Parametrically assisted nonparametric estimation of a density in the deconvolution problem

Aurore Delaigle and Peter Hall
Department of Mathematics and Statistics, University of Melbourne, Australia

Abstract: Nonparametric estimation of a density from contaminated data is a difficult problem, for which convergence rates are notoriously slow. We introduce parametrically assisted nonparametric estimators which can dramatically improve on performance of standard nonparametric estimators when the assumed model is close to the true density, without degrading much the quality of purely nonparametric estimators in other cases. We establish optimal convergence rates for our problem, and discuss estimators that attain these rates. The very good numerical properties of the methods are illustrated via a simulation study.

Key words: bandwidth, kernel density estimator, measurement errors, nonparametric estimation, parametric estimation.

1 Introduction

Nonparametric estimation of a density f_X in the case of measurement error models has been studied widely. For an excellent introduction to measurement error models, see Carroll et al. (2006). One of the most popular estimation techniques is the deconvolution kernel method of Carroll and Hall (1988) and Stefanski and Carroll (1990). See also Liu and Taylor (1989), Fan (1991a,b), Diggle and Hall (1993), Fan (1993) and Masry (1993). Although this estimator is nonparametrically optimal, its convergence rates can be quite slow for some types of errors, and its practical performance is not completely satisfactory. Moreover, even when critical information about the target density is given, the information cannot readily be incorporated to improve performance. A semiparametric spline method has been considered by Hazelton and Turlach (2010), where the ratio of f_X and the contaminated density is modelled parametrically. However, their estimator is not consistent unless their parametric assumption is correct. As a result, while it can work well when their 5-knot spline model is a good approximation to the true ratio, it can perform significantly

worse than standard deconvolution estimators in other cases. Carroll et al. (2011) propose an estimator that includes qualitative constraints (e.g. unimodality), but their method is computationally involved and can offer only modest improvements; for example, it does not improve convergence rates of conventional nonparametric estimators. In this paper, we consider estimation of a density f_X based on error contaminated data, especially when f_X is known or believed to be close to a parametric family of densities $f(\cdot\,|\,\theta)$ where θ is a vector-valued parameter. We wish to construct an estimator that is consistent even when the parametric assumption is incorrect.

Consistently estimating f_X in a partially parametric setting has been considered before in the literature, but, to our knowledge, only in the context of zero measurement error; see Olkin and Spiegelman (1987), Hjort and Glad (1995), Jones et al. (1995) and Hjort and Jones (1996).

Even in an error-free context, a convincing case for existing methods has arguably not been made in the literature. The existing treatment, especially from a theoretical viewpoint, has not focused on cases where the density f_X is close, as a function, to the density $f(\cdot\,|\,\theta)$, for example in the sense that the supremum of the difference between derivatives of f_X and $f(\cdot\,|\,\theta)$ is small. In order for our theoretical analysis to have the correct focus we have to quantify what we mean by "small," and the simplest way of doing that is to ask that it converge to zero as sample size increases. (Looking ahead to section 3, this corresponds to taking η, in (3.8), to decrease to zero as n diverges.) If the difference does not decrease with increasing n then, in asymptotic terms, there are relatively few advantages to borrowing from a parametric model, since the convergence rate remains unaltered.

Specifically, we take up the problem in cases where measurement error is present, and show that in this more difficult and more general setting the convergence rate can be improved significantly, to an optimal extent. The use of performance enhancing methods in this context has substantial potential advantages because the performance of conventional estimators is relatively poor.

We propose three consistent density estimators which incorporate this information and can produce order-of-magnitude improvements in the convergence rate of standard deconvolution methods when $f(\cdot\,|\,\theta)$ is not too far from f_X, without degrading the rates otherwise. On the theoretical side, we demonstrate that asymptotic improvements in convergence rates are generally possible when the distance between f_X and the assumed parametric model decreases with sample size. We prove that, if this distance tends to zero, then pronounced rate improvements can be obtained using our suggested estimators, and those improvements result in asymptotically optimal performance.

2 Model and Methodology

We observe independent data $W_1, \ldots, W_n$, identically distributed as

$$W = X + U\,, \tag{2.1}$$

where X and U are independent and the density f_U of U is known and is symmetric about 0. We wish to estimate f_X, the density of X.

Let $h > 0$ and K denote a bandwidth and kernel function, respectively, and write $\phi_K(t) = \int e^{itx}\, K(x)\, dx$ for the Fourier transform of K, where $i = \sqrt{-1}$. (Throughout this paper, we use ϕ_f to denote the Fourier transform of a function f.) The deconvolution kernel estimator of f_X, introduced by Carroll and Hall (1988) and Stefanski and Carroll (1990), is given by

$$\hat{f}_{\rm dec}(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\left(\frac{x - W_j}{h}\right) \quad \text{where} \quad K_U(u) = \frac{1}{2\pi} \int e^{-itu}\, \frac{\phi_K(t)}{\phi_U(t/h)}\, dt\,. \tag{2.2}$$

(Here and throughout we assume that ϕ_U never vanishes and that the integral in (2.2) is well defined.) This nonparametric estimator is consistent under mild conditions and has optimal nonparametric convergence rates; see Carroll and Hall (1988) and Fan (1991a). However, because of the intrinsic difficulty of nonparametric deconvolution,

these rates can be quite slow, which is reflected by somewhat disappointing finite sample performance.

Nevertheless, in some cases we may have some (correct or incorrect) parametric information about f_X. Clearly, in such instances we could improve the estimator if we could modify it in a way that incorporates the parametric guess only when it is correct, or close to being correct. Let $f(\cdot\,|\,\theta)$ denote a parametric model for the density of X, where $\theta \in \Theta$ is a finite-dimensional parameter. In the next sections, we construct parametrically assisted estimators that can improve the performance of $\hat{f}_{\rm dec}$ when f_X is approximately of the form $f(\cdot\,|\,\theta)$, and do not significantly degrade its quality otherwise.

2.1 The ratio estimator

First we consider an extension to the measurement error case of the method of Hjort and Glad (1995), developed in the error-free case. There, $X_1, \ldots, X_n$ are observed without error, and the authors modify the standard kernel density estimator $\check{f}_X(x) = (nh)^{-1} \sum_{j=1}^n K\{(x-X_j)/h\}$ of f_X by taking $\bar{f}_X(x\,|\,\theta) = n^{-1}\, f(x\,|\,\theta) \sum_{j=1}^n \mathcal{K}_x(X_j\,|\,\theta, h)\,,$ where $\mathcal{K}_x(u\,|\,\theta, h) = K\{(x-u)/h\}/\{h f(u\,|\,\theta)\}\,.$

To adapt $\bar{f}_X$ to the error context, where we observe only $W_1, \ldots, W_n$, we suggest using an estimator of the form

$$\hat{f}_{\rm rat}(x\,|\,\theta) = f(x\,|\,\theta)\, \frac{1}{n} \sum_{j=1}^n \mathcal{L}_x(W_j\,|\,\theta, h)\,, \tag{2.3}$$

where the function $\mathcal{L}_x$, which needs to be determined, is such that

$$E\big\{\mathcal{L}_x(W_j\,|\,\theta, h)\,|\,X_j\big\} = \mathcal{K}_x(X_j\,|\,\theta, h)\,. \tag{2.4}$$

Indeed, with $\mathcal{L}_x$ constructed in this way, our estimator is guaranteed to have the same bias as the error-free estimator $\bar{f}_X$. While the idea is simple, the difficulty is to find a function $\mathcal{L}_x$ that satisfies (2.4). However, as Delaigle et al. (2009) suggested, deconvolution problems are often easier to solve when considered in the Fourier domain. In

Appendix A.1 we compute the Fourier version of (2.4), and prove that the solution can be written as

$$\mathcal{L}_x(u \,|\, \theta, h) = \frac{1}{2\pi} \int e^{-itu} \, \phi_{\mathcal{K}_x(\,\cdot\,|\,\theta,h)}(t) / \phi_U(t) \, dt \,, \tag{2.5}$$

where $\phi_{\mathcal{K}_x(\,\cdot\,|\,\theta,h)}$ denotes the Fourier transform of the function $\mathcal{K}_x(\,\cdot\,|\,\theta, h)$.

In practice, θ would typically be estimated from the data, to minimise an estimator of a distance between f_X and $f(\,\cdot\,|\,\theta)$, or to maximise a likelihood. Let $\hat{\theta}$ denote a value of θ computed from the data $W_1, \ldots, W_n$. Our estimator of f_X, based on $\hat{\theta}$, is defined by

$$\hat{f}_{\rm rat}(x \,|\, \hat{\theta}) = f(x \,|\, \hat{\theta}) \, \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}_x(W_j \,|\, \hat{\theta}, h) \,. \tag{2.6}$$

In section 3 we prove that $\hat{f}_{\rm rat}$ has optimal convergence rates in a variety of settings, when the parametric assumption $f(\,\cdot\,|\,\theta)$ is reasonably accurate; our proof is also valid in the error-free case, where our results are also new. We shall see too that the improvements of $\hat{f}_{\rm rat}$ over the standard deconvolution estimator $\hat{f}_{\rm dec}$, defined at (2.2), are in terms of reduced bias, and that it asymptotically has the same variance as $\hat{f}_{\rm dec}$.

Remark 1. Using the technique employed to derive $\mathcal{L}_x$ in (2.5), the work of Naito (2004), which introduces a general class of multiplicatively adjusted density estimators that includes estimators suggested by Hjort and Jones (1996), based on local likelihood ideas, as well as the approach of Hjort and Glad (1995), can be generalised to the case of measurement error. There $\hat{f}_{\rm rat}$, defined at (2.3), would be altered to $\hat{f}_\alpha(x \,|\, \hat{\theta}, h) = f(x \,|\, \hat{\theta}) \, \xi_\alpha(x \,|\, \hat{\theta}, h)$, where

$$\begin{aligned}
\xi_\alpha(x \,|\, \theta, h) &= \frac{n^{-1} \, \sum_j \, \mathcal{L}_{\alpha,x}(W_j \,|\, \theta, h)}{\int f(u \,|\, \theta) \, \mathcal{K}_{\alpha,x}(u \,|\, \theta, h) \, du} \,, \\
\mathcal{L}_{\alpha,x}(u \,|\, \theta, h) &= \frac{1}{2\pi} \int e^{-itu} \, \phi_{\mathcal{K}_{\alpha,x}(\,\cdot\,|\,\theta,h)}(t) \, / \phi_U(t) \, dt \,, \\
\mathcal{K}_{\alpha,x}(u \,|\, \theta, h) &= h^{-1} K\{(x-u)/h\} \, f(u \,|\, \theta)^{1-\alpha} \,,
\end{aligned}$$

and α is a user-chooseable parameter. Similar extensions are possible for the density versions of parametrically guided regression estimators suggested by Fan, Wu and

Feng (2009). However, we feel that the latter approach loses some of its appeal when developed in the setting of density estimation, since there the density approximation being used as the "start" is often not itself a density, and so some of the motivation is lost. Of course, this is not an issue in the setting of regression, but in the context of density estimation it extends to measurement error problems.

2.2 The bias corrected estimator

Despite its good theoretical properties, in practice the estimator $\hat{f}_{\text{rat}}$ suffers from a number of drawbacks, such as the difficulty of implementation (no analytic formula) and numerical problems caused by division by $f(u \,|\, \theta)$; see section 4.3. As an alternative to reducing the bias of $\hat{f}_{\text{dec}}$ when $f(\,\cdot \,|\, \theta)$ is close to f_X, without increasing it much otherwise, we suggest instead subtracting from $\hat{f}_{\text{dec}}$ a parametric estimator of its bias. Since $\text{Bias}(\hat{f}_{\text{dec}}) = f_X * K_h - f_X$, where $K_h = h^{-1}K(x/h)$ and $*$ denotes convolution, this means that we suggest using the estimator

$$\hat{f}_{\text{bco}}(x \,|\, \theta) = \hat{f}_{\text{dec}}(x) - \{f(\,\cdot \,|\, \theta) * K_h\}(x) + f(x \,|\, \theta) \tag{2.7}$$

in cases where θ is fixed. When θ is estimated from the data, as $\hat{\theta}$, we take $\hat{f}_{\text{bco}}(x \,|\, \hat{\theta}) = \hat{f}_{\text{dec}}(x) - \{f(\,\cdot \,|\, \hat{\theta}) * K_h\}(x) + f(x \,|\, \hat{\theta})$. This estimator is much simpler to compute than $\hat{f}_{\text{rat}}$, and, as we shall see in section 5, it works particularly well in practice. In addition, it shares the optimality property of $\hat{f}_{\text{rat}}$; see section 3.

2.3 The weighted estimator

As an alternative we could also adapt the error-free estimator of Olkin and Spiegelman (1987) to the measurement error case. In their error-free context, they suggest using the estimator $wf(x \,|\, \theta) + (1 - w)\check{f}_X(x)$, where $\check{f}_X$ is the standard error-free kernel estimator of f_X, and $w \in [0, 1]$ is a parameter which ideally is close to 1 when $f(\,\cdot \,|\, \theta)$ is close to f_X, and close to zero otherwise. In practice, Olkin and Spiegelman (1987) suggest choosing w and θ by maximum likelihood (ML).

For w and θ fixed, their estimator is straightforward to adapt to the measurement error context:

$$\hat{f}_{\text{wgt}}(x \,|\, \theta) = w\, f(x \,|\, \theta) + (1 - w)\, \hat{f}_{\text{dec}}(x)\,, \tag{2.8}$$

and in principle we could, like them, choose w and θ by maximum likelihood. Of course, in the measurement error context, since the X_is are not available, ML should be based on the density $g(x \,|\, \theta) = \{f_X(\,\cdot\,|\,\theta) * f_U\}(x)$ corresponding to the contaminated data $W_1, \ldots, W_n$. However, working with the density of the W_is instead of that of the X_is increases considerably the variability of ML estimators, especially when the error variance is large (see section 4.1).

We experimented with this ML approach for choosing w and found it gave poor results. Let $\|\cdot\|_2$ denote the L_2-norm. In our errors-in-variables context, we suggest instead choosing w from the data by taking

$$\hat{w} = \|F_n - \check{F}_W\|_2 / \{\|F_n - \check{F}_W\|_2 + \|F_n - G(\,\cdot\,|\,\hat{\theta})\|_2\}\,, \tag{2.9}$$

where $\hat{\theta}$ is the ML estimator of θ, $G(w \,|\, \hat{\theta}) = \int_{-\infty}^{w} g(x \,|\, \hat{\theta})\, dx$, $\check{F}_W(w) = \int_{-\infty}^{w} \check{f}_W(x)\, dx$, with $\check{f}_W$ denoting the standard kernel estimator of f_W constructed from the W_is, and F_n the empirical cdf of W. Here we compute $\check{f}_W$ with a bandwidth h_w of the same order as the deconvolution bandwidth (it can be proved that this choice is optimal, but $\hat{f}_{\text{wgt}}$ does not generally enjoy optimal convergence rates; see section 3).

3 Theoretical Properties

3.1 Error types and convergence rates of standard deconvolution estimator

As in other nonparametric deconvolution problems, the rates of convergence of the estimators introduced in section 2 depend on the rate of decay of the tails of the characteristic function ϕ_U. Two classes of error are usually considered in the nonparametric deconvolution literature: ordinary smooth errors and supersmooth errors;

see Fan (1991a). If ϕ_U never vanishes then the distribution of U is ordinary smooth, of order $\alpha > 0$, when

$$d_0\,(1+|t|)^{-\alpha} \le |\phi_U(t)| \le d_1\,(1+|t|)^{-\alpha} \quad \text{for all } t\,, \tag{3.1}$$

for constants $d_0 > 0$ and $d_1 > 0$; and the distribution is supersmooth, of order $\alpha > 0$, if

$$d_0\,(1+|t|)^{\alpha_1}\,\exp\bigl(-|t|^\alpha/\gamma\bigr) \le |\phi_U(t)| \le d_1\,(1+|t|)^{\alpha_2}\,\exp\bigl(-|t|^\alpha/\gamma\bigr) \quad \text{for all } t\,, \tag{3.2}$$

for constants $d_0 > 0$, $d_1 > 0$, $\gamma > 0$, α_1 and α_2.

As a prelude to summarising the properties of the standard deconvolution estimator $\hat{f}_{\rm dec}$, defined at (2.2), we introduce the following standard assumptions:
(A1) f_X has β bounded derivatives on the real line, $\mathbb{R}$;
(A2) K is real valued and symmetric, $\int(1+|u|^\beta)\,|K(u)|\,du < \infty$, $\int K = 1$, and $\int u^j\,K(u)\,du = 0$ for $j = 1, \ldots, \beta-1$;
(A3) when ϕ_U satisfies (3.1), $\int\{|t|^\alpha|\phi_K(t)| + |t|^{2\alpha}|\phi_K(t)|^2\}\,dt < \infty$, and when ϕ_U satisfies (3.2), $\phi_K(t) = \bigl(1-t^2\bigr)^\kappa\,I(-1 \le t \le 1)$, where $\kappa > 0$.

Under (A1)–(A3), it is well known (see Fan, 1991a) that, in the ordinary smooth case (3.1), for each $x \in \mathbb{R}$,

$$\text{if we take } h \asymp n^{-1/(2\alpha+2\beta+1)}, \text{ then } \hat{f}_{\rm dec}(x) - f_X(x) = O_p(n^{-\beta/(2\alpha+2\beta+1)}); \tag{3.3}$$

and, in the supersmooth case (3.2), for each $x \in \mathbb{R}$,

$$\text{if } h = c\,(\ln n)^{-1/\alpha} \text{ with } c > (2/\gamma)^{1/\alpha}, \text{ then } \hat{f}_{\rm dec}(x) - f_X(x) = O_p\{(\ln n)^{-\beta/\alpha}\}. \tag{3.4}$$

3.2 Parametric model

To derive convergence rates for our estimators we need to impose conditions on the parametric model. In this section we state and discuss our assumptions, of which the most basic is:

$$f_X \text{ is in a class } \mathcal{F} \text{ of densities that are uniformly bounded.} \tag{3.5}$$

To make our analysis realistic we shall permit f_X to depend on n, but take $\mathcal{F}$, in (3.5), and the model $f(\,\cdot\,|\,\theta)$ to be fixed. See Appendix A.6 for discussion. For definiteness we assume that Θ is a nondegenerate, closed ball in $\mathbb{R}^p$, although of course other shapes are possible.

The estimator $\hat{\theta}$, computed from data drawn from the distribution with density f_X, rather than with density $f(\,\cdot\,|\,\theta)$ for some θ, should be viewed as an approximation to a particular value, θ_1 say, of θ. Since f_X may depend on n then θ_1 can too. We shall assume that $\hat{\theta}$ comes from Θ and differs from θ_1 by only $O_p(n^{-1/2})$:

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{f_X\in\mathcal{F}} P_{f_X}(\|\hat{\theta}-\theta_1\| > C\,n^{-1/2}) \to 0 \text{ as } C\to\infty, \text{ where } \theta_1\in\Theta\\ &\text{can depend on } n \text{ and } f_X, \text{ and moreover, } P_{f_X}(\hat{\theta}\in\Theta) = 1.\end{aligned} \tag{3.6}$$

Here P_{f_X} denotes the probability measure P when the data have density f_X, and $\|\,\cdot\,\|$ is the conventional Euclidean norm in p-variate space. For example, if the family $f(\,\cdot\,|\,\theta)$ contains the density f_X and $\hat{\theta}$ is computed by maximum likelihood, then θ_1 is the true value of θ, and (3.6) states merely that the maximum likelihood estimator is $\sqrt{n}$-consistent.

Since f_X and θ_1 can depend on n, so too can the function

$$\psi = f_X - f(\,\cdot\,|\,\theta_1)\,, \tag{3.7}$$

which necessarily satisfies $\int \psi = 0$. We can also view θ_1 as a function of ψ, using the notation $\theta_1(\psi)$. The simplest case is that where ψ does not depend on n, in which case $\theta_1 = \theta_1(\psi)$ is also fixed. The theory developed by Hjort and Glad (1995) addresses this case, although those authors treat only the error-free case. Of significantly greater interest, because it goes to the heart of the motivation for using a parametric model to assist a nonparametric estimator, is the setting where ψ in (3.7), or an appropriate derivative of it, converges to zero as n diverges.

To reflect this, recall that in section 1 we stressed that, to make a convincing theoretical case for the advantages of using a parametric start, we would have to treat cases where the difference between derivatives of the densities f_X and $f(\,\cdot\,|\,\theta)$

becomes small as n increases. In (3.8) below we assume that the difference between the highest order derivatives is of order $\eta = \eta(n)$; as we have just indicated, the most interesting case is that where $\eta = o(1)$. See also Appendix A.5, where we discuss related conditions imposed in section 3.6. We could ask that the left-hand side of the first identity in (3.8) equal $O(\eta)$, but that turns out to be an unnecessary strengthening of the assumption when deriving the upper bounds in (3.9) and (3.10):

$$\sup_{0\le j\le \ell_0} \sup_{x\in\mathbb{R}} |\psi^{(j)}(x)| = O(1) \quad \text{and} \quad \sup_{x\in\mathbb{R}} |\psi^{(\ell_0)}(x)| = O(\eta)\,, \tag{3.8}$$

where $\eta = \eta(n) = O(1)$. When $\eta = o(1)$ the density f_X is demonstrably close to the model, and it is reasonable to ask whether, and in what manner, the size of η is reflected in the convergence rates of estimators. A similar problem, although in a different setting, was explored by Eguchi and Copas (1998).

3.3 Convergence rates for the bias corrected estimator

To study theoretical properties of the bias corrected estimator $\hat{f}_{\rm bco}$, given by (2.7), with θ there replaced by the estimator $\hat{\theta}$, we assume that (A2) and (A3) hold, and that:

(A4) $\sup_{x\in\mathbb{R}} \sup_{\theta\in\Theta} \|\partial f(x\,|\,\theta)/\partial\theta\| < \infty$.

If (3.5) and (3.6) hold, if ψ satisfies (3.8) with $\ell_0 = \beta$ (this replaces condition (A1) of section 3.1), if (A2)–(A4) hold, and if $\hat{\theta} = \theta_1 + O_p(n^{-1/2})$, then it can be proved (see Appendix A.2) that when ϕ_U satisfies (3.1), i.e. the distribution of U is in the ordinary smooth class of errors (see section 3.1), and we take $h \asymp \min\{1,(n\eta^2)^{-1/(2\alpha+2\beta+1)}\}$, then for each $x \in \mathbb{R}$ we have

$$\hat{f}_{\rm bco}(x\,|\,\hat{\theta}) - f_X(x) = \begin{cases} O_p(n^{-1/2}) & \text{if } \eta = O(n^{-1/2}) \\ O_p\bigl\{(\eta^{2\alpha+1}/n^{\beta})^{1/(2\alpha+2\beta+1)}\bigr\} & \text{if } n^{1/2}\eta \to \infty\,; \end{cases} \tag{3.9}$$

and when ϕ_U satisfies (3.2), i.e. the distribution of U is in the supersmooth class of errors (again, see section 3.1), if $h = \min\{1, c\,[\ln(n\eta^2)]^{-1/\alpha}\}$, with $c > (2/\gamma)^{1/\alpha}$, then

for each $x \in \mathbb{R}$ we have

$$\hat{f}_{\rm bco}(x \,|\, \hat{\theta}) - f_X(x) = \begin{cases} O_p(n^{-1/2}) & \text{if } \eta = O(n^{-1/2}) \\ O_p\big[\eta \,\{\ln(n\eta^2)\}^{-\beta/\alpha}\big] & \text{if } n^{1/2}\eta \to \infty \,. \end{cases} \tag{3.10}$$

Comparing (3.9) and (3.10) with (3.3) and (3.4), we conclude that $\hat{f}_{\rm bco}(x)$ converges to $f_X(x)$ at a rate (sometimes considerably) faster than the standard rate in this measurement-error problem. For finite sample size, as long as the parametric model $f(\,\cdot\,|\,\theta)$ is a reasonable approximation to the true density f_X (i.e. η is relatively small), it can be expected that the estimator $\hat{f}_{\rm bco}$ will improve on $\hat{f}_{\rm dec}$. In section 3.6 we complete the picture by showing that these convergence rates are optimal. In Appendix A.2 we derive the bias and variance of our estimator (see (A.1)–(A.3)). These show that the improvements offered by $\hat{f}_{\rm bco}$ over $\hat{f}_{\rm dec}$ are in terms of bias; the two estimators have the same asymptotic variance.

3.4 Convergence rates for the weighted estimator

Broadly similar results can be derived for the estimator $\hat{f}_{\rm wgt}$, defined at (2.8). Specifically, let us assume that (3.5), (3.6), (3.8) for $\ell_0 = 0$, and (A1)–(A4), hold. If $\widehat{w}$ is given by (2.9), where $\check{F}_W$ is computed using a bandwidth h_w, then it can be proved that, when ϕ_U satisfies (3.1) and we take $h_w \asymp n^{-1/(2\alpha+2\beta+1)}$, and when ϕ_U satisfies (3.2) and $h_w = c\,(\ln n)^{-1/\alpha}$, with $c > (2/\gamma)^{1/\alpha}$, we have: $\widehat{w} = O_p\big\{h_w^\beta/(\eta \vee n^{-1/2} + h_w^\beta)\big\}$ and $1 - \widehat{w} = O_p\big\{(\eta \vee n^{-1/2})/(\eta \vee n^{-1/2} + h_w^\beta)\big\}$.

Hence, under (3.1), if we compute $\hat{f}_{\rm dec}$ with a bandwidth $h \asymp n^{-1/(2\alpha+2\beta+1)}$ and under (3.2), if we take $h = c\,(\ln n)^{-1/\alpha}$ with $c > (2/\gamma)^{1/\alpha}$, then, for each $x \in \mathbb{R}$

$$\hat{f}_{\rm wgt}(x \,|\, \hat{\theta}) - f_X(x) = \begin{cases} O_p(\eta \vee n^{-1/2}) & \text{if } \eta \le h^\beta \\ O_p(h^\beta) & \text{if } \eta > h^\beta \,. \end{cases} \tag{3.11}$$

It is straightforward to show that, except in pathological cases where biases are of unusually small order, the convergence rates at (3.11) cannot be improved using other choices of $\widehat{w}$. Note that the convergence rate in the second part of (3.11) is the rate of $\hat{f}_{\rm dec}$ (see (3.3)–(3.4)), and is slower than that of $\hat{f}_{\rm bco}$ unless η does not tend to

zero. In particular, $\hat{f}_{\rm wgt}$ is not optimal, as will be reflected by our numerical results in section 5.

3.5 Convergence rates for the ratio estimator

Next we show that the convergence rates of the estimator $\hat{f}_{\rm rat}$ are identical to those for $\hat{f}_{\rm bco}$. However, we shall see later that $\hat{f}_{\rm bco}$ often enjoys better performance in practice. This is due at least partly to the inherent construction of $\hat{f}_{\rm rat}$, which involves one random variable divided by another. Fluctuations of the random variable in the denominator cause problems that are particularly difficult to remove. Related to this, the convergence rates for the estimator $\hat{f}_{\rm rat}$ are awkward to derive, and so we shall give details only in the case of ordinary smooth errors.

To frame our regularity conditions we put

$$g_k(x \,|\, \theta) = \left\{ \frac{\partial^k}{\partial u^k} \, \frac{f(x \,|\, \theta)}{f(x-u \,|\, \theta)} \right\}_{u=0},$$

and we define the function ω in terms of the remainder in a short Taylor expansion:

$$\frac{f(x \,|\, \theta)}{f(x-u \,|\, \theta)} = \sum_{k=0}^{r+1} \frac{u^k}{k!} \, g_k(x \,|\, \theta) + \omega(u, x \,|\, \theta) \, u^{r+1}, \tag{3.12}$$

where the integer $r \geq 0$ is fixed. Put $\gamma_{h,x}(u \,|\, \theta) = \omega(hu, x \,|\, \theta) \, u^{r+1} \, K(u)$ and let $\gamma_{h,x}^{(\ell)}(u \,|\, \theta)$ be the ℓth partial derivative of $\gamma_{h,x}(u \,|\, \theta)$ with respect to u. Write $\phi_K^{(k)}$ for the kth derivative of ϕ_K.

We make the following assumptions, where $\mathcal{I}$ denotes a compact interval on which we wish to estimate f_X:

(B1) f_X satisfies (3.5) and ψ satisfies (3.8) for $\ell_0 = 0, \ldots, \beta$;

(B2) K satisfies (A2), $\int (1+|t|)^{\alpha+1} \, |\phi_K^{(k)}(t)| \, dt < \infty$ and $\int \{(1+|t|)^{\alpha} \, |\phi_K^{(k)}(t)|\}^2 \, dt < \infty$ for $k = 0, \ldots, r$, where, here and below, the fixed integer r is as in (3.12);

(B3) $h = h(n)$ satisfies $h = O(n^{-c})$, for some $c \geq 0$; $h \asymp 1$ if $c = 0$; and $nh^{2\alpha+1} \to \infty$ if $c > 0$;

(B4) $f(x \,|\, \theta)$ is bounded away from zero uniformly in $x \in \mathcal{I}$ and $\theta \in \Theta$;

(B5) for $k = 1, \ldots, r$, each first derivative of $g_k(x \,|\, \theta)$, with respect to θ, is bounded uniformly in $x \in \mathcal{I}$ and $\theta \in \Theta$;
(B6) $\gamma_{h,x}^{(\ell)}(u \,|\, \theta)$ exists for $\ell = 1, \ldots, s$, and $\int |\gamma_{h,x}^{(\ell)}(u \,|\, \theta)|\, du$ is bounded uniformly in $0 \le h \le C_1$, $x \in \mathcal{I}$, $\theta \in \Theta$ and $\ell = 0, s$, where $s > \alpha + 1$ is an integer and $C_1 > 0$ is a finite constant.

These conditions are discussed in Appendix A.4, where we also give examples illustrating (B6). Assume (3.1), (3.6) and (B1)–(B6), and take $h \asymp \min\{1, (n\eta^2)^{-1/(2\alpha+2\beta+1)}\}$. Under these conditions, if r is large enough so that $h^{r-s+1} = O\{(\eta^{2\alpha+1}/n^{\beta})^{1/(2\alpha+2\beta+1)}\}$, then for each $x \in \mathcal{I}$ we have

$$\hat{f}_{\mathrm{rat}}(x) - f_X(x) = \begin{cases} O_p(n^{-1/2}) & \text{if } \eta = O(n^{-1/2}) \\ O_p\{(\eta^{2\alpha+1}/n^{\beta})^{1/(2\alpha+2\beta+1)}\} & \text{if } n^{1/2}\eta \to \infty\,. \end{cases} \tag{3.13}$$

A derivation of (3.13) in the case $n^{1/2}\eta \to \infty$ is given in section B.2 of the supplementary file. A proof in the case $\eta = O(n^{-1/2})$ is relatively straightforward, and so is omitted.

The estimator $\hat{f}_{\mathrm{rat}}(x)$ has properties similar to $\hat{f}_{\mathrm{bco}}$: it converges to $f_X(x)$ at a faster rate than the standard deconvolution estimator as soon as $\eta = \eta(n) \to 0$ (see (3.13)), and the improvements offered by $\hat{f}_{\mathrm{rat}}$ over $\hat{f}_{\mathrm{dec}}$ are principally in terms of lower bias (see section B.2). Moreover, in a variety of settings and up to terms that are asymptotically negligible, $\hat{f}_{\mathrm{rat}}$ and the standard deconvolution estimator $\hat{f}_{\mathrm{dec}}$ have identical standard deviation; see section B.2.

3.6 Optimality

The statistically challenging aspect of this problem is accommodating the function ψ in (3.7), rather than estimating the parameter θ in the model $f(\,\cdot\,|\,\theta)$. Indeed, a value of θ having the property that $f(\,\cdot\,|\,\theta)$ approximates f_X typically can be estimated root-n consistently. To address optimality it is sufficient to take the model density $f(\,\cdot\,|\,\theta)$ to be a fixed density f_0, not depending on θ, and to consider estimating $f_X = f_0 + \psi$

at a fixed point x_0. We assume that:

$$f(x \,|\, \theta) \equiv f_0(x), \text{ not depending on } \theta, \text{ where } f_0(x) \text{ is bounded away from zero for } x \text{ in a neighborhood of } x_0. \tag{3.14}$$

In order for our main result to be relevant to an account of optimality for smooth densities f_X, it should address properties that hold uniformly in a class of functions ψ which includes cases of greatest statistical difficulty. This class will be denoted by Ψ_n, and enjoys the following properties; here and below, $C_2, C_3, \ldots$ denote positive constants:

$$\begin{aligned} &\text{defining } s(\psi) = \max_{0 \le j \le \beta} \sup_x |\psi^{(j)}(x)|, \text{ where } \beta \ge 1, \text{ we have} \\ &\sup_{\psi \in \Psi_n} s(\psi) \le C_2\, \eta \text{ for each } n, \text{ and in particular, each } \psi \in \Psi_n \text{ has } \beta \\ &\text{derivatives; and } \int |\psi| < \infty, \int \psi = 0, \text{ and } f_0 + \psi \text{ is a proper probability} \\ &\text{density function whenever } \psi \in \Psi_n \text{ and } \theta_1(\psi) \in \Theta. \end{aligned} \tag{3.15}$$

Moreover, we assume that $\eta = \eta(n)$ satisfies

$$C_3\, n^{-1/2} \le \eta \le C_4\,. \tag{3.16}$$

See Appendix A.5 for a discussion of these conditions.

It can be proved that the upper bounds to convergence rates, given earlier at (3.9) and (3.13), hold uniformly over $\psi \in \Psi_n$, for each $x \in \mathcal{I}$. In particular, with $\hat{f}$ denoting either $\hat{f}_{\text{bco}}$ or $\hat{f}_{\text{rat}}$, the second parts of each of (3.9) and (3.13) can be extended to:

$$\lim_{C \to \infty} \limsup_{n \to \infty} \sup_{\psi \in \Psi_n} P\Big\{ \big| \hat{f}(x \,|\, \hat{\theta}) - f_X(x) \big| > C \left(\eta^{2\alpha+1} / n^{\beta} \right)^{1/(2\alpha+2\beta+1)} \Big\} = 0 \tag{3.17}$$

for each $x \in \mathcal{I}$. Let $\mathcal{S}$ denote the set of all measurable functions of the data $W_1, \ldots, W_n$ ($\mathcal{S}$ represents the class of all possible estimators of $f_X(x_0)$ that can be computed from the data).

The following result implies that, in the context of ordinary-smooth errors, no estimator can converge at a rate faster than the upper bound at (3.17); see Appendix A.3 for a proof. Note that the conditions on the derivatives of ϕ_U imposed in Theorem 1

slightly restrict the class of ordinary smooth errors; these conditions are the same as those imposed in Fan's (1991a) Theorem 5, and are used to control terms that depend on ϕ_U''. These, and similar conditions on ϕ_U appearing in Theorem 2 below, can be replaced by another set of conditions, for example those used in Delaigle and Meister's (2008) Theorem 2.1.

Theorem 1. *Assume that the error distribution satisfies* $|\phi_U^{(j)}(t)| \le C_5\,(1+|t|)^{-j-\alpha}$ *for* $j = 0, 1, 2$, *where* $\alpha > 0$. *If* $f_X = f_0 + \psi$, *where* f_0 *satisfies* (3.14), *and if* (3.16) *holds and* C_4 *there is sufficiently small, then there exist two candidates* ψ_1, ψ_2 *for* ψ, *each depending on* n *and each satisfying* (3.15), *such that*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \inf_{\phi(W_1,\ldots,W_n)\in\mathcal{S}} \sup_{\psi\in\{\psi_1,\psi_2\}} P_{f_X}\Big\{\big|\phi(W_1,\ldots,W_n) - f_X(x_0)\big| > C\left(\eta^{2\alpha+1}/n^{\beta}\right)^{1/(2\alpha+2\beta+1)}\Big\} = 0\,. \quad (3.18)$$

Together, (3.17) and (3.18) show that the rates given at (3.9) and (3.13), for the ordinary smooth error case, are optimal.

Finally in this section we state a version of Theorem 1 in the supersmooth case.

Theorem 2. *Assume that* $\max_{j=0,1,2} |\phi_U^{(j)}(t)| \le C_6\,(1+|t|)^{\alpha_3}\,\exp(-|t|^{\alpha}/\gamma)$, *where* $\alpha, \gamma > 0$ *and* α_3 *are constants. If* $f_X = f_0 + \psi$, *where* f_0 *satisfies* (3.14), *and if* (3.16) *holds and* C_4 *there is sufficiently small, then there exist two candidates* ψ_1, ψ_2 *for* ψ, *each depending on* n *and each satisfying* (3.15), *such that*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \inf_{\phi(W_1,\ldots,W_n)\in\mathcal{S}} \sup_{\psi\in\{\psi_1,\psi_2\}} P_{f_X}\Big[\big|\phi(W_1,\ldots,W_n) - f_X(x_0)\big| > C\,\eta\left\{\log\left(1+n\eta^2\right)\right\}^{-\beta/\alpha}\Big] = 0\,.$$

4 Details of implementation

4.1 Maximum likelihood estimator of θ

To choose θ from the data, a natural approach is to maximise the likelihood of the contaminated observations $W_1, \ldots, W_n$. Under the parametric assumption that $f_X =$

$f(\cdot\,|\,\theta)$, the parametric form of the density f_W of the W_js is given by $f_W(\cdot\,|\,\theta) = f(\cdot\,|\,\theta) * f_U$. Therefore, the ML estimator of θ is given by $\hat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\theta\in\Theta} \sum_{j=1}^{n} \log \int f(W_j - u\,|\,\theta) f_U(u)\,du$, and can be computed by the EM algorithm. In theory, $\hat{\theta}_{\mathrm{ML}}$ has the usual $n^{-1/2}$ convergence rate under mild conditions, but because $f_W(\cdot\,|\,\theta)$ is obtained by convolution of $f(\cdot\,|\,\theta)$ and f_U, its variance is larger than that of the ML estimator based on error-free data. In particular, the smoother the errors (i.e., the faster ϕ_U tends to zero in the tails), the larger the variance of $\hat{\theta}_{\mathrm{ML}}$. In practice, this means that if the errors are supersmooth or the variance of U is large, then $\hat{\theta}_{\mathrm{ML}}$ can be too variable unless n is large, and this can significantly affect performance of function estimators.

4.2 Minimum distance estimator of θ

Another possibility is to use the Minimum Distance (MD) estimator studied by Cao et al. (1995), and defined, in their error-free setting, by $\hat{\theta}_{\mathrm{MD}} = \mathrm{argmin}_{\theta\in\Theta} D\{\check{f}_X, f(\cdot\,|\,\theta)\}$, where D is a distance, for example the L_2-norm, and $\check{f}_X$ is the standard error-free kernel estimator of f_X. In our errors-in-variables setting, we could apply their method to the data W_i, taking $\hat{\theta}$ to minimise $D\{\check{f}_W, f(\cdot\,|\,\theta) * f_U\}$, where $\check{f}_W$ is the standard error-free kernel estimator of f_W computed from the W_is. In theory, this estimator has $n^{-1/2}$ convergence rates when the bandwidth used to compute $\check{f}_W$ is sufficiently small, but in practice, we found that, for small n, it had difficulties estimating the variance components of $f(\cdot\,|\,\theta)$ in cases where $f(\cdot\,|\,\theta)$ was a mixture. These difficulties come from the fact that the fluctuations of $f(\cdot\,|\,\theta)$, when θ varies, are smoothed out in the convolution $f(\cdot\,|\,\theta) * f_U$, which makes it hard to properly identify peaks in $f(\cdot\,|\,\theta)$.

To overcome this problem, in the errors-in-variables context, we define the MD estimator by

$$\hat{\theta}_{\mathrm{MD}} = \underset{\theta\in\Theta}{\mathrm{argmin}} \int \left\{\hat{f}_{\mathrm{dec}}(x) - f(x\,|\,\theta)\right\}^2 dx, \tag{4.1}$$

where, on this occasion, $\hat{f}_{\mathrm{dec}}$ denotes the deconvolution estimator of f_X computed

with the sinc kernel, defined via $\phi_K(t) = I\{-1 \le t \le 1\}$. It can be proved that when $\hat{f}_{\rm dec}$ in (4.1) is computed with this kernel and an appropriate bandwidth, $\hat{\theta}_{\rm MD}$ converges to θ at a $n^{-1/2}$ rate as long as f_X is sufficiently smooth relative to f_U.

To apply this method, we need to choose a bandwidth for computing $\hat{f}_{\rm dec}$ in (4.1). We can easily compute a good bandwidth h_f for estimating f_X by $\hat{f}_{\rm dec}$ with the sinc kernel, using for example the normal reference (NR) bandwidth $h_{\rm NR}$ for this kernel (see section C.3 in the supplementary file). However, a good bandwidth h_θ for estimating θ through $\hat{f}_{\rm dec}$ is not necessarily close to h_f.

Let $c = h_\theta/h_f$. We suggest estimating h_θ by $\hat{h}_\theta = \hat{c}\, h_{\rm NR}$, where we compute the estimator $\hat{c}$ of c by the following bootstrap method:

1. Generate B bootstrap samples of size n from $f(\,\cdot\,|\,\hat{\theta}_P) * f_U$, conditionally on $W_1, \ldots, W_n$, where $\hat{\theta}_P$ is a pilot estimator of θ obtained by minimising the right hand side of (4.1), with $\hat{f}_{\rm dec}$ computed using a second order kernel and the plug-in bandwidth of Delaigle and Gijbels (2002, 2004).

2. For $b = 1, \ldots, B$, compute the NR bandwidth $h_{{\rm NR},b}$ of section C.3, using the bth bootstrap sample. For each c on a grid, and for $b = 1, \ldots, B$, estimate $\hat{\theta}_P$ by
$$\hat{\theta}^*_b(c\, h_{{\rm NR},b}) = \operatorname*{argmin}_{\theta\in\Theta} \int \left\{\hat{f}^*_{\rm dec,b}(x; c\, h_{{\rm NR},b}) - f(x\,|\,\theta)\right\}^2 dx,$$
where $\hat{f}^*_{\rm dec,b}(x; c\, h_{{\rm NR},b})$ denotes the deconvolution estimator computed from the bth bootstrap sample, using the bandwidth $c\, h_{{\rm NR},b}$ and the sinc kernel.

3. Choose $\hat{c} = \operatorname{argmin}_c B^{-1} \sum_{1\le b\le B} \{\hat{\theta}^*_b(c\, h_{{\rm NR},b}) - \hat{\theta}_P\}^2$.

In step 1, we use a second order kernel and a bandwidth slightly larger than the optimal bandwidth for estimating θ, because it leads to an estimator $\hat{\theta}_P$ that is not too variable, which in turns prevents the estimator of the bandwidth from being too variable.

4.3 Computing the ratio estimator

In general, there is no closed form for the Fourier transform $\phi_{\mathcal{K}_x}$ of $\mathcal{K}_x$, which needs to be approximated by numerical integration, although in the particular case where U has a Laplace distribution, or the distribution of a sum of Laplace random variables, $\mathcal{L}_x$ can be written explicitly as a linear combination of derivatives of $\mathcal{K}_x$ (see section C.1 in the supplementary file).

This makes this estimator rather unattractive, despite its good asymptotic properties. Another difficulty when computing $\hat{f}_{\text{rat}}$ is that it involves dividing by $f(u \,|\, \theta)$, which causes problems when $f(u \,|\, \theta)$ is too small, even when the estimator can be written explicitly. To avoid this, we can replace a too small denominator by a small ridge parameter. In our numerical work we developed such a procedure in the particular Laplace error case; see section C.1 in the supplementary file for details. As we shall see in section 5.2, this gave good results, but more generally, given that our other estimators are simpler to calculate, and perform very well in practice, it seems too cumbersome to develop an effective version of $\hat{f}_{\text{rat}}$ for each possible combination of error and kernel.

4.4 Choice of the bandwidths and the kernel

Except for step 2 in the bootstrap algorithm of section 4.2, to compute $\hat{f}_{\text{dec}}$ we used a second order kernel and the plug-in bandwidth of Delaigle and Gijbels (2004). As usual in the deconvolution kernel literature, in the ordinary smooth error case we use the standard normal kernel, and in the supersmooth case we use the kernel defined via $\phi_K(t) = (1-t^2)^3 I\{-1 \le t \le 1\}$. Each of our three estimators $\hat{f}_{\text{rat}}$, $\hat{f}_{\text{bco}}$ and $\hat{f}_{\text{wgt}}$ requires a bandwidth h. In general, a bandwidth should be chosen to minimise some distance between the estimator and the true unknown density. For each estimator, we experimented with data-driven versions of the bandwidth that minimises the L_2 distance between the estimator and f_X, and compared the results with those obtained

when using the plug-in bandwidth of Delaigle and Gijbels (2004) to compute $\hat{f}_{\rm dec}$. For $\hat{f}_{\rm bco}$ and $\hat{f}_{\rm wgt}$, we found we could simply use this plug-in bandwidth, but for $\hat{f}_{\rm rat}$ we found we could get better results by using a more sophisticated simulation-extrapolation (SIMEX) bandwidth of the type suggested by Delaigle and Hall (2008). See section C.2 in the supplementary file for details of how to compute this bandwidth.

5 Numerical study

5.1 Simulation settings

We examined the practical performance of our methods by applying them to simulated examples. We generated 100 samples of independent variables $X_1, \ldots, X_n$, where $n = 200$, 300 and 700, from six densities: (i) $0.5\,\phi_{-2.2,1} + 0.5\,\phi_{2.2,1}$; (ii) $0.5\,\phi_{-1.5,1} + 0.5\,\phi_{1.5,1}$; (iii) $0.75\,\phi_{-21} + 0.25\,\phi_{1,0.5}$; (iv) Gamma(3, 1); (v) -2+0.25 Gamma(5, 1) + 0.75 Gamma(24, 0.5); (vi) 3+0.25 Gamma(5, .5) + 0.75 Gamma(24, 0.5), where $\phi_{\mu,\sigma}$ denotes the density of a normal random variable with mean μ and variance σ^2, and Gamma(k, θ) denotes the density of a Gamma random variable with shape parameter k and scale θ. We obtained the contaminated data $W_1, \ldots, W_n$ by taking $W_i = X_i + U_i$, where the $U_i \sim f_U$ were independent random variables, and independent of the X_is. For f_U, in each case, we considered Laplace and normal densities such that the noise to signal ratio NSR $= \mathrm{Var}(U)/\mathrm{Var}(X)$ equals 10% or 25%.

When computing our estimators, for densities (i)–(iii) we used the correct model for $f(\cdot\,|\,\theta)$, that is, we took $f(\cdot\,|\,\theta)$ to be a normal mixture of the form $p\,\phi_{\mu_1,\sigma_1} + (1-p)\,\phi_{\mu_2,\sigma_2}$, and $\theta = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)$. However, since in practice it can be expected that we often use an approximate model, we used a wrong parametric form for the gamma mixture densities (iv) to (vi). More precisely, we modelled (iv) by a normal density, and we modelled (v) and (vi) by the normal mixture form introduced above. Our theory indicates that, in finite sample, using approximate models should lead to

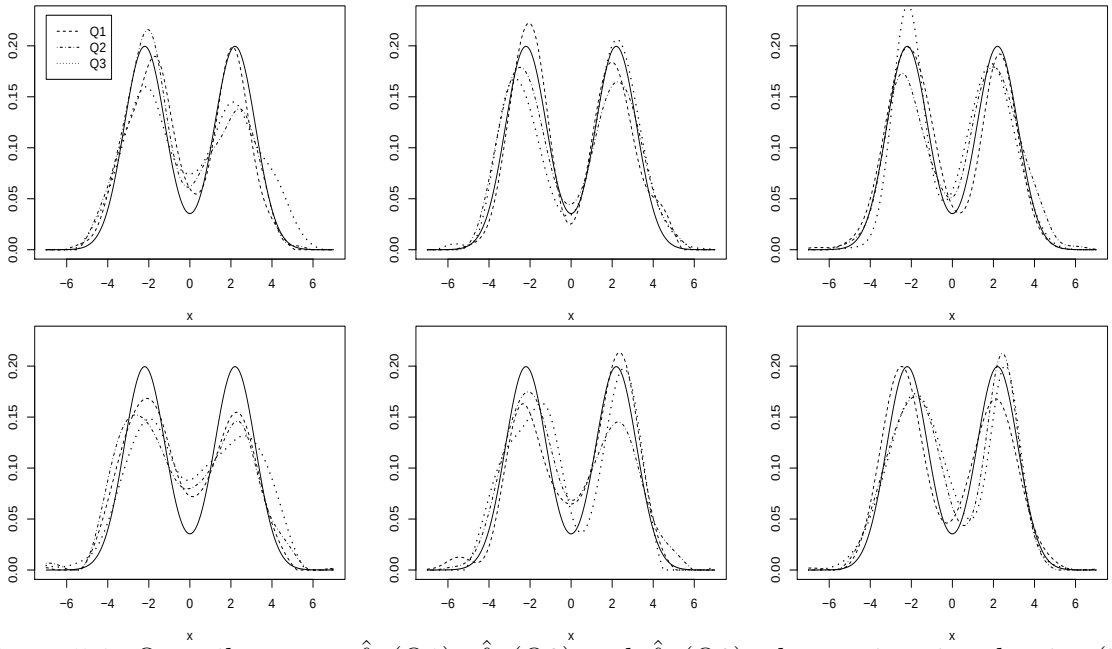

Figure 5.1: Quartile curves $\hat{f}_1$ (Q1), $\hat{f}_2$ (Q2) and $\hat{f}_3$ (Q3) when estimating density (i) with $\hat{f}_{\text{dec}}$ (left), $\hat{f}_{\text{bco}}$ (middle) or $\hat{f}_{\text{wgt}}$ (right), when $n = 200$ and $U \sim$ Laplace with NSR = 10% (1st row) or NSR = 25% (2nd row).

improvements over $\hat{f}_{\text{dec}}$, and we shall see that our simulation results confirm this.

5.2 Simulation results

We applied the estimators $\hat{f}_{\text{dec}}$, $\hat{f}_{\text{rat}}$, $\hat{f}_{\text{bco}}$ and $\hat{f}_{\text{wgt}}$ to the contaminated samples generated for each setting described above. We implemented our three estimators with θ estimated by $\hat{\theta}_P$ (the pilot estimator of section 4.2), $\hat{\theta}_{\text{MD}}$ and $\hat{\theta}_{\text{ML}}$, and found that, due to the variability caused by its ratio form, $\hat{f}_{\text{rat}}$ performed best with $\hat{\theta}_P$, whereas $\hat{f}_{\text{bco}}$ and $\hat{f}_{\text{wgt}}$ performed best with, respectively, $\hat{\theta}_{\text{MD}}$ and $\hat{\theta}_{\text{ML}}$. In the tables below we report results only for these three best combinations.

Let $\hat{f}$ denote any one of these estimators; for each, we computed the 100 integrated squared errors $\text{ISE} = \int (\hat{f} - f_X)^2$ corresponding to the 100 generated samples. In the tables we report the median and inter-quartile range (IQR) of the 100 ISEs for each case. In the figures, for a given method and density f_X, we show the three estimated

curves $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ that resulted in, respectively, the first, second and third quartile of these 100 ISE values, and we refer to them as "the quartile curves". Additional tables, reporting the integrated squared bias of estimators, are provided in Tables D.1 and D.3 in section D of the supplementary file. We also implemented the method of Hazelton and Turlach (2010) with 5 knots, as recommended there. We found that, depending on the cases, their method could either improve on $\hat{f}_{\text{dec}}$, or worsen it significantly. This is not unexpected, since their method is not generally consistent. Given this lack of robustness, we do not discuss their method further.

Table 5.1: Median (IQR) of 100 values of $10^3 \times$ISE obtained when estimating densities (i) to (iii) in the Laplace error case, when NSR = 10% or 25% and $n = 200$ or 700, using the estimators $\hat{f}_{\text{dec}}$, $\hat{f}_{\text{rat}}(\cdot; \hat{\theta}_P)$, $\hat{f}_{\text{bco}}(\cdot; \hat{\theta}_{\text{MD}})$ and $\hat{f}_{\text{wgt}}(\cdot; \hat{\theta}_{\text{ML}})$.

	Density (i)		Density (ii)		Density (iii)	
	$n = 200$	$n = 700$	$n = 200$	$n = 700$	$n = 200$	$n = 700$
	NSR = 10%					
$\hat{f}_{\text{dec}}$	7.24 (4.83)	3.55 (1.55)	4.63 (2.89)	2.44 (1.42)	7.79 (5.61)	3.91 (2.38)
$\hat{f}_{\text{rat}}(\cdot; \hat{\theta}_P)$	4.86 (2.86)	2.03 (1.29)	3.93 (3.06)	1.33 (1.11)	5.77 (4.24)	2.49 (1.76)
$\hat{f}_{\text{bco}}(\cdot; \hat{\theta}_{\text{MD}})$	4.24 (3.57)	1.91 (1.56)	4.33 (2.58)	1.56 (1.25)	6.57 (3.78)	2.96 (1.96)
$\hat{f}_{\text{wgt}}(\cdot; \hat{\theta}_{\text{ML}})$	3.66 (3.07)	1.07 (1.22)	3.18 (2.76)	1.03 (1.04)	5.58 (4.94)	1.44 (1.36)
	NSR = 25%					
$\hat{f}_{\text{dec}}$	11.3 (9.14)	5.78 (3.39)	6.26 (4.29)	3.68 (1.99)	10.9 (7.41)	5.76 (3.27)
$\hat{f}_{\text{rat}}(\cdot; \hat{\theta}_P)$	8.78 (6.32)	4.63 (2.42)	5.41 (4.20)	2.77 (1.62)	9.79 (6.87)	4.64 (2.55)
$\hat{f}_{\text{bco}}(\cdot; \hat{\theta}_{\text{MD}})$	6.38 (5.89)	2.67 (2.17)	5.27 (3.81)	3.02 (1.75)	10.2 (6.39)	5.09 (3.27)
$\hat{f}_{\text{wgt}}(\cdot; \hat{\theta}_{\text{ML}})$	6.77 (7.26)	1.95 (2.27)	4.07 (4.60)	1.54 (1.80)	9.04 (9.15)	2.66 (2.92)

Table 5.1 shows the median and IQR of the 100 ISEs for estimating densities (i) to (iii) with $\hat{f}_{\text{dec}}$, $\hat{f}_{\text{rat}}(\cdot; \hat{\theta}_P)$, $\hat{f}_{\text{bco}}(\cdot; \hat{\theta}_{\text{MD}})$ and $\hat{f}_{\text{wgt}}(\cdot; \hat{\theta}_{\text{ML}})$, when the errors are Laplace, $n = 200$ or $n = 700$, and NSR = 10% or NSR = 25%. These results show that, when the parametric model is correctly specified, the parametrically assisted estimators can considerably improve the purely nonparametric estimator $\hat{f}_{\text{dec}}$, sometimes reducing the median ISE value by more than 50%. Depending on the case, each of $\hat{f}_{\text{bco}}$ and $\hat{f}_{\text{wgt}}$ can be the most competitive estimator, but in all cases presented in the table, $\hat{f}_{\text{rat}}$ performs similarly to, or worse than, $\hat{f}_{\text{bco}}$. Therefore, given the complexity of

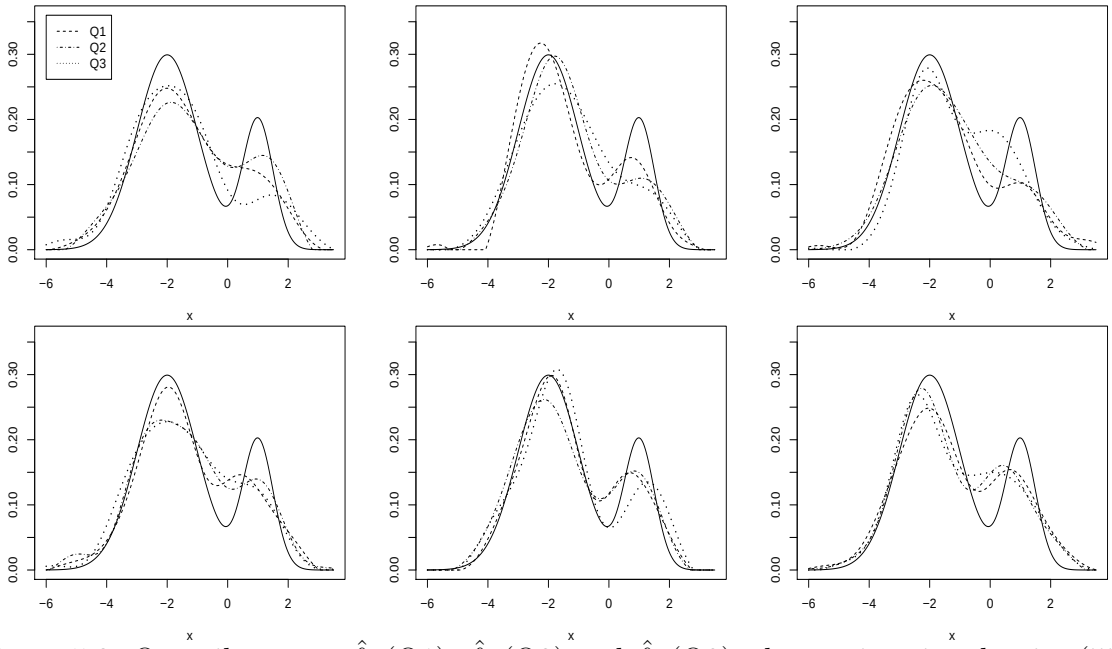

Figure 5.2: Quartile curves $\hat{f}_1$ (Q1), $\hat{f}_2$ (Q2) and $\hat{f}_3$ (Q3) when estimating density (iii) with $\hat{f}_{\text{dec}}$ (left), $\hat{f}_{\text{bco}}$ (middle) or $\hat{f}_{\text{wgt}}$ (right), when $U \sim N(0, \sigma^2)$ with NSR = 25% and $n = 200$ (1st row) or $n = 300$ (2nd row).

computing $\hat{f}_{\text{rat}}$ (see section 4.3), even in the Laplace error case, this estimator seems much less attractive than the other two, and we did not implement it for other errors.

Table 5.2 shows the results for the estimators $\hat{f}_{\text{dec}}$, $\hat{f}_{\text{bco}}(\,\cdot\,; \hat{\theta}_{\text{MD}})$ and $\hat{f}_{\text{wgt}}(\,\cdot\,; \hat{\theta}_{\text{ML}})$ when the errors are normal. It leads to similar conclusions to Table 5.1, although in this case, for density (iii), $\hat{f}_{\text{wgt}}(\,\cdot\,; \hat{\theta}_{\text{ML}})$ was outperformed by $\hat{f}_{\text{dec}}$. This is because estimating θ is much more difficult for supersmooth errors (e.g. normal errors) than for ordinary smooth errors, particularly when f_X has peaks of unequal size, like density (iii). Now, poorer estimates of θ affect $\hat{f}_{\text{wgt}}$ in two ways: the parametric part $f(x \,|\, \hat{\theta}_{\text{ML}})$ is less good, and the weight $\widehat{w}$, which depends on $\hat{\theta}_{\text{ML}}$, is closer to zero. This implies that, in the normal error case (and more generally in supersmooth error cases), $\hat{f}_{\text{wgt}}$ can perform more poorly, and can be even a little worse than $\hat{f}_{\text{dec}}$. See also Table D.2 in section D of the supplementary file, where we compare the weight $\widehat{w}$ used by $\hat{f}_{\text{wgt}}$ in the Laplace and normal error cases, for densities (i) to (iii).

Table 5.2: Median (IQR) of 100 values of $10^3 \times$ISE obtained when estimating densities (i) to (iii) in the normal error case, when NSR = 10% or 25% and $n = 200$ or 700, using the estimators $\hat{f}_{\rm dec}$, $\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$ and $\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$.

	Density (i)		Density (ii)		Density (iii)	
	$n = 200$	$n = 700$	$n = 200$	$n = 700$	$n = 200$	$n = 700$
	NSR = 10%					
$\hat{f}_{\rm dec}$	8.74 (5.56)	4.84 (2.77)	5.03 (3.62)	2.53 (1.57)	8.98 (4.94)	5.42 (2.50)
$\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$	4.51 (3.61)	2.11 (1.71)	4.31 (4.12)	1.61 (1.45)	6.40 (4.34)	2.98 (2.04)
$\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$	5.59 (4.83)	2.84 (2.13)	3.60 (3.19)	1.76 (1.03)	9.51 (5.35)	5.59 (2.53)
	NSR = 25%					
$\hat{f}_{\rm dec}$	19.0 (11.0)	12.6 (6.32)	8.01 (4.59)	5.61 (3.33)	15.9 (9.12)	10.7 (5.84)
$\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$	15.2 (10.8)	4.98 (6.22)	7.34 (6.18)	4.14 (3.68)	14.2 (11.0)	6.24 (4.79)
$\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$	11.3 (7.64)	6.20 (4.15)	5.20 (3.81)	2.93 (2.54)	16.6 (7.82)	11.2 (5.01)

Table 5.3: Median (IQR) of 100 values of $10^3 \times$ISE obtained when estimating densities (iv) to (vi) in the normal error case, when NSR = 10% or 25% and $n = 200$ or 700, using the estimators $\hat{f}_{\rm dec}$, $\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$ and $\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$.

	Density (iv)		Density (v)		Density (vi)	
	$n = 200$	$n = 700$	$n = 200$	$n = 700$	$n = 200$	$n = 700$
	NSR = 10%					
$\hat{f}_{\rm dec}$	4.96 (3.97)	2.74 (1.43)	3.17 (2.25)	1.61 (0.95)	4.94 (2.10)	2.81 (1.23)
$\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$	3.68 (3.16)	2.16 (1.39)	1.80 (1.58)	0.88 (0.74)	4.32 (2.66)	1.82 (1.21)
$\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$	5.10 (3.99)	2.78 (1.44)	3.20 (2.31)	1.66 (0.97)	5.30 (2.33)	2.95 (1.25)
	NSR = 25%					
$\hat{f}_{\rm dec}$	7.42 (5.49)	5.54 (2.85)	6.25 (3.96)	3.80 (2.46)	7.46 (3.32)	5.38 (2.03)
$\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$	5.56 (5.10)	4.02 (2.09)	3.93 (3.87)	1.65 (1.26)	7.45 (3.48)	4.32 (2.23)
$\hat{f}_{\rm wgt}(\,\cdot\,;\hat{\theta}_{\rm ML})$	7.49 (5.20)	5.69 (2.80)	6.20 (4.74)	3.63 (2.43)	8.26 (3.10)	5.64 (2.00)

In Figure 5.1 we show the quartile curves $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ obtained when estimating density (i) with $\hat{f}_{\rm dec}$, $\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$ when $n = 200$, and NSR = 10% or 25%. Figure 5.2 depicts those quartile curves in the case where f_X is density (iii), the errors are normal, NSR = 25% and $n = 200$ or $n = 300$. The figures illustrate the significant improvement that the estimator $\hat{f}_{\rm bco}(\,\cdot\,;\hat{\theta}_{\rm MD})$ can offer over $\hat{f}_{\rm dec}$. For example, it is able to recover the peaks of f_X more successfully than $\hat{f}_{\rm dec}$. As expected from our theory, the estimator $\hat{f}_{\rm wgt}$ also improves on $\hat{f}_{\rm dec}$, but in a less impressive manner than $\hat{f}_{\rm bco}$. As usual, and as also illustrated in the tables, we see that all estimators improve as the sample size n increases and when the NSR decreases.

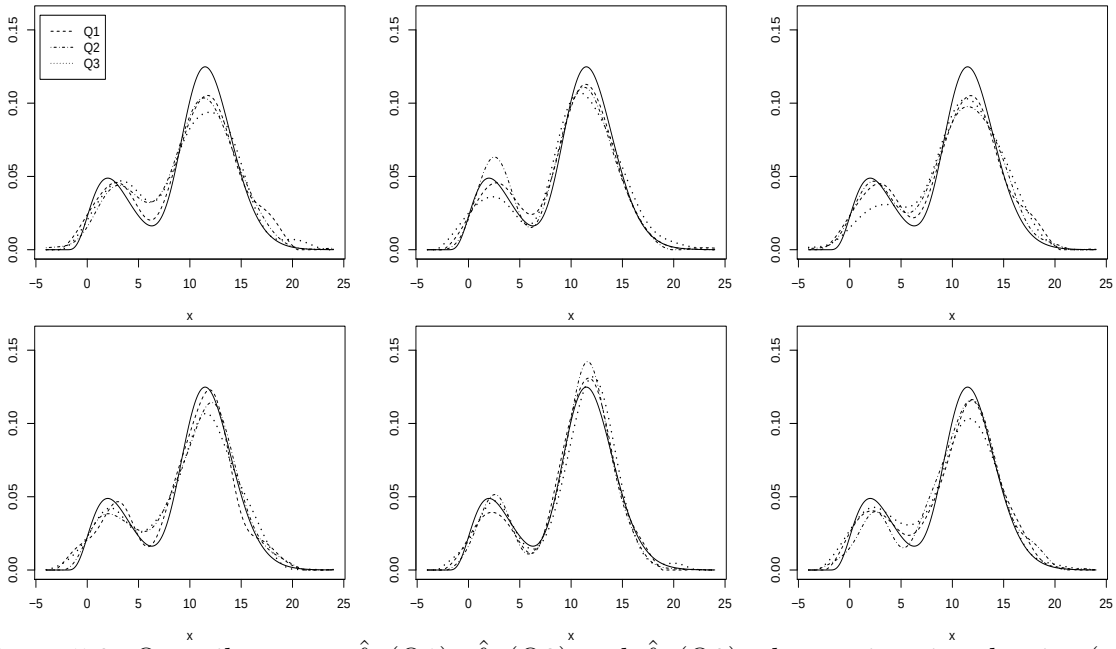

Figure 5.3: Quartile curves $\hat{f}_1$ (Q1), $\hat{f}_2$ (Q2) and $\hat{f}_3$ (Q3) when estimating density (v) with $\hat{f}_{\text{dec}}$ (left), $\hat{f}_{\text{bco}}$ (middle) or $\hat{f}_{\text{wgt}}$ (right), when $U \sim N(0, \sigma^2)$ with NSR = 10% and $n = 200$ (1st row) or $n = 300$ (2nd row).

Finally, in Table 5.3 we illustrate the robustness of our estimators against misspecification of the parametric model $f(\cdot \,|\, \theta)$. The table reports results for estimating the gamma density (density (iv)) and the mixtures of two gamma densities (densities (v) and (vi)), assuming that $f(\cdot \,|\, \theta)$ is equal to, respectively, a normal density or a mixture of two normal densities. Clearly, a normal density, which is symmetric, is not a very good approximation to density (iv), which is asymmetric, but a normal mixture is a reasonable approximation to densities (v) and (vi). In other words, to make the connection with our theory, η is rather large for density (iv), whereas it is rather small for densities (v) and (vi). In both cases η is fixed, but in finite samples, a small η can be interpreted as tending to zero fast, and a large η can be interpreted as tending to zero slowly. Therefore, from the theory, we can expect that in case (iv), $\hat{f}_{\text{bco}}$ will perform significantly better than $\hat{f}_{\text{wgt}}$, and this is confirmed by the results shown in the table. Moreover, even though $f(\cdot \,|\, \theta)$ is misspecified, $\hat{f}_{\text{bco}}$ can still im-

prove on $\hat{f}_{\text{dec}}$ quite significantly in many cases, without degrading it much in other cases. As in Table 5.2, and for the same reasons as there, here $\hat{f}_{\text{wgt}}$ was often not able to improve the results of $\hat{f}_{\text{dec}}$.

To illustrate these properties visually, Figure 5.3 shows the quartile curves obtained in the case where f_X was density (v), the errors were normally distributed with NSR = 10%, and sample size n was 200 or 300. We can see that, even though $f(\cdot \,|\, \theta)$ is misspecified, the estimated normal mixtures help $\hat{f}_{\text{bco}}$ recover the location and height of the peaks of f_X, which in turns leads $\hat{f}_{\text{bco}}$ to noticeably improve $\hat{f}_{\text{dec}}$. In this misspecified case, as anticipated by our theoretical results, the estimator $\hat{f}_{\text{wgt}}$ does not improve on $\hat{f}_{\text{dec}}$. Of course, performance improves as n increases.

5.3 Conclusion of simulations

To summarise, our numerical results illustrate that the estimators $\hat{f}_{\text{bco}}(\,\cdot\,; \hat{\theta}_{\text{MD}})$ and $\hat{f}_{\text{wgt}}(\,\cdot\,; \hat{\theta}_{\text{ML}})$ can both considerably improve on the deconvolution estimator $\hat{f}_{\text{dec}}$. Depending on the case, each of those two estimators can outperform the other, but, as anticipated by our theory, $\hat{f}_{\text{bco}}$ is more robust to misspecification of the parametric model $f(\cdot \,|\, \theta)$; this is our favoured estimator. Although it has good asymptotic properties, $\hat{f}_{\text{rat}}$ is not as competitive as $\hat{f}_{\text{bco}}$, in part due to the difficulty of computing it.

6 Supplemental Materials

Appendix B contains technical details, appendix C contains details of implementation, and appendix D contains additional simulation results.

Acknowledgements

We thank Dr Mihee Lee for her help with early versions of codes and for useful discussion. We also thank the Australian Research Council for grant support, and

the Institute for Mathematical Sciences of the National University of Singapore for their support.

References

Burman, P. and Chaudhuri, P. (2011). On a hybrid approach to parametric and nonparametric regression. In Jiang, J., Roussas, G.G. and Samaniego, F.J. (Eds.) *Nonparametric statistical methods and related topics : a Festschrift in honor of Professor P K Bhattacharya on the occasion of his 80th Birthday.* World Scientific.

Cao, R., Cuevas, A. and Fraiman, R. (1995). Minimum distance density-based estimation. *Comput. Stat. Data Anal.*, **20** 611–631.

Carroll, R.J., Delaigle, A., Hall, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *J. Amer. Statist. Assoc.*, **106** 191–202.

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

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models*. 2nd ed. Chapman & Hall.

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., 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). Practical bandwidth selection in deconvolution kernel density estimation. *Comput. Stat. Data Anal.*, **45** 249–267.

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

Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, **14**, 562–579.

Diggle, P.J. and Hall, P. (1993). Fourier approach to nonparametric deconvolution of a density estimate. *J. R. Statist. Soc. B*, **55** 523–531.

Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. R. Statist. Soc.* Ser. B, **60** 709–724.

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

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

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

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

Fan, Y. and Ullah, A. (1999). Asymptotic normality of a combined regression estimator. *J. Mult. Anal.*, **71**, 191–240.

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.

Hazelton, M. L. and Turlach, B. A. (2010). Semiparametric density deconvolution. *Scand. J. Statist.*, **37** 91–108.
Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.*, **23** 882–904.
Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.*, **24** 1619–1647.
Jones, M.C., Linton, O. and Nielsen, J.P. (1995). A simple effective bias reduction method for density and regression estimation. *Biometrika*, **82** 327–338.
Liu, M.C. and Taylor, R.L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Can. J. Statist.*, **17** 427–438.
Masry, E. (1993). Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *J. Mul. Anal.*, **44** 47–68.
Naito, K. (2004). Semiparametric density estimation by local L_2-fitting. *Ann. Statist.*, **32** 1162–1191.
Olkin, I. and Spiegelman, C. H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.*, **82** 858–865.
Staudenmayer, J. and Ruppert, D. (2004). Local polynomial regression and simulationextrapolation. *J. R. Statist. Soc.* Ser. B, **66**, 17–30.
Staudenmayer, J. and Ruppert, D. and Buonaccorsi, J. (2008). Density estimation in the presence of heteroscedastic measurement error. *J. Amer. Statist. Assoc.*, **103**, 726–736.
Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21** 169–184.

A Appendix

A.1 Details of construction of the ratio estimator $\hat{f}_{\text{rat}}$

We wish to find the function $\mathcal{L}_x$ that satisfies (2.4). Since f_U is symmetric, we have

$$E[\mathcal{L}_x(W_j \,|\, \theta, h)|X_j] = \int \mathcal{L}_x(X_j + u \,|\, \theta, h) f_U(u)\, du = \{\mathcal{L}_x(\,\cdot\,|\,\theta, h) * f_U\}(X_j)\,.$$

To satisfy (2.4), we need this to be equal to $\mathcal{K}_x(X_j \,|\, \theta, h)$. In other words, for all x and y we need to find $\mathcal{L}_x$ such that $\{\mathcal{L}_x(\cdot\,|\,\theta, h) * f_U\}(y) = \mathcal{K}_x(y \,|\, \theta, h)$, which is equivalent to

$$\{\phi_{\mathcal{L}_x(\,\cdot\,|\,\theta,h)*f_U}\}(t) = \phi_{\mathcal{L}_x(\,\cdot\,|\,\theta,h)}(t)\,\phi_U(t) = \phi_{\mathcal{K}_x(\,\cdot\,|\,\theta,h)}(t)$$

for all t. We deduce from the Fourier inversion theorem that

$$\mathcal{L}_x(y\,|\,\theta, h) = \frac{1}{2\pi}\int e^{-ity}\frac{\phi_{\mathcal{K}_x(\,\cdot\,|\,\theta,h)}(t)}{\phi_U(t)}\,dt = \frac{1}{2\pi}\int \frac{e^{-ity}}{\phi_U(t)}\int e^{itw}\frac{K\left(\frac{x-w}{h}\right)}{h f_X(w\,|\,\theta)}\,dw\,dt\,,$$

which proves that $\mathcal{L}_x$ is of the form given in (2.5). For simplicity, it is assumed here that the Fourier transforms and inverse Fourier transforms are all well defined; see section 3.5 for more detailed assumptions.

A.2 Proof of (3.9) and (3.10)

We shall derive only the part of (3.9) and (3.10) pertaining to the case where $n\eta^2$ diverges. Settings where $n\eta^2$ is bounded are relatively straightforward to treat, and instances that overlap both cases can be dealt with using a subsequence argument.

To derive these results, observe that if (3.6) and (A2)–(A4) hold,

$$\hat{f}_{\mathrm{bco}}(x \,|\, \hat{\theta}) = \hat{f}_{\mathrm{bco}}(x \,|\, \theta_1) + O_p\big(n^{-1/2}\big)\,; \tag{A.1}$$

note that if (3.8) obtains for $\ell_0 = \beta$,

$$E\big\{\hat{f}_{\mathrm{bco}}(x \,|\, \theta_1)\big\} - f_X(x) = \int \{\psi(x - hu) - \psi(x)\}\, K(u)\, du = O\big(h^\beta\, \eta\big)\,; \tag{A.2}$$

and observe that, using standard arguments,

$$\mathrm{Var}\,\big\{\hat{f}_{\mathrm{bco}}(x \,|\, \theta_1)\big\} = \mathrm{Var}\,\big\{\hat{f}_{\mathrm{dec}}(x \,|\, \theta_1)\big\}\,. \tag{A.3}$$

In the ordinary smooth case, result (3.9) follows directly from (A.1)–(A.3), since in that case we have $\mathrm{Var}\,\big\{\hat{f}_{\mathrm{dec}}(x \,|\, \theta_1)\big\} = O\big\{(nh^{2\alpha+1})^{-1}\big\}$; see Fan (1991a).

In the supersmooth case, condition (A3) can be used to prove that $\mathrm{Var}\{\hat{f}_{\mathrm{dec}}(x \,|\, \theta_1)\} = O\{n^{-1}\, h^{2(\kappa+1)\alpha+2\alpha_1}\, \exp(2h^{-\alpha}/\gamma)\}$, where α_1 is as in (3.2) and κ is as in (A3); see section **??** in the supplementary file for a proof. Combining these bounds for bias and variance, respectively, we deduce that the mean squared error of $\hat{f}_{\mathrm{bco}}(x \,|\, \theta_1)$ is given by

$$\begin{aligned} E\big\{\hat{f}_{\mathrm{bco}}(x \,|\, \theta_1) - f_X(x)\big\}^2 &= O\Big\{h^{2\beta}\, \eta^2 + n^{-1}\, h^{2(\kappa+1)\alpha+2\alpha_0}\, \exp\big(2h^{-\alpha}/\gamma\big)\Big\} \\ &= O\Big[\eta^2 \Big\{h^{2\beta} + \big(n\eta^2\big)^{-1}\, h^{2(\kappa+1)\alpha+2\alpha_0}\, \exp\big(2h^{-\alpha}/\gamma\big)\Big\}\Big]\,. \end{aligned}$$

Since $n\eta^2$ diverges as $n \to \infty$, $h^{2\beta} + (n\eta^2)^{-1}\, h^{2(\kappa+1)\alpha+2\alpha_0}\, \exp(2h^{-\alpha}/\gamma)$ is asymptotic to the mean squared error of the deconvolution kernel estimator $\hat{f}_{\mathrm{dec}}$ for sample size

$n\eta^2$, which implies that the optimal rate of convergence to zero of this quantity is obtained by taking a bandwidth $h = c\,\{\log(n\eta^2)\}^{-1/\alpha}$, where $c > (2/\gamma)^{1/\alpha}$. It follows that the square root of the minimum of our bound to mean squared error is of order $\eta\,\{\log(n\eta^2)\}^{-\beta/\alpha}$, which proves (3.10).

A.3 Proof of Theorem 1

The result can be proved along the lines of the proof of Theorem 5 in Fan (1991a), taking, as there, $f_0(x) = C_r\,(1 + x^2)^{-r}$ for some $r > 1/2$, and replacing Fan's f_n by $f_n = f_0 + c\,\eta\,\delta^\beta\,H(x/\delta)$, with H as in Fan (1991a). In particular, H is a fixed, non-degenerate function that has β bounded derivatives, and is such that $\int H = 0$, $\sup_x\,(1 + |x|)^{m+1}\,|H(x)| < \infty$ for some $m \geq 2$, and ϕ_H (the Fourier transform of H) vanishes outside $[1, 2]$. The quantities $(\beta, \alpha, m, m_0, \delta_n)$ in Fan (1991a) are, in our case, $(\alpha, 0, \beta, m + 1, \delta)$, respectively.

As in Fan (1991a), without loss of generality, we take $x_0 = 0$. Since $f_0(x) > 0$ in a neighborhood of x_0, then for all sufficiently small values of $\eta\,\delta^\beta$, and hence for sufficiently small C_4 in (3.16); and for all sufficiently small c; $f_n \geq 0$ on the real line. Therefore f_n is a proper density function. The functions ψ_1 and ψ_2 referred to in Theorem 1 are identical to 0 and $c\,\eta\,\delta^\beta\,H(x/\delta)$, respectively. They satisfy (3.15). The argument used to prove Theorem 5 of Fan (1991a) gives the bounds $I_1 = O(\eta^2\,\delta^{2\alpha})$ and $I_2 = O(\eta^2\,\delta^{2\alpha})$ in place of the bounds derived by Fan (1991a); the quantities I_1 and I_2 are as defined there. Hence, $\delta^{2\beta+1}\,(I_1 + I_2) = O(\delta^{2\alpha+2\beta+1}\,\eta^2) = O(n^{-1})$ if we take $\delta = \delta(n) = (n\,\eta^2)^{-1/(2\alpha+2\beta+1)}$. Theorem 1 now follows as in Fan (1991a).

A.4 Discussion of assumptions in section 3.5, and examples illustrating (B6)

The properties assumed in (B2) are satisfied by kernels used in practice, which are such that either ϕ_K is bounded and compactly supported, or K is a Gaussian density.

The assumptions $h \to 0$ and $nh^{2\alpha+1} \to \infty$, imposed in (B3) except in the special case $c = 0$, are generally necessary and sufficient for the bias and variance, respectively, of the standard deconvolution estimator $\hat{f}_{\rm dec}$, to converge to zero, and so are mild. Since (B4) and (B5) refer to the parametric model $f(\,\cdot\,|\,\theta)$, not the actual density of X; and since a degree of smoothness, with respect to both x and θ, is often appropriate for the model; then (B4) and (B5) are reasonable. We discuss (B6) below. Assumptions and results for the error-free case are obtained by simply taking $\alpha = 0$ throughout.

In the following examples we define $\omega^{(\ell)}(w, x\,|\,\theta) = (\partial/\partial w)^\ell\, \omega(w, x\,|\,\theta)$, and show that (B6) holds in a wide variety of cases.

EXAMPLE 1. Assume that

$$\max_{\ell=0,\ldots,s} \left|\omega^{(\ell)}(w, x\,|\,\theta)\right| \le C_7\,(1 + |w|)^{\alpha_1}\,, \tag{A.4}$$

uniformly in $-\infty < w < \infty$, $x \in \mathcal{I}$ and $\theta \in \Theta$, where $C_7, \alpha_1 > 0$. Condition (A.4) typically holds when $f(x\,|\,\theta)$ is a smooth function of x and has tails that are polynomially light. In such cases we can take K to be a conventional kernel function used in deconvolution problems, for which ϕ_K is compactly supported and

$$\max_{\ell=0,\ldots,s} \left|K^{(\ell)}(w)\right| \le C_8\,(1 + |w|)^{-\alpha_2}\,, \tag{A.5}$$

for all w, where $C_8, \alpha_2 > 0$ and α_2 can be taken as large as needed by choosing K appropriately. If (A.4) and (A.5) hold and $\alpha_2 > \alpha_1 + r + 2$ then, for $\ell = 0, \ldots, s$,

$$\sup_{x\in\mathcal{I}} \sup_{\theta\in\Theta} \left|\gamma_{h,x}^{(\ell)}(u\,|\,\theta)\right| \le C_9\,(1 + |u|)^{\alpha_1+r+1-\alpha_2} = C_9\,(1 + |u|)^{-1-c}\,,$$

where $c > 0$. Therefore (B6) holds.

EXAMPLE 2. In place of (A.4), assume that

$$\max_{\ell=0,\ldots,s} \left|\omega^{(\ell)}(w, x\,|\,\theta)\right| \le C_9\, \exp\left(\alpha_3\, w^2\right), \tag{A.6}$$

uniformly in $-\infty < w < \infty$, $x \in \mathcal{I}$ and $\theta \in \Theta$, where $C_9, \alpha_3 > 0$. Condition (A.6) holds when $f(\,\cdot\,|\,\theta)$ is a Gaussian density, and more generally if $f(\,\cdot\,|\,\theta)$ represents an exponential family. Here we take K to be the normal $\mathrm{N}(0, \tau^2)$ density, i.e. $K(w) = \tau^{-1}\,(2\pi)^{-1/2}\,\exp(-\frac{1}{2}\,\tau^{-2}\,w^2)$. In this case, (B6) is valid provided that $\tau^{-2} > 2\,\alpha_3$.

A.5 Discussion of conditions for Theorems 1 and 2

In framing the lower bounds in Theorems 1 and 2 we assume that the model is known completely, not only up to an unknown parameter θ. Assumption (3.14) merely asserts that fact, taking the density to be f_0 and adding the minor additional constraint that f_0 is bounded away from zero in the vicinity of the point x_0 at which we are going to estimate f_X. (This serves only to ensure that if $\psi = f_X - f_0$ is small in absolute terms, then it is also small relative to the value taken by f_0 near x_0.) Taking the model to be known completely does not weaken the impact of the lower bounds, since any statistically interesting formulation of the minimax lower bounds would have to encompass that case.

Next we discuss conditions (3.15) and (3.16). Recall that our goal is to estimate the density f_X using guidance from f_0. If f_0 is close to f_X ($\psi = f_X - f_0$ is small), then we expect that f_0 can help us improve purely nonparametric estimators, and we hope to be able to construct an estimator that has better convergence rates than a purely nonparametric estimator. However, if f_0 is far from f_X (ψ is large), then we expect that knowing f_0 won't help us improve nonparametric estimators of f_X, but we hope that incorporating f_0 in the estimation procedure won't degrade the convergence rates of a purely nonparametric estimator.

The sequence $\eta = \eta(n)$, in equations (3.15) and (3.16), is used to bound the order of magnitude of the function ψ. The discussion above shows that we should consider both the case where $\eta \to 0$ as $n \to \infty$ (ψ small) and the case where $\eta \not\to 0$ as $n \to \infty$ (ψ large), whence condition (3.16). Clearly, the convergence rates of an estimator

that uses f_0 in the right way should depend on η: the smaller η, the closer f_0 is to f_X, and so the faster the rates should be; this is what Theorem 1 expresses. Since f_X is assumed to have β derivatives, then in (3.15) we do not assume bounds just on ψ, but also on its first β derivatives. This merely strengthens the first property in (3.8) by asking that the bound to the derivative of largest order also applies to derivatives of lower order. (As noted in discussion prior to (3.8), this strengthening is not necessary when deriving upper bounds to convergence rates, but it turns out to be useful for the lower bounds in Theorems 1 and 2.) The additional conditions imposed in (3.15), specifically $\int |\psi| < \infty$ and $\int \psi = 0$, are mentioned only for clarity; they follow directly from the fact that f_0 and $f_0 + \psi$ are both proper density functions.

Note that (3.16) asserts that η should be bounded and, in addition, not any smaller than a constant multiple of $n^{-1/2}$. The upper bound here is merely a reflection of the fact that we want the densities f_0 and $f_0 + \psi$ to have uniformly bounded derivatives up to the βth; in this case there is clearly no need to permit η to be unbounded. The lower bound asserts only that we are not treating problems of superefficiency, where we would estimate densities at rates that are strictly faster than the conventional parametric one of $n^{-1/2}$. For example, if we were to allow η to converge to zero more rapidly than $n^{-1/2}$, the quantity $(\eta^{2\alpha+1}/n^{\beta})^{1/(2\alpha+2\beta+1)}$, appearing in the probability statement on the left-hand side of (3.18), would converge to zero faster than $n^{-1/2}$, and so the difference between the estimator $\phi(W_1, \ldots, W_n)$ and its target $f_X(x_0)$ would also converge to zero at this superefficient rate.

Finally, the class Ψ_n is the class of functions ψ discussed above; it depends on n through η. Since each ψ corresponds to a function f_X, Ψ_n also corresponds to a class of functions f_X. Minimax rates are usually derived uniformly over a class of functions, which is why we have introduced Ψ_n.

A.6 Discussion of condition on f_X

If the data distribution is kept fixed as n diverges then the asymptotic theory is degenerate, and in particular is uninformative. There are just two possible asymptotic regimes: either the data distribution coincides exactly with the parametric assumption, and the estimator is root-n consistent, or the data distribution is different from that assumption, and the estimator has exactly the same first-order theoretical properties as a conventional nonparametric estimator. However, this sharply dichotomous performance does not correctly reflect performance in practice, where as the data distribution moves steadily closer to the model, for fixed sample size, the performance of the parametrically assisted estimator steadily and gradually improves; it does not change sharply.

Of course, this evolutionary change is not unexpected. The challenge is not to explain it, and why conventional theory does not predict it. (Since conventional theory is predicated on the parametric model being either correct or wrong, it can be expected to imply sharply dichotomous performance.) Rather, the challenge is to develop theory that correctly reflects the continuous evolution in performance observed in practice, as the data distribution moves closer to the parametric model. Therefore we allow the data distribution to depend on n, and to converge to the model at a certain rate as sample size increases. The performance of our method then depends on that rate of convergence. See Fan and Ullah (1999), Hall and Simar (2002), Staudenmayer and Ruppert (2004), Staudenmayer, Ruppert and Buonaccorsi (2008) and Burman and Chaudhuri (2011) for related work of this type.