ON DECONVOLUTION WITH REPEATED MEASUREMENTS

Aurore Delaigle[1,2] Peter Hall[2,3] Alexander Meister[2,4]

ABSTRACT. In a large class of statistical inverse problems it is necessary to suppose that the transformation that is inverted is known. Although, in many applications, it is unrealistic to make this assumption, the problem is often insoluble without it. However, if additional data are available then it is possible to estimate consistently the unknown error density. Data are seldom available directly on the transformation, but repeated, or replicated, measurements increasingly are becoming available. Such data consist of "intrinsic" values that are measured several times, with errors that are generally independent. Working in this setting we treat the nonparametric deconvolution problems of density estimation with observation errors, and regression with errors in variables. We show that, even if the number of repeated measurements is quite small, it is possible for modified kernel estimators to achieve the same level of performance they would if the error distribution were known. Indeed, density and regression estimators can be constructed from replicated data so that they have the same first-order properties as conventional estimators in the known-error case, without any replication, but with sample size equal to the sum of the numbers of replicates. Practical methods for constructing estimators with these properties are suggested, involving empirical rules for smoothing-parameter choice.

KEYWORDS. Bandwidth choice, density estimation, errors in variables, Fourier inversion, kernel methods, nonparametric regression, rates of convergence, ridge parameter, replication, statistical smoothing.

SHORT TITLE. Deconvolution.

AMS 2000 SUBJECT CLASSIFICATIONS. Primary 62G07, 62G08; secondary 65R32.

[1] Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

[2] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[3] Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia

[4] Institut für Stochastik und Anwendungen, Universität Stuttgart, D-70569 Stuttgart, Germany

1. INTRODUCTION

Statistical deconvolution problems arise in a great many settings, and typically have the form: $g = T(f)$, where g is a function about which we have data, T is a transformation, and $f = T^{-1}(g)$ is a function we wish to estimate. In a large class of such problems, including density deconvolution and errors-in-variables regression, it is common to assume that T is known. Indeed, the nature of the data usually precludes any other approach.

In this paper we consider cases where there is a small number replications of each intrinsically different observation, the observation errors being independent and the intrinsic parts of the observations being the same among replicates. Data of this type are numerous, and increasingly are becoming available in various fields. Examples include work of Jaech (1985), who describes an experiment where the concentration of uranium is measured for several fuel pellets; of Biemer et al. (1991), who discuss repeated observations in a social science context; of Andersen et al. (2003), on nuclear magnetic reasonance; of Bland and Altman (1986), on lung function; of Eliasziw et al. (1994), on physiotherapy for the knee; of Oman et al. (1999), relating to kidney function; and of Dunn (1989), a brain-related study. For further medical examples, see Carroll et al. (1995) and Dunn (2004).

When data of this type are available, it is usually possible to construct consistent estimators of the function f of interest, without making parametric assumptions about the transformation T. We treat both density deconvolution and errors-in-variables regression, focusing on cases where the convergence rate, and first-order properties more generally, are the same when the error distribution is known and when it is not known but is estimated from repeated measurements. In section 2 we construct a relatively simple density estimator and generalise it to the regression case.

Theoretical properties of our estimators are taken up in section 3. We show that a sufficient condition for first-order properties of estimators, in the cases of known and unknown error distributions, to be equivalent, is that, colloquially speaking, "the target density is smoother than half a derivative of the error density." Instances

where this condition is violated are those where the convergence rate is relatively poor, even when the error density is known.

We direct attention to examples where the number of replications of each observation is relatively small. (We use the terms "replications" and "repeated measurements" synonymously.) In theoretical terms, this means that the number of replications is uniformly bounded. That is generally the case in practice, since gathering large numbers of replications is expensive. Moreover, particularly in cases where statistical performance is the same when the error density is known or unknown, it is seldom advantageous to have large numbers of replications.

For instance, we show that if the total number of data is $M = np$, where $p \geq 2$ equals the number of times that each of n intrinsically different observations is replicated, then first-order properties of nonparametric estimators depend only on M, not on the separate values of n and p. We prove this result rigorously when p is bounded, but a similar argument shows that it is also valid if p diverges sufficiently slowly as M increases. More generally, the result holds if $M = \sum_j N_j$ where N_j is the number of replicates of the jth intrinsically different observation. Properties of the estimator depend, to first order, only on M, provided that each $N_j \geq 2$.

In section 4 we develop an adaptive, data driven procedure for smoothing-parameter choice, and show that it enjoys good performance for real and simulated datasets.

Related work in the context of density estimation includes that of Li and Vuong (1998), who derived upper bounds to convergence rates in the measurement-error problem when replications are present. Li and Vuong's results are important; they comprise some of the first contributions to density deconvolution in cases where the error distribution is not known. Nevertheless, the properties reported by Li and Vuong (1998), and bounds given also by Susko and Nadon (2002), are too coarse to permit it to be shown that convergence rates can be identical in the cases of known and unknown error distributions. Further discussion is given in section 3.5.

Recent, related research in the regression setting, and in the econometrics literature, includes that of Li (2002), Li and Hsiao (2004) and Schennach (2004a,b), who

demonstrated that replications can be used to good effect in regression problems with measurement error. See also work of Horowitz and Markatou (1996) on error estimation from panel data, and the extensive literature, accessible through work of Newey and Powell (2003), on inference in the context of instrumental variables. However, except in parametric contexts, this and related work is not sufficiently detailed to show that the convergence rates familiar in problems where the error distribution is known can also be enjoyed when the distribution is accessible only via repeated measurements.

The problem of density estimation with unknown error density, estimated from a sample of the error, has been considered by Diggle and Hall (1993), Barry and Diggle (1995) and Neumann (1997). Madansky (1959), Carroll et al. (1993) and Huang and Yang, among others, have discussed linear regression with replicated data, when at least some of the predictors are measured with error. Early work on the problem of density deconvolution, under the assumption of known distribution of measurement error, includes that of Carroll and Hall (1988), Stefanski and Carroll (1990) and Fan (1991). More recent contributions, including surveys of earlier research, include the papers of Delaigle and Gijbels (2002, 2004) and van Es and Uh (2005). The literature on kernel methods for errors-in-variables regression is particularly large, and is surveyed by Carroll et al. (1995).

2. MODELS AND METHODOLOGY

2.1. Density deconvolution. Suppose we observe

$$W_{jk} = X_j + U_{jk} \quad \text{for} \quad 1 \le k \le N_j \quad \text{and} \quad 1 \le j \le n\,, \tag{2.1}$$

where the random variables X_j are identically distributed as X, the U_{jk}'s are identically distributed as U, and the X_j's and U_{jk}'s are totally independent. We wish to estimate the density of X. In the context of our discussion in section 1, (2.1) indicates that there are n subsets of "intrinsically different" data, and, within the jth of these subsets, N_j repeated, or replicated, measurements of the variable X_j.

Let f_U and f_X denote the respective densities of U and X, and write $f_U^{\rm Ft}$ and $f_X^{\rm Ft}$ for the respective characteristic functions (i.e. the Fourier transforms of those

densities). Provided that

$$f_U^{\mathrm{Ft}} \text{ is real-valued and does not vanish at any point on the real line}\,, \tag{2.2}$$

a consistent estimator of f_U^{Ft} is given by

$$\hat{f}_U^{\mathrm{Ft}}(t) = \Big| \frac{1}{N} \sum_{j=1}^{n} \sum_{(k_1,k_2)\in\mathcal{S}_j} \cos\{t(W_{jk_1} - W_{jk_2})\} \Big|^{1/2}\,, \tag{2.3}$$

where $\mathcal{S}_j$ denotes the set of $\frac{1}{2}\,N_j(N_j-1)$ distinct pairs (k_1,k_2) with $1 \le k_1 < k_2 \le N_j$, $N = N(n) = \frac{1}{2}\,\sum_{j\le n}\, N_j\,(N_j-1)$, and we ignore values of j for which $N_j = 1$. Assumption (2.2) is conventional when using kernel methods for density deconvolution; see Stefanski and Carroll (1990) and Fan (1991), for example.

An estimator of f_X is given by

$$\hat{f}_X(x) = \frac{1}{Mh}\, \sum_{j=1}^{n}\, w_j \sum_{k=1}^{N_j} \widehat{L}\Big(\frac{x - W_{jk}}{h}\Big)\,,$$

where $M = \sum_j\, N_j$, the weights w_j are nonnegative and satisfy $\sum_j\, w_j\, N_j = M$,

$$\widehat{L}(u) = \frac{1}{2\pi} \int e^{-itu}\, \frac{K^{\mathrm{Ft}}(t)}{\hat{f}_U^{\mathrm{Ft}}(t/h) + \rho}\, dt\,, \tag{2.4}$$

K is a symmetric kernel function with compactly supported Fourier transform K^{Ft}, $h > 0$ is a bandwidth, and $\rho \ge 0$ is a ridge parameter.

We introduce the ridge only so we can take expectation without concern for fluctuations of the denominator in the integral at (2.4). The ridge would not be necessary if our aim were to develop limit theory for $\hat{f}_X$ that did not involve taking expected values. See section 3.1 for discussion and theory in the case $\rho = 0$.

If f_U were known then, instead of $\hat{f}_X$, we would use the following generalization of the conventional deconvolution estimator:

$$\tilde{f}_X(x) = \frac{1}{Mh}\, \sum_{j=1}^{n}\, w_j \sum_{k=1}^{N_j} L\Big(\frac{x - W_{jk}}{h}\Big)$$

(see e.g. Carroll and Hall, 1988), where

$$L(u) = \frac{1}{2\pi} \int e^{-itu} \, \frac{K^{\mathrm{Ft}}(t)}{f_U^{\mathrm{Ft}}(t/h)} \, dt \, .$$

The bias of $\tilde{f}_X$ does not depend on choice of the weights, and it can readily be shown that the asymptotic variance is minimised by taking each $w_j = 1$. Optimality of this choice persists in the case of regression deconvolution, which we consider in section 2.2.

Therefore, we take each $w_j = 1$ in the work below. In particular, $\hat{f}_X$ and $\tilde{f}_X$ henceforth denotes the estimators

$$\hat{f}_X(x) = \frac{1}{Mh} \sum_{j=1}^{n} \sum_{k=1}^{N_j} \widehat{L}\Big(\frac{x - W_{jk}}{h}\Big) \, , \quad \tilde{f}_X(x) = \frac{1}{Mh} \sum_{j=1}^{n} \sum_{k=1}^{N_j} L\Big(\frac{x - W_{jk}}{h}\Big) \, .$$

Section 3.3 demonstrates that $\hat{f}_X$ is first-order equivalent to $\tilde{f}_X$. For this result, and in the setting of "ordinary-smooth errors" (see (3.1)), the main assumption needed is that f_X be sufficiently smooth relative to f_U. See condition (3.12). Properties of $\tilde{f}_X$ are summarised in section 3.4.

2.2. Errors-in-variables regression. Here the model at (2.1) is extended, so that it addresses data (W_{jk}, Y_j) generated as

$$W_{jk} = X_j + U_{jk} \, , \quad Y_j = g(X_j) + V_j \, , \quad \text{for} \quad 1 \le k \le N_j \quad \text{and} \quad 1 \le j \le n \, , \tag{2.5}$$

where the X_j's, U_{jk}'s and V_j's are identically distributed as X, U and V, respectively, $E(V) = 0$, $E(V^2) < \infty$, and the X_j's, U_{jk}'s and V_j's are totally independent. We wish to estimate the function g.

Define

$$\hat{a}(x) = \frac{1}{Mh} \sum_{j=1}^{n} \sum_{k=1}^{N_j} Y_j \, \widehat{L}\Big(\frac{x - W_{jk}}{h}\Big) \, , \quad \tilde{a}(x) = \frac{1}{Mh} \sum_{j=1}^{n} \sum_{k=1}^{N_j} Y_j \, L\Big(\frac{x - W_{jk}}{h}\Big) \, . \tag{2.6}$$

In the classical case, where f_U is known and each $N_j = 1$, the standard kernel estimator of g is $\tilde{g} = \tilde{a}/\tilde{f}_X$, and of course $\tilde{g}$ is also appropriate in the case of replicated data.

The intuition behind $\tilde{g}$ is that $\tilde{a}$ is a consistent estimator of the function $a = f_X g$. When f_U is not known we can estimate a by $\hat{a}$, and so we can modify $\tilde{g}$ in the manner of section 2.1, estimating g by $\hat{g} = \hat{a}/\hat{f}_X$. We show in section 3.6 that $\hat{g}$ is first-order equivalent to $\tilde{g}$.

3. THEORETICAL PROPERTIES

3.1. Density deconvolution. First we state assumptions. We ask that, for constants $\alpha > 0$ and $C_1 > 1$, and all real t,

$$C_1^{-1} (1 + |t|)^{-\alpha} \le \left| f_U^{\mathrm{Ft}}(t) \right| \le C_1 (1 + |t|)^{-\alpha} . \tag{3.1}$$

This is often referred to as the case of ordinary-smooth errors. The importance of the lower bound in (3.1), in addition to the upper bound (which is conventional when deriving convergence rates), are discussed in section 3.3.

Given $\beta, C_2 > 0$, let $\mathcal{F}(\beta, C_2)$ denote the class of densities f_X for which

$$\sup_{-\infty < t < \infty} (1 + |t|)^{\beta} \left| f_X^{\mathrm{Ft}}(t) \right| \le C_2 .$$

(The class $\mathcal{F}(\beta, C_2)$ is a Fourier analogue of Fan's class $\mathcal{C}_{m,\alpha,B}$ of functions; his $m + \alpha + 1$ is our β.) Let K have the property:

$$\sup |K^{\mathrm{Ft}}| < \infty \text{ and, for some } c > 0,\ K^{\mathrm{Ft}}(t) = 0 \text{ for all } |t| > c. \tag{3.2}$$

The kernels used in deconvolution commonly have this property, and so, while our results can be derived under weaker conditions, there is little motivation for that generalisation.

The theorem below gives an upper bound to pointwise mean-squared distance between $\hat{f}_X$ and $\tilde{f}_X$, uniformly in all points and all densities $f_X \in \mathcal{F}(\beta, C_2)$. In section 3.3 we use that result to show that, if the bandwidth h is chosen so that it gives optimal performance of $\hat{f}_X$, and if a relation (3.12) on the relative smoothnesses of f_U and f_X holds, then the difference between $\hat{f}_X$ and $\tilde{f}_X$ is negligible relative to the distance between either estimator and the true density, f_X.

Theorem 3.1. *Let $C_1 > 1$ and $C_2, \alpha, \beta > 0$. Assume that (i) $1 \le N_j \le C_1$ for each j; (ii) $N(n) \ge C_1^{-1} n$ for each $n \ge 1$; (iii) $f_U^{\rm Ft}$ satisfies (3.1); (iv) $\alpha > \frac{1}{2}$; (v) $K^{\rm Ft}$ satisfies (3.2); (vi) $h_1(n) \le h \le h_2(n)$, where $h_2(n) \to 0$ and, for some $\delta > 0$, $n^{(1-\delta)/4\alpha}\, h_1(n)$ is bounded away from zero; and (vii) $c_1\, n^{-c_2} \le \rho \le c_3 \min\{h_1(n)^{4\alpha+2}, n^{-1}\}$, where $c_1, c_2, c_3 > 0$. Then, for each integer $k \ge 1$,*

$$\sup_{f_X \in \mathcal{F}(\beta, C_2)} \sup_{-\infty < x < \infty} E\big\{\hat{f}_X(x) - \tilde{f}_X(x)\big\}^2 \le \text{const.}\, p_n\,, \tag{3.3}$$

where

$$\begin{aligned} p_n = p_n(k) = n^{-1} \big\{h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + (\log n)^2\big\} \\ + n^{-2} \big(h^{2(\beta-4\alpha)-2} + h^{-6\alpha-1}\big) + n^{-k}\, h^{-4(k+2)\alpha-2} \end{aligned} \tag{3.4}$$

and the constant in (3.3) depends on k but not on $h \in [h_1(n), h_2(n)]$ or on n.

Proofs of Theorems 3.1–3.3 are given in section 5.

We argued in section 2 that, if we were to develop limit theory that did not involve taking expected values, the ridge parameter ρ could be taken equal to zero. In that setting we should replace uniform pointwise error, at (3.9), by error at a single point, or by a global metric such as integrated squared error. Otherwise we incur a logarithmic penalty on the right-hand side of (3.9). (This is to be expected, since the same penalty arises in more conventional problems; see e.g. Bickel and Rosenblatt (1973).) We should also remove the supremum over densities $f_X \in \mathcal{F}(\beta, C_2)$, since the uniformity implied by the supremum is not meaningful if we remove the expectation.

For the sake of definiteness, when working with $\rho = 0$ we measure accuracy in terms of squared error at a particular point, or integrated squared error. To treat the latter, note that (3.3) implies that, for each pair x_1, x_2 for which $-\infty < x_1 < x_2 < \infty$,

$$\sup_{f_X \in \mathcal{F}(\beta, C_2)} \int_{x_1}^{x_2} E\big\{\hat{f}_X(x) - \tilde{f}_X(x)\big\}^2\, dx = O(p_n)\,. \tag{3.5}$$

Let $\hat{f}_X^0(x)$ denote the version of $\hat{f}_X$ constructed with $\rho = 0$. We claim that (3.5) continues to apply to $\hat{f}_X^0$, provided the expectation and supremum over f_X are

removed from the left-hand side, and the right-hand side is interpreted in an "in probability" sense. Moreover, squared error at each fixed point x converges at the same rate:

$$\bigl|\hat{f}_X^0(x) - \tilde{f}_X(x)\bigr| = O_p\bigl(p_n^{1/2}\bigr)\,, \quad \int_{x_1}^{x_2} \bigl\{\hat{f}_X^0(x) - \tilde{f}_X(x)\bigr\}^2\,dx = O_p(p_n)\,. \tag{3.6}$$

Theorem 3.2. *Let $C_1 > 1$, let $C_2, \alpha, \beta > 0$, let $-\infty < x_1 < x_2 < \infty$, and take $\rho = 0$ in the definition of $\widehat{L}$, at (2.4), and hence also in the definition of $\hat{f}_X$, obtaining the estimator $\hat{f}_X^0$. Assume that conditions (i)–(vi) in Theorem 3.1 hold. Then (3.6) holds for each $f_X \in \mathcal{F}(\beta, C_2)$, each $x \in (-\infty, \infty)$ and each pair x_1, x_2 for which $-\infty < x_1 < x_2 < \infty$.*

3.2. Asymptotic optimality. The size of bandwidth that minimises pointwise mean squared error, when using $\tilde{f}_X$ to estimate f_X, is $h \asymp h_0 \equiv n^{-1/\{2(\alpha+\beta)-1\}}$; and, for such a bandwidth, pointwise mean squared error of $\tilde{f}_X$ is of size q_n, where

$$q_n = n^{-2(\beta-1)/\{2(\alpha+\beta)-1\}}\,. \tag{3.7}$$

The same result holds if we replace $\tilde{f}_X$ by the errors-in-variables regression estimator, $\tilde{g}$, which we define in section 3.6. See Fan (1991) and Fan and Truong (1993) for discussion of theory in these respective cases, and also for proofs of lower bounds which show that the rate q_n is minimax optimal, in an L_2 sense.

However, these results address only the case where there is no replication, i.e. each $N_j = 1$. In the case of upper bounds, generalisation to settings where each $N_j \geq 2$ is relatively straightforward. See section 3.4 for details. Below we generalise lower bounds in the setting of density deconvolution.

Theorem 3.3. *Assume that $\alpha, \beta > \frac{1}{2}$. Let $\mathcal{F}(\beta, C)$ denote the class of densities f_X defined in section 3.1, and write $\breve{\mathcal{F}}$ for the class of all measurable functionals of the data. Assume that $2 \leq N_j \leq B$ for each j, where $2 \leq B < \infty$. Then, for each fixed x and each sufficiently large $C > 0$, there exists $D > 0$ such that, for all sufficiently large n,*

$$\inf_{\breve{f} \in \breve{\mathcal{F}}} \sup_{f_X \in \mathcal{F}(\beta, C)} E_{f_X}\bigl\{\breve{f}(x) - f_X(x)\bigr\}^2 \geq D\,q_n\,. \tag{3.8}$$

3.3. Equivalence of $\hat{f}_X$ and $\tilde{f}_X$. In view of the results derived in section 3.2, and in order to establish that $\hat{f}_X$ is asymptotically equivalent to $\tilde{f}_X$ when the latter is performing optimally, it is instructive to show that when $h \asymp h_0$,

$$\sup_{f_X \in \mathcal{F}(\beta, C_2)} \sup_{-\infty < x < \infty} E\big\{\hat{f}_X(x) - \tilde{f}_X(x)\big\}^2 = o(q_n)\,, \tag{3.9}$$

if the ridge-prameter ρ is taken to be nonzero; or, if the ridge is zero, that

$$\big|\hat{f}^0_X(x) - \tilde{f}_X(x)\big| = o_p\big(q_n^{1/2}\big)\,, \int_{x_1}^{x_2} \big\{\hat{f}^0_X(x) - \tilde{f}_X(x)\big\}^2\,dx = o_p(q_n)\,. \tag{3.10}$$

Compare with (3.6). In fact, (3.9) and (3.10) follow from Theorems 3.1 and 3.2, respectively, if we prove that

$$q_n = o(p_n)\,. \tag{3.11}$$

Provided

$$\beta > \alpha + \tfrac{1}{2}\,, \tag{3.12}$$

it is straightforward to show that if $h \asymp h_0$ then

$$n^{-1}\left\{h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + (\log n)^2\right\} + n^{-2}\left(h^{2(\beta-4\alpha)-2} + h^{-6\alpha-1}\right) = o(q_n)\,, \tag{3.13}$$

and also that if k is sufficiently large and $h \asymp h_0$ then $n^{-k}\, h^{-4(k+2)\alpha-2} = o(q_n)$. This result and (3.13) imply (3.11).

Therefore condition (3.12), which can be characterised colloquially as the assertion that "f_X is smoother than half a derivative of f_U," is sufficient to ensure that, in deconvolution problems, there is no first-order loss of performance in using replicated data to estimate the error density when the latter is not known. Intuition behind (3.12) is given in section 3.5.

Of course, (3.12) fails if α is too large; that is, if f_U is too smooth. This is the reason for placing the lower bound on $|f_U^{\mathrm{Ft}}(t)|$ in (3.1). Without that bound, f_U can be arbitrarily smooth.

3.4. Properties of $\tilde{f}_X$. Let $\check{f}_X$ denote the "standard" version of $\tilde{f}_X$, obtained by taking $N_j = 1$ for each j, but with sample size M rather than n. Theorem 3.4,

which is given below and is straightforward to derive, argues that the bias of $\tilde{f}_X$ is identical to that of $\check{f}_X$, and that the variance of $\tilde{f}_X$ equals that of $\check{f}_X$, to first order.

Recall that U and X have the distributions of U_{jk} and X_j, respectively, that $W = X + U$, and that $N = \frac{1}{2} \sum_{j \le n} N_j \, (N_j - 1)$. Put

$$
\begin{aligned}
m_n(x) &= \int K(u) \, f_X(x - hu) \, du \,, \\
v_n(x) &= \frac{1}{M} \left\{ \frac{1}{h} \int L(u)^2 \, f_W(x - hu) \, du - m_n(x)^2 \right\}, \\
w_n(x) &= \frac{2N}{M^2} \left\{ \frac{1}{h} \int K(u)^2 \, f_X(x - hu) \, du - m_n(x)^2 \right\}.
\end{aligned}
$$

Theorem 3.4. *The mean and variance of $\check{f}_X(x)$ equal $m_n(x)$ and $v_n(x)$, respectively; the mean of $\tilde{f}_X(x)$ equals $m_n(x)$; and the variance of $\tilde{f}_X(x)$ equals $v_n(x) + w_n(s)$.*

The quantity w_n is generally of strictly smaller order than v_n, since $\int K^2$ remains fixed but $\int L^2$ diverges as h decreases. Therefore, in terms of first-order properties of mean and variance, $\tilde{f}_X$ and $\check{f}_X$ have identical performance. In view of this property, and bearing in mind the asymptotic equivalence of $\hat{f}_X$ and $\tilde{f}_X$ derived in section 3.3, we can fairly say that:

> to first order, $\hat{f}_X$ has the same properties as a conventional deconvolution density estimator, computed when the error density is known and the sample size is M but without any replication. (3.14)

Of course, this assertion requires (3.9) and hence needs (3.12).

Together, (3.8), (3.9) and (3.14) demonstrate minimax optimality of the estimator $\hat{f}_X$. Of course, this property necessitates the supremum being taken over f_X in (3.9). That requirement motivated our introduction of the ridge parameter in our definition of $\hat{f}_X$.

3.5. Discussion of different approaches to density deconvolution. Let (2.2)′ denote the version of (2.2) where the assumption that f_U^{Ft} is real-valued is omitted. For cases where (2.2)′ holds but (2.2) fails, Li and Vuong (1998) suggest an estimator of f_U^{Ft} quite different from our $\hat{f}_U^{\mathrm{Ft}}$. However, from a practical viewpoint the condition

that $f_U^{\rm Ft}$ be real-valued is mild. In particular, in the nonparametric literature on density deconvolution and errors-in-variables regression where f_U is assumed known, that quantity is invariably taken to be symmetric, in which case $f_U^{\rm Ft}$ is real-valued.

The alternative estimator suggested by Li and Vuong (1998) in the context of (2.2)′ requires the distributions of both U and X to have characteristic functions that do not vanish anywhere (see Li and Vuong's condition A3) and also to be compactly supported (see their assumption A4). We are not aware of a distribution which enjoys both these properties. Certainly, none of the standard, compactly-supported distributions satisfy A3. This, and the numerical complexity of Li and Vuong's estimator, discouraged us from considering their technique.

If α is sufficiently less than β then the problem of estimating f_U from the differences $W_{jk_1} - W_{jk_2}$ is more difficult statistically, although more straightforward numerically, than the problem of estimating f_U from the raw data W_{jk}. This indicates why condition (3.12) is required. For values of α that are large relative to β, alternative deconvolution methods may possibly give better theoretical performance, although we are not aware of any that are attractive computationally.

3.6. Errors-in-variables regression. The results in this section are closely analogous to those in earlier sections, so we give only an outline. Recall from section 2.2 that, under the model (2.5), our estimator of g is $\hat g = \hat a / \hat f_X$, where $\hat a$ is an estimator, defined at (2.6), of $a = f_X g$. Properties of $\hat g$ follow directly from those of the numerator and denominator in the ratio $\hat a / \hat f_X$. The denominator is treated in Theorems 3.1 and 3.2; here we address the numerator.

Given $f_X \in \mathcal{F}(\beta, C_2)$, let $\mathcal{G}(\beta, C_2 \,|\, f_X)$ denote the class of functions g for which

$$\sup_{-\infty<t<\infty} (1+|t|)^\beta \left| \int e^{itx} f_X(x)\, g(x)\, dx \right| \le C_2 \,.$$

Recall that conditions associated with the errors-in-variables model (2.5) include the assumption that $E(V) = 0$ and $E(V^2) < \infty$.

Theorem 3.5. *Let $C_1 > 1$ and $C_2, \alpha, \beta > 0$. Assume (i)–(vii) in Theorem 3.1.*

Then, for each integer $k \geq 1$,

$$\sup_{f_X \in \mathcal{F}(\beta,C_2),\, g \in \mathcal{G}(\beta,C_2 \,|\, f_X)} \sup_{-\infty<x<\infty} E\big\{\hat{a}(x) - \tilde{a}(x)\big\}^2 \leq \text{const.}\, p_n\,, \tag{3.15}$$

where p_n *is as at* (3.4) *and the constant in* (3.15) *depends on* k *but not on* $h \in [h_1(n), h_2(n)]$ *or on* n.

We know from section 3.3 that, if α and β satisfy (3.12), and if h is of the same size as the bandwidth that minimises mean squared error of $\tilde{f}_X$ (this is also the size of the optimal bandwidth for $\tilde{a}$ and $\tilde{g}$), then $p_n = o(q_n)$. (Recall that q_n is given by (3.7), and that $q_n^{1/2}$ equals the minimum order of magnitude of error for estimators of f_X, a and g.) It then follows from Theorems 3.1 and 3.5, and (3.11), that if conditions (i)–(vii) hold, $\hat{f}_X(x) - \tilde{f}_X(x) = o_p(q_n^{1/2})$ and $\hat{a}(x) - \tilde{a}(x) = o_p(q_n^{1/2})$. Therefore, provided $f_X(x) > 0$, we have:

$$\hat{g}(x) = \frac{\hat{a}(x)}{\hat{f}_X(x)} = \frac{\tilde{a}(x)}{\tilde{f}_X(x)} + o_p\big(q_n^{1/2}\big) = \tilde{g}(x) + o_p\big(q_n^{1/2}\big)\,. \tag{3.16}$$

That is, if the bandwidth is chosen so that it is optimal for estimating g by $\tilde{g}$, then $\hat{g}$ is first-order equivalent to $\tilde{g}$.

It is straightforward to state and prove the analogue of Theorem 3.4 for the estimator $\tilde{a}$ instead of $\tilde{f}_X$. This leads directly to the analogue of (3.14), where the only change necessary is to replace $\hat{f}_X$ by $\hat{g}_X$ and alter "density estimator" to "regression estimator."

The proof of Theorem 3.5 is omitted, since it closely parallels that of Theorem 3.1, given in section 5.1. An argument similar to that used in section 5.2 to derive Theorem 3.2 can be employed to show that (3.16) holds even if the ridge parameter, ρ, is taken as zero. Therefore, (3.14) applies in the ridge-free case.

3.7. Supersmooth error case. All our discussion in the previous paragraphs was based on the assumption that the error distribution is ordinary smooth, and in particular satisfies (3.1). It is also of interest to treat the case of supersmooth errors, so named because there the error density is infinitely differentiable. In that context the following condition is imposed in place of (3.1): for constants $\alpha > 0$,

$\gamma > 0$ and $C_1 > 1$, and all real t,

$$C_1^{-1} \exp\bigl(-\gamma\,|t|^\alpha\bigr) \le |\psi(t)| \le C_1 \exp\bigl(-\gamma\,|t|^\alpha\bigr)\,. \tag{3.17}$$

For such error distributions, pointwise mean squared error, when employing $\tilde f_X$ to estimate f_X, is of optimal order when using a bandwidth $h = D\,(\log n)^{-1/\alpha}$, where $D > (4\gamma)^{1/\alpha}$ denotes a constant. In this case, pointwise mean squared error of $\tilde f_X$ is of size $q_n = (\log n)^{-2(\beta-1)/\alpha}$. Here, the rate of convergence of the estimator $\tilde f_X$ is so slow that the loss of performance incurred by estimating f_U from the data, and using $\hat f_X$ instead of $\tilde f_X$, is negligible, regardless of restrictions such as (3.12). In particular, the following theorem holds. Its proof follows the lines of that of Theorem 3.1, but is more straightforward and hence is omitted.

Theorem 3.6. *Let $C_1 > 1$ and $\alpha, \beta, \gamma > 0$. Assume that (i) $1 \le N_j \le C_1$ for each j; (ii) $N(n) \ge C_1^{-1}\,n$ for each $n \ge 1$; (iii) $f_U^{\rm Ft}$ satisfies (3.17); (iv) $K^{\rm Ft}$ satisfies (3.2) with $c = 1$; (v) $h = D\,(\log n)^{-1/\alpha}$, with $D > (4\gamma)^{1/\alpha}$; and (vi) $\rho = C_1\,n^{-\kappa}$, with $\kappa > \frac14$. Then, for some $\epsilon > 0$,*

$$\sup_{f_X \in \mathcal{F}(\beta, C_2)} \sup_{-\infty < x < \infty} E\bigl\{\hat f_X(x) - \tilde f_X(x)\bigr\}^2 \le \text{const.}\, n^{-\epsilon}\,.$$

This result is readily generalised to the estimator $\hat g$, provided h is chosen so that the optimal convergence rate for $\tilde g$ as an estimator of g is attained. In particular, if $h = D\,(\log n)^{-1/\alpha}$ where $D > (4\gamma)^{1/\alpha}$, then $\hat g$ is first-order equivalent to $\tilde g$.

4. NUMERICAL PROPERTIES

4.1. Simulated examples. We study numerical properties of the estimators $\hat f_X$ and $\hat g$ in several simulated examples. In the density case, and following model (2.1), we generate 500 random samples of replicated observations for n individuals, W_{ij} where $i = 1, \ldots, n$ and $j = 1, \ldots, N_i$. We take the noise-to-signal ratio σ_U^2/σ_X^2 equal to 25%, except in the case of density (iii) below, where we take $\sigma_U^2/\sigma_X^2 = 10\%$. The notation σ_T^2 denotes the variance of a random variable T. The error density f_U is chosen to be a Laplace or a centred normal density.

We consider four target densities f_X: (i) $X \sim 0.5\,\mathrm{N}(-3,1)+0.5\,\mathrm{N}(2,1)$, (ii) $X \sim \chi^2(3)$, (iii) $X \sim \sum_{\ell=0}^{5}(2^{5-\ell}/63)\,\mathrm{N}\{65-96\,2^{-\ell}/21,(32/63)^2/2^{2\ell}\}$ and (iv) $X \sim \mathrm{N}(0,1)$. Density (i) is bimodal and symmetric, density (ii) is asymmetric and density (iii) is the smooth comb density discussed by Marron and Wand (1992). Note that, even in the error-free case, the latter density is particularly hard to estimate because of its numerous features.

In the regression case we generate 500 datasets of randomly-sampled vectors (W_{ij}, Y_i), $i = 1,\ldots,n$, $j = 1,\ldots,N_i$, according to the model (2.5). The density f_X is chosen to be a uniform $U[0,1]$ or a normal $\mathrm{N}(0.5,\sigma_X^2)$ density, with σ_X^2 chosen so that 0 and 1 are respectively the 0.025 and 0.975 quantiles of f_X. The error density f_U is a Laplace or centred normal density, and the noise-to-signal ratio σ_U^2/σ_X^2 equals 10%. Except for our Bernoulli regression example (see case (iii) below), the error density f_V is a centred normal density such that the noise-to-signal ratio $\sigma_V^2/\sigma^2(g)$ equals 10%, where $\sigma^2(g)$ denotes the mean squared deviation of g from its average value.

We consider three regression curves: (i) $g(x) = x^2\,(1-x)^2$, (ii) $g(x) = 3x + 20\,(2\pi)^{-1/2}\,\exp\{-100\,(x-\frac{1}{2})^2\}$, (iii) $Y|X = x \sim \mathrm{Bernoulli}\{g(x)\}$, with $g(x) = 0.45\,\sin(2\pi x) + 0.5$. Note that curve (i) is unimodal and symmetric around 0.5, curve (ii) is a mixture of a straight line and an exponential curve, and curve (iii) is an asymmetric sinusoid.

We sought an automatic way of choosing the bandwidth, h. In the density case, we suggest using $\hat{h}_{\mathrm{PI}}$, the plug-in bandwidth of Delaigle and Gijbels (2002, 2004), where the characteristic function of the error is replaced by (2.3). This procedure is justified by the discussion in section 3.3. In the regression case, a bandwidth-choice procedure could also be based on a data-driven selector for the known error case. However, since, to our knowledge, there does not exist such a method, we must first propose one.

A cross-validation (CV) criterion for selecting h would choose

$$h_{\mathrm{CV}} = \mathrm{argmin}_h \sum_{k=1}^{n} \left(\frac{Y_k - \sum_{j=1}^{n} Y_j S_j(X_k)}{1 - S_k(X_k)} \right)^2 ,$$

where, for $j = 1, \ldots, n$,

$$S_j(x) = \sum_{\ell=1}^{N_j} L\Big(\frac{x - W_{j\ell}}{h}\Big) \Big/ \sum_{J=1}^{n} \sum_{\ell=1}^{N_J} L\Big(\frac{x - W_{J\ell}}{h}\Big)\,.$$

Since the observations X_k are not available, we need to replace all quantities of the form

$$L\Big(\frac{X_k - W_{j\ell}}{h}\Big) = \frac{1}{2\pi} \int \exp(-itX_k/h)\, \exp(itW_{j\ell}/h)\, \frac{K^{\mathrm{Ft}}(t)}{f_U^{\mathrm{Ft}}(t/h)}\, dt\,,$$

by empirical estimators. We suggest replacing $\exp(-itX_k/h)$ by an estimator of its expected value, $f_X^{\mathrm{Ft}}(-t/h)$, based on the replications of the kth intrinsic observation. Such an estimator can be defined by $\hat{f}_W^{\mathrm{Ft}}(-t/h)/f_U^{\mathrm{Ft}}(-t/h)$, where $\hat{f}_W^{\mathrm{Ft}}(t) = K^{\mathrm{Ft}}(ht) \sum_{m=1}^{N_k} \exp(itW_{km})$ is a kernel estimator of f_W^{Ft}. Proceeding that way, our CV criterion becomes:

$$\tilde{h}_{\mathrm{CV}} = \mathrm{argmin}_h \sum_{k=1}^{n} \left(\frac{Y_k - \sum_{j=1}^{n} Y_j \widehat{S_j(X_k)}}{1 - \widehat{S_k(X_k)}} \right)^2, \tag{4.1}$$

where

$$\widehat{S_j(X_k)} = \sum_{m=1}^{N_k} \sum_{\ell=1}^{N_j} L_2\Big(\frac{W_{km} - W_{j\ell}}{h}\Big) \Big/ \sum_{J=1}^{n} \sum_{m=1}^{N_k} \sum_{\ell=1}^{N_J} L_2\Big(\frac{W_{km} - W_{J\ell}}{h}\Big)\,, \tag{4.2}$$

with $L_2(x) = (2\pi)^{-1} \int \exp(-itx/h)\, |K^{\mathrm{Ft}}(t)|^2\, |f_U^{\mathrm{Ft}}(t/h)|^{-2}\, dt$.

In the case of unknown error density, we define $\hat{h}_{\mathrm{CV}}$ as in (4.1) but we replace L_2 in (4.2) by

$$\widehat{L}_2(x) = (2\pi)^{-1} \int \exp(-itx/h)\, |K^{\mathrm{Ft}}(t)|^2\, |\hat{f}_U^{\mathrm{Ft}}(t/h)|^{-2}\, dt\,,$$

with $\hat{f}_U^{\mathrm{Ft}}(t)$ as in (2.3). As in the error-free case, the computations needed to calculate this bandwidth can be reduced considerably by binning the data. See, for example, Fan and Gijbels (1996), page 96. We suggest placing the W_{ij}'s into 200 equi-spaced bins between their empirical 0.025 and 0.975 quantiles.

The selection of a ridge parameter can be avoided if, instead of using $\hat{f}_U^{\mathrm{Ft}}(t) + \rho$ in $\hat{L}$, we employ $\tilde{f}_U^{\mathrm{Ft}}(t) = \hat{f}_U^{\mathrm{Ft}}(t)\, I(t \in A) + \hat{f}_P^{\mathrm{Ft}}(t)\, I(t \notin A)$, where A denotes the

largest interval around 0 in which the estimator $\hat{f}_U^{\mathrm{Ft}}(t)$ is not erratic, i.e. does not oscillate, and $\hat{f}_P^{\mathrm{Ft}}(t)$ is a parametric function estimated from the observations and defined by $\hat{f}_P^{\mathrm{Ft}}(t) = (1 + A_U t^2)^{-B_U}$, with A_U and B_U chosen so as to match the empirical second and fourth moments of the error with those of $\hat{f}_P$. In the event that these moments are negative, we set $B_U = 1$ and take A_U equal to half the empirical variance of the error, which corresponds to $\hat{f}_P$ being a Laplace density. This method gives very good results in practice, sometimes even better than in the case of known error density.

In our simulations, we consider samples of sizes $n = 50$, 100 and 250, and fix the number of replications, N_j, at 2 or 4. In each case we generate 500 datasets, for each of which we calculate an estimate of the target curve by using the bandwidth $\hat{h}_{\mathrm{PI}}$ (density case) or the bandwidth $\hat{h}_{\mathrm{CV}}$ (regression case). We take $K^{\mathrm{Ft}} = (1 - t^2)^3\, I(t \in [-1, 1])$; this kernel is commonly used in deconvolution problems. To evaluate performance, we calculate the integrated squared error (ISE) distance of each estimate, where ISE $= \int_I (\hat{m} - m)^2$, with $m = f_X$ or $m = g$, and where I is the whole real line (density case) or $I = [0, 1]$ (regression case). In the graphs, we present the three estimates that resulted in the first, second and third quartiles of the 500 calculated ISE's, and we denote them by, respectively, q_1, q_2 and q_3. We report only part of the simulations, although our conclusions are similar for the other, non-reported results.

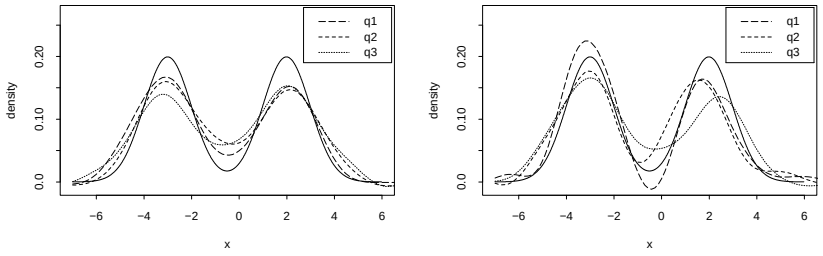

Figure 1. *Quantile curves of 500 estimates $\hat{f}_X$ of density (i) in the Laplace error case, when $M = 500$ and $N_j = 2$ (left panel) or $N_j = 4$ (right panel).*

In Figure 1, we illustrate the effect of increasing the number of replications by comparing the quartile curves for $N_j = 2$ and $N_j = 4$, obtained from 500

samples from density (i) contaminated by Laplace errors when $M = 500$. These and related results indicate better performance when $N_j = 2$ than when $N_j = 4$. As suggested in the introduction, for the same total number of observations, M, it is more advantageous to have a large number of intrinsically different observations, n, than a large number of replications, N_j.

Figure 2 examines the loss incurred through estimation of the error density. We show boxplots of 500 ISE's calculated for 500 estimates of densities (i) and (iv), for sample sizes $n = 50$, 100 or 250, with $N_j = 2$ and a normal error density. The boxplots are grouped by sample size. In each set of two boxplots, the first shows the results for the estimator $\hat{f}_X$ (unknown error) and the second, the results obtained, for the same 500 samples, using the estimator $\tilde{f}_X$ (known error). The graphs show that the performance lost by estimating the error density is minor. In some cases the results are even better for $\hat{f}_X$ than for $\tilde{f}_X$.

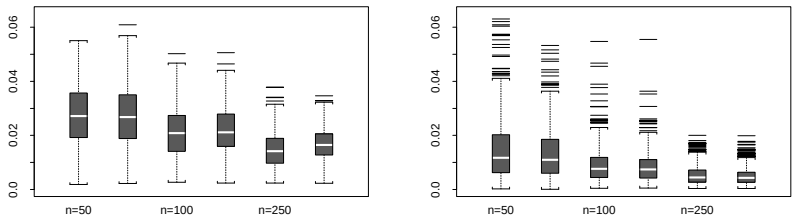

Figure 2. *Boxplots of the ISE of 500 estimates of density (i) (left panel) or density (iv) (right panel), in the normal error case for $N_j = 2$ and $n = 50$, 100 or 250. In each group of two boxplots, the first is for $\hat{f}_X$ (unknown error), and the second, for $\tilde{f}_X$ (known error).*

In Figure 3, we show the quartile curves obtained for 500 samples from density (iii) contaminated by Laplace error when $n = 250$, with $N_j = 2$, together with boxplots of the calculated ISE's for $n = 50$, 100 and 250 in the known and unknown error cases. The results show that, as in the error-free case, it is difficult to recover all the modes of this density, and that knowing the error density brings only minor improvements.

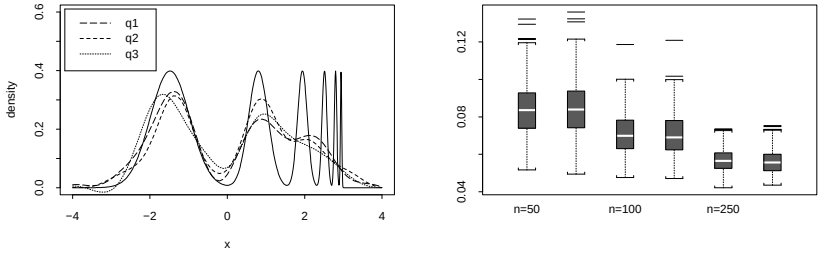

Figure 3. *Quartile curves of 500 estimates* $\hat{f}_X$ *of density* (*iii*) *in the Laplace error case, for* $N_j = 2$ *and* $n = 250$ *(left panel), together with boxplots (right panel) of the 500 calculated ISE's when* $n = 50$, 100 *or* $n = 250$. *In each group of two boxplots, the first is for* $\hat{f}_X$ *(unknown error), and the second, for* $\tilde{f}_X$ *(known error).*

In Figure 4, we show the quartile curves obtained from 500 samples for regression function (ii), when the error U is Laplace, $n = 250$ and $N_j = 2$, in the case where $X \sim \mathrm{N}(0.5, \sigma_X^2)$, using $\hat{g}$ (unknown error) or $\tilde{g}$ (known error). We see that the results are quite good in all cases, and that, in this example, knowing the error density does not seem to improve the results.

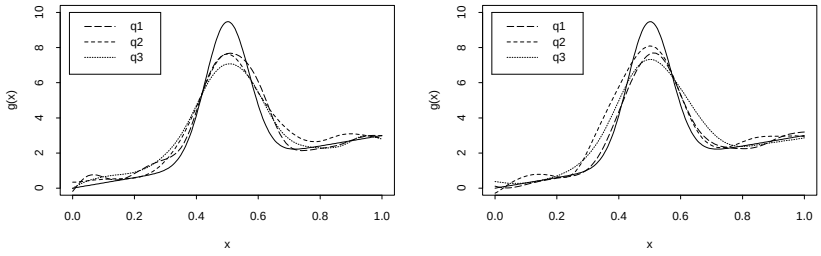

Figure 4. *Quartile curves of 500 estimates* $\hat{g}$ *(left panel) or* $\tilde{g}$ *(right panel) of the regression function* (*ii*) *in the Laplace error case, for* $N_j = 2$ *and* $n = 250$, *when* $X \sim \mathrm{N}(0.5, \sigma_X^2)$.

Figure 5 shows the quartile curves obtained from 500 samples in the case of regression function (*i*) for $n = 100$ and $N_j = 2$, when the error U is normal and $X \sim U[0, 1]$. We also show boxplots of the 500 calculated ISE's in the case of Laplace and normal error U and $n = 100$ or 250, using $\hat{g}$ (unknown error) or $\tilde{g}$ (known error). We see that the estimated curves are quite good and the results are slightly better when the error density is known.

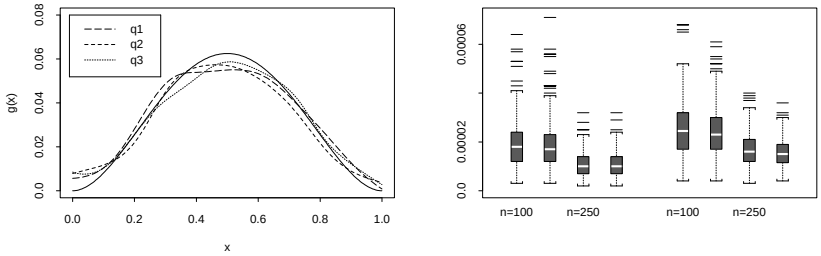

Figure 5. *Quartile curves of 500 estimates $\hat{g}$ of the regression function (i) in the normal error case for $N_j = 2$, $n = 100$ and $X \sim U[0,1]$ (left panel); and boxplots of 500 ISE's for the same regression curve in the case of Laplace error (first group of four) or normal error (last group of four), for $n = 100$ or 250 (right panel). In each group of two boxplots, the first is for $\hat{g}$ and the second is for $\tilde{g}$.*

Finally, Figure 6 shows the quartile curves in the case of regression curve (iii), when the error U is Laplace, $X \sim U[0,1]$, $N_j = 2$ and $n = 100$ or 250. In this case, too, we see that the results are quite good and improve as sample size increases.

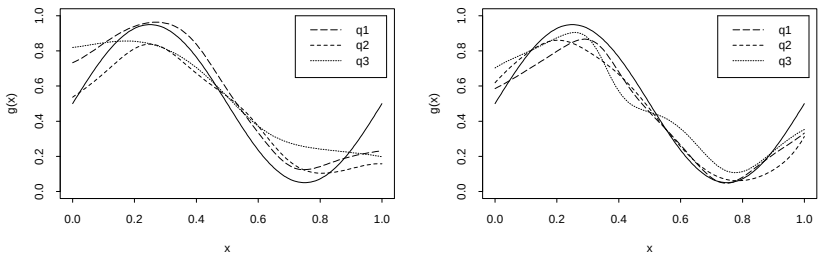

Figure 6. *Quartile curves of 500 estimates $\hat{g}$ of the regression function (iii) in the Laplace error case for $N_j = 2$, $X \sim U[0,1]$ and $n = 100$ (left panel) or $n = 250$ (right panel).*

4.2. Real-data examples. We apply our methods to two medical examples. The first dataset, described by Bland and Altman (1986), was collected to compare two methods for measuring peak expiratory flow rate (PEFR). Two replicated measurements of PEFR were made on 17 individuals, using each of two different methods: a Wright peak flow meter and a mini Wright meter. As described by Bland and Altman (1986), when evaluating a new method for measuring a clinical quantity, usually the true values remain unknown and a common practice is to compare the

new method with the established method, rather than with the true quantities. The goal is thus to check whether the mini meter and the Wright meter are in agreement.

To this end, we define X_i as the average of all possible readings on the mini meter for individual i, and define Y_i similarly for the 'regular' Wright meter. The latter gives more stable (less variable) readings than the mini meter, and, therefore, for each individual i, we set Y_i equal to the average of the two Wright readings. Since readings from the mini meter are more variable then there we need to incorporate measurement errors. For $j = 1, 2$, we take W_{ij} to be the jth replicated mini Wright measurement.

The regression estimate is shown in the left panel of Figure 7, together with the observations (W_{ij}, Y_i). The unusual shape of the estimate, deviant from a straight line, suggests that the two PERF measurement methods might not be in good agreement and that further investigation should be carried out. Bland and Altman (1986) note that a standard parametric analysis of these data, not taking the noise into account, indicates agreement between the two methods. The alternative analysis they propose concludes that the two methods are not in good agreement.

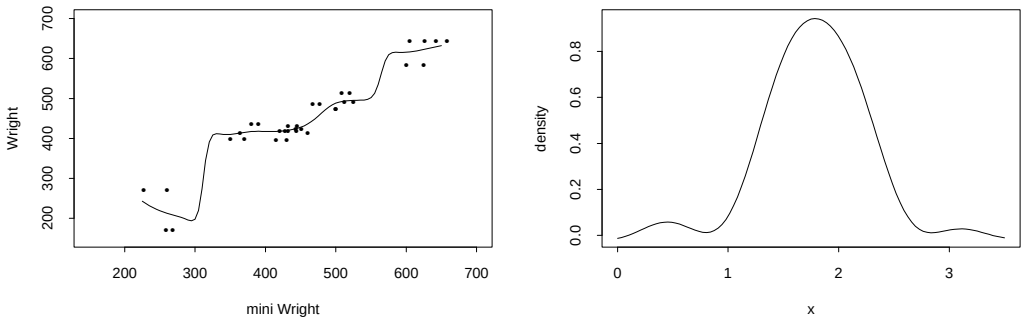

Figure 7. *Regression estimate for the PEFR data (left panel) and density estimate for the CAT data (right panel).*

The second dataset concerns two replicated measurements derived from CAT scans of the heads of 50 psychiatric patients. More precisely, the ventricule-brain ratio (VBR) was measured twice for each patient, using a hand-held planimeter. See Turner et al. (1986) and Dunn (2004). The logarithm of the VBR can be described

by model (2.1), and for the ith patient we set $W_{ij} = \log(\mathrm{VBR}_{ij})$, $j = 1, 2$, where VBR_{ij} denotes the jth contaminated replication of the measurement of VBR for patient i. The density estimate of the non-contaminated log VBR is plotted in Figure 7, which shows a smooth and symmetric density.

5. TECHNICAL ARGUMENTS

5.1. Proof of Theorem 3.1. Without loss of generality, $c = 1$ in (3.2). Put $\psi = f_U^{\mathrm{Ft}}$, $\phi = \psi^2$ and

$$\Delta(t) = \frac{1}{N} \sum_{j=1}^{n} \sum_{(k_1,k_2)\in\mathcal{S}_j} \left[\cos\{t(W_{jk_1} - W_{jk_2})\} - \phi(t) \right].$$

In this notation,

$$\begin{aligned} \left(\hat{f}_U^{\mathrm{Ft}} + \rho\right)^{-1} &= \left\{\psi\,(1 + \phi^{-1}\,\Delta)^{1/2} + \rho\right\}^{-1} \\ &= \psi^{-1}\,I(\psi > \rho) + \sum_{\ell=1}^{k} c_\ell\,\psi^{-2\ell-1}\,\Delta^\ell + \chi_1 + \chi_2\,, \end{aligned} \tag{5.1}$$

where the constants c_ℓ are derived from binomial coefficients, $|\chi_1| \le \rho^{-1}\,I(\psi \le \rho)$,

$$\begin{aligned} |\chi_2| \le \text{const.} \Big\{ \frac{\rho}{\psi + \rho} \left(\psi^{-3}\,|\Delta| + \psi^{-(2k+1)}\,|\Delta|^k \right) + \psi^{-(2k+3)}\,|\Delta|^{k+1} \\ + \rho\,\psi^{-2}\,I(\psi > \rho) + \rho^{-1}\,I\big(|\Delta| > \tfrac{1}{2}\,\phi\big) \Big\}, \end{aligned}$$

and "const.", here and below, denotes a generic positive constant depending only on k, f_U and the parameters α and C_2 of $\mathcal{F}(\beta, C_2)$.

Result (5.1) implies that

$$\hat{f}_X(x) - \tilde{f}_X(x) = \sum_{\ell=1}^{k} c_\ell\,\delta_{1\ell}(x) + \delta_{01}(x) + \delta_{02}(x) - \delta_2(x)\,, \tag{5.2}$$

where, for $\ell = 1, 2$ in the case of $\delta_{0\ell}$, and $1 \le \ell \le k$ for $\delta_{1\ell}$,

$$\begin{aligned} \delta_{0\ell}(x) &= \frac{1}{2\pi} \int e^{-itx}\,\hat{f}_W^{\mathrm{Ft}}(t)\,\chi_\ell(t)\,K^{\mathrm{Ft}}(ht)\,dt\,, \\ \delta_{1\ell}(x) &= \frac{1}{2\pi} \int e^{-itx}\,\hat{f}_W^{\mathrm{Ft}}(t)\,\psi(t)^{-2\ell-1}\,\Delta(t)^\ell\,K^{\mathrm{Ft}}(ht)\,dt\,, \\ \delta_2(x) &= \frac{1}{2\pi} \int e^{-itx}\,\hat{f}_W^{\mathrm{Ft}}(t)\,\psi(t)^{-1}\,K^{\mathrm{Ft}}(ht)\,I\{\psi(t) \le \rho\}\,dt \end{aligned}$$

and $\hat{f}_W^{\mathrm{Ft}}(t) = M^{-1} \sum_j \sum_k e^{itW_{jk}}$.

We claim that, for a constant $n_0 \geq 1$, the functions δ_{01} and δ_2 vanish identically whenever $n \geq n_0$. This is a consequence of the fact that (a) $K^{\mathrm{Ft}}(ht)$ vanishes if $h|t| \geq 1$, and (b) if $n \geq n_0$, the indicator functions $I\{\psi(t) \leq \rho\}$ and $I(|t| \leq h^{-1})$ cannot both equal 1. To appreciate why (b) is true, note that, in view of (3.1), (c) $I\{\psi(t) \leq \rho\}$ implies that $|t| \geq \text{const.}\, \rho^{-1/\alpha}$; that, by assumption (vi) in the theorem, (d) $h \geq h_1(n)$; and, by assumption (vii) in the theorem, $\rho \leq c_3\, h_1^{4\alpha+1}$, whence (e) $\rho^{-1/\alpha} \geq \text{const.}\, h_1(n)^{-4}$. Together, (c)–(e) entail: (f) $I\{\psi(t) \leq \rho\}$ implies that $|t| \geq \text{const.}\, h^{-4}$. Result (b) follows from (f).

Therefore, assuming $n \geq n_0$, we deduce from (5.2) that

$$\hat{f}_X(x) - \tilde{f}_X(x) = \sum_{\ell=1}^{k} c_\ell\, \delta_{1\ell}(x) + \delta_{02}(x)\,.$$

This formula, and the fact that $\hat{f}_W^{\mathrm{Ft}} = \psi\, f_X^{\mathrm{Ft}} + \Delta_1$ where $\Delta_1 = \hat{f}_W^{\mathrm{Ft}} - E(\hat{f}_W^{\mathrm{Ft}})$, imply that

$$\begin{aligned}
&\sup_{-\infty<x<\infty} E\big\{\hat{f}_X(x) - \tilde{f}_X(x)\big\}^2 \\
&\qquad \leq \text{const.} \left[\max_{r=2,3} \max_{1\leq\ell\leq k} \sup_{-\infty<x<\infty} E\big\{\delta_{r\ell}(x)^2\big\} + \sup_{-\infty<x<\infty} E\big\{\delta_{02}(x)^2\big\} \right], \qquad (5.3)
\end{aligned}$$

where

$$\begin{aligned}
\delta_{2\ell}(x) &= \frac{1}{2\pi} \int e^{-itx}\, f_X^{\mathrm{Ft}}(t)\, \psi(t)^{-2\ell}\, \Delta(t)^\ell\, K^{\mathrm{Ft}}(ht)\, dt\,, \\
\delta_{3\ell}(x) &= \frac{1}{2\pi} \int e^{-itx}\, \psi(t)^{-2\ell-1}\, \Delta_1(t)\, \Delta(t)^\ell\, K^{\mathrm{Ft}}(ht)\, dt\,.
\end{aligned}$$

Crude bounds show that for $\ell \geq 1$,

$$\begin{aligned}
E\big\{\delta_{2\ell}(x)^2\big\} &\leq \text{const.}\, n^{-\ell} \left\{ \left(\int_1^{1/h} |t|^{2\ell\alpha-\beta}\, dt \right)^2 + 1 \right\} \\
&\leq \text{const.}\, n^{-\ell} \left\{ h^{2\beta-4\ell\alpha-2} + (\log n)^2 \right\}, \qquad (5.4) \\
E\big\{\delta_{3\ell}(x)^2\big\} &\leq \text{const.}\, n^{-\ell-1} \left\{ \left(\int_1^{1/h} |t|^{(2\ell+1)\,\alpha}\, dt \right)^2 + 1 \right\} \\
&\leq \text{const.}\, n^{-\ell-1}\, h^{-2\,(2\ell+1)\,\alpha-2}\,. \qquad (5.5)
\end{aligned}$$

Here and below, terms in $\log n$ take account of instances where integrals either converge or just fail to converge, in the latter case being of the form $\int_1^{1/h} |t|^{-1}\,dt$. Consider, for example, the case $\beta = 2\ell\alpha + 1$ in (5.4).

A longer argument gives $E\{\delta_{21}(x)^2\} \le \text{const.}\, n^{-1}\,\{T(h) + \log n\}$, where

$$\begin{aligned}
T(h) &= \int_1^{1/h}\int_1^{1/h} \bigl|f_X^{\mathrm{Ft}}(t_1)\, f_X^{\mathrm{Ft}}(t_2)\bigr| \left\{\frac{\psi(t_1 - t_2)}{\psi(t_1)\,\psi(t_2)}\right\}^2 dt_1\,dt_2 \\
&\le \text{const.} \int_1^{1/h}\int_1^{1/h} (1 + |t_1|)^{2\alpha-\beta}\,(1 + |t_2|)^{2\alpha-\beta}\,(1 + |t_1 - t_2|)^{-2\alpha}\,dt_1\,dt_2 \\
&\le \text{const.} \int_1^{1/h}\int_1^{1/h} (1 + |t_1|)^{2\alpha-\beta}\,\bigl\{(1 + |t_1|)^{2\alpha-\beta} \\
&\qquad + (1 + |t_1 - t_2|)^{2\alpha-\beta} + 1\bigr\}\,(1 + |t_1 - t_2|)^{-2\alpha}\,dt_1\,dt_2 \\
&\le \text{const.} \int_1^{1/h} \bigl\{(1 + |t_1|)^{2\alpha-\beta} + (1 + |t_1|)^{2(2\alpha-\beta)}\bigr\}\,dt_1 \\
&\le \text{const.}\,\bigl(h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + \log n\bigr)\,.
\end{aligned}$$

Here we have used the fact that $\alpha > \frac{1}{2}$. Therefore,

$$E\bigl\{\delta_{21}(x)^2\bigr\} \le \text{const.}\, n^{-1}\,\bigl(h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + \log n\bigr)\,. \tag{5.6}$$

Using (5.4) for $2 \le \ell \le k$, and (5.6) for $\ell = 1$, we obtain:

$$\begin{aligned}
\max_{1\le\ell\le k}\, \sup_{-\infty<x<\infty} E\bigl\{\delta_{2\ell}(x)^2\bigr\} \le \text{const.}\,\Bigl[n^{-1}\,\bigl\{h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + (\log n)^2\bigr\} \\
+ n^{-2}\,h^{2(\beta-4\alpha)-2} + n^{-k}\,h^{2\beta-4k\alpha-2}\Bigr\}\,.
\end{aligned} \tag{5.7}$$

The argument leading to (5.6) allows us to increase by 1 the exponent of h in the term $h^{2\beta-4\alpha-2}$ on the right-hand side of (5.5), in the case $\ell = 1$. Analogously, for each $\ell \ge 1$, the argument leading to (5.6) can be used to sharpen (5.5) to:

$$E\bigl\{\delta_{3\ell}(x)^2\bigr\} \le \text{const.}\, n^{-\ell-1}\,h^{-2\,(2\ell+1)\,\alpha-1}\,.$$

Hence,

$$\max_{1\le\ell\le k}\, \sup_{-\infty<x<\infty} E\bigl\{\delta_{3\ell}(x)^2\bigr\} \le \text{const.}\,\bigl(n^{-2}\,h^{-6\alpha-1} + n^{-(k+1)}\,h^{-2(2k+1)\alpha-1}\bigr)\,. \tag{5.8}$$

Combining (5.7) and (5.8) we deduce that

$$\begin{aligned}
&\max_{r=2,3} \max_{1\le\ell\le k} \sup_{-\infty<x<\infty} E\{\delta_{r\ell}(x)^2\} \\
&\quad \le \text{const.} \Big[n^{-1} \left\{ h^{\beta-2\alpha-1} + h^{2(\beta-2\alpha)-1} + (\log n)^2 \right\} \\
&\qquad + n^{-2} \left(h^{2(\beta-4\alpha)-2} + h^{-6\alpha-1} \right) + n^{-k}\, h^{2\beta-4k\alpha-2} + n^{-(k+1)}\, h^{-2(2k+1)\alpha-1} \Big] \\
&\quad = O(p_n)\,. \qquad (5.9)
\end{aligned}$$

This bounds the first term on the right-hand side of (5.3). Next we address the second term there. Since $h \ge h_1(n)$, where $n^{(1-\delta)/4\alpha}\, h_1(n)$ is bounded away from zero, then $n^{1-\delta} h^{4\alpha}$ is bounded away from zero. It follows from this property, and Markov's inequality, that for each $B > 0$,

$$\begin{aligned}
\sup_{t\,:\,K^{\mathrm{Ft}}(ht)\ne 0} P\{|\Delta(t)| > \tfrac{1}{2}\,\phi(t)\} &\le \sup_{t\,:\,K^{\mathrm{Ft}}(ht)\ne 0} E\{2\,\phi(t)^{-1}\,|\Delta(t)|\}^{2B} \\
&\le \text{const.} \sup_{t\,:\,K^{\mathrm{Ft}}(ht)\ne 0} \left(|t|^{4\alpha}\, n^{-1} \right)^B \\
&= O\{(nh^{4\alpha})^{-B}\} = O\left(n^{-B\delta}\right). \qquad (5.10)
\end{aligned}$$

Hence, since ρ^{-1} and h^{-1} are only polynomially large in n, then for each $B > 0$,

$$\rho^{-2}\, E\left[\int I\{|\Delta(t)| > \tfrac{1}{2}\,\phi(t)\}\, \left|K^{\mathrm{Ft}}(ht)\right| dt \right]^2 = O\left(n^{-B}\right) \qquad (5.11)$$

for each $B > 0$. Moreover,

$$\begin{aligned}
&E\Bigg(\int \Bigg[\frac{\rho}{\psi(t)+\rho} \left\{ \psi(t)^{-3}\,|\Delta(t)| + \psi(t)^{-(2k+1)}\,|\Delta(t)|^k \right\} + \psi(t)^{-(2k+3)}\,|\Delta(t)|^{k+1} \\
&\qquad\qquad + \rho\,\psi(t)^{-2} \Bigg] \left|K^{\mathrm{Ft}}(ht)\right| \left\{ \psi(t)\left|f_X^{\mathrm{Ft}}(t)\right| + |\Delta_1(t)| \right\} dt \Bigg)^2 \\
&\le \text{const.} \Big\{ \Big(\rho\, h^{-(4\alpha+1)}\, n^{-1/2} + h^{-(2k+1)\alpha-1}\, n^{-k/2} + h^{-(2k+3)\alpha-1}\, n^{-(k+1)/2} \\
&\qquad\qquad + \rho\, h^{-2\alpha-1} \Big)^2 \left(h^{\alpha+\beta} + n^{-1/2} \right)^2 + n^{-1}\,(\log n)^2 \Big\}\,, \qquad (5.12)
\end{aligned}$$

where the term in $n^{-1}\,(\log n)^2$ is added to deal with the case of particularly large values of β, in which instance multiplying through by $(h^{\alpha+\beta} + n^{-1/2})^2$ gives too

small an order of magnitude. Combining (5.11) and (5.12), and using the fact that, by assumption, $\rho \le$ const. $\min\{h_1(n)^{4\alpha+2}, n^{-1}\}$, we deduce that

$$\sup_{-\infty<x<\infty} E\{\delta_{02}(x)^2\} = O(p_n)\,. \tag{5.13}$$

Together, (5.3), (5.9) and (5.13) imply (3.3).

5.2. Proof of Theorem 3.2. For brevity we derive only the second part of (3.6). Since

$$\left|\{\hat f_U^{\mathrm{Ft}}(t) + \rho\}^{-1} - \hat f_U^{\mathrm{Ft}}(t)^{-1}\right| \le \rho / \hat f_U^{\mathrm{Ft}}(t)^2$$

then

$$\left|\widehat L(u) - \widehat L^0(u)\right| \le \frac{\rho h}{2\pi} \int \hat f_U^{\mathrm{Ft}}(t)^{-2}\, K^{\mathrm{Ft}}(ht)\, dt\,, \tag{5.14}$$

where $\widehat L^0$ denotes the version of $\widehat L$ constructed with $\rho = 0$. We know from (5.10) that, with probability π_n, say, equal to $1 - O(n^{-B})$ for each $B > 0$, $\frac{1}{2}\, f_U^{\mathrm{Ft}}(t)^2 \le \hat f_U^{\mathrm{Ft}}(t)^2$ for all t such that the integrand at (5.14) does not vanish. Therefore, with probability at least π_n,

$$\begin{aligned}\sup_{-\infty<u<\infty} \left|\widehat L(u) - \widehat L^0(u)\right| &\le \frac{\rho h}{\pi} \int f_U^{\mathrm{Ft}}(t)^{-2}\, K^{\mathrm{Ft}}(ht)\, dt \\ &\le \frac{C_1^2 \rho h s}{\pi} \int_{-1/h}^{1/h} (1+|t|)^{2\alpha}\, dt \le C_3\, \rho h^{-2\alpha}\,, \end{aligned} \tag{5.15}$$

where $s = \sup |K^{\mathrm{Ft}}|$ and $C_3 > 0$. (We continue to take $c = 1$ in (3.2).)

Result (5.15) implies that, with probability at least π_n,

$$\sup_{-\infty<x<\infty} \left|\hat f_X(x) - \hat f_X^0(x)\right| \le C_3\, \rho h^{-2\alpha-1}\,. \tag{5.16}$$

By assumption (vi) in Theorem 3.1, $h \ge C_4\, n^{-C_5}$, and so $h^{-2\alpha-1} \le C_6\, n^{C_7}$, where $C_4, \ldots, C_7$ denote positive constants. Therefore, taking $\rho = n^{-C_8}$ where $C_8 \ge C_7 + 1$ (this choice satisfies condition (vii) in Theorem 3.1), we conclude from (5.16) that, with probability at least π_n,

$$\sup_{-\infty<x<\infty} \left|\hat f_X(x) - \hat f_X^0(x)\right| \le C_3\, n^{-1}\,. \tag{5.17}$$

We know from (3.3) in Theorem 3.1 that, for the choice of ρ above, (3.5) holds. That result and (5.17) imply the send part (3.6).

5.3. Proof of Theorem 3.3. Without loss of generality, each $N_j = 2$. Then the data at (2.1) are independent in pairs, the jth pair being (W_{j1}, W_{j2}). Suppose there are two options for f_X, the first a fixed density, f_{X0}, and the second, f_{Xn}, varying with n through a perturbation g_n: $f_{Xn} = f_{X0} + g_n$, where g_n integrates to zero. The corresponding characteristic function is $\chi_n = \chi_0 + \gamma_n$, where χ_0 is the characteristic function for f_{X0}, and $\gamma_n(t) = \int e^{itx}\, g_n(x)\, dx$. Choose $f_{X0} \in \mathcal{F}(\beta, C_2)$, and select f_U so that f_U^{Ft} satisfies (3.1).

Let H denote the perturbation function introduced by Fan (1991, pp. 1268–1269) in his case $\ell = 0$, and put $g_n(x) = c\, \delta_n^{\beta-1}\, H(x/\delta_n)$, where $c, \delta_n > 0$ and $\delta_n \to 0$. Then, if $D_1 = D_1(C_2, c)$ is sufficiently large, $f_{Xn} \in \mathcal{F}(\beta, D_1)$ for all n. Note too, for future reference, that $H(0) \neq 0$.

In this notation, the joint density of (W_1, W_2) is given by $f_{W_1W_2}$ when the density of X is f_{X0}, and by $f_{W_1W_2,n}$ when the density of X is f_{Xn}:

$$f_{W_1W_2,n}(w_1, w_2) - f_{W_1W_2}(w_1, w_2) = c\, \delta_n^{\beta-1}\, a_n(w_1, w_2)\,, \tag{5.18}$$

where

$$\begin{aligned} f_{W_1W_2}(w_1, w_2) &= \int f_{U0}(w_1 - x)\, f_{U0}(w_2 - x)\, f_{X0}(x)\, dx\,, \\ a_n(w_1, w_2) &= \int f_{U0}(w_1 - x)\, f_{U0}(w_2 - x)\, H(x/\delta_n)\, dx\,. \end{aligned}$$

The Fourier transform of a_n is

$$\iint a_n(w_1, w_2)\, \exp(isw_1 + itw_2)\, dw_1\, dw_2 = \delta_n\, \psi(s)\, \psi(t)\, \phi_H\{(s+t)\, \delta_n\}\,,$$

where $\phi_H(t) = \int e^{itx}\, H(x)\, dx$. Hence, by Parseval's identity, the integrated squared distance between $f_{W_1W_2}(w_1, w_2)$ and $f_{W_1W_2,n}(w_1, w_2)$ equals a constant multiple of:

$$\begin{aligned} &\delta_n^{2\beta} \iint |\psi(s)|^2\, |\psi(t)|^2\, |\phi_H\{(s+t)\, \delta_n\}|^2\, ds\, dt \\ &\quad = \delta_n^{2\beta-1} \int_{-\infty}^{\infty} |\psi(t)|^2\, dt \int_1^2 \left|\psi\big(u\delta_n^{-1} - t\big)\right|^2 |\phi_H(u)|^2\, du \asymp \delta_n^{2\alpha+2\beta-1}\,. \end{aligned} \tag{5.19}$$

Results (5.18) and (5.19), and the arguments of Fan (1991), imply that, if we define δ_n by $\delta_n^{2\alpha+2\beta-1} = n^{-1}$; and if we take $\bar{\mathcal{F}}$ to be the class of all empirical rules for discriminating between f_{X0} and f_{Xn}, using only the data $\mathcal{W} = \{(W_{j1}, W_{j2}) : 1 \le j \le n\}$; then, for all sufficiently large c (in the definition of g_n),

$$D_2 \equiv \liminf_{n\to\infty} \inf_{\bar{f}\in\bar{\mathcal{F}}} \left\{P_{f_{X0}}(\bar{f} = f_{Xn}) + P_{f_{Xn}}(\bar{f} = f_{X0})\right\} > 0\,. \tag{5.20}$$

Let $\breve{f}$ denote any estimator of f_X, and, in the calculations below, take $\bar{f} = f_{X0}$ if $|\breve{f}(0) - f_{X0}(0)| \le |\breve{f}(0) - f_{Xn}(0)|$, and $\bar{f} = f_{Xn}$ otherwise. Then, in view of (5.20),

$$\begin{aligned} 2 \sup_{f_X\in\{f_{X0},f_{Xn}\}} & E_{f_X}\left\{\breve{f}(0) - f_X(0)\right\}^2 \\ &\ge E_{f_{X0}}\left\{\breve{f}(0) - f_{X0}(0)\right\}^2 + E_{f_{Xn}}\left\{\breve{f}(0) - f_{Xn}(0)\right\}^2 \\ &\ge \tfrac{1}{4}\left\{f_{Xn}(0) - f_{X0}(0)\right\}^2 \left\{P_{X0}(\bar{f} = f_{Xn}) + P_{Xn}(\bar{f} = f_{X0})\right\} \\ &\ge \tfrac{1}{8}\, D_2 \left\{f_{Xn}(0) - f_{X0}(0)\right\}^2, \end{aligned} \tag{5.21}$$

the latter inequality holding for all sufficiently large n.

Let $\breve{\mathcal{F}}$ denote the class of all measurable functionals of the data $\mathcal{W}$. Since $f_{Xn}(0) - f_{X0}(0) = g_n(0) = c\,\delta_n^{\beta-1}\,H(0)$, and $H(0) \ne 0$, then the far right-hand side of (5.21) equals $D_3\,\delta_n^{2(\beta-1)}$ where $D_3 > 0$. Therefore (5.21) implies that

$$\inf_{\breve{f}\in\breve{\mathcal{F}}} \sup_{f_X\in\{f_{X0},f_{Xn}\}} E_{f_X}\left\{\breve{f}(0) - f_X(0)\right\}^2 \ge D_3\, n^{-2(\beta-1)/(2\alpha+2\beta-1)}\,. \tag{5.22}$$

Since both f_{X0} and f_{Xn} are (for all n) in $\mathcal{F}(\beta, C)$ if C is sufficiently large, then (3.8) follows from (5.22)

REFERENCES

ANDERSEN, C.M., BRO, R. AND BROCKHOFF, P.B. (2003). Quantifying and handling erros in instrumental measurements using the measurement error theory. *J. Chemometrics* **17**, 621–629.

BARRY, J. AND DIGGLE, P. (1995). Choosing the smoothing parameter in a Fourier approach to nonparametric deconvolution of a density function. *J. Nonparametric Statist.* **4**, 223–232.

BICKEL, P.J. AND ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1**, 1071–1095.

BIEMER, P., GROVES, R., LYBERG, L., MATHIOWETZ, N. AND SUDMAN, S. (1991). *Measurement Errors in Surveys*. John Wiley and Sons, New York

BLAND, J.M AND ALTMAN, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310.

CARROLL, R.J., ELTINGE, J.L. AND RUPPERT, D. (1993). Robust linear regression in replicated measurement error models. *Statist. Probab. Lett.* **19**, 169–175.

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

CARROLL, R.J., RUPPERT, D. AND STEFANSKI, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

DELAIGLE, A. AND GIJBELS, I. (2002), Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc.* Ser. B **64**, 869–86.

DELAIGLE, A. AND GIJBELS, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Comput. Statist. Data Anal.* **45**, 249–267.

DIGGLE, P. AND HALL, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *J. Roy. Statist. Soc.* Series B **55**, 523–531.

DUNN, G. (1989). *Design and Analysis of Reliability Studies*. Arnold, London.

DUNN, G. (2004). *Statistical Evaluation of Measurement Errors, Design and Analysis of Reliability Studies*, 2nd Edn. Arnold, London.

ELIASZIW, M., YOUNG, S. L., WOODBURY, M. G. AND FRYDAY-FIELD, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys. Therapy* **74**, 777–788.

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

FAN, J. AND GIJBELS, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

FAN, J. AND TRUONG, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

HOROWITZ, J.L. AND MARKATOU, M. (1996). Semiparametric estimation of regression models for panel data. *Rev. Econom. Stud.* **63**, 145–168.

HUWANG, L. AND YANG, J. (2000). Trimmed estimation in the measurement error model when the covariate has replicated observations. *Proc. Nat. Sci. Council ROC(A)* **24**, 405–412.

JAECH, J. (1985). *Statistical Analysis of Measurement Errors*. Wiley, New York.

LI, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables, models. *J. Econometrics* **110**, 1–26.

LI, T. AND HSIAO, C. (2004). Robust estimation of generalised linear models with measurement errors. *J. Econometrics* **118**, 51–65.

LI, T. AND VUONG, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivar. Anal.* **65**, 139–165.

MADANSKY, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Assoc.* **54**, 173–205.

MARRON, J.S. AND WAND, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.

NEUMANN, M.H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametric Statist.* **7**, 307–330.

NEWEY, W.K. AND POWELL, J.L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

OMAN, S.D., MEIR, N. AND HAIM, N. (1999). Comparing two measures of creatinine clearance: an application of errors-in-variables and bootstrap techniques. *Appl. Statist.* **48**, 39–52.

SCHENNACH, S.M. (2004a). Estimation of nonlinear models with measurement error. *Econometrica* **72**, 33–75.

SCHENNACH, S.M. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory* **20**, 1046–1093.

STEFANSKI, L.A. AND CARROLL, R.J. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 169–184.

SUSKO, E. AND NADON, R. (2002). Estimation of a residual distribution with small numbers of repeated measurements. *Canad. J. Statist.* **30**,383–400.

TURNER, S. W., TOONE, B. K.AND BRETT-JONES, J. R. (1986). Computerized tomographic scan changes in early schizophrenia — preliminary findings. *Psychol. Med.* **16**, 219–225.

VAN ES, B. AND HU, H.W. (2005). Asymptotic normality of kernel-type deconvolution estimators. *Scand. J. Statist.* **32**, 467–483.