated zeros of f_δ^{ft}.

Remark 3. *(Generalized supersmooth case).* Our estimator can also be applied in the generalized supersmooth case, with for f_δ as in (2.3). Using techniques similar to the ordinary smooth case, it can be proved that, as long as $h \sim c\,(\ln n)^{-1/\alpha}$ with $c > (2d)^{1/\alpha}$, the estimator converges to f at the rate $(\log n)^{-2\beta/\alpha}$. As in the generalized ordinary smooth case, this rate is optimal and is the same as the rate derived by Fan (1991, 1993) in the supersmooth case where f_δ^{ft} has no zeros. However this result is less interesting since this rate can already be achieved by the ridge parameter method of Hall and Meister (2007).

Remark 4. *(White Noise Model).* A third application of our general deconvolution estimator is the white noise model in which one observes the stochastic process $Y(x)$, $x \in \mathbb{R}$, driven by the stochastic differential equation

$$dY(x) \;=\; \big[f * f_\delta\big](x)\,dx \;+\; n^{-1/2}\,dW(x)\,,$$

where W denotes a standard Wiener process. The goal is to reconstruct the true signal or image f from the functional observation Y, that is affected by both pointspread effects and random noise. The model has applications in the field of image reconstruction; see Qiu (2005) for an introduction to those topics. We assume that the support of the target function f and the pointspread function f_δ is contained in some intervals $[a, b]$ and $[-R, R]$, respectively, so that we obtain an empirical version of $\xi^{\mathrm{ft}}(t)$ by

$$\widehat{\Phi}_\xi(t) \;=\; \int_{a-R}^{b+R} \exp(itx)\,dY(x)\,,$$

where the integral is to be understood in the Ito sense. Then (4.22) is satisfied due to Ito's lemma; and Theorem 4.1 provides the standard deconvolution rates.

4.2 Consistency of density and regression estimators

Consistency of the density estimator follows from general results rather straightforwardly. In particular, elementary calculations show that (4.22) holds, and thus

consistency of $\widehat{f}_X$ and associated rates of convergence are as described by Theorem 4.1.

For the Nadaraya-Watson estimator, note that our estimator $\widehat{m}$ is obtained by taking

$$\widehat{\Phi}_\xi(t) \;=\; \frac{1}{n}\sum_{j=1}^{n} Y_j \exp(itW_j)$$

in (4.24), and thus (4.22) is satisfied for $\xi = m * f_\delta$, as $f_\delta \in L_2(\mathbb{R})$. Under the smoothness assumptions $m, f_X \in \mathcal{F}^S_{C,\beta}$, Theorem 4.1 can be applied to the estimators $\widehat{m}$ and $\widehat{f}_X$. If the class $\mathcal{F}'_{C,D,\beta}$ contains all (g, f_X) so that $m, f_X \in \mathcal{F}^S_{C,\beta}$, $\max\{\|g\|_\infty, \|f_X\|_\infty\} \le D$, and $\mathcal{I} \subseteq [a,b]$ where $\mathcal{I}$ denotes the support of g and f_X, these results can then be easily combined to prove that

$$\begin{aligned}
&\sup_{(g,f_X)\in\mathcal{F}'_{C,D,\beta}} E\int_{\{f_X(x)\ge\rho\}} \big|\widehat{g}(x) - g(x)\big|^2 dx \\
&\le\; O(1)\cdot\Big(\sup_{f_X\in\mathcal{F}^S_{C,\beta}} E\int_a^b \big|\widehat{f}_X(x) - f_X(x)\big|^2 dx \;+\; \sup_{m\in\mathcal{F}^S_{C,\beta}} E\int_a^b \big|\widehat{m}(x) - m(x)\big|^2 dx\Big),
\end{aligned} \tag{4.27}$$

so that the weighted MISE converges to zero with the rates given in Theorem 4.1.

Consistency of the local polynomial estimator (3.21) is more difficult to obtain. In the next theorem, we derive the convergence rate for the pointwise risk of the estimator (3.21); its proof does not follow straightforwardly from the general results, and is given in Section 7.

Theorem 4.2. *Suppose that* $\|g^{(j)}\|_\infty \le$ *const. for all* $j = 0,\dots,\beta$, *integer* $\beta > 0$, *that* f_X *is a bounded density, that is continuous at some* $x \in (a,b)$, *that the support of* g *and* f_X *are contained in* $[a,b]$, *and that* f_δ *satisfies* (2.2). *Assume that* $0 < \rho < f_X^p(x)\,\det(\boldsymbol{S})$ *where* $\boldsymbol{S} = \big(\int y^{j+k}K(y)dy\big)_{0\le j,k\le p}$, *and let* K *be a symmetric kernel that is a continuous and compactly supported density with* $|(K^{\mathrm{ft}})^{(k)}(t)| \le$ *const.* $\cdot\,|t|^{-\alpha-2}$ *for all* $k = 0,\dots,p$. *Then, for* $h \asymp n^{-2\beta/(2\beta+2\alpha+1)}$, *where* $\beta = p+1$, *we have* $E\big|\widehat{g}_p^{(l)}(x) - g^{(l)}(x)\big|^2 = O\big(n^{-2(\beta-l)/(2\beta+2\alpha+1)}\big)$.

Note that the conditions with respect to the kernel function K in Theorem 4.2

are satisfied for any compactly supported density K that is at least $(\alpha+2)$-fold continuously differentiable on the whole real line.

Theorem 4.2 gives us the individual convergence rates for the local polynomial regression estimator, which can be achieved without the assumption of f_X being β-fold continuously differentiable, unlike the Nadaraya-Watson estimator.

5 A general bandwidth selector

5.1 General cross-validation procedure

As usual for nonparametric estimators, in order to reach the convergence rates established in Theorem 4.1, the order of the underlying bandwidth depends on the smoothness degree β of f, that is unknown in practice. In this section we develop a data-driven cross-validation (CV) bandwidth selection procedure for the estimator (4.24); it can be used for both density deconvolution and the Nadaraya-Watson regression estimator. In the case of the local polynomial estimator of order $p>0$, we can use the SIMEX procedure of Section 5.2. Cross-validation procedures for density deconvolution have been studied for kernel methods by Stefanski and Carroll (1990), Hesse (1999) and, for ridge parameter approaches, by Hall and Meister (2007). For related methods in wavelet deconvolution, see Pensky and Vidakovic (1999).

Unlike error-free contexts, since no direct data coming from the curve of interest are available, CV procedures in deconvolution problems are usually based on estimators of the MISE calculated in its Fourier domain. In our context, we suggest using the rate-efficient upper bound on the MISE derived in the proof of Theorem 4.1, that is,

$$\begin{aligned} R_n(h) =& \frac{c}{\pi n}\int \Big|\sum_{k=0}^{J}\eta_k e^{2itk\pi/\lambda}\Big|^2 \big|K^{\mathrm{ft}}(ht)\big|^2 \big|\big(e^{2it\pi/\lambda}-1\big)^{\nu}/f_\delta^{\mathrm{ft}}(t)\big|^2\,dt \\ &+ \frac{1}{\pi}\int \Big|\sum_{k=0}^{J}\eta_k e^{2itk\pi/\lambda}\Big|^2 \frac{\big|e^{2it\pi/\lambda}-1\big|^{2\nu}}{|f_\delta^{\mathrm{ft}}(t)|^2}\big|\xi^{\mathrm{ft}}(t)\big|^2 K^{\mathrm{ft}}(ht)(K^{\mathrm{ft}}(ht)-2)\,dt \end{aligned}$$

plus a term which does not depend on h. Here, the constant c comes from (4.22),

$\sup_{t\in\mathbb{R}} E|\widehat{\Phi}_\xi(t) - \xi^{ft}(t)|^2 \le c/n$ (see also Section 6.1).

Since this bound is rate-efficient, any choice of h under which R_n is minimized leads to the convergence rates derived in Theorem 4.1. In practice, R_n needs to be estimated since it depends on g^{ft}. We suggest choosing h to minimize

$$\begin{aligned}\widehat{R}_n(h) =& \frac{c}{\pi n}\int \Big|\sum_{k=0}^{J}\eta_k e^{2itk\pi/\lambda}\Big|^2 \big|K^{\mathrm{ft}}(ht)\big|^2 \big|\big(e^{2it\pi/\lambda}-1\big)^{\nu}/f_\delta^{\mathrm{ft}}(t)\big|^2 dt \\ &+ \frac{1}{\pi}\int \Big|\sum_{k=0}^{J}\eta_k e^{2itk\pi/\lambda}\Big|^2 \frac{\big|e^{2it\pi/\lambda}-1\big|^{2\nu}}{|f_\delta^{\mathrm{ft}}(t)|^2}\,\widehat{\Psi}_\xi(t)K^{\mathrm{ft}}(ht)(K^{\mathrm{ft}}(ht)-2)\,dt\,,\end{aligned}$$

where $\widehat{\Psi}_\xi(t)$ is an empirical version of $|\xi^{\mathrm{ft}}(t)|^2$.

In the errors-in-variables regression problem, with the Nadaraya-Watson estimator of Section 3.1, we choose two bandwidths: one for the numerator $\widehat{m}$ and one for the denominator $\widehat{f}_X$. For $\widehat{m}$ we take

$$\widehat{\Psi}_\xi(t) \;=\; \frac{1}{n(n-1)}\sum_{j\neq k} Y_jY_k \exp\big(it(W_j-W_k)\big)\,, \tag{5.28}$$

and for $\widehat{f}_X$ we take $\widehat{\Psi}_\xi(t) = \{n(n-1)\}^{-1}\sum_{j\neq k}\exp\big(it(W_j-W_k)\big)$, which we also use for choosing the bandwidth in the problem of density deconvolution.

To prove asymptotic properties of the data-driven estimator of the regression curve g, take the grid $H = \big\{1/k : k = 1,\dots,\lfloor n^{1/(2\alpha+1)}\rfloor\big\}$ and let

$$\widehat{h} \;=\; \mathrm{argmin}_{h\in H}\;\widehat{R}_n(h)\,. \tag{5.29}$$

For $\mathcal{F}'_{C,D,\beta}$ as in Section 3, let

$$\begin{aligned}\mathcal{F}^A_{C,\beta,c_0,c_1} \;=\; \Big\{(g,f_X)\in\mathcal{F}'_{C,D,\beta} :\; &\int_\omega^\infty \big|\varphi^{\mathrm{ft}}(t)\big|^2 dt \;\ge\; c_0\omega^{-2\beta},\ \|\varphi\|_1\le c_1\,,\\ &\big|\varphi^{\mathrm{ft}}(\omega)\big| \;\le\; c_0^{-1}|\omega|^{-\beta-1/2}\forall\omega\ge 1 \text{ for } \varphi = f_X, m\Big\}\,.\end{aligned}$$

In density deconvolution, the class $\mathcal{F}'_{C,D,\beta}$ can be replaced by $\mathcal{F}^S_{C,\beta}$ in the above definition and we may put $\varphi = g = f_X$. The next theorem establishes that the bandwidth selector (5.29) is rate-adaptive in errors-in-variables regression and density deconvolution, the latter following from the results on f_X. In the sequel, we write $\widehat{f}_h = \widehat{f}$ where h denotes the bandwidth used in the construction of the estimator $\widehat{f}$.

Theorem 5.1. *Consider the Nadaraya-Watson estimator of section for the errors-in-variables regression problem, and suppose that the conditions of Theorem 4.1 are satisfied. Assume that* $E\varepsilon_1^s < \infty$ *for all integer* $s \geq 0$*, and take* $K(x) = (\sin x)/(\pi x)$*. Take* $\widehat{h}$ *as in* (5.29)*. Then for* b *an arbitrary but fixed constant,*

$$\sup_{(f,f_X)\in\mathcal{F}^A_{C,\beta,c_0,c_1}} E\int_a^b \left|\widehat{m}_{\widehat{h}}(x) - m(x)\right|^2 dx \;=\; O\big(n^{-2\beta/(2\beta+2\alpha+1)}\big)\,,$$

$$\sup_{(f,f_X)\in\mathcal{F}^A_{C,\beta,c_0,c_1}} E\int_a^b \left|\widehat{f}_{X;\widehat{h}}(x) - f_X(x)\right|^2 dx \;=\; O\big(n^{-2\beta/(2\beta+2\alpha+1)}\big)\,.$$

5.2 Alternative SIMEX bandwidth selectors

The cross-validation method introduced in Section 5.1 is attractive because it can be applied in a variety of contexts, including density and regression, and our theoretical results show that it is rate-adaptive. However, as our numerical results illustrate, in finite samples the CV method suffers from the usual problems encountered in practice. As an alternative, we suggest using methods based on SIMEX (Simulation-Extrapolation) ideas. SIMEX was introduced by Cook and Stefanski (1994) and Stefanski and Cook (1995) to estimate a parametric regression curve. Here, we use SIMEX to select the bandwidth of a nonparametric curve estimator.

SIMEX is related to the bootstrap: estimate all unknown quantities via samples of data generated from the observations, but differently from the bootstrap method. In SIMEX, the idea is to learn the effect that adding noise has on the target quantities by creating samples (SIMulation step) that contain more and more noise, and extrapolate (EXtrapolation step) this effect back to the original problem. Note that bandwidths selected by such a SIMEX approach rely more on approximation ideas than on rate-adaptivity, but have very good practical performances. Their use can also be justified theoretically, but the proofs are long and require the use of non-standard theoretical arguments; see Delaigle and Hall (2008) and their supplemental material.

SIMEX bandwidth for regression. In the regression case, we use the SIMEX bandwidth proposed by Delaigle and Hall (2008) adapted to our context. The adap-

tation is straightforward: it consists in replacing their K_δ by our K_δ. Therefore we refer to that paper for implementation details, Section 6.2 for numerical performance.

SIMEX bandwidth for density. In the density case, we suggest a SIMEX bandwidth, which can be summarized as follows.

1) If we knew the density f_X, we could select the bandwidth for estimating f_X as $h_X = \text{argmin}_h \int \{\widehat{f}_X(x;h) - f_X(x)\}^2 \, dx$, where $\widehat{f}_X$ is the deconvolution estimator constructed from the contaminated data $W_i = X_i + U_i$.

2) (SIMulation step). Consider the problem with one (respectively, two) additional level(s) of error, where the goal is to estimate f_W (resp. $f_W^\star$) from contaminated data $W_i^\star = W_i + U_i^\star$ (resp. $W_i^{\star\star} = W_i^\star + U_i^{\star\star}$), with $U_i^\star \sim f_\delta$ (resp. $U_i^{\star\star} \sim f_\delta$). If we knew f_W (resp. $f_W^\star$) we could select the bandwidth h_W (resp. $h_W^\star$) for estimating f_W (resp. $f_W^\star$) as $h_W = \text{argmin}_h \int \{\widehat{f}_W(x;h) - f_W(x)\}^2 \, dx$ (resp. $h_W^\star = \text{argmin}_h \int \{\widehat{f}_W^\star(x;h) - f_W^\star(x)\}^2 \, dx$), where $\widehat{f}_W$ (resp. $\widehat{f}_W^\star$) is our deconvolution estimator of f_W (resp. $f_W^\star$) constructed from the contaminated data $W_i^\star$ (resp. $W_i^{\star\star}$).

3) Here f_W and $f_W^\star$ are unknown, but unlike the original problem where we do not have direct data on f_X, we have direct data $W_1, \ldots, W_n \sim f_W$ and $W_1^\star, \ldots, W_n^\star \sim f_W^\star$. Thus we can construct standard (non deconvolution) kernel estimators $\tilde{f}_W$ and $\tilde{f}_W^\star$ of f_W and $f_W^\star$, and then approximate h_W and $h_W^\star$ by $\widehat{h}_W = \text{argmin}_h \int \{\widehat{f}_W(x;h) - \tilde{f}_W(x)\}^2 \, dx$ and $\widehat{h}_W^\star = \text{argmin}_h \int \{\widehat{f}_W^\star(x;h) - \tilde{f}_W^\star(x)\}^2 \, dx$, respectively.

4) (EXtrapolation step). The idea is that $W_i^\star$ measures W_i in the same way as $W_i^{\star\star}$ measures $W_i^\star$ and W_i measures X_i. Thus we can expect that h_W approximates h_X in roughly the same way as $h_W^\star$ approximates h_W. With this motivation, we take the SIMEX bandwidth to be $\widehat{h}_X = \widehat{h}_W^2 / \widehat{h}_W^\star$.

5) To avoid strong dependence on the particular samples generated, we repeat step 2) B times to generate B resamples; we then replace the integrated squared errors of step 3) by the average of the corresponding B integrated squared errors.

6 Finite sample results

6.1 Details of implementation of the method

In the density case, where the estimator is given by (2.12), to apply the CV method in practice we can take $c = 1$. In the regression case, where we use the estimator (3.16), we apply the method to $\widehat{f}_X$ with $c = 1$ to find a bandwidth h_1, say, then apply the method to $\widehat{m}$ with $\widehat{c} = n^{-1}\sum_{j=1}^{n} Y_j^2$ to find the bandwidth h_2. As usual in CV procedures, in case of multiple minima we take the second smallest local minimum. To select ρ, we employ the method used by Delaigle and Hall (2008) at page 282, but we replace their K_δ by our K_δ.

The lower bound a of the left endpoint L of the support of f need not be very close to L, and it suffices to have a rough bound. In many practical cases, the experimenter has an idea of the range of X, but otherwise a can be selected from the data. For example, if f_δ is compactly supported we can take $\widehat{a} = \min\{W_1, \ldots, W_n\}$ as an empirical version of a. In more general contexts we can take the $\widehat{a} = \widehat{L} - |\widehat{L}|$, where $\widehat{L}$ is estimated by methods similar to those of Delaigle and Gijbels (2006) or Meister (2006).

6.2 Simulation results

We applied our method to estimate various densities and regression curves. In each case we took the error δ to be a convolution of a Laplace and the uniform $U[-1, 1]$ density (Lap∗Uni), or a convolution of two uniforms (Uni∗Uni), such that $\mathrm{Var}(\delta)/\mathrm{Var}(X) = 0.1$ or 0.25. The three densities we considered were (i) $f_X(x) = \beta_{4,4}((x + 7)/15)/15$, (ii) $f_X(x) = 0.5\beta_{9,4}((x + 7)/15)/15 + 0.5\beta_{9,9}((x + 7)/9)/9$, (iii) $f_X(x) = 0.5\phi_{-3,1}(x) + 0.5\phi_{2,1}(x)$, where $\beta_{a,b}(x)$ denotes the density of a beta random variable with parameters a and b and $\phi_{\mu,\sigma}(x)$ denotes the density of a normal random variable with mean μ and variance σ^2. In case (iii), f_X does not have a finite left endpoint but, as we will see, our estimator worked well in this case too. For regression functions we took (iv) $g(x) = 2\sin(x)\exp(-x^2/10)$, $f_X(x) = \beta_{4,4}((x+7)/15)/15$,

Table 1: Comparison of the various methods — (2.12) with h_{CV} (CV), (2.12) with h_{SIMEX} (SIM) and $\widehat{f}_{\mathrm{naive}}$ — for estimation of density (i): values of the quantiles $q_{0.1}$, $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, and $q_{0.9}$, of $10^3 \times ISE$, when $\delta \sim$ Lap$*$Uni.

	$n = 100$		$n = 250$		$n = 500$	
$\widehat{f}_X$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$
			$\mathrm{Var}\, U = 10\%\, \mathrm{Var}\, X$			
SIM	2.23	[1.37,3.80]–[0.78,5.29]	1.29	[0.77,2.17]–[0.56,2.86]	0.81	[0.56,1.21]–[0.40,1.55]
CV	1.70	[0.87,3.24]–[0.50,6.10]	0.66	[0.34,1.40]–[0.18,3.44]	0.34	[0.18,0.74]–[0.10,1.42]
$\widehat{f}_{\mathrm{naive}}$	2.33	[1.53,4.15]–[0.81,5.49]	1.56	[0.93,2.18]–[0.66,2.94]	0.99	[0.67,1.37]–[0.42,1.70]
			$\mathrm{Var}\, U = 25\%\, \mathrm{Var}\, X$			
SIM	3.17	[1.89,4.86]–[0.91,6.72]	1.72	[1.12,2.61]–[0.74,3.93]	1.46	[1.01,2.16]–[0.78,2.84]
CV	2.52	[1.31,3.95]–[0.67,5.60]	0.89	[0.46,1.59]–[0.25,2.72]	0.46	[0.22,0.89]–[0.13,1.75]
$\widehat{f}_{\mathrm{naive}}$	3.51	[2.30,5.35]–[1.34,7.20]	2.53	[1.69,3.33]–[1.11,4.43]	1.91	[1.43,2.39]–[1.01,3.00]

and $\epsilon \sim N(0, 0.2)$; (v) $g(x) = \exp(-0.25x^2)$, $f_X(x) = \beta_{4,4}((x+7)/15)/15$, and $\epsilon \sim N(0, 0.05)$.

In each case, we generated 200 contaminated samples, applied our methods with the CV bandwidth selector h_{CV} and the (infinite order) sinc kernel, or the SIMEX bandwidth h_{SIMEX} with the (second order) kernel $K^{\mathrm{ft}}(t) = (1-t^2)^3 \cdot 1_{[-1,1]}(t)$. We also calculated the naive estimator $\widehat{f}_{\mathrm{naive}}$, that ignores the error present in the data. In the density case, $\widehat{f}_{\mathrm{naive}}$ is the usual kernel density estimator (with standard normal kernel and plug-in bandwidth) applied to the contaminated data; in the regression case, $\widehat{f}_{\mathrm{naive}}$ is the usual Nadaraya-Watson estimator (with standard normal kernel and cross-validation bandwidth) applied to the contaminated data. Thus $\widehat{f}_{\mathrm{naive}}$ is not a consistent estimator of f, but rather of its contaminated version, and it illustrates the importance of taking the error into account when estimating f. For each method we calculated the corresponding 200 values of ISE $= \int(\widehat{f} - f)^2$, where $\widehat{f}$ denotes the estimator of f. Figures 1-3 present three curves q_1, q_2, and q_3 corresponding to the first, second, and third quartiles of these 200 ISE's. In each graph the target curve is shown as a solid line.

Table 2: Comparison of the various methods — (2.12) with h_{CV} (CV), (2.12) with h_{SIMEX} (SIM) and $\widehat{f}_{\mathrm{naive}}$ — for estimation of density (ii): values of the quantiles $q_{0.1}$, $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, and $q_{0.9}$, of $10^3 \times ISE$, when $\delta \sim$ Lap$*$Uni.

	$n = 100$		$n = 250$		$n = 500$	
$\widehat{f}_X$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$	$q_{0.5}$	$[q_{0.25}, q_{0.75}]$–$[q_{0.1}, q_{0.9}]$
			$\mathrm{Var}\, U = 10\% \,\mathrm{Var}\, X$			
SIM	9.04	[5.85,12.2]–[4.49,15.5]	5.99	[4.11,7.93]–[2.98,9.79]	4.25	[3.16,5.38]–[2.33,7.25]
CV	9.74	[7.46,15.4]–[5.86,24.4]	5.66	[4.36,7.61]–[3.61,12.2]	3.95	[3.07,4.97]–[2.23,9.04]
$\widehat{f}_{\mathrm{naive}}$	11.5	[8.67,13.9]–[6.82,16.4]	8.66	[6.77,10.2]–[5.38,11.7]	6.78	[5.56,8.14]–[4.78,9.63]
			$\mathrm{Var}\, U = 25\% \,\mathrm{Var}\, X$			
SIM	13.8	[9.92,18.2]–[6.65,22.4]	12.9	[9.95,15.4]–[7.92,19.2]	6.16	[4.92,8.69]–[3.48,11.6]
CV	14.3	[10.9,26.5]–[8.38,76.8]	8.55	[6.59,11.8]–[5.05,21.1]	6.86	[4.94,9.46]–[4.01,19.7]
$\widehat{f}_{\mathrm{naive}}$	18.2	[15.6,21.2]–[12.5,23.5]	15.0	[12.8,17.0]–[10.6,19.6]	12.7	[11.2,14.9]–[10.1,17.2]

Tables 1 and 2 give the quantiles $0.1, 0.25, 0.5, 0.75$, and 0.9 of these ISE's when estimating densities (i) and (ii) with $\mathrm{Var}(\delta)/\,\mathrm{Var}(X) = 0.1$ or 0.25, and with sample sizes $n = 100$ to 500. We compare the results of our consistent estimator using h_{CV} or h_{SIMEX}, and $\widehat{f}_{\mathrm{naive}}$. Figure 1 shows some of the estimated curves for density (ii). These tables and graphs illustrate familiar properties of kernel smoothing. For example, the results improve as the sample size increases or as the error variance decreases. They also show that h_{CV} combined with a sinc kernel manages to recover the peaks of the target curve better than h_{SIMEX} with a finite order kernel, but is often more variable and more wiggly in the tails. Note that, as usual, this wiggliness is caused by both the sinc kernel and the fact that CV tends to select too small bandwidths. In each case, $\widehat{f}_{\mathrm{naive}}$ was strongly biased.

Figure 2 shows the results for density (iii). We took $\delta \sim$ Lap$*$Uni, $\delta \sim$ Uni$*$Uni, or $\delta \sim N(0, \mathrm{Var}(\delta))$. In the latter case, when applying the estimator we pretended that δ was Uni$*$Uni. Of course, the true error was normal, but the idea was to see if the method is relatively robust against error misspecification. In each case, we compare our estimator using a SIMEX bandwidth, with $\widehat{f}_{\mathrm{naive}}$. The graphs show

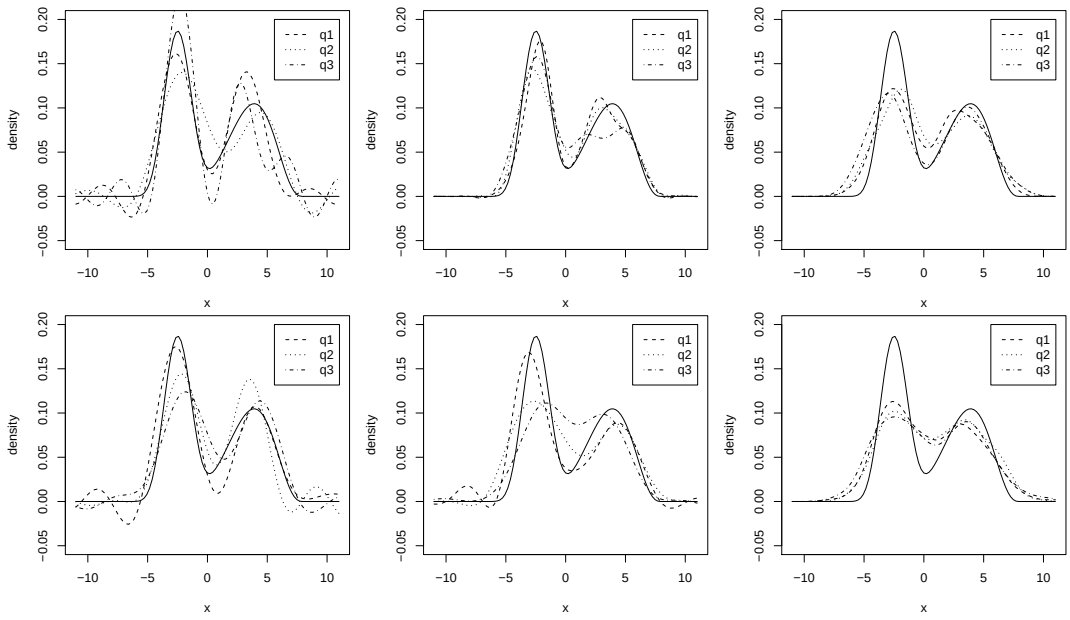

Figure 1: Estimates of f_X for curve (ii) when $\delta \sim \text{Lap}*\text{Uni}$, $n = 250$, and $\text{Var}(\delta) = 0.10\,\text{Var}(X)$ (top) or $\text{Var}(\delta) = 0.25\,\text{Var}(X)$ (bottom), using the estimator (2.12) with h_{CV} (left) h_{SIMEX} (middle), or using $\widehat{f}_{\text{naive}}$ (right).

clearly that even if the error is misspecified, taking the error into account produces estimators that are considerably less biased than $\widehat{f}_{\text{naive}}$.

In the regression case we found that the CV approach had similar drawbacks and h_{SIMEX} often gave better results. For example, in Figure 3 we show the quartile curves of the estimators of curve (iv), using (3.16) when $n = 250$ and the bandwidth was h_{CV} or h_{SIMEX}. Although, as in the density case, the CV method with sinc kernel recovered the peaks better than the method (in this case SIMEX) that used a finite order kernel, it also produced unattractive wiggly curves away from the peaks. $\widehat{f}_{\text{naive}}$ clearly targeted the wrong curve.

Figure 4 shows the quartile curves for estimators of curve (v) from samples of size $n = 100$ to $n = 500$. We compare our estimator using h_{SIMEX}, with $\widehat{f}_{\text{naive}}$. Again, we see that $\widehat{f}_{\text{naive}}$ is strongly biased and clearly outperformed by our method. In this case too, we found that using h_{CV} helped recover the peaks better, but overall gave more variable estimators than h_{SIMEX}.

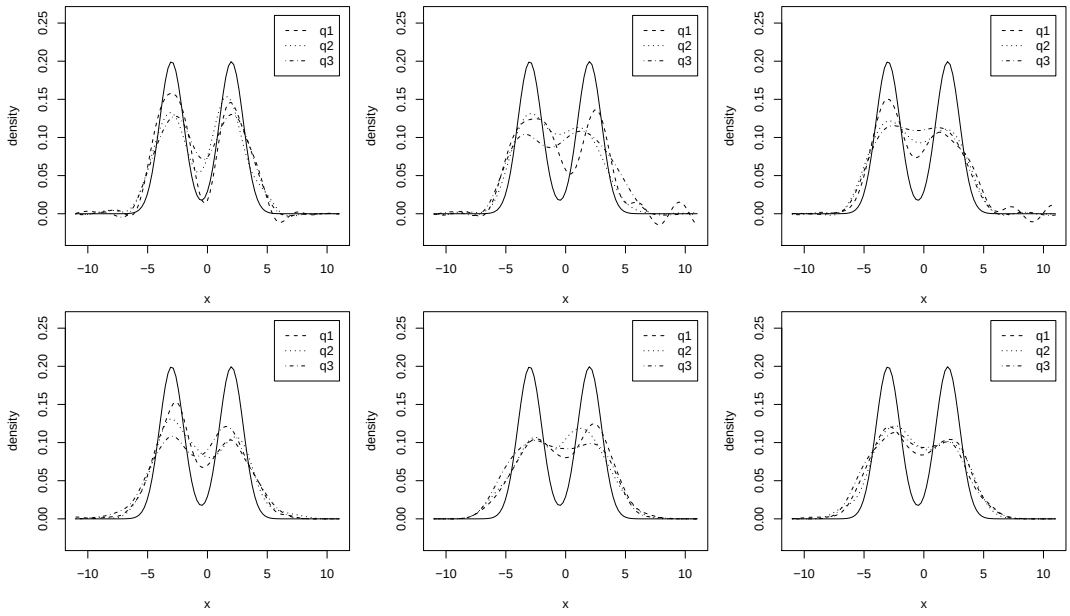

Figure 2: Estimates of f_X for curve (iii) when $\mathrm{Var}(\delta) = 0.25\,\mathrm{Var}(X)$ and $n = 250$, in the case where $\delta \sim$ Lap$*$Uni (left), $\delta \sim$ Uni$*$Uni (center), or $\delta \sim N(0, \mathrm{Var}(\delta))$ but we pretend it is Uni$*$Uni, using the estimator (2.12) with h_{SIMEX} (top), or using $\widehat{f}_{\mathrm{naive}}$ (bottom).

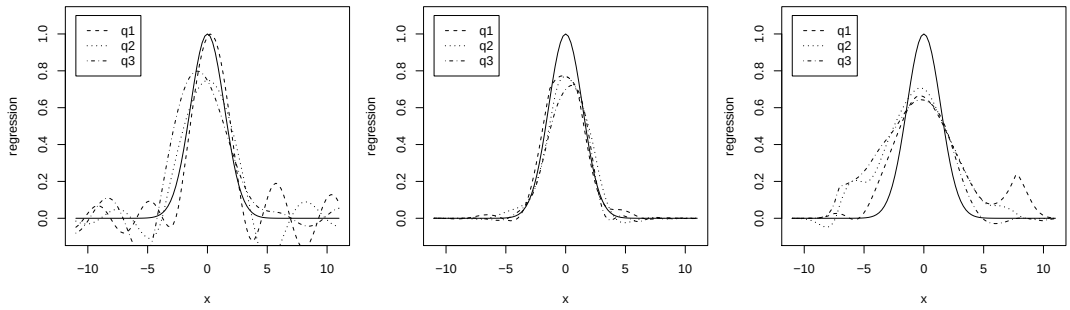

Figure 3: Estimates of g for curve (iv) when $\delta \sim$ Lap$*$Uni, $\mathrm{Var}(\delta) = 0.25\,\mathrm{Var}(X)$, and $n = 250$, using the estimator (3.16) with h_{CV} (left) or h_{SIMEX} (middle), or using $\widehat{f}_{\mathrm{naive}}$ (right).

6.3 A data application

In an unpublished work by Sun, Wang and Woodroofe (2009), the authors suggest that, when δ is very smooth (e.g. normal), better numerical results could be ob-

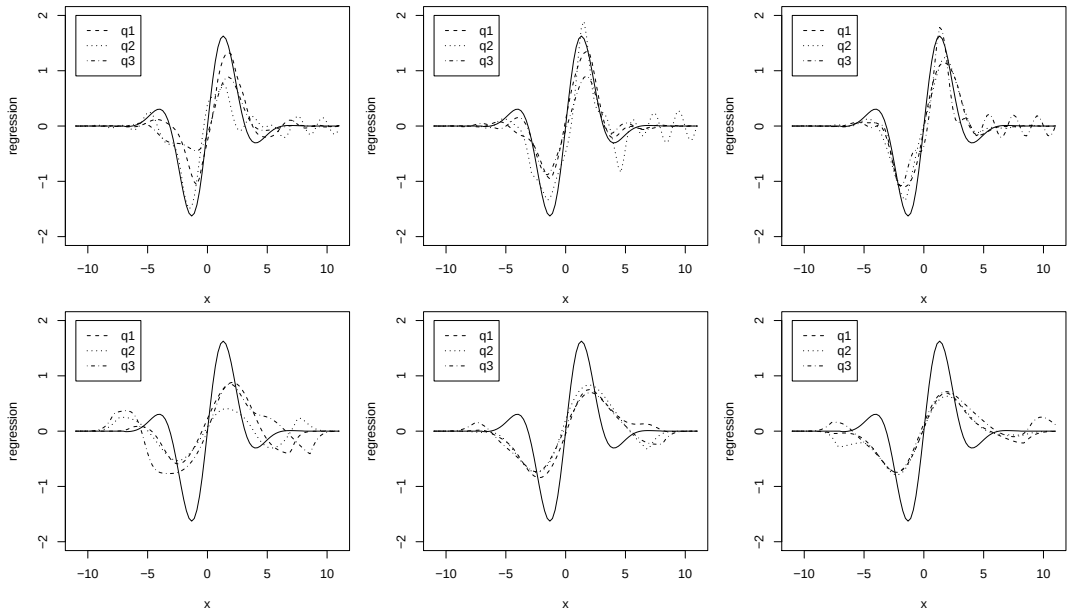

Figure 4: Estimates of g for curve (v), when $\delta \sim \text{Lap}*\text{Uni}$, $\text{Var}(\delta) = 0.25\,\text{Var}(X)$ and $n = 100$ (left), $n = 250$ (middle) and $n = 500$ (right), using the estimator (3.16) with the SIMEX method (top), or using $\widehat{f}_{\text{naive}}$ (bottom).

tained by approximating its distribution by that of a sum of a few uniform random variables (i.e., pretend that δ is a sum of uniforms even though it is not). Approximating a normal by a sum of uniforms can be justified by the Central Limit Theorem; and simple algorithms generate a normal variable by taking the sum of a few i.i.d. uniform random variables; see for example Ahrens and Dieter (1972); see also our illustration in Figure 2. More general approximations could also be justified via small error variance approximation, as in Carroll and Hall (2004) and Delaigle (2008). An interesting application of Sun and Wang's suggestion is that when we do not know what f_δ is, we could apply deconvolution pretending that δ was a convolution of a small number of uniforms. This approach is attractive because such a convolution is not very smooth (in the deconvolution terminology), and thus is less likely to cause dramatic approximation errors. See the related work by Carroll and Hall (2004) and Delaigle (2008).

Our method can be used to deconvolve a sum of uniform random variables, and

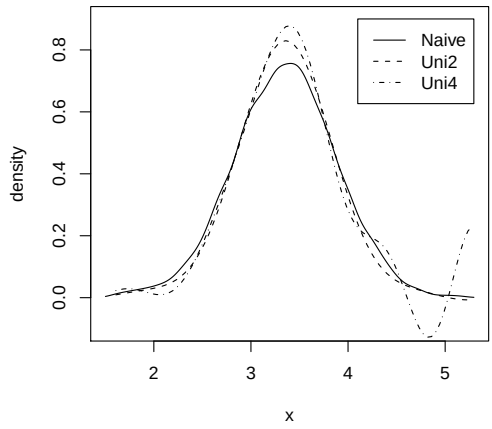

Figure 5: Estimator of the density f_X for the Nhanes data, using $\widehat{f}_{\text{naive}}$ or using our estimator with $\widehat{h}_{\text{SIMEX}}$, assuming that δ is a convolution of two uniforms (Uni2) or a convolution of four uniforms (Uni4).

we applied it to data from the second National Health And Nutrition Examination Survey (NHANES) study. The goal was to estimate the density of the long-term log daily saturated fat intake based on a sample of size 4708, consisting of women aged between 25 and 50 years. We employed the transformation used by Carroll et al. (2006), Chapter 4, for the first NHANES, that is log(5+saturated fat). In this study, the long-term intake was approximated by a 24 hour recall, causing large measurement errors. The variance of δ could be approximated from the recalls, giving $\text{Var}(\delta) \approx 0.5\,\text{Var}(X)$. However, f_δ itself is unknown and we applied our estimator assuming that δ was a convolution of two or of four identically distributed uniforms, each time with $\text{Var}(\delta)$ as above. We compared the results with $\widehat{f}_{\text{naive}}$.

The three estimators of the density are plotted in Figure 5. As is usual in deconvolution, especially with this high level of noise, $\widehat{f}_{\text{naive}}$ seems to oversmooth the data. The estimators using the two convolutions of uniforms seem to correct for some error and give close results, although assuming a convolution of four uniforms gives an estimator that seems visually less attractive – it is well known that deconvolving smoother errors is more difficult in practice, especially if the noise level is high, see Stefanski and Carroll (1990).

7 Proofs

Proof of Theorem 4.1: By (2.10) and Fourier inversion of $p \in L_2(\mathbb{R})$, we have

$$f(x) \,=\, \frac{1}{2\pi}\int \exp(-itx)\Big(\sum_{k=0}^{J}\eta_k \exp(2k\pi it/\lambda)\Big)p^{\mathrm{ft}}(t)dt\,,$$

for almost all $x \in [a,b]$. From (2.5) we deduce that, for such x,

$$f(x) \,=\, \frac{1}{2\pi}\int \exp(-itx)\Big(\sum_{k=0}^{J}\eta_k \exp(2k\pi it/\lambda)\Big)\big(\exp(2it\pi/\lambda)-1\big)^{\nu} f^{\mathrm{ft}}(t)dt. \quad (7.30)$$

Therefore

$$\begin{aligned} E\int_a^b \big|\widehat{f}(x)-f(x)\big|^2 dx \,\le\, E\int \Big|\frac{1}{2\pi}\int e^{-itx}\Big(\sum_{k=0}^{J}\eta_k e^{2k\pi it/\lambda}\Big)\big(e^{2it\pi/\lambda}-1\big)^{\nu} \\ \times\big[K^{\mathrm{ft}}(th)\widehat{\Phi}_\xi(t)/f_\delta^{\mathrm{ft}}(t)-f^{\mathrm{ft}}(t)\big]dt\Big|^2 dx \\ =\frac{1}{2\pi}\int \Big|\sum_{k=0}^{J}\eta_k e^{2k\pi it/\lambda}\Big|^2 \big|e^{2it\pi/\lambda}-1\big|^{2\nu} E\big|K^{\mathrm{ft}}(th)\widehat{\Phi}_\xi(t)/f_\delta^{\mathrm{ft}}(t)-f^{\mathrm{ft}}(t)\big|^2 dt\,, \quad (7.31)\end{aligned}$$

where the inequality in the first step comes from the extension of the integration domain from $[a,b]$ to $\mathbb{R}$. In the second step, we have used Parseval's identity. Furthermore, we have

$$\begin{aligned} &\big|e^{2it\pi/\lambda}-1\big|^{2\nu} E\big|K^{\mathrm{ft}}(th)\widehat{\Phi}_\xi(t)/f_\delta^{\mathrm{ft}}(t)-f^{\mathrm{ft}}(t)\big|^2 \\ &\quad\le\, 2\big|e^{2it\pi/\lambda}-1\big|^{2\nu}\big|K^{\mathrm{ft}}(th)-1\big|^2\big|f^{\mathrm{ft}}(t)\big|^2 \\ &\qquad\quad +\, 2\big|K^{\mathrm{ft}}(th)\big|^2\big|e^{2it\pi/\lambda}-1\big|^{2\nu}\big|f_\delta^{\mathrm{ft}}(t)\big|^{-2}E\big|\widehat{\Phi}_\xi(t)-\xi^{\mathrm{ft}}(t)\big|^2\,. \quad (7.32)\end{aligned}$$

Now from (2.2), we have, for almost all $t \in \mathbb{R}$,

$$\big|\exp(2it\pi/\lambda)-1\big|^{2\nu}\big|f_\delta^{\mathrm{ft}}(t)\big|^{-2} \,\le\, \text{const.}\cdot(1+|t|)^{2\alpha}\,.$$

where here and below, const. denotes a generic positive constant which may take different values at different lines. Also, we employ (4.22) so that

$$E\big|\exp(2it\pi/\lambda)-1\big|^{2\nu}\big|K^{\mathrm{ft}}(th)\widehat{\Phi}_\xi(t)/f_\delta^{\mathrm{ft}}(t)-f^{\mathrm{ft}}(t)\big|^2$$

$$\leq \text{ const.} \cdot \left(\left| K^{\text{ft}}(th) - 1 \right|^2 \left| f^{\text{ft}}(t) \right|^2 + \left| K^{\text{ft}}(th) \right|^2 (1 + |t|)^{2\alpha} / n \right).$$

Inserting that inequality into the integral in (7.31), we get the upper bound

$$\text{const.} \cdot \left(\int \left| K^{\text{ft}}(th) - 1 \right|^2 \left| f^{\text{ft}}(t) \right|^2 + \frac{1}{n} \int \left| K^{\text{ft}}(th) \right|^2 (1 + |t|)^{2\alpha} dt \right)$$

on the MISE where const. does not depend on f. Applying the Sobolev condition $f \in \mathcal{F}^S_{C,\beta}$ and the conditions imposed on the kernel K, we obtain $O\big(h^{2\beta} + n^{-1}h^{-1-2\alpha}\big)$ as a uniform upper bound on the MISE. Inserting the proposed bandwidth h completes the proof of the theorem. ■

Proof of Theorem 4.2: We define

$$\begin{aligned} \big[B_q f_X\big](x) &= \int y^q K(y) f_X(x + yh) dy \,, \\ \big[B_q m\big](x) &= \int y^q K(y) \Big(\sum_{l=0}^{p} (yh)^l g^{(l)}(x)/l! \Big) f_X(x + yh) dy \,, \end{aligned}$$

$\mathbf{T_B} = \big(\big[B_0 m\big](x), \ldots, \big[B_p m\big](x) \big)^T$, $\mathbf{F_B} = \big(f(x), hg'(x), \ldots, h^p g^{(p)}(x)/p! \big)^T$, and the matrix $\mathbf{S_B} = \big\{ \big[B_{j+k} f_X\big](x) \big\}_{0 \leq j,k \leq p}$. Since $\mathbf{F_B} = \mathbf{S_B}^{-1} \mathbf{S_B} \mathbf{F_B} = \mathbf{S_B}^{-1} \mathbf{T_B}$, we have $\mathbf{S_B F_B} = \mathbf{T_B}$. Moreover, as $h \to 0$, the matrix $\mathbf{S_B}$ converges to $f_X(x) \cdot \mathbf{S}$, with $\mathbf{S}$ defined in the theorem, for almost all $x \in [a, b]$. Hence, we have

$$\det(\mathbf{S_B}) \to f_X^p(x) \cdot \det(\mathbf{S}) > \rho > 0 \,,$$

so that $\mathbf{S_B}$ is invertible and satisfies $\det(\mathbf{S_B}) > \rho$ for n sufficiently large. By Cramer's rule, we derive that $g^{(l)}(x) = l! h^{-l} \det(\mathbf{A_{B,l}}) / \det(\mathbf{S_B})$, for any $l = 0, \ldots, p$ where $\mathbf{A_{B,l}}$ is the matrix that has its lth column equal to $\mathbf{T_B}$ and all other columns equal to those of $\mathbf{S_B}$. Elementary calculations yield

$$\big| \widehat{g}_p^{(l)}(x) - g^{(l)}(x) \big| \leq l! h^{-l} \rho^{-1} \big(\big| \det(\widehat{\mathbf{A}}_{n,l}) - \det(\mathbf{A_{B,l}}) \big| + O(h^l) \cdot \big| \det(\widehat{\mathbf{S}}_n) - \det(\mathbf{S_B}) \big| \big) \tag{7.33}$$

for n large enough.

Since we restrict our consideration to $x \in (a, b)$, we have $x + yh \in [a, b]$ for n large enough, and thus it follows from (7.30) that $E\widehat{S}_{n,k}(x) = \big[B_k f_X\big](x)$ and

$\big|E\widehat{T}_{n,k}(x) - \big[B_k m\big](x)\big| \;=\; O\big(h^{\beta}\big)$, by Taylor approximation of f (remember that $p = \beta - 1$).

For the distance between the determinants of two $(p+1)\times(p+1)$-matrices $\mathbf{X}$ and $\mathbf{Y}$, we obtain, by the definition of the determinant, that

$$|\det(\mathbf{X}) - \det(\mathbf{Y})| \;\le\; C_p \cdot \max_{\sigma\in S_p} \max_{A\subseteq\{0,\ldots,p\},A\neq\emptyset} \prod_{j\in A} |X_{j,\sigma(j)}| \prod_{j\notin A} |Y_{j,\sigma(j)} - X_{j,\sigma(j)}|\,, \quad (7.34)$$

with a constant $C_p > 0$, where S_p denotes the collection of all permutations of $\{0,\ldots,p\}$, $(\sigma(0),\ldots,\sigma(p))$ is a permutation of $\{0,\ldots,p\}$, and $X_{j,k}$, $Y_{j,k}$, $j,k = 0,\ldots,p$, denote the components of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

We obtain that $\det(E\widehat{\mathbf{S}}_n) = \det(\mathbf{S_B})$ and, by (7.34), that $\big|\det\big(E\widehat{\mathbf{A}}_{n,l}\big) - \det\big(\mathbf{A_{B,l}}\big)\big| = O\big(h^{\beta}\big)$, where the expectation of a matrix is equal to the matrix consisting of its expected components.

In view of (7.34), since there exists a uniform upper bound on all components of $E\widehat{\mathbf{A}}_{n,l}$ and $E\widehat{\mathbf{S}}_n$, it remains to be shown that, for any coefficients $k_0 \in \{0,\ldots,p\}$, $k_1,\ldots,k_p \in \{0,\ldots,2p\}$, and any set $L \subseteq \{1,\ldots,p\}$, we have

$$E\Big[|\widehat{T}_{n,k_0} - E\widehat{T}_{n,k_0}|^2 \prod_{l\in L} |\widehat{S}_{n,k_l} - E\widehat{S}_{n,k_l}|^2\Big] \;=\; O\big(n^{-1}h^{-1-2\alpha}\big)\,, \quad (7.35)$$

$$E\Big[|\widehat{S}_{n,k_0} - E\widehat{S}_{n,k_0}|^2 \prod_{l\in L} |\widehat{S}_{n,k_l} - E\widehat{S}_{n,k_l}|^2\Big] \;=\; O\big(n^{-1}h^{-1-2\alpha}\big)\,. \quad (7.36)$$

The left side of these expressions can be written as

$$n^{-2(\#L+1)} \sum_{j_0=1}^{n}\sum_{j_0'=1}^{n} \cdots \sum_{j_{\#L}=1}^{n}\sum_{j_{\#L}'=1}^{n} E\big(Z_{p,j_0,k_0} Z_{p,j_0',k_0'} \prod_{l\in L} Z_{0,j_l,k_l} Z_{0,j_l',k_l'}\big)\,,$$

where

$$Z_{q,j,k} \;=\; Y_j^q \frac{1}{h} K_{\delta,k_l}\Big(\frac{x - W_j}{h}\Big) \;-\; EY_j^q \frac{1}{h} K_{\delta,k_l}\Big(\frac{x - W_j}{h}\Big)\,.$$

Taking into account that the $Z_{q,j,k}$ are centered, we obtain, as in the proof of Lemma 5.6 of Delaigle and Gijbels (2006), that

$$n^{-2(\#L+1)} \sum_{i=0}^{\#L+1} n^i h^{i-(2\alpha+2)(\#L+1)} \;\le\; O\big(n^{-1-\#L}h^{-(2\alpha+1)(\#L+1)}\big) \;\le\; O\big(n^{-1}h^{-(2\alpha+1)}\big),$$

which proves (7.35) and (7.36).

Applying (7.35) and (7.36) to the squared version of (7.33) gives us the desired rate-efficient upper bound. ■

Proof of Theorem 5.1: We restrict to rate-adaptivity of $\widehat{m}_{\widehat{h}}$, as that of $\widehat{f}_{X;\widehat{h}}$ is included as a special case by taking $Y_j \equiv 1$ a.s..

Let $Q(h) = n^{-1}h^{-1-2\alpha} + h^{2\beta}$, which corresponds to the MISE of estimator (4.24) up to a positive constant factor. We write $h^* \in H$ for the bandwidth that minimizes $R_n(h)$, and hence

$$M_n(h) \;=\; R_n(h) \;+\; \frac{1}{\pi}\int \Big|\sum_{k=0}^{J} \eta_k \exp(2itk\pi/\lambda)\Big|^2 \big|p^{\mathrm{ft}}(t)\big|^2 dt$$

under the constraint $h \in H$. Note that $M_n(h)$ is the rate-efficient upper bound derived in the proof of Theorem 4.1 and that the ratio $M_n(h)/Q(h)$ is bounded above and below by uniform positive constants. Moreover, the set $\{1/h \,:\, h \in H\}$ is sufficiently dense in the interval $[1, n^{1/(1+2\alpha)}]$ so that there exists $h \in H$, $h \asymp n^{1/(1+2\alpha+2\beta)}$. Hence, we have $Q(h^*) \asymp n^{-2\beta/(1+2\beta+2\alpha)}$ and $\widehat{m}_{h^*}$ reaches the optimal convergence rate. Thus,

$$\begin{aligned} E\int_a^b |\widehat{m}_{\widehat{h}}(x) - m(x)|^2 dx \;&=\; \sum_{k=1}^{\lfloor n^{1/(1+2\alpha)}\rfloor} E\big[\,1_{\{\widehat{h}=1/k\}} \cdot \|\widehat{m}_{1/k} - m\|_2^2\big] \\ &\leq\; \text{const.} \cdot \sum_{k=1}^{\lfloor n^{1/(1+2\alpha)}\rfloor} [P(\widehat{h} = 1/k)]^{1/2} \cdot Q(1/k)\,, \qquad (7.37) \end{aligned}$$

by the Cauchy-Schwarz inequality, and where we have used the fact that $E\varepsilon_1^4 < \infty$ and $m \in \mathcal{F}^A_{\beta,C,c_0,c_1}$ imply $\big(E\|\widehat{m}_{1/k} - m\|_2^4\big)^{1/2} \;\leq\; O\big(Q(1/k)\big)$. Furthermore, we have $R_n(\widehat{h}) - R_n(h^*) \;\geq\; C \cdot Q(\widehat{h}) \;-\; D \cdot Q(h^*)$ a.s., for some appropriate constants $C, D > 0$. On the other hand, we have $\widehat{R}_n(\widehat{h}) - \widehat{R}_n(h^*) \;\leq\; 0$ a.s., by definition. From there, we conclude that for any random variable S_n which does not depend on h, one of

$$\begin{aligned} \big|\widehat{R}_n(\widehat{h}) - R_n(\widehat{h}) - S_n\big| \;&\geq\; (C/2) \cdot Q(\widehat{h}) \;-\; (D/2) \cdot Q(h^*)\,, \\ \big|\widehat{R}_n(h^*) - R_n(h^*) - S_n\big| \;&\geq\; (C/2) \cdot Q(\widehat{h}) \;-\; (D/2) \cdot Q(h^*)\,, \end{aligned}$$

holds a.s.

We can show that under appropriate selection of S_n we have

$$\sup_{m\in\mathcal{F}^A_{\beta,C,c_0,c_1}} E\big|\widehat{R}_n(h) - R_n(h) - S_n\big|^{2s} \;\le\; D_s\cdot n^{-\gamma s}\cdot Q^{2s}(h)\,, \tag{7.38}$$

for all $h \in H$ where $\gamma > 0$, D_s are some constants and $s \ge 1$ is an integer. The proof of (7.38), which is given in the long version of this paper, leans on partitionning techniques and some combinational methods; therein, the condition $m, f_\delta \in L_2(\mathbb{R})$ is essential.

We put $K = \big\{k = 1,\ldots,\lfloor n^{1/(1+2\alpha)}\rfloor \,:\, Q(1/k) \;\ge\; (2D/C)\cdot Q(h^*)\big\}$. By Markov's inequality, we conclude from (7.37) that

$$\begin{aligned} E\|\widehat{m}_{\widehat{h}} - m\|_2^2 \;&\le\; \text{const.}\cdot\Big(Q(h^*) \;+\; \sum_{k\in K} Q^{1-s}(1/k)\cdot\big(E\big|\widehat{R}_n(1/k) - R_n(1/k) - S_n\big|^{2s} \\ &\qquad\qquad + E\big|\widehat{R}_n(h^*) - R_n(h^*) - S_n\big|^{2s}\big)^{1/2}\Big) \\ &\le\; \text{const.}\cdot\big(Q(h^*) \;+\; n^{1/(1+2\alpha)-\gamma s/2}\big)\,, \end{aligned}$$

where we used (7.38) in the last step, and where all the constants are independent of m. Taking $s > 0$ sufficiently large, the proof is completed. ■

Acknowledgments

Delaigle's research was supported by a grant and a fellowship from the Australian Research Council.

References

Ahrens, J. H. and Dieter, U. (1972). Computer methods for sampling from the exponential and normal distributions. *Cummications of the ACM* **15**, 873–882.

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. B* **66**, 31–46.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, second edition. Chapman and Hall CRC Press, Boca Raton.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statist. Sinica* **18**, 1025–1045.

Delaigle, A., Fan, J., and Carroll, R.J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104**, 348–359.

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

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

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.* **17**, 235–239.

Dosso, S.E., Brooke, G.H., Kilistoff, S.J., Sotirin, B.J., McDonald, V.K, Fallat, M.R., and Collison, N.E. (1998). High-Precision Array Element Localization for Vertical Line Arrays in the Arctic Ocean. *IEEE J. of Oceanic Engineering*, **23**, 365–379.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *J. Amer. Statist. Assoc.* **92**, 526–535.

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

Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.* **21**, 600–610.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, Boca Raton.

Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

Groeneboom, P. and Jongbloed, G. (2003). Density estimation in the uniform deconvolution model. *Statist. Neerlandica* **57**, 136–157.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35**, 1535-1558.

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

Knyupfer, A.P. (1966). Static errors of analog-digital converters. *Measurement Techniques* **10**, 37–41.

Meister, A. (2006). Support estimation via moment estimation in presence of noise. *Statistics* **40**, 259–275.

Meister, A. (2008). Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions. *Inverse Problems* **24**, 015003 (14 pages).

Neumann, M.H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparam. Stat.* **7**, 307–330.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27**, 2033–2053.

Qiu, P.. (2005). *Image Processing and Jump Regression Analysis.* Wiley, New York.

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

Sun, J., Morrison, H., Harding, P., and Woodroofe, M. (2002). Density and Mixture Estimation from Data with Measurement Errors. *Technical report.*

Sun, J., Wang, X. and Woodroofe, M. (2009). Non-Fourier deconvolution density estimation in measurement error problems. *Manuscript.*