8

Kernel methods and minimum contrast estimators for empirical deconvolution

Aurore Delaigle[a] and Peter Hall[b]

Abstract

We survey classical kernel methods for providing nonparametric solutions to problems involving measurement error. In particular we outline kernel-based methodology in this setting, and discuss its basic properties. Then we point to close connections that exist between kernel methods and much newer approaches based on minimum contrast techniques. The connections are through use of the sinc kernel for kernel-based inference. This 'infinite order' kernel is not often used explicitly for kernel-based deconvolution, although it has received attention in more conventional problems where measurement error is not an issue. We show that in a comparison between kernel methods for density deconvolution, and their counterparts based on minimum contrast, the two approaches give identical results on a grid which becomes increasingly fine as the bandwidth decreases. In consequence, the main numerical differences between these two techniques are arguably the result of different approaches to choosing smoothing parameters.

Keywords bandwidth, inverse problems, kernel estimators, local linear methods, local polynomial methods, minimum contrast methods, nonparametric curve estimation, nonparametric density estimation, nonparametric regression, penalised contrast methods, rate of convergence, sinc kernel, statistical smoothing

AMS subject classification (MSC2010) 62G08, 62G05

[a] Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia; A.Delaigle@ms.unimelb.edu.au
[b] Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia; halpstat@ms.unimelb.edu.au

1 Introduction

1.1 Summary

Our aim in this paper is to give a brief survey of kernel methods for solving problems involving measurement error, for example problems involving density deconvolution or regression with errors in variables, and to relate these 'classical' methods (they are now about twenty years old) to new approaches based on minimum contrast methods. Section 1.1 motivates the treatment of problems involving errors in variables, and section 1.2 describes conventional kernel methods for problems where the extent of measurement error is so small as to be ignorable. Section 2.1 shows how those standard techniques can be modified to take account of measurement errors, and section 2.2 outlines theoretical properties of the resulting estimators.

In section 3 we show how kernel methods for dealing with measurement error are related to new techniques based on minimum contrast ideas. For this purpose, in section 3.1 we specialise the work in section 2 to the case of the sinc kernel. That kernel choice is not widely used for density deconvolution, although it has previously been studied in that context by Stefanski and Carroll (1990), Diggle and Hall (1993), Barry and Diggle (1995), Butucea (2004), Meister (2004) and Butucea and Tsybakov (2007a,b). Section 3.2 outlines some of the properties that are known of sinc kernel estimators, and section 3 points to the very close connection between that approach and minimum contrast, or penalised contrast, methods.

1.2 Errors in variables

Measurement errors arise commonly in practice, although only in a minority of statistical analyses is a special effort made to accommodate them. Often they are minor, and ignoring them makes little difference, but in some problems they are important and significant, and we neglect them at our peril.

Areas of application of deconvolution, and regression with measurement error, include the analysis of seismological data (e.g. Kragh and Laws, 2006), financial analysis (e.g. Bonhomme and Robin, 2008), disease epidemiology (e.g. Brookmeyer and Gail, 1994, Chapter 8), and nutrition.

The latter topic is of particular interest today, for example in connection with errors-in-variables problems for data gathered in food frequency questionnaires (FFQs), or dietary questionnaires for epidemiological studies (DQESs). Formally, an FFQ is 'A method of dietary assessment in which subjects are asked to recall how frequently certain foods were consumed during a specified period of time,' according to the Nutrition Glossary of the European Food Information Council. An FFQ seeks detailed information about the nature and quantity of food eaten by the person filling in the form, and often includes a query such as, "How many of the above servings are from fast food outlets (McDonalds, Taco Bell, etc.)?" (Stanford University, 1994). This may seem a simple question to answer, but nutritionists interested in our consumption of fat generally find that the quantity of fast food that people admit to eating is biased downwards from its true value. The significant concerns in Western society about fat intake, and about where we purchase our oleaginous food, apparently influences our truthfulness when we are asked probing questions about our eating habits.

Examples of the use of statistical deconvolution in this area include the work of Stefanski and Carroll (1990) and Delaigle and Gijbels (2004b), who address nonparametric density deconvolution from measurement-error data, obtained from FFQs during the second National Health and Nutrition Examination Survey (1976–1980); Carroll *et al.* (1997), who discuss design and analysis aspects of linear measurement-error models when data come from FFQs; Carroll *et al.* (2006), who use measurement-error models, and deconvolution methods, to develop marginal mixed measurement-error models for each nutrient in a nutrition study, again when FFQs are used to supply the data; and Staudenmayer *et al.* (2008), who employ a dataset from nutritional epidemiology to illustrate the use of techniques for nonparametric density deconvolution. See Carroll *et al.* (2006, p. 7) for further discussion of applications to data on nutrition.

How might we correct for errors in variables? One approach is to use methods based on deconvolution, as follows. Let us write Q for the quantity of fast food that a person admits to eating, in a food frequency questionnaire; let Q_0 denote the actual amount of fast food; and put $R = Q/Q_0$. We expect that the distribution of R will be skewed towards values greater than 1, and we might even have an idea of the shape of the distribution responsible for this effect, i.e. the distribution of $\log R$. Indeed, we typically work with the logarithm of the formula $Q = Q_0\, R$, and in that context, writing $W = \log Q$, $X = \log Q_0$ and $U = \log R$, the equation defining the variables of interest is:

$$W = X + U\,. \tag{1.1}$$

We have data on W, and from that we wish to estimate the distribution of X, i.e. the distribution of the logarithm of fast-food consumption.

It can readily be seen that this problem is generally not solvable unless the distribution of U, and the joint distribution of X and U, are known. In practice we usually take X and U to be independent, and undertake empirical deconvolution (i.e. estimation of the distribution, or density, of X from data on W) for several candidates for the distribution of U. If we are able to make repeated measurements of X, in particular to gather data on $W^{(j)} = X + U^{(j)}$ for $1 \leq j \leq m$, say, then we have an opportunity to estimate the distribution of U as well.

It is generally reasonable to assume that X, $U^{(1)}$, $\ldots$, $U^{(M)}$ are independent random variables. The distribution of U can be estimated whenever $m \geq 2$ and the distribution is uniquely determined by $|\phi_U|^2$, where ϕ_U denotes the characteristic function of U. The simplest example of this type is arguably that where U has a symmetric distribution for which the characteristic function does not vanish on the real line. One example of repeated measurements in the case $m = 2$ is that where a food frequency questionnaire asks at one point how many times we visited a fast food outlet, and on a distant page, how many hamburgers or servings of fried chicken we have purchased.

The model at (1.1) is simple and interesting, but in examples from nutrition science, and in many other problems, we generally wish to estimate the response to an explanatory variable, rather than the distribution of the explanatory variable. Therefore the proper context for our food frequency questionnaire example is really regression, not distribution or density estimation. In regression with errors in variables we observe data pairs (W, Y), where

$$W = X + U\,, \quad Y = g(X) + V\,, \tag{1.2}$$

$g(x) = E(Y \,|\, X = x)$, and the random variable V, denoting an experimental error, has zero mean. In this case the standard regression problem is altered on account of errors that are incurred when measuring the value of the explanatory variable. In (1.2) the variables U, V and X are assumed to be independent.

The measurement error U, appearing in (1.1) and (1.2), can be interpreted as the result of a ‘laboratory error’ in determining the ‘dose’ X which is applied to the subject. For example, a laboratory technician might use the dose X in an experiment, but in attempting to determine the dose after the experiment they might commit an error U, with the result that the actual dose is recorded as $X + U$ instead of X. Another way of modelling the effect of measurement error is to reverse the roles of X and W, so that we observe (W, Y) generated as

$$X = W + U\,, \quad Y = g(X) + V\,. \tag{1.3}$$

Here a precise dose W is specified, but when measuring it prior to the experiment our technician commits an error U, with the result that the actual dose is $W + U$. In (1.3) it assumed that U, V and W are independent.

The measurement error model (1.2) is standard. The alternative model (1.3) is believed to be much less common, although in some circumstances it is difficult to determine which of (1.2) and (1.3) is the more appropriate. The model at (1.3) was first suggested by Berkson (1950), for whom it is named.

1.3 Kernel methods

If the measurement error U were very small then we could estimate the density f of X, and the function g in the model (1.2), using standard kernel methods. For example, given data $X_1, \ldots, X_n$ on X we could take

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{1.4}$$

to be our estimator of $f(x)$. Here K is a kernel function and h, a positive quantity, is a bandwidth. Likewise, given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ on (X, Y) we could take

$$\hat{g}(x) = \frac{\sum_i Y_i\, K\{(x - X_i)/h\}}{\sum_i K\{(x - X_i)/h\}} \tag{1.5}$$

to be our estimator of $g(x)$, where g is as in the model at (1.2).

The estimator at (1.4) is a standard kernel density estimator, and is itself a probability density if we take K to be a density. It is consistent under particularly weak conditions, for example if f is continuous and $h \to 0$ and $nh \to \infty$ as n increases. Density estimation is discussed at length by Silverman (1986) and Scott (1992). The estimator $\hat{g}$, which we generally also compute by taking K to be a probability density, is often referred to as the ‘local constant’ or Nadaraya–Watson estimator of g. The first of these names follows from the fact that $\hat{g}(x)$ is the result of fitting a constant to the data by local least squares:

$$\hat{g}(x) = \operatorname*{argmin}_{c} \sum_{i=1}^{n} (Y_i - c)^2 \, K\Big(\frac{x - X_i}{h}\Big) . \tag{1.6}$$

The estimator $\hat{g}$ is also consistent under mild conditions, for example if the variance of the error, V, in (1.2) is finite, if f and g are continuous, if $f > 0$ at the point x where we wish to estimate g, and if $h \to 0$ and $nh \to \infty$ as n increases. General kernel methods are discussed by Wand and Jones (1995), and statistical smoothing is addressed by Simonoff (1996).

Local constant estimators have the advantage of being relatively robust against uneven spacings in the sequence $X_1, \ldots, X_n$. For example, the ratio at (1.5) never equals a nonzero number divided by zero. However, local constant estimators are particularly susceptible to boundary bias. In particular, if the density of X is supported and bounded away from zero on a compact interval, then $\hat{g}$, defined by (1.5) or (1.6), is generally inconsistent at the endpoints of that interval. Issues of this type have motivated the use of local polynomial estimators, which are defined by $\hat{g}(x) = \hat{c}_0(x)$ where, in a generalisation of (1.6),

$$(\hat{c}_0(x), \ldots, \hat{c}_p(x)) = \operatorname*{argmin}_{(c_0,\ldots,c_p)} \sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=0}^{p} c_j \, (x - X_i)^j \Big\}^2 \, K\Big(\frac{x - X_i}{h}\Big) . \tag{1.7}$$

See, for example, Fan and Gijbels (1996). In (1.7), p denotes the degree of the locally fitted polynomial. The estimator $\hat{g}(x) = \hat{c}_0(x)$, defined by (1.7), is also consistent under the conditions given earlier for the estimator defined by (1.5) and (1.6).

In the particular case $p = 1$ we obtain a local-linear estimator of $g(x)$:

$$\hat{g}(x) = \frac{S_2(x)\, T_0(x) - S_1(x)\, T_1(x)}{S_0(x)\, S_2(x) - S_1(x)^2} , \tag{1.8}$$

where

$$\begin{aligned} S_r(x) &= \frac{1}{nh} \sum_{i=1}^{n} \Big(\frac{x - X_i}{h}\Big)^r \, K\Big(\frac{x - X_i}{h}\Big) , \\ T_r(x) &= \frac{1}{nh} \sum_{i=1}^{n} Y_i \Big(\frac{x - X_i}{h}\Big)^r \, K\Big(\frac{x - X_i}{h}\Big) , \end{aligned} \tag{1.9}$$

h denotes a bandwidth and K is a kernel function.

Estimators of all these types can be quickly extended to cases where errors in variables are present, for example as in the models at (1.1) and (1.2), simply by altering the kernel function K so that it acts to cancel out the influence of the errors. We shall give details in section 2. Section 3 will discuss recently introduced methodology which, from some viewpoints looks quite different from, but is actually almost identical to, kernel methods.

2 Methodology and theory

2.1 Definitions of estimators

We first discuss a generalisation of the estimator at (1.4) to the case where there are errors in the observations of X_i, as per the model at (1.1). In particular, we assume that we observe data $W_1, \ldots, W_n$ which are independent and identically distributed as $W = X + U$, where X and U are independent and the distribution of U has known characteristic function ϕ_U which does not vanish anywhere on the real line. Let K be a kernel function, write $\phi_K = \int e^{itx}\, K(x)\, dx$ for the associated Fourier transform, and define

$$K_U(x) = \frac{1}{2\pi} \int e^{-itx}\, \frac{\phi_K(t)}{\phi_U(t/h)}\, dt\,. \tag{2.1}$$

Then, to construct an estimator $\hat{f}$ of the density $f = f_X$ of X, when all we observe are the contaminated data $W_1, \ldots, W_n$, we simply replace K by K_U, and X_i by W_i, in the definition of $\hat{f}$ at (1.4), obtaining the estimator

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

Here the subscript 'decon' signifies that $\hat{f}_{\text{decon}}$ involves empirical deconvolution. The adjustment to the kernel takes care of the measurement error, and results in consistency in a wide variety of settings. Likewise, if data pairs $(W_1, Y_1), \ldots, (W_n, Y_n)$ are generated under the model at (1.2) then, to construct the local constant estimator at (1.5), or the local linear estimator defined by (1.8) and (1.9), all we do is replace each X_i by W_i, and K by K_U. Other local polynomial estimators can be calculated using a similar rule, replacing $h^{-r}(x - X_i)^r K\{(x - X_i)/h\}$ in S_r and T_r by $K_{U,r}\{(x - W_i)/h\}$, where

$$K_{U,r}(x) = \frac{1}{2\pi i^r} \int e^{-itx}\, \frac{\phi_K^{(r)}(t)}{\phi_U(t/h)}\, dt\,.$$

The estimator at (2.2) dates from work of Carroll and Hall (1988) and

Stefanski and Carroll (1990). Deconvolution-kernel regression estimators in the local-constant case were developed by Fan and Truong (1993), and extended to the general local polynomial setting by Delaigle *et al.* (2009).

The kernel K_U is deliberately constructed to be the function whose Fourier transform is ϕ_K/ϕ_U. This adjustment permits cancellation of the influence of errors in variables, as discussed at the end of section 1.3. To simplify calculations, for example computation of the integral in (1.2), we generally choose K not to be a density function but to be a smooth, symmetric function for which ϕ_K vanishes outside a compact interval. The commonly-used candidates for ϕ_K are proportional to functions that are used for K, rather than ϕ_K, in the case of regular kernel estimation discussed in section 1.3. For example, kernels K for which $\phi_K(t) = (1-|t|^r)^s$ for $|t| \le 1$, and $\phi_K(t) = 0$ otherwise, are common; here r and s are integers. Taking $r = 2s = 2$, $r = s = 2$ and $r = \frac{2}{3}\,s = 2$ corresponds to the Fourier inverses of the biweight, quartic and triweight kernels, respectively. Taking $s = 0$ gives the inverse of the uniform kernel, i.e. the sinc kernel, which we shall meet again in section 3. Further information about kernel choice is given by Delaigle and Hall (2006).

These kernels, and others, have the property that $\phi_K(t) = 1$ when $t = 0$, thereby guaranteeing that $\int K = 1$. The latter condition ensures that the density estimator, defined at (2.2) and constructed using this kernel, integrates to 1. (However, the estimator defined by (2.2) will generally take negative values at some points x.) The normalisation property is not so important when the kernel is used to construct regression estimators, where the effects of multiplying K by a constant factor cancel from the 'deconvolution' versions of formulae (1.5) and (1.8), and likewise vanish for all deconvolution-kernel estimators based on local polynomial methods.

Note that, as long as ϕ_K and ϕ_U are supported either on the whole real line or on a symmetric compact domain, the kernel K_U, defined by (2.1), and its generalised form $K_{U,r}$, are real-valued. Indeed, using properties of the complex conjugate of Fourier transforms of real-valued functions, and the change of variable $u = -t$, we have, using the notation $\overline{a}(t)$ for the complex conjugate of a complex-valued function a of a real variable t,

$$\begin{aligned}\overline{K}_{U,r}(x) &= (-1)^{-r}\frac{1}{2\pi i^r}\int e^{itx}\,\frac{\overline{\phi_K^{(r)}}(t)}{\overline{\phi}_U(t/h)}\,dt\\ &= (-1)^{-r}\frac{1}{2\pi i^r}\int e^{itx}\,\frac{(-1)^{-r}\phi_K^{(r)}(-t)}{\phi_U(-t/h)}\,dt\end{aligned}$$

$$= \frac{1}{2\pi i^r} \int e^{-iux} \, \frac{\phi_K^{(r)}(u)}{\phi_U(u/h)} \, du = K_{U,r}(x).$$

In practice it is almost always the case that the distribution of U is symmetric, and in the discussion of variance in section 2.2, below, we shall make this assumption. We shall also suppose that K is symmetric, again a condition which holds almost invariably in practice.

The estimators discussed above were based on the assumption that the characteristic function ϕ_U of the errors in variables is known. This enabled us to compute the deconvolution kernel K_U at (2.1). In cases where the distribution of U is not known, but can be estimated from replicated data (see section 1.2), we can replace ϕ_U by an estimator of it and, perhaps after a little regularisation, compute an empirical version of K_U. This can give good results, in both theory and practice. In particular, in many cases the resulting estimator of the density of X, or the regression mean g, can be shown to have the same first-order properties as estimators computed under the assumption that the distribution of U is known. Details are given by Delaigle *et al.* (2008).

Methods for choosing the smoothing parameter, h, in the estimators discussed above have been proposed by Hesse (1999), Delaigle and Gijbels (2004a,b) and Delaigle and Hall (2008).

2.2 Bias and variance

The expected value of the estimator at (2.2) equals

$$\begin{aligned} E\{\hat{f}_{\text{decon}}(x)\} &= \frac{1}{2\pi h} \int E\big[e^{-it\{x-W\}/h}\big] \, \frac{\phi_K(t)}{\phi_U(t/h)} \, dt \\ &= \frac{1}{2\pi} \int e^{-itx} \frac{\phi_K(ht)}{\phi_U(t)} \, \phi_X(t) \, \phi_U(t) \, dt \\ &= \frac{1}{2\pi} \int e^{-itx} \phi_K(ht) \, \phi_X(t) \, dt = \frac{1}{h} \int K(u/h) \, f(x-u) \, du \\ &= E\{\hat{f}(x)\} \,, \end{aligned} \tag{2.3}$$

where the first equality uses the definition of K_U, and the fourth equality uses Plancherel's identity. Therefore the deconvolution estimator $\hat{f}_{\text{decon}}(x)$, calculated from data contaminated by measurement errors, has exactly the same mean, and therefore the same bias, as $\hat{f}(x)$, which would be computed using values of X_i observed without measurement error. This confirms that using the deconvolution kernel estimator does indeed allow for cancellation of measurement errors, at least in terms of their presence in the mean.

Of course, variance is a different matter. Since $\hat{f}_{\text{decon}}(x)$ equals a sum of independent random variables then

$$\begin{aligned}&\text{var}\{\hat{f}_{\text{decon}}(x)\}\\ &\quad= \big(nh^2\big)^{-1}\,\text{var}\Big\{K_U\Big(\frac{x-W}{h}\Big)\Big\}\\ &\quad\sim (nh)^{-1}\,f_W(x)\int K_U^2 = \frac{f_W(x)}{2\pi nh}\int \phi_K(t)^2\,|\phi_U(t/h)|^{-2}\,dt\,. \qquad (2.4)\end{aligned}$$

(Here the relation $\sim$ means that the ratio of the left- and right-hand sides converges to 1 as $h \to 0$.) Thus it can be seen that the variance of $\hat{f}_{\text{decon}}(x)$ depends intimately on tail behaviour of the characteristic function ϕ_U of the measurement-error distribution.

If ϕ_K vanishes outside a compact set, which, as we noted in section 2.1, is generally the case, and if $|\phi_U|$ is asymptotic to a positive regularly varying function ψ (see Bingham *et al.*, 1989), in the sense that $|\phi_U(t)| \asymp \psi(t)$ (meaning that the ratio of both sides is bounded away from zero and infinity as $t \to \infty$), then the integral on the right-hand side of (2.3) is bounded between two constant multiples of $\psi(1/h)^{-2}$ as $h \to 0$. Therefore by (2.4), provided that $f_W(x) > 0$,

$$\text{var}\{\hat{f}_{\text{decon}}(x)\} \asymp (nh)^{-1}\,\psi(1/h)^{-2} \qquad (2.5)$$

as n increases and h decreases. Recall that we are assuming that f_U and K are both symmetric functions.

If the density f of X has two bounded and continuous derivatives, and if K is bounded and symmetric and satisfies $\int x^2\,|K(x)|\,dx < \infty$, then the bias of $\hat{f}_{\text{decon}}$ can be found from (2.3), using elementary calculus and arguments familiar in the case of standard kernel estimators:

$$\begin{aligned}\text{bias}(x) &= E\{\hat{f}_{\text{decon}}(x)\} - f(x) = E\{\hat{f}(x)\} - f(x)\\ &= \int K(u)\,\{f(x-hu) - f(x)\}\,du = \tfrac{1}{2}\,h^2\,\kappa\,f''(x) + o\big(h^2\big) \qquad (2.6)\end{aligned}$$

as $h \to 0$, where $\kappa = \int x^2\,K(x)\,dx$. Therefore, provided that $f''(x) \neq 0$, the bias of the conventional kernel estimator $\hat{f}(x)$ is exactly of size h^2 as $h \to 0$. Combining this property, (2.3) and (2.5) we deduce a relatively concise asymptotic formula for the mean squared error of $\hat{f}_{\text{decon}}(x)$:

$$E\{\hat{f}_{\text{decon}}(x) - f(x)\}^2 \asymp h^4 + (nh)^{-1}\,\psi(1/h)^{-2}\,. \qquad (2.7)$$

For a given error distribution we can work out the behaviour of $\psi(1/h)$

as $h \to 0$, and then from (2.7) we can calculate the optimal bandwidth and determine the exact rate of convergence of $\hat{f}_{\text{decon}}(x)$ to $f(x)$, in mean square. In many instances this rate is optimal, in a minimax sense; see, for example, Fan (1991). It is also generally optimal in the case of the errors-in-variables regression estimators discussed in section 2.1, based on deconvolution-kernel versions of local polynomial estimators. See Fan and Truong (1993).

Therefore, despite their almost naive simplicity, deconvolution-kernel estimators of densities and regression functions have features that can hardly be bettered by more complex, alternative approaches. The results derived in the previous paragraph, and their counterparts in the regression case, imply that the estimators are limited by the extent to which they can recover from the data. (This is reflected in the fact that the rate of decay of the tails of ϕ_U drives the results on convergence rates.) However, the fact that the estimators are nevertheless optimal, in terms of their rates of convergence, implies that this restriction is inherent to the problem, not just to the estimators; no other estimators would have a better convergence rate, at least not uniformly in a class of problems.

3 Relationship to minimum contrast methods

3.1 Deconvolution kernel estimators based on the sinc kernel

The sinc, or Fourier integral, kernel is given by

$$L(x) = \begin{cases} (\pi x)^{-1} \sin(\pi x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \,. \end{cases} \tag{3.1}$$

Its Fourier transform, defined as a Riemann integral, is the 'boxcar function', $\phi_L(t) = 1$ if $|t| \leq 1$ and $\phi_L(t) = 0$ otherwise. In particular, ϕ_L vanishes outside a compact set, which property, as we noted in section 2.1, aids computation. The version of K_U, at (2.1), for the sinc kernel is

$$L_U(x) = \frac{1}{2\pi} \int_{-1}^{1} e^{-itx}\, \phi_U(t/h)^{-1}\, dt = \frac{1}{\pi} \int_0^1 \cos(tx)\, \phi_U(t/h)^{-1}\, dt \,,$$

where the second identity holds if the distribution of U is symmetric and has no zeros on the real line.

The kernel L is sometimes said to be of 'infinite order', in the sense that if a is any function with an infinite number of bounded, integrable derivatives then

$$\int \left[\int \{a(x+hu)-a(x)\}\, L(u)\, du \right]^2 dx = O\big(h^r\big) \qquad (3.2)$$

as $h \downarrow 0$, for all $r > 0$. If K were of finite order then (3.2) would hold only for a finite range of values of r, no matter how many derivatives the function a enjoyed. For example, if K were a symmetric function for which $\int u^2\, K(u)\, du \neq 0$, and if we were to replace L in (3.2) by K, then (3.2) would hold only for $r \leq 4$, not for all r. In this case we would say that K was of second order, because

$$\int \{a(x+hu)-a(x)\}\, K(u)\, du = O\big(h^2\big)\, .$$

If we take a to be the density, f, of the random variable X, and take K in the definition of $\hat{f}$ at (1.4) to be the sinc kernel, L, then (3.2) equals the integral of the squared bias of $\hat{f}$. Therefore, in the case of a very smooth density, the 'infinite order' property of the sinc kernel ensures particularly small bias, in an average sense.

Properties of conventional kernel density estimators, but founded on the sinc kernel, for data without measurement errors, have been studied by, for example, Davis (1975, 1977). Glad *et al.* (1999)have provided a good survey of properties of sinc kernel methods for density estimation, and have argued that those estimators have received an unfairly bad press. Despite criticism of sinc kernel estimators (see e.g. Politis and Romano, 1999), the approach is "more accurate for quite moderate values of the sample size, has better asymptotics in non-smooth cases (the density to be estimated has only first derivative), [and] is more convenient for bandwidth selection etc" than its conventional competitors, suggest Glad *et al.* (1999).

The property of greater accuracy is borne out in both theoretical and numerical studies, and derives from the infinite-order property noted above. Indeed, if f is very smooth then the low level of average squared bias can be exploited to produce an estimator $\hat{f}$ with particularly low mean squared error, in fact of order n^{-1} in some cases. The most easily seen disadvantage of sinc-kernel density estimators is their tendency to suffer from spurious oscillations, inherited from the infinite number of oscillations of the kernel itself.

These properties can be expected to carry over to density and regression estimators based on contaminated data, when we use the sinc kernel. To give a little detail in the case of density estimation from data contaminated by measurement errors, we note that if the density f of X is infinitely differentiable, but we observe only the contaminated data $W_1, \ldots, W_n$ distributed as W, generated as at (1.1); if we use the density estimator at (1.4), but computed using $K = L$, the sinc kernel; and if $|\phi_U(t)| \geq C\,(1+|t|)^{-\alpha}$ for constants C, $\alpha > 0$; then, in view of (2.3), (2.4) and (3.2), we have for all $r > 0$,

$$\begin{aligned}
&\int \{\hat{f}_{\text{decon}}(x) - f(x)\}^2\,dx \\
&\qquad = \int \{E\hat{f}(x) - f(x)\}^2 + \big(nh^2\big)^{-1} \int \text{var}\Big\{L_U\Big(\frac{x-W}{h}\Big)\Big\}\,dx \\
&\qquad \leq \int \bigg[\int \{f(x+hu) - f(x)\}\,L(u)\,du\bigg]^2\,dx + (nh)^{-1}\int L_U^2 \\
&\qquad = O\bigg\{h^r + (nh)^{-1}\int_{-1}^{1} |\phi_U(t/h)|^{-2}\,dt\bigg\} \\
&\qquad = O\Big\{h^r + \big(nh^{2\alpha+1}\big)^{-1}\Big\}. \qquad (3.3)
\end{aligned}$$

It follows that, if f has infinitely many integrable derivatives and if the tails of $\phi_U(t)$ decrease at no faster than a polynomial rate as $|t| \to \infty$, then the bandwidth h can be chosen so that the mean integrated squared error of a deconvolution kernel estimator of f, using the sinc kernel, converges at rate $O(n^{\epsilon-1})$ for any given $\epsilon > 0$.

This very fast rate of convergence contrasts with that which occurs if the kernel K is of only finite order. For example, if K is a second-order kernel, in which case (3.2) holds only for $r \leq 4$ when L is replaced by K, the argument at (3.3) gives:

$$\int \{\hat{f}_{\text{decon}}(x) - f(x)\}^2\,dx = O\Big\{h^4 + \big(nh^{2\alpha+1}\big)^{-1}\Big\}.$$

The fastest rate of convergence of the right-hand side to zero is attained with $h = n^{-1/(2\alpha+5)}$, giving

$$\int \{\hat{f}_{\text{decon}}(x) - f(x)\}^2\,dx = O\big(n^{-4/(2\alpha+5)}\big).$$

In fact, this is generally the best rate of convergence of mean integrated squared error that can be obtained using a second-order kernel when the characteristic function ϕ_U decreases like $|t|^{-\alpha}$ in the tails, even if the density f is exceptionally smooth. Nevertheless, second-order kernels are often preferred to the sinc kernel in practice, since they do not suffer from the unwanted oscillations that afflict estimators based on the sinc kernel.

3.2 Minimum contrast estimators, and their relationship to deconvolution kernel estimators

In the context of the measurement error model at (1.1), Comte *et al.* (2007) suggested an interesting minimum contrast estimator of the density f of X. Their approach has applications in a variety of other settings (see Comte *et al.*, 2006, 2008; Comte and Taupin, 2007), including to the regression model at (1.2), and the conclusions we shall draw below apply in these cases too. Therefore, for the sake of brevity we shall treat only the density deconvolution problem.

To describe the minimum contrast estimator in that setting, define

$$\hat{a}_{k\ell} = \frac{1}{2\pi n} \sum_{j=1}^{n} \int \exp(it\, W_j)\, \phi_{L_{k\ell}}(t)\, \phi_U(t)^{-1}\, dt\,,$$

where $\phi_{L_{k\ell}}$ denotes the Fourier transform of the function $L_{k\ell}$ defined by $L_{k\ell}(x) = \ell^{1/2}\, L(\ell\, x - k)$, k is an integer and $\ell > 0$. In this notation the minimum contrast nonparametric density estimator is

$$\tilde{f}(x) = \sum_{k=-k_0}^{k_0} \hat{a}_{k\ell}\, L_{k\ell}(x)\,.$$

There are two tuning parameters, k_0 and ℓ. Comte *et al.* (2007) suggest choosing ℓ to minimise a penalisation criterion.

The resulting minimum contrast estimator is called a penalised contrast density estimator. The penalisation criterion suggested by Comte *et al.* (2007) for choosing ℓ is related to cross-validation, although its exact form, which involves the choice of additional terms and multiplicative constants, is based on simulation experiments. It is clear on inspecting the definition of $\tilde{f}$ that ℓ plays a role similar to that of the inverse of bandwidth in a conventional deconvolution kernel estimator. In particular, ℓ should diverge to infinity with n. Comte *et al.* (2007) suggest taking $k_0 = 2^m - 1$, where $m \geq \log_2(n+1)$ is an integer. In numerical experiments they use $m = 8$, which gives good performance in the cases they consider. More generally, k_0/ℓ should diverge to infinity as sample size increases.

The minimum contrast density estimator of Comte *et al.* (2007) is actually very close to the standard deconvolution kernel density estimator at (1.4), where in the latter we use the sinc kernel at (3.1). Indeed, as the theorem below shows, the two estimators are exactly equal on a grid, which becomes finer as the bandwidth, h, for the sinc kernel density estimator decreases. However, this relationship holds only for values of x

for which $|x| \leq k_0/\ell$; for larger values of $|x|$ on the grid, $\tilde{f}(x)$ vanishes. (This property is one of the manifestations of the fact that, as noted earlier, k and ℓ generally should be chosen to depend on sample size in such a manner that $k_0/\ell \to \infty$ as $n \to \infty$.)

Theorem *Let $\hat{f}_{\mathrm{decon}}$ denote the deconvolution kernel density estimator at (1.4), constructed using the sinc kernel and employing the bandwidth $h = \ell^{-1}$. Then, for any point $x = hk$ with k an integer, we have*

$$\tilde{f}(x) = \begin{cases} \hat{f}_{\mathrm{decon}}(x) & \text{if } |x| \leq k_0/\ell \\ 0 & \text{if } |x| > k_0/\ell\,. \end{cases}$$

A proof of the theorem will be given in section 3.3. Between grid points the estimator $\tilde{f}$ is a nonstandard interpolation of values of the kernel estimator $\hat{f}_{\mathrm{decon}}$. Note that, if we take $h = \ell^{-1}$, the weights $L(\ell x - k) = L\{(x - hk)/h\}$ used in the interpolation decrease quickly as k moves further from x/h, and, except for small k, neighbour weights are close in magnitude but differ in sign. (Here L is the sinc kernel defined at (3.1).) In effect, the interpolation is based on rather few values $\hat{f}_{\mathrm{decon}}(k/\ell)$ corresponding to those k for which k is close to x/h.

In practice the two estimators are almost indistinguishable. For example, Figure 3.1 compares them using the bandwidth that minimises the integrated squared difference between the true density and the estimator, for one generated sample in the case where X is normal $\mathrm{N}(0,1)$, U is Laplace with $\mathrm{var}(U)/\mathrm{var}(X) = 0.1$, and $n = 100$ or $n = 1000$. In the left graphs the two estimators can hardly be distinguished. The right graphs show magnifications of these estimators for $x \in [-\frac{1}{2}, 0]$. Here it can be seen more clearly that the minimum contrast estimator is an approximation of the deconvolution kernel estimator, and is exactly equal to the latter at $x = 0$.

These results highlight the fact that the differences in performance between the two estimators derive more from different tuning parameter choices than from anything else. In their comparison, Comte *et al.* (2007) used a minimum contrast estimator with the sinc kernel L and a bandwidth chosen by penalisation, whereas for the deconvolution kernel estimator they employed a conventional second-order kernel K and a different bandwidth-choice procedure. Against the background of the theoretical analysis in section 3.1, the different kernel choices (and different ways of choosing smoothing parameters) explain the differences observed between the penalised contrast density estimator and the deconvolution kernel density estimator based on a second-order kernel.

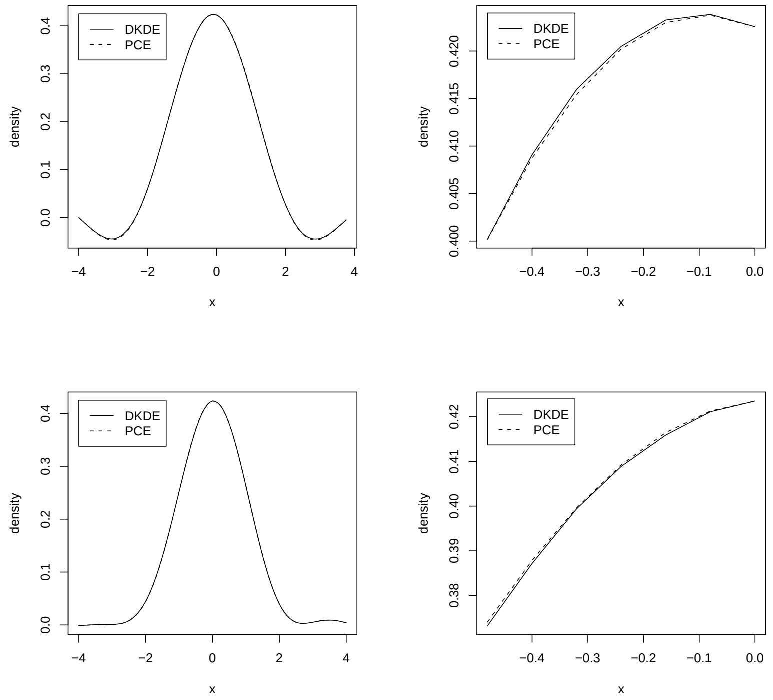

Figure 3.1 Deconvolution kernel density estimator (DKDE) and minimum contrast estimator (PCE) for a particular sample of size $n = 100$ (upper panels) or $n = 1000$ (lower panels) in the case $\mathrm{var}(U)/\mathrm{var}(X) = 0.1$. Right panels show magnifications of the estimates for $x \in [-0.5, 0]$ in the respective upper panels.

3.3 Proof of Theorem

Note that $\phi_{L_{k\ell}}(t) = \ell^{-1/2} \exp(itk/\ell)\, \phi_L(t/\ell)$ and

$$\hat{a}_{k\ell} = \frac{1}{2n\pi\ell^{1/2}} \sum_{j=1}^{n} \int_{-\ell\pi}^{\ell\pi} \exp\bigl\{ -it\,\bigl(k\,\ell^{-1} - W_j\bigr)\bigr\}\, \frac{\phi_L(t/\ell)}{\pi_U(t)}\, dt\,.$$

Therefore,

$$\tilde{f}(x)$$

$$= \frac{1}{2n\pi} \sum_{k=-k_0}^{k_0} L(\ell x - k) \sum_{j=1}^{n} \int_{-\ell\pi}^{\ell\pi} \exp\bigl\{ -it\bigl(k\ell^{-1} - W_j\bigr)\bigr\} \frac{\phi_L(t/\ell)}{\pi_U(t)}\, dt$$

$$= \sum_{k=-k_0}^{k_0} L(\ell x - k)\, \hat{f}_{\mathrm{decon}}(k/\ell)\,. \tag{3.4}$$

If r is a nonzero integer then $L(r) = 0$. Therefore, if $x = kh = s/\ell$ for an integer s then $L(\ell x - k) = 0$ whenever $k \neq s$, and $L(\ell x - k) = 1$ if $k = s$. Hence, (3.4) implies that $\tilde{f}(x) = \hat{f}_{\mathrm{decon}}(x)$ if $|k| \leq k_0$, and $\tilde{f}(x) = 0$ otherwise.

References

Barry, J., and Diggle, P. J. 1995. Choosing the smoothing parameter in a Fourier approach to nonparametric deconvolution of a density estimate. *J. Nonparametr. Stat.*, **4**, 223–232.

Berkson, J. 1950. Are there two regression problems? *J. Amer. Statist. Assoc.*, **45**, 164–180.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. 1989. *Regular Variation*, revised ed. Encyclopedia Math. Appl., vol. 27. Cambridge: Cambridge Univ. Press.

Bonhomme, S., and Robin, J.-M. 2008. *Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics*. University College London, Centre for Microdata Methods & Practice working paper 3/08; `http://www.cemmap.ac.uk/wps/cwp308.pdf`.

Brookmeyer, R., and Gail, M. H. 1994. *AIDS Epidemiology: a Quantitative Approach*. Oxford: Oxford Univ. Press.

Butucea, C. 2004. Deconvolution of supersmooth densities with smooth noise. *Canad. J. Statist.*, **32**, 181–192.

Butucea, C., and Tsybakov, A. B. 2007a. Sharp optimality for density deconvolution with dominating bias, I. *Theory Probab. Appl.*, **52**, 111–128.

Butucea, C., and Tsybakov, A. B. 2007b. Sharp optimality for density deconvolution with dominating bias, II. *Theory Probab. Appl.*, **52**, 336–349.

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., Freedman, L. S., and Pee, D. 1997. Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics*, **53**, 1440–1457.

Carroll, R. J., Midthune, D., Freedman, L. S., and Kipnis, V. 2006. Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, **62**, 75–84.

Comte, F., Rozenholc, Y., and Taupin, M.-L. 2006. Penalized contrast estimator for adaptive density deconvolution. *Canad. J. Statist.*, **34**, 431–452.

Comte, F., Rozenholc, Y., and Taupin, M.-L. 2007. Finite sample penalization in adaptive density deconvolution. *J. Stat. Comput. Simul.*, **77**, 977–1000.

Comte, F., Rozenholc, Y., and Taupin, M.-L. 2008. Adaptive density estimation for general ARCH models. *Econometric Theory*, **24**, 1628–1662.

Comte, F., and Taupin, M.-L. 2007. Nonparametric estimation of the regression function in an errors-in-variables model. *Statist. Sinica*, **17**, 1065–1090.

Davis, K. B. 1975. Mean square error properties of density estimates. *Ann. Statist.*, **3**, 1025–1030.

Davis, K. B. 1977. Mean integrated square error properties of density estimates. *Ann. Statist.*, **5**, 530–535.

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(485)**, 348–359.

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

Delaigle, A., and Gijbels, A. 2004b. Practical bandwidth selection in deconvolution kernel density estimation. *Comput. Statist. Data Anal.*, **45**, 249–267.

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

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

Delaigle, A., Hall, P., and Meister, A. 2008. On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–685.

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

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.* London: Chapman and Hall.

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

Glad, I. K., Hjort, N. L., and Ushakov, N. 1999. *Density Estimation Using the Sinc Kernel.* Department of Mathematical Sciences, Norwegian University of Science & Technology, Trondheim, Statistics Preprint No. 2/2007; `http://www.math.ntnu.no/preprint/statistics/2007/S2-2007.pdf`.

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

Kragh, E., and Laws, R. 2006. Rough seas and statistical deconvolution. *Geophysical Prospecting*, **54**, 475–485.

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

Politis, D. N., and Romano, J. P. 1999. Multivariate density estimation with general flat-top kernels of infinite order. *J. Multivariate Anal.*, **68**, 1–25.

Scott, D. W. 1992. *Multivariate Density Estimation. Theory, Practice, and Visualization.* New York: John Wiley & Sons.

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

Simonoff, J. S. 1996. *Smoothing Methods in Statistics.* New York: Springer-Verlag.

Stanford University. 1994. *Food Frequency Questionnaire #1 2 3 4*. Available at `http://www.permanente.net/homepage/kaiser/pdf/6116.pdf`.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. P. 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*, **2**, 169–184.

Wand, M. P., and Jones, M. C. 1995. *Kernel Smoothing.* London: Chapman and Hall.