Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes.

Aurore Delaigle
Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia.

Abstract: Estimating a curve nonparametrically from data measured with error is a difficult problem that has been studied by many authors. Constructing a consistent estimator in this context can sometimes be quite challenging, and in this paper we review some of the tools that have been developed in the literature for kernel-based approaches, founded on the Fourier transform and a more general unbiased score technique. We use those tools to rederive some of the existing nonparametric density and regression estimators for data contaminated by classical or Berkson errors, and discuss how to compute these estimators in practice. We also review some errors sometimes encountered in the area, and highlight a number of problems with an existing R package `decon`.

Keywords: bandwidth, deconvolution, density estimation, Matlab code for deconvolution estimator, Matlab code for nonparametric regression with errors-in variables, measurement errors, nonparametric curve estimation, R package `decon`, regression estimation.

Running head: Review of kernel methods with errors-in-variables.

1 Introduction

We consider nonparametric estimation of a density or a regression curve from data observed with errors. Measurement errors arise in data coming from a wide variety of applications (e.g. astronomy, nutrition, epidemiology and chemistry), because most quantities are measured with error, for instance due to imperfect measurement devices, inadequate reagent quality, or the impossibility to directly access the variables of interest. For example, in nutrition studies the long term saturated fat intake of patients is often approximated by the fat intake of one or two 24-hour recalls. Ignoring the errors present in the data when constructing estimators produces biased estimators, and as a result the topic of nonparametric curve estimation with errors-in-variables has received a great deal of attention over the last two decades. We review the main techniques suggested in the literature, with a focus on kernel estimators, which are by far the most popular and the most developed nonparametric errors-in-variables techniques.

Measurement errors are often classified in two types, called classical errors and Berkson errors. In the classical error case, the errors arise from an imprecise measurement of a quantity. For example, imprecision can be due to the inaccuracy of a measuring device (e.g. measurements in a lab), or the intrinsic difficulty in measuring the quantity of interest (e.g. cholesterol level or systolic blood pressure). In the Berkson error case, we do not observe directly the variable of interest, and measure instead a proxy that is linearly related to it. This type of errors arise typically in exposure studies, where, instead of measuring exposure of an individual to a toxic substance, we are only able to measure exposure at some fixed stations. Classical errors are those that have received the most attention in the literature; they are more common in practice, and are often simpler to deal with. Roughly speaking, methods that exist and are valid in the error-free case can be extended to the classical error context, using arguments similar to those employed in the error-free case. The sit-

uation is quite different for Berkson errors, where estimating densities is trivial, but estimating a regression curve is much more challenging.

Deriving consistent estimators in the errors-in-variables setting can be rather subtle, and arguments paralleling existing results in more standard contexts can sometimes lead to erroneous conclusions. We review some of the difficulties that can be encountered in this process, and explain a general technique based on unbiased scores which can be used to construct consistent estimators in a variety of settings. A key to solving many errors-in-variables problems is often to take the Fourier transform of the various functions involved. Once in the Fourier domain, equations generally become much simpler to solve; for example, convolutions become products, which are much easier to deal with.

We introduce the error models in section 2. In section 3, we explain techniques for deriving consistent estimators, and illustrate these approaches by constructing errors-in-variables density and regression estimators. Section 4 focuses on the practical implementation of estimators, provides links for Matlab (MathWorks, Inc., 2012) codes that compute the estimators, and points to a number of problems with the R (R Development Core Team, 2011) package `decon` of Wang and Wang (2011). Finally, in section 5 we expose some of the errors in reasoning that are sometimes encountered in the area.

2 Error models

2.1 Introduction

In the standard nonparametric density estimation problem, we wish to construct a nonparametric estimator of a density f_X, using a sample of independent and identically distributed (i.i.d.) data $X_1, \ldots, X_n$, where $X_i \sim f_X$. In the standard nonparametric regression problem, the goal is to estimate a regression curve m nonparamet-

rically, using a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d. observations modelled according to

$$Y_i = m(X_i) + \epsilon_i, \tag{2.1}$$

where $X_i \sim f_X$ and, for all x, $E(\epsilon_i | X_i = x) = 0$ and $\text{var}(\epsilon_i | X_i = x) < \infty$.

In the errors-in-variables context, we wish to estimate the same quantities, but instead of observing the X_i's, we observe a contaminated version of them. In this section we review the two main error types (classical and Berkson), and introduce a model that combines both types of errors.

2.2 Classical errors

In the classical measurement error context, we are interested in estimating the density f_X of a variable X, but we only observe a sample of i.i.d. data $W_1, \ldots, W_n$, where

$$W_i = X_i + U_i, \tag{2.2}$$

with $X_i \sim f_X$, $U_i \sim f_U$, and where the U_i's are i.i.d. and independent of the X_i's. The error density f_U is traditionally assumed to be known, and for simplicity we shall make that assumption throughout. When f_U is unknown, it can be easily estimated from replicated observations, with little impact on the conclusions drawn in this paper; see Li and Vuong (1998) and Delaigle, Hall and Meister (2008). Let $\phi_R(t) = \int e^{itx} f_R(x)\, dx$ denote the characteristic function of a generic random variable R. We assume throughout, as is commonly done in the literature, that f_U is symmetric and that ϕ_U satisfies $\phi_U(t) \neq 0$ for all t. Estimating f_X from the W_i's at (2.2) is often referred to as a deconvolution problem.

In the classical errors-in-variables regression context, we wish to estimate a regression curve m, but we only observe a sample $(W_1, Y_1), \ldots, (W_n, Y_n)$ of i.i.d. observations modelled according to

$$Y_i = m(X_i) + \epsilon_i, \quad W_i = X_i + U_i, \tag{2.3}$$

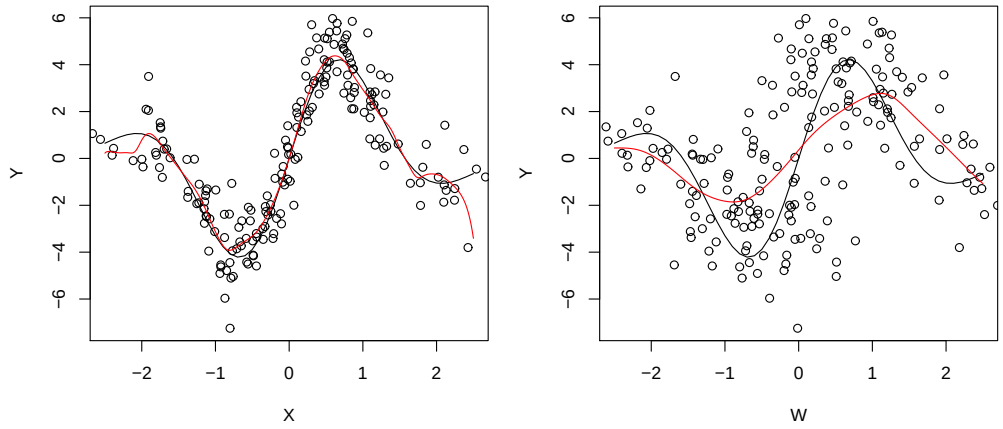

Figure 1: Scatterplot of the data (X_i, Y_i) (left) and (W_i, Y_i) (right) in the classical error case. The red curve is the standard local linear estimator of m computed from the data (X_i, Y_i) (left) or (W_i, Y_i) (right). The true m is depicted by the black line.

where $X_i \sim f_X$, $U_i \sim f_U$, the U_i's are i.i.d. and independent of the (X_i, ϵ_i)'s, and, for all x and for $i = 1, \ldots, n$, $E(\epsilon_i | X_i = x) = 0$ and $\text{var}(\epsilon_i | X_i = x) < \infty$. Moreover, the U_i's satisfy all the assumptions of model (2.2).

When the data are contaminated in this way, we cannot apply standard estimation techniques designed for error-free data: when computed with contaminated data, they produce biased estimators, which often perform rather poorly. To illustrate this, we generated a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size $n = 200$ from the model at (2.1), and then, following (2.3), added normal errors U_i to the X_i's, taking the noise to signal ratio, $\text{var}(U)/\text{var}(X)$, equal to 20%. We then computed the standard local linear estimator of m using the data (X_i, Y_i) (see section 3.1), and then computed it again using the data (W_i, Y_i). The resulting estimators are shown in Figure 1, together with scatterplots of both versions of the data. Clearly, the estimator computed from the (W_i, Y_i)'s is biased and oversmooths the data. This is typically the effect that errors have on estimators, and when the data are contaminated, we need to use techniques that can take the measurement error into account.

2.3 Berkson errors

It is the dependence structure between X and W that determines the type of errors. Classical errors typically arise when X is measured by an experimenter, who, by measuring X incorrectly, adds a noise U to the true value of X. Berkson errors, which were first considered by Berkson (1950), are of a completely different nature. There, we are interested in a variable X, but can only observe i.i.d. data $W_1, \ldots, W_n$, where

$$X_i = W_i + V_i, \tag{2.4}$$

with $X_i \sim f_X$, $V_i \sim f_V$, and the Berkson errors V_i are i.i.d. and independent of the W_i's. Although (2.4) can be written as $W_i = X_i - V_i$, which appears to be of the same form as (2.2) with $U_i = -V_i$, $-V_i$ cannot be treated as a classical error because it is not independent of X_i. In the Berkson case, often the variable W_i is a proxy for X_i. It is not a version of X_i corrupted by an error due to inaccurate measurements; rather it is a variable genuinely different from, but linearly related to, X_i. Throughout we make the usual assumption that f_V is symmetric.

In the Berkson errors-in-variables regression context, we observe a sample of i.i.d. data $(W_1, Y_1), \ldots, (W_n, Y_n)$ modelled according to

$$Y_i = m(X_i) + \epsilon_i, \quad X_i = W_i + V_i, \tag{2.5}$$

where $X_i \sim f_X$, $V_i \sim f_V$, and the V_i's are i.i.d. and independent of the W_i's. We assume throughout that ϵ_i is independent of W_i and V_i; moreover, $E(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) < \infty$ and the V_i's satisfy all the assumptions of model (2.4).

As in the classical error case, applying standard estimators designed for error-free data, to data contaminated by Berkson errors, leads to inconsistent, biased estimators. This is illustrated in Figure 2, where we show scatterplots of data (X_i, Y_i) and (W_i, Y_i), generated according to the model at (2.5), with a noise to signal ratio, $\mathrm{var}(V)/\mathrm{var}(X)$, of 20%, and standard local linear estimators of m constructed from these data. As in

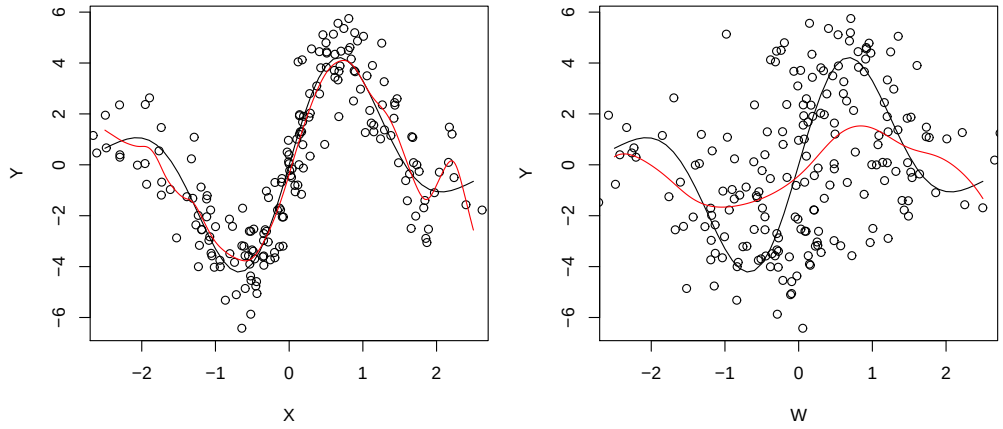

Figure 2: Scatterplot of the data (X_i, Y_i) (left) and (W_i, Y_i) (right) in the Berkson error case. The red curve is the standard local linear estimator of m computed from the data (X_i, Y_i) (left) or (W_i, Y_i) (right). The true m is depicted by the black line.

section 2.2, the estimator based on the contaminated data is biased. More generally, when data are observed with Berkson errors, standard estimators do not usually work and we have to use methods especially designed for dealing with those errors.

2.4 Berkson and classical errors

In the Berkson model, the proxy itself can be observed with errors. In such cases, the observed data are a sample $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ modelled according to

$$Y_i = m(X_i) + \epsilon_i, \quad X_i = W_i + V_i, \quad Z_i = W_i + U_i, \tag{2.6}$$

where $X_i \sim f_X$, $V_i \sim f_V$, $U_i \sim f_V$, and the Berkson errors V_i are i.i.d., the classical errors U_i are i.i.d., and the U_i's and the V_i's are independent, and independent of the W_i's and of the ϵ_i's. We assume that ϵ_i is independent of W_i, $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) < \infty$. We also assume throughout that f_U and f_V are known and symmetric, and $\phi_U(t) \neq 0$ for all t.

3 Deriving consistent estimators

3.1 Error-free observations

Before showing how to derive density and regression estimators in those error settings, we recall how these quantities can be estimated in the error-free case, where we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ coming from model (2.1). There exist a variety of nonparametric methods which consistently estimate m, but here we focus on kernel-based approaches. The simplest of these is the Nadaraya-Watson estimator

$$\widehat{m}(x) = \frac{(nh)^{-1} \sum_{j=1}^{n} Y_j K\left(\frac{x - X_j}{h}\right)}{(nh)^{-1} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)}, \tag{3.1}$$

where K is a smooth and symmetric function called kernel, and $h > 0$ is a smoothing parameter called bandwidth. The denominator on the right hand side of (3.1) is an estimator of $f_X(x)$, called kernel density estimator and denoted by $\widehat{f_X}(x)$; the numerator is an estimator of $m(x) f_X(x)$.

More generally, the local polynomial estimator of $m(x)$, of order p, is obtained by fitting, locally around x, a polynomial of order p. It is defined by $\widehat{m}(x) = \widehat{\beta}_{0,x}$, where $\widehat{\beta}_{0,x}$ is obtained through the following minimisation problem:

$$(\widehat{\beta}_{0,x}, \ldots, \widehat{\beta}_{p,x}) = \text{argmin}_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \left(X_j - x\right)^j \right\}^2 K\left(\frac{x - X_j}{h}\right). \tag{3.2}$$

It can be proved that the local polynomial estimator of order $p = 0$ corresponds to the Nadaraya-Watson estimator. Another popular particular case is the local linear estimator, which is obtained by taking $p = 1$.

3.2 Errors-in-variables: using Fourier transforms

3.2.1 Summary

Because of the additive structure of the errors in each of the models in (2.3), (2.4), (2.5) and (2.6), progress can often be made by considering the problem in the Fourier

domain. To illustrate this, in this section we show three examples of how the Fourier transform can be used to easily construct estimators. Section 3.2.2 shows how to estimate f_X in the classical error case. In section 3.2.3 we derive of an estimator of m in the Berkson error case. In section 3.2.4 we construct estimators of f_X when the data are contaminated by Berkson errors, or by a mixture of Berkson and classical errors.

3.2.2 Estimating f_X in the classical error model

Suppose we wish to estimate the density f_X from data $W_1, \ldots, W_n$ coming from the model at (2.2). It is easily proved that $f_W(w) = f_X * f_U(w)$, where $f * g(x) = \int f(x-u)g(u)\,du$ denotes the convolution product of two functions f and g. Therefore, to express f_X as a function of f_W and f_U, we need to deconvolve this convolution equation, which, a priori, does not seem to be an easy task.

However, using standard arguments, it can be proved that $\phi_W(t) = \phi_X(t)\phi_U(t)$. In other words, the convolution product in the original domain becomes a product in the Fourier domain. Using the Fourier inversion theorem, if $\phi_X \in L_1$ (i.e. if $|\phi_X|$ is integrable) we deduce that

$$f_X(x) = \frac{1}{2\pi} \int e^{-itx} \phi_W(t)/\phi_U(t)\,dt. \tag{3.3}$$

Expression (3.3) suggests a simple procedure for estimating f_X from $W_1, \ldots, W_n$. First, estimate $\phi_W(t)$ by the empirical characteristic function $\widehat{\phi}_W(t) = n^{-1} \sum_{j=1}^{n} e^{itW_j}$. Then, plug $\widehat{\phi}_W$ into (3.3). This, however, requires some modification because $\widehat{\phi}_W(t)$ is inaccurate in its tails, and is multiplied by $1/\phi_U(t)$, which tends to infinity as $|t| \to \infty$ (remember that ϕ_U is a characteristic function).

To prevent $\widehat{\phi}_W(t)$ from having too much influence for $|t|$ large, where it is unreliable, we can multiply it by a damping factor $d(t)$, which tends to zero sufficiently

fast as $|t|$ increases, and take

$$\widehat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx} \widehat{\phi}_W(t)\, d(t)/\phi_U(t)\, dt.$$

The most popular approach is to take $d(t) = \phi_K(ht)$, where $\phi_K(t) = \int e^{itx} K(x)\, dx$ is the Fourier transform of a kernel K and $h > 0$ is a bandwidth. Interestingly, the resulting estimator of $\phi_W(t)$, $\widetilde{\phi}_W(t) = \widehat{\phi}_W(t)\phi_K(ht)$, is the Fourier transform of the kernel density estimator of f_W constructed from the W_i's. An alternative approach based on ridging was suggested by Hall and Meister (2007).

Plugging $\widetilde{\phi}_W$ into (3.3), and choosing K so that $\phi_K(h\cdot)/\phi_U(\cdot) \in L_1$, we get

$$\widehat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx} \widehat{\phi}_W(t)\phi_K(ht)/\phi_U(t)\, dt\,, \tag{3.4}$$

which is the deconvolution kernel estimator of Carroll and Hall (1988) and Stefanski and Carroll (1990); see also Diggle and Hall (1993) for results with the sinc kernel $K(x) = \sin x/(\pi x)$. The estimator can be expressed as

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\Big(\frac{x - W_j}{h}\Big), \tag{3.5}$$

where

$$K_U(x) = \frac{1}{2\pi} \int e^{-itx} \phi_K(t)/\phi_U(t/h)\, dt. \tag{3.6}$$

Under sufficient smoothness conditions on f_X and appropriate conditions on the kernel K, $\widehat{f}_X$ is a consistent estimator of f_X; see Carroll and Hall (1988) and Stefanski and Carroll (1990). The errors have no effect on the bias of the estimator, which is identical to that of the standard kernel density estimator in the error-free case. However, the variance of $\widehat{f}_X$ is considerably larger than that of the standard kernel density estimator. As a result, unlike the error-free case, the convergence rates of $\widehat{f}_X$ are not driven only by the smoothness of f_X, but also by the rate of decay of ϕ_U in its tails. This can be understood from the definition of the estimator at (3.4), which involves dividing by $\phi_U(t)$. If $|\phi_U(t)|$ decreases exponentially fast as $|t| \to \infty$, then

U is called a supersmooth error, and $\widehat{f_X}$ converges to f_X at a logarithmic rate. If $|\phi_U(t)|$ decreases polynomially fast as $|t| \to \infty$, then U is called an ordinary smooth error, and $\widehat{f_X}$ converges to f_X at a polynomial rate. Its has been proved by Carroll and Hall (1988) and Fan (1991,1993) that these rates are optimal. That is, one cannot construct a nonparametric estimator that has faster convergence rates than the deconvolution estimator.

3.2.3 Estimating f_X and m in the Berkson error case

Suppose now that we want to estimate f_X from data $W_1, \ldots, W_n$ modelled as in (2.4). In this case, we have $f_X(x) = f_W * f_V(x) = \int f_V(x-w) f_W(w)\, dw = E\{f_V(x-W)\}$, which can simply be estimated by $\widehat{f_X}(x) = \sum_{j=1}^n f_V(x - W_j)$. This estimator is unbiased and converges to f_X at the fast parametric $\sqrt{n}$ rate; see Delaigle (2007). Estimating f_X in the Berkson error case is thus considerably easier than in the classical error case.

The situation is very different in the regression case, where the task is to estimate the curve m from data (W_j, Y_j) modelled according to (2.5). Since $Y_i = m(W_i+V_i)+\epsilon_i$, we can write

$$g(w) \equiv E(Y_i|W_i = w) = \int m(w+u) f_V(u)\, du = m * f_V(w).$$

Assuming that $\phi_V(t) \neq 0$ for all t, that $\phi_m(t) = \int e^{itx} m(x)\, dx$ is well defined, and letting $\phi_g(t) = \int e^{itx} g(x)\, dx$, we deduce that $\phi_g(t) = \phi_m(t)\phi_V(t)$. If $\phi_m \in L_1$, the Fourier inversion theorem implies that

$$m(x) = \frac{1}{2\pi} \int e^{-itx} \phi_g(t)/\phi_V(t)\, dt.$$

Using the same ideas as in section 3.2.2, we can estimate m by

$$\widehat{m}(x) = \frac{1}{2\pi} \int e^{-itx} \widehat{\phi}_g(t) \phi_K(ht)/\phi_V(t)\, dt,$$

where $\widehat{\phi}_g(t)$ denotes the Fourier transform (assuming it exists) of a standard nonparametric estimator of $g(w) = E(Y|W = w)$ constructed from the data (W_i, Y_i). In Delaigle, Hall and Qiu (2006), $\widehat{g}$ is constructed through a discrete Fourier series, whereas Carroll, Delaigle and Hall (2007) use a local polynomial regression estimator. See also Delaigle and Meister (2011) for a simpler approach in the particular case where $\phi_V^{-1}(t)$ is a polynomial. We refer to those papers for more details and for specific conditions that guarantee that the estimators are well defined and consistent.

While it might have been thought that, like in the density case, the presence of Berkson errors would make it possible to construct a $\sqrt{n}$ rate nonparametric estimator of m, the effect of those errors is completely different in the regression case. This can be understood from the fact that the estimator $\widehat{m}$ involves dividing by $\phi_V(t)$. As a result, the convergence rate of $\widehat{m}$ is of the same type as the rate of convergence of $\widehat{f}_X$ in the classical error case, with a similar distinction of rates depending on whether V is a supersmooth or ordinary smooth error.

3.2.4 Estimating f_X in the model with both Berkson and classical errors

The same ideas as those employed in section 3.2.2 can be used for constructing an estimator of f_X from data $Z_1, \ldots, Z_n$ modelled according to (2.6). In this case, we have $f_X(x) = f_W * f_V(x)$ and $f_Z(x) = f_W * f_U(x)$. In the Fourier domain, this can be written as $\phi_X(t) = \phi_W(t)\phi_V(t)$ and $\phi_Z(t) = \phi_W(t)\phi_U(t)$, from which we deduce that $\phi_X(t) = \phi_Z(t)\phi_V(t)/\phi_U(t)$. As in the classical error case, $\phi_Z(t)$ can be estimated by the empirical characteristic function of the observed data, that is $\widehat{\phi}_Z(t) = n^{-1}\sum_{j=1}^n e^{itZ_j}$. Using the Fourier inversion theorem, if $\phi_V/\phi_U \in L_1$, we can estimate $f_X(x)$ by

$$\widehat{f}_X(x) = \frac{1}{2\pi}\int e^{-itx}\widehat{\phi}_Z(t)\phi_V(t)/\phi_U(t)\,dt.$$

If $\phi_V/\phi_U \notin L_1$, we need to use a damping factor as in section 3.2.2. For example

(see Delaigle, 2007), we can take

$$\widehat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx} \widehat{\phi}_Z(t) \phi_V(t) \phi_K(ht) / \phi_U(t)\, dt.$$

Rates of convergence of this estimator are of the same type as those for the classical error case. In particular, depending on whether $|\phi_V(t)/\phi_U(t)|$ decreases exponentially fast or algebraically fast as $|t| \to \infty$, the convergence rates are of logarithmic or polynomial order. However, the presence of Berkson errors V_i implies that the convergence rates of $\widehat{f}_X$ are better than in the classical error setting, where $V_i \equiv 0$.

3.3 A general method based on unbiased scores

Adapting to the error case an estimator that exists in the error-free case can be much more difficult than the three examples studied above, and in such cases, Fourier transforms alone do not suffice to solve the problem. Different strategies have to be used in different situations, but in the classical errors-in-variables model, there exists an approach which can often be used to derive, from an estimator valid in the error-free case, an estimator valid in the error case.

We consider the general problem of estimating a function g from data $(W_1, T_1), \ldots, (W_n, T_n)$, with W_i as at (2.2), and where the T_i's represent data observed without noise. For example, in the regression case described by the model in (2.3), $T_i = Y_i$. Suppose that, in the error-free case, $g(x)$ can be estimated by an estimator $\widehat{g}(x; X_1, \ldots, X_n, T_1, \ldots, T_n, h)$ that depends on the data (X_i, T_i) and on a parameter h, for example a bandwidth. Suppose too that the bias and the variance of this estimator tend to zero as $n \to \infty$. To construct an estimator of g in the error case, a possible approach is to determine a function $\widetilde{g}$ which is such that

$$E[\widetilde{g}(x; W_1, \ldots, W_n, T_1, \ldots, T_n, h) | X_1, \ldots, X_n, T_1, \ldots, T_n] = \widehat{g}(x; X_1, \ldots, X_n, T_1, \ldots, T_n, h). \quad (3.7)$$

If $\widetilde{g}$ satisfies this property, then $\widetilde{g}(x; W_1, \ldots, W_n, T_1, \ldots, T_n, h)$ will be an estimator of $g(x)$ that has the same bias as $\widehat{g}(x; X_1, \ldots, X_n, T_1, \ldots, T_n, h)$, and thus will be asymptotically unbiased. Therefore, as long as the variance of $\widetilde{g}(x; W_1, \ldots, W_n, T_1, \ldots, T_n, h)$ tends to zero, this is a consistent estimator of $g(x)$. In the errors-in-variables literature, a function $\widetilde{g}$ that satisfies (3.7) is often referred to as an unbiased score for $\widehat{g}$; see Stefanski and Carroll (1987), Stefanski (1989) and Nakamura (1990) for early references.

While the idea might seem simple, it remains to see how we can find the function $\widetilde{g}$. The answer depends on the problem at hand, but we detail here a procedure that is often successful in the nonparametric context. In the error-free case, many nonparametric estimators can be written as a combination of quantities of the form

$$n^{-1} \sum_{j=1}^{n} M(x; X_j, T_j, h) \tag{3.8}$$

for some function M and a bandwidth h. For example, in the regression case, the Nadaraya-Watson estimator in (3.1) is a ratio of two such quantities: $M(x; X_j, T_j, h) = T_j K\{(x - X_j)/h\}$ with $T_j = Y_j$ at the numerator, and with $T_j = 1$ at the denominator.

To construct an unbiased score for the quantity at (3.8), take

$$n^{-1} \sum_{j=1}^{n} L(x; W_j, T_j, h),$$

where the function L is chosen so that it satisfies

$$E\{L(x; W_j, T_j, h) | X_j, T_j\} = M(x; X_j, T_j, h).$$

This equation can be quite difficult to solve, but it is often simpler when formulated in the Fourier domain (here, as above, we assume that the Fourier transforms we write are all well defined), that is, when written as

$$\int e^{itx} E\big\{L(x; W_j, T_j, h) | X_j, T_j\big\}\, dx = \int e^{itx} M(x; X_j, T_j, h)\, dx.$$

3.4 Application of the unbiased scores method

In this section, to illustrate how the unbiased scores method of section 3.3 can be used, we rederive the deconvolution kernel density estimator of section 3.2, and show how to obtain the local polynomial estimator of Delaigle, Fan and Carroll (2009).

3.4.1 Estimating f_X in the classical error model

Consider again the problem of estimating the density f_X from data $W_1, \ldots, W_n$ coming from the classical error model (2.2). In the error-free case, where we observe $X_1, \ldots, X_n$, the kernel density estimator of f_X is defined by

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{x - X_j}{h}\Big).$$

To define an estimator in the classical error case, take

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} L\Big(\frac{x - W_j}{h}\Big), \tag{3.9}$$

and choose the function L so that it satisfies

$$E\Big[L\Big(\frac{x - W_j}{h}\Big)\Big|X_j\Big] = K\Big(\frac{x - X_j}{h}\Big).$$

To determine L, we write the above equation in the Fourier domain:

$$\int e^{itx} E\Big[L\Big(\frac{x - W_j}{h}\Big)\Big|X_j\Big]\, dx = \int e^{itx} K\Big(\frac{x - X_j}{h}\Big)\, dx.$$

Equivalently, assuming here and below that we can interchange integral and expectation,

$$E\Big[\int e^{itx} L\Big(\frac{x - X_j - U_j}{h}\Big)\, dx\Big|X_j\Big] = \int e^{itx} K\Big(\frac{x - X_j}{h}\Big)\, dx,$$

or again, making a change of variables,

$$E\Big[e^{itX_j} e^{itU_j} \int e^{ithv} L(v)\, dv\Big|X_j\Big] = e^{itX_j} \int e^{ithv} K(v)\, dv.$$

This implies that

$$\phi_U(t) \int e^{ithv} L(v)\, dv = \int e^{ithv} K(v)\, dv,$$

which implies that $\phi_L(t) = \phi_K(t)/\phi_U(t/h)$. Using again the Fourier inversion theorem, we deduce that $L(x) = K_U(x)$, with K_U as in (3.6). Plugging $L = K_U$ into (3.9), we find the deconvolution estimator derived in section 3.2.

3.4.2 Local polynomial regression in the classical error model

The same idea can be used to construct local polynomial estimators in the classical error model at (2.3). In the error-free case (Fan and Gijbels, 1996), where the data come from model (2.1), the local polynomial estimator of order p of m introduced in section 3.1 can be written as $\widehat{m}(x) = \mathbf{e}_1^T \widehat{\boldsymbol{S}}_n^{-1} \widehat{\boldsymbol{T}}_n$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, $\widehat{\boldsymbol{S}}_n = (\widehat{S}_{n,k,k'})_{0 \le k,k' \le p}$ is a $(p+1) \times (p+1)$ matrix, with

$$\widehat{S}_{n,k,k'} = \frac{1}{nh^{k+k'+1}} \sum_{j=1}^{n} K\Big(\frac{X_j - x}{h}\Big)(X_j - x)^{k+k'}, \quad k, k' = 0, \ldots, p, \tag{3.10}$$

and where $\widehat{\boldsymbol{T}}_n = (\widehat{T}_{n,0}, \ldots, \widehat{T}_{n,p})^T$, with

$$\widehat{T}_{n,k} = \frac{1}{nh^{k+1}} \sum_{j=1}^{n} Y_j K\Big(\frac{X_j - x}{h}\Big)(X_j - x)^{k}. \tag{3.11}$$

To construct a consistent estimator of m that can be computed from data (W_i, Y_i) modelled according to (2.3), Delaigle, Fan and Carroll (2009) suggest taking $\widetilde{m}(x) = \mathbf{e}_1^T \widetilde{\boldsymbol{S}}_n^{-1} \widetilde{\boldsymbol{T}}_n$, where $\widetilde{\boldsymbol{S}}_n$ and $\widetilde{\boldsymbol{T}}_n$ are defined as $\widehat{\boldsymbol{S}}_n$ and $\widehat{\boldsymbol{T}}_n$, except that the $\widehat{S}_{n,k,k'}$'s and $\widehat{T}_{n,k}$'s are replaced by unbiased scores $\widetilde{S}_{n,k,k'}$ and $\widetilde{T}_{n,k}$. As indicated by those authors, to find these unbiased scores, for $\ell = 0, \ldots, 2p$, it suffices to find a function L_ℓ that satisfies

$$E\Big[L_\ell\Big(\frac{W_j - x}{h}\Big)(W_j - x)^{\ell}\Big|X_j\Big] = K\Big(\frac{X_j - x}{h}\Big)(X_j - x)^{\ell}. \tag{3.12}$$

In what follows we assume that we can interchange integration and expectation, and derivative and integration, and that the expressions we write are all well defined. For details and conditions, see Delaigle, Fan and Carroll (2009).

Consider the Fourier version of (3.12):

$$E\Big[\int e^{itx} L_\ell\Big(\frac{W_j - x}{h}\Big)(W_j - x)^\ell \, dx \Big| X_j\Big] = \int e^{itx} K\Big(\frac{X_j - x}{h}\Big)(X_j - x)^\ell \, dx.$$

After a change of variable, this can be written as

$$E\Big[e^{itX_j} e^{itU_j} \int e^{-ithv} L_\ell(v) v^\ell \, dv \Big| X_j\Big] = e^{itX_j} \int e^{-ithv} K(v) v^\ell \, dv,$$

which implies that $\phi_{L_\ell}^{(\ell)}(-ht) = \phi_K^{(\ell)}(-ht)/\phi_U(t)$. We deduce from the Fourier inversion theorem that

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

The unbiased scores $\widetilde{S}_{n,k,k'}$ are obtained by replacing $K\{(X_j - x)/h\}(X_j - x)^{k+k'}$ by $L_{k+k'}\{(W_j - x)/h\}(W_j - x)^{k+k'}$ in the definition of $\widehat{S}_{n,k,k'}$. As in Delaigle, Fan and Carroll (2009), the unbiased scores $\widetilde{T}_{n,k'}$ are obtained similarly from $\widehat{T}_{n,k}$. We also refer to that paper for consistency of the resulting local polynomial estimators. Their convergence rates are of the same type as the convergence rates of the estimator of f_X in the classical error case derived in section 3.2.2.

The local constant estimator, obtained by taking $p = 0$ in the above calculations, can be written as

$$\widetilde{m}(x) = \frac{(nh)^{-1} \sum_{j=1}^n Y_j K_U\big(\frac{x - W_j}{h}\big)}{(nh)^{-1} \sum_{j=1}^n K_U\big(\frac{x - W_j}{h}\big)}, \tag{3.13}$$

with K_U as in (3.6). It corresponds to the estimator proposed by Fan and Truong (1993). These authors established optimality of this estimator.

3.5 Other methods in the classical error case

The functions m and f_X can also be estimated by other nonparametric estimators. These include the orthogonal series method of Hall and Qiu (2006) for cases where the density f_X is supported on a compact interval, and the wavelet estimators of Pensky and Vidakovic (1999), Fan and Koo (2002) and Pensky (2002). Moreover, spline

techniques have been considered by Mendelsohn and Rice (1982) in the density case, and by Carroll, Maca and Ruppert (1999), Berry, Carroll and Ruppert (2002), Ganguli, Staudenmayer and Wand (2005) and Marley and Wand (2010) in the regression case. Some other techniques have also been developed, which provide approximations and are consistent only under the assumptions that the variance of U tends to zero as $n \to \infty$. These include the nonparametric SIMEX method of Staudenmayer and Ruppert (2004) and the TAYLEX method of Carroll and Hall (2004).

4 Computing estimators

4.1 Numerical integration

Computing the estimators in practice requires particular care. Here we explain the difficulties in the classical error case, but they are similar in the Berkson setting. Calculating $\widehat{f}_X$ at (3.5) requires one to compute K_U at (3.6). In most cases there is no analytic expression for the integral at (3.6), which needs to be approximated numerically. However, the integrand oscillates, sometimes quite heavily, which causes standard fast numerical integration algorithms to fail. As indicated by Delaigle and Gijbels (2007), one way to overcome this problem is to use the fast Fourier transform.

A simpler but more time consuming alternative is to approximate the integral at (3.6) by the simple trapezoidal rule using a fine grid (fine enough to include enough points in each cycle of oscillation). At the time of writing, Matlab code for computing the density and regression estimators in the classical error case in this way is available at the author's webpage `http://www.ms.unimelb.edu.au/~aurored/links.html`. If speed is an issue, one should preferably compute the integral using the fast Fourier transform. In `C`, this can be done using the routine `dftint` from Press et al. (1992).

Although the deconvolution kernel K_U integrates to 1, it is not a density, and takes negative values in large parts of its domain. For an illustration, see Figure 3,

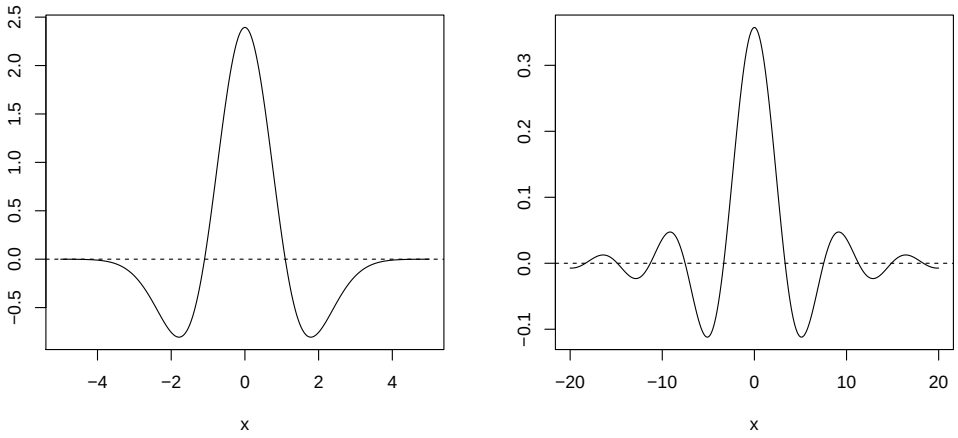

Figure 3: The function K_U when U has a Laplace distribution, K is the standard normal density, and $\mathrm{var}(U)/h^2 = 10$ (left) or when $K = K_2$, U is normal and $\mathrm{var}(U)/h^2 = 10$ (right). The horizontal dotted line indicates 0 for reference.

where we depict K_U when U has a Laplace distribution, K is the standard normal density, and $\mathrm{var}(U)/h^2 = 10$, and when U has a normal distribution, K is the kernel K_2 defined in section 5.5 and $\mathrm{var}(U)/h^2 = 12$. As a result, in finite sample the deconvolution kernel density estimator $\widehat{f}_X(x)$ at (3.4) often takes negative values at points x where the true density $f_X(x)$ is close to zero. In such cases, it is common to replace $\widehat{f}_X(x)$ by

$$\widetilde{f}_X(x) = \widehat{f}_X(x) \cdot 1\{\widehat{f}_X(x) > 0\} \Big/ \int \widehat{f}_X(y) \cdot 1\{\widehat{f}_X(y) > 0\}\, dy\,. \tag{4.1}$$

This approach, which has been studied by Hall and Murison (1993), is standard in the error-free case when density estimators are computed with higher order kernels.

4.2 Choosing the bandwidth in the density case with classical errors

It is easy to prove that the deconvolution kernel density estimator $\widehat{f}_X(x)$ at (3.4) satisfies $\widehat{f}_W(x) = \widehat{f}_X * f_U(x)$, where

$$\widehat{f}_W(x) = \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{x - W_j}{h}\Big)$$

is the standard kernel density estimator of $f_W(x)$. Despite this relation, $\widehat{f}_X$ cannot be computed with a bandwidth h of the size usually employed for computing $\widehat{f}_W$, for example the plug-in bandwidth h_{SJ} of Sheather and Jones (1991).

Under the standard assumption that f_X has two derivatives, the optimal bandwidth for computing $\widehat{f}_W$ is of order $n^{-1/5}$. However, the variance of $\widehat{f}_X$ is much larger than that of $\widehat{f}_W$, and it can even tend to infinity if we take h of order $n^{-1/5}$. To compute $\widehat{f}_X$, we should use a larger bandwidth. Assuming again two derivatives, in the ordinary smooth error case the optimal bandwidth is of order $n^{-1/(2\beta+5)}$, and in the supersmooth case, we can take h of order $(\ln n)^{-1/\beta}$, where β is a parameter dictating the rate of decay of $|\phi_U|$ to zero; see Fan (1991).

A data-driven choice of h that works well in practice is the 2-stage plug-in rule suggested by Delaigle and Gijbels (2002, 2004); see section 4.4. At the time of writing, Matlab codes for computing this bandwidth, as well as the cross-validation bandwidth of Stefanski and Carroll (1990), are available at
`http://www.ms.unimelb.edu.au/~aurored/links.html`.

4.3 Choosing the bandwidth in the regression case with classical errors

In the regression context, for the same reasons as in the density case, we cannot simply compute errors-in-variables regression estimators with the bandwidth that

would be used with error-free data. In the classical error case, Delaigle and Hall (2008) developed an effective SIMEX algorithm. At the time of writing, a Matlab code for computing their bandwidth is available at
`http://www.ms.unimelb.edu.au/~aurored/links.html`.

As noted by those authors, in practice errors-in-variables regression estimators can be numerically unstable. Let $\widehat{m}$ denote the local polynomial estimator of Delaigle, Fan and Carroll (2009), or the local constant version of Fan and Truong (1993). These estimators take the form $\widehat{m}(x) = N(x)/D(x)$, where N and D are functions that depend on the data. For example, in the local constant case, $N(x) = (nh)^{-1}\sum_{j=1}^{n} Y_j K_U\{(x-W_j)/h\}$ and $D(x) = (nh)^{-1}\sum_{j=1}^{n} K_U\{(x-W_j)/h\}$. Of course, estimators also have this form in the error-free setting, but a difficulty in the error case is that the function K_U vanishes at some points (see Figure 3). This implies that in finite samples, $D(x)$ can be too small at some points x, causing $\widehat{m}(x)$ to be quite poor. To overcome this difficulty, Delaigle and Hall (2008) suggested to change $\widehat{m}$ to

$$\widehat{m}(x) = N(x)/\max\{D(x), \rho\},$$

where $\rho > 0$ is a ridge parameter.

However, sometimes, difficulties can also be caused by the numerator, and occasionally, even if N and D are apparently neither too large nor too small, the ratio $|N|/|D|$ can still be unusually large; see Achilleos (2011) for practical illustrations of this problem. One way to overcome this difficulty is to prevent the estimator from taking aberrant values. Motivated by the correction at (4.1) in the density case, which is applied because we know that a density should always be positive and integrate to one, Delaigle and Hall (2013) suggested an empirical band within which one can reasonably expect a regression estimator to lie. Their idea is that, in most regression problems, we can expect that $m(x) \in [q_{0.05,x}, q_{0.95,x}]$, where $q_{\alpha,x}$ denotes the α-level quantile of the Y_i's whose $X_i \in A_x$, with A_x denoting a neighborhood of x. To obtain a version of this that can be computed from the W_i's, they take

$A_x = [x - \sigma_U, x + \sigma_U]$, where $\sigma_U^2 = \text{var}(U)$, and replace $q_{\alpha,x}$ by the α-level quantile of the Y_i's whose $W_i \in A_x$. Finally, they change $\widehat{m}(x)$ into $\widetilde{m}(x) = \widehat{m}(x) \cdot 1\{\widehat{m}(x) \in [q_{0.05,x}, q_{0.95,x}]\} + q_{0.05,x} \cdot 1\{\widehat{m}(x) < q_{0.05,x}\} + q_{0.95,x} \cdot 1\{\widehat{m}(x) > q_{0.95,x}\}$.

4.4 Problems with the R package `decon` of Wang and Wang (2011)

An R package `decon` developed by Wang and Wang (2011) aims at implementing the deconvolution estimator and some data-driven bandwidth selectors. However, at the time of writing, version 1.2-4 of this package suffers from a number of problems which are also present in Wang and Wang (2011). We describe these problems in this section, and number them by (1) to (4).

(1) In kernel density estimation, the rule of thumb bandwidth is another name given to the well known normal reference bandwidth. In the errors-in-variables setting, the normal reference bandwidth was introduced by Delaigle and Gijbels (2004). It is obtained by minimising the asymptotic mean squared error of $\widehat{f}_X$,

$$\text{AMISE}(h) = (2\pi nh)^{-1} \int |\phi_K(t)|^2 |\phi_U(t/h)|^{-2}\, dt + \frac{h^4}{4}\, R(f_X'')\, \mu_{K,2}^2,$$

where $\mu_{K,2} = \int x^2 K(x)\, dx$, and where $R(f_X'') = \int \{f_X''(x)\}^2\, dx$ is estimated by assuming that $X \sim N(\mu_X, \sigma_X^2)$. That is, $\widehat{R}(f_X'') = 0.375\, \widehat{\sigma}_X^{-5} \pi^{-1/2}$, with $\widehat{\sigma}_X^2 = \widehat{\sigma}_W^2 - \sigma_U^2$ and where $\widehat{\sigma}_W^2$ is the empirical variance of the W_i's. An advantage of this bandwidth is that it is simple to calculate, but as in the error-free case, it is well known that it rarely competes with more sophisticated bandwidths such as, for example, the plug-in bandwidth. In particular, it tends to be too large and to oversmooth the data.

We show in section A.1 that, when the errors follow a Laplace(σ) distribution, K is the standard normal kernel, and lower order terms of the AMISE are neglected, the normal reference bandwidth is given by

$$\widehat{h} = (5\, \sigma^4 \widehat{\sigma}_X^5)^{1/9} n^{-1/9}. \tag{4.2}$$

In section 3.1 of Wang and Wang (2011), the authors discuss "the rule of thumb" bandwidth for Laplace and normal errors, which is implemented by their function `bw.dnrd`. In the Laplace case, their rule of thumb bandwidth is defined by $\widehat{h} = (5\,\sigma^4)^{1/9} n^{-1/9}$; see equation (14) in Wang and Wang (2011). Comparing with the rule of thumb bandwidth at (4.2), we can see that the term $\widehat{\sigma}_X^5$ is missing from their rule of thumb bandwidth. In particular, their bandwidth does not depend on the data, and cannot generally be expected to give good results. Likewise, in the normal error case, their rule of thumb bandwidth (see formula (13) of Wang and Wang, 2011) does not depend on the data and cannot generally be expected to give good practical results.

(2) The plug-in bandwidth was suggested by Delaigle and Gijbels (2002, 2004). It is obtained by minimising the AMISE expression given above, where $R(f_X'')$ is estimated by $\int \{\widehat{f}_X''(x)\}^2\,dx$, with $\widehat{f}_X''$ denoting the second derivative of the deconvolution kernel estimator of f_X, computed with a pilot bandwidth g. In order for the procedure to give good performance, this pilot bandwidth has to be chosen with a lot of care, and Delaigle and Gijbels (2002) suggest a sophisticated 2-stage procedure. In numerical work, a plug-in bandwidth computed in this way often gives the best practical results among existing data-driven bandwidths.

Section 3.2 in Wang and Wang (2011) is titled "plug-in bandwidth", but the bandwidth derived there is the normal reference bandwidth of Delaigle and Gijbels (2004) introduced in point (1). Likewise, the associated function `bw.dmise` in the package `decon` is listed as a function computing the plug-in bandwidth, but it actually computes the normal reference bandwidth.

(3) The bootstrap bandwidth was suggested by Delaigle and Gijbels (2004). In general it performs similarly to, but is slightly worse than, the plug-in method. As for the plug-in bandwidth described above, computing the bootstrap bandwidth requires the choice of a pilot bandwidth g, and the success of the procedure depends on an appropriate choice of g; see Delaigle and Gijbels (2004) for a practical 2-stage rule.

The package `decon` includes a function called `bw.dboot1`, which aims at computing the bootstrap bandwidth. However, the details provided in section 3.3 in Wang and Wang (2011) indicate that instead of using the 2-stage pilot bandwidth g, the authors take g equal to the rule of thumb bandwidth. The latter does not have the correct theoretical order of magnitude, and in our experience, in practice it does not usually compete with the 2-stage pilot bandwidth. The package includes another bootstrap bandwidth, computed by a function `bw.dboot2`, but it is not clear to us what exactly this function is computing.

(4) Bandwidth for regression: in their package, Wang and Wang (2011) propose a function `DeconNpr`, which computes the errors-in-variables local constant estimator. Care is needed with that function too: the default bandwidth used by the function `DeconNpr` is the bandwidth computed by the function `bw.dboot1`, thus a bandwidth for density estimation, but not a bandwidth appropriate for regression estimation.

5 Common errors

We conclude this article by reviewing some of the invalid derivations that are sometimes encountered in the area.

5.1 Nadaraya-Watson estimator with classical errors

Suppose we observe data $(W_1, Y_1), \ldots, (W_n, Y_n)$ modelled according to (2.3), and let $g(w) = E(Y|W = w)$ and $\eta = Y - g(W)$. It can be easily proved that $E\{g(W)|X = x\} = g * f_U(x)$. Thus, since $Y = g(W) + \eta$, we have

$$m(x) = E(Y|X = x) = g * f_U(x) + E(\eta|X = x).$$

It is tempting to conclude from there that $m(x) = g * f_U(x)$. If this were true, then we could simply estimate m by $\widehat{m} = \widehat{g} * f_U$, where $\widehat{g}$ denotes a consistent estimator

of g, for example the Nadaraya-Watson estimator (recall that $g(w) = E(Y|W = w)$, and so can be directly estimated from the observed contaminated data).

However, while it is true that $E(\eta|W = w) = 0$, we do not generally have that $E(\eta|X = x) = 0$. Therefore, in general it is not true that $m = g * f_U$. A subtlety here, which might explain some of the confusion, is that the residual ϵ in the model at (2.3), and which satisfies $E(\epsilon|X) = 0$, also satisfies $E(\epsilon|W) = 0$. The relation between m and g can be shown to be $g(w) = (mf_X) * f_U(w)/f_W(w)$; see appendix A.2.

5.2 Local polynomial estimators with classical errors

Suppose we observe data $(W_1, Y_1), \ldots, (W_n, Y_n)$ modelled according to (2.3). The local constant estimator at (3.13) can also be found by solving, at each x,

$$\widehat{m}(x) = \text{argmin}_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_U\Big(\frac{x - W_j}{h}\Big). \tag{5.1}$$

That is, the local constant estimator can be defined as in the error-free case discussed in section 3.1, but replacing there X_j by W_j and K by K_U (compare with (3.2)). By analogy, it could be thought that, in the classical error case, a pth order local polynomial estimator of m can be defined by $\widehat{m}(x) = \widehat{\beta}_{0,x}$, where

$$(\widehat{\beta}_{0,x}, \ldots, \widehat{\beta}_{p,x}) = \text{argmin}_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \Big\{Y_i - \sum_{j=0}^{p} \beta_j (W_j - x)^j\Big\}^2 K_U\Big(\frac{x - W_j}{h}\Big). \tag{5.2}$$

However, for $p > 0$ this does not provide a consistent estimator of m. A consistent estimator was derived in section 3.4.2.

5.3 Nadaraya-Watson estimator with Berkson errors

Suppose we have observations $(W_1, Y_1), \ldots, (W_n, Y_n)$ coming from the model at (2.5). We have seen in section 3.2.4 that a consistent estimator f_X can be defined by $\widehat{f}_X(x) = n^{-1} \sum_{j=1}^{n} f_V(x - W_j)$. Motivated by this, and by analogy with the construction of the

regression estimators at (3.1) and at (3.13), it may be thought that a good estimator of m is

$$\widehat{m}(x) = \frac{n^{-1}\sum_{j=1}^{n} Y_j f_V(x - W_j)}{n^{-1}\sum_{j=1}^{n} f_V(x - W_j)}\,. \tag{5.3}$$

However $\widehat{m}$ at (5.3) is not generally a consistent estimator of m. Indeed, the numerator is not a consistent estimator of $m(x)f_X(x)$, but rather of $\iint m(w+v)f_V(x-w)f_W(w)f_V(v)\,dv\,dw$; see appendix A.3 for a proof. A consistent estimator of m was introduced in section 3.2.3, and is (unfortunately) much more complex.

5.4 Nadaraya-Watson estimator with Berkson and classical errors

Suppose we observe data $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ modelled according to (2.6). Consider first the particular situation where f_U and f_V are identical. In this case, the Z_i's have the same distribution as the X_i's, which implies that we can construct a kernel density estimator of f_X by taking

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{x - Z_j}{h}\Big).$$

Similarly as in section 5.3, motivated by this fact, it may be though that an estimator of m can be defined by

$$\widehat{m}(x) = \frac{(nh)^{-1}\sum_{j=1}^{n} Y_j K\big(\frac{x-Z_j}{h}\big)}{(nh)^{-1}\sum_{j=1}^{n} K\big(\frac{x-Z_j}{h}\big)}\,.$$

However, as in section 5.3, this estimator is not consistent because the numerator is not a consistent estimator of $m(x)f_X(x)$, but rather of $\iint m(w+v)f_W(w)f_V(v)f_U(z-w)\,dw\,dv$. See appendix A.4 for a proof.

5.5 Unboundedness of the function K_U as $n \to \infty$

Unlike the error-free case, where $\|K\|_\infty$ is bounded by a constant independent of n, in the classical error case, $\|K_U\|_\infty \to \infty$ as $n \to \infty$. Indeed, we have

$$K_U(0) = \frac{1}{2\pi} \int \phi_K(t)/\phi_U(t/h)\, dt,$$

and following the arguments of Fan (1991), it can be proved that this quantity is unbounded as $n \to \infty$.

We conclude this section by a last note to indicate that kernels employed in the error case are quite different from those used in the error-free setting. Especially in the supersmooth error case, to guarantee that K_U is well defined, the kernel is often chosen so that its Fourier transform is compactly supported. For example, a kernel that is often employed in numerical work is the kernel K_2 whose Fourier transform is equal to $\phi_{K_2}(t) = (1-t^2)^3 \cdot 1_{[-1,1]}(t)$. Unlike kernels often employed in the conventional error-free setting, such kernels can take negative values (hence are not densities) and are not compactly supported.

Acknowledgments

Research supported by a fellowship and a grant from the Australian Research Council.

A Technical details

A.1 Computing the rule of thumb bandwidth

In the case where the errors follow a Laplace distribution with scale parameter σ, and K is the standard normal kernel, we have that

$$\text{AMISE}(h) = \frac{1}{2\sqrt{\pi} n h} + \frac{\sigma^2}{2\sqrt{\pi} n h^3} + \frac{3\,\sigma^4}{8\sqrt{\pi} n h^5} + \frac{h^4}{4} R(f_X'').$$

Neglecting smaller order terms, we can approximate the AMISE by

$$\frac{3\,\sigma^4}{8\sqrt{\pi}nh^5} + \frac{h^4}{4}\,R(f_X'').$$

Minimising this quantity w.r.t. h, we find

$$h = \Big\{\frac{15\,\sigma^4}{8\sqrt{\pi}n\,R(f_X'')}\Big\}^{1/9},$$

and replacing $R(f_X'')$ by $0.375\,\widehat{\sigma}_X^{-5}\pi^{-1/2}$, we deduce the following bandwidth

$$\widehat{h} = \Big(\frac{5\,\sigma^4\widehat{\sigma}_X^5}{n}\Big)^{1/9}.$$

A.2 Details for section 5.1

We have

$$\begin{aligned}
g(w) &= E(Y|W=w) \\
&= E\{m(X)|W=w\} + E(\epsilon|W=w) \\
&= \int m(x) f_{X|W}(x|w)\,dx \\
&= f_W^{-1}(w) \int m(x) f_X(x) f_{W|X}(w|x)\,dx \\
&= f_W^{-1}(w) \int m(x) f_X(x) f_U(w-x)\,dx \\
&= (m f_X) * f_U(w)/f_W(w).
\end{aligned}$$

A.3 Details for section 5.3

We have

$$\begin{aligned}
E\Big[n^{-1}\sum_{j=1}^{n} Y_j f_V(x-W_j)\Big] =& E\Big[m(X_j) f_V(x-W_j)\Big] \\
=& E\Big[m(W_j+V_j) f_V(x-W_j)\Big] \\
=& \iint m(w+v) f_V(x-w) f_W(w) f_V(v)\,dv\,dw.
\end{aligned}$$

For the first equality, we used the independence of ϵ_j and W_j.

A.4 Details for section 5.4

The numerator is a consistent estimator of

$$\begin{aligned} f_Z(z)\, E(Y|Z=z) =& f_Z(z)\, E\{m(X)|Z=z\} \\ =& f_Z(z)\, E\{m(W+V)|Z=z\} \\ =& f_Z(z) \int m(w+v) f_{W,V|Z}(w,v|z)\, dw\, dz \\ =& \iint m(w+v) f_{W,V,Z}(w,v,z)\, dw\, dv \\ =& \iint m(w+v) f_W(w) f_V(v) f_U(z-w)\, dw\, dv. \end{aligned}$$

References

Achilleas, A. (2011). Deconvolution kernel density and regression estimation; on local bandwidth selection in the presence of measurement error. *PhD thesis*, University of Bristol.

Berry, S., Carroll, R., and Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems, *J. Amer. Statist. Assoc.*, **97**, 160–169.

Carroll, R.J., Delaigle, A., Hall, P. (2007). Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *J. Roy. Statist. Soc.* Series B, **69**, 859–878.

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

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

Carroll, R.J., Maca, J.D. and Ruppert, D. (1999). Nonparametric Regression in the Presence of Measurement Error. *Biometrika*, **86**, 541–554.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*, 2nd Edn. Chapman and Hall CRC Press, Boca Raton.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of Berkson and classical errors. *Canad. J. Statist.*, **35**, 89–104.

Delaigle, A., Fan, J. and Carroll, R.J. (2009). A Design-adaptive Local Polynomial Estimator for the Errors-in-Variables Problem. *J. Amer. Statist. Assoc.*, **104**, 348–359.

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc.* Series B, **64**, 869–886.

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

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

Delaigle, A. and Gijbels, I. (2007). Frequent problems in calculating integrals and optimizing objective functions: a case study in density deconvolution. *Statis. Comput.*, **17**, 349–355.

Delaigle, A. and Hall, P. (2013). Methodology for nonparametric deconvolution when the error distribution is unknown. *Manuscript.*

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

Delaigle, A. and Meister, A. (2011). Rate-optimal nonparametric estimation in classical and Berkson errors-in-variables problems. *J. Statist. Plan. Infer.*, **141**, 102–114.

Diggle, P.J. and Hall, P. (1993). A Fourier approach to non-parametric deconvolution of a density estimate. *J. Roy. Statist. Soc.* Series B, **55**, 523–531.

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

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

Fan, J. and Gijbels, I. (1996). Local polynomial modeling and its applications. *Chapman and Hall*, London.

Fan, J. and Koo, J.-Y. (2002). Wavelet deconvolution. *IEEE T. Inform. Theory*, **48**, 734-747.

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

Ganguli, B., Staudenmayer, J. and Wand, M.P. (2005). Additive models with predictors subject to measurement error. *Aust. N. Z. J. Statist.*, **47**, 193–202.

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

Hall, P. and Murison, R.D. (1993). Correcting the Negativity of High-Order Kernel Density Estimators. *J. Multivar. Anal.*, **47**, 103-122.

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

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

Marley, J.K. and Wand, M.P. (2010). Non-standard semiparametric egression via BRugs. *J. Statist. Soft.*, **37**, 1–30.

MathWorks, Inc. (2012). MATLAB Release 2012b. Natick, Massachusetts, United States.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of microfluorometric histograms with B splines. *J. Amer. Statist. Assoc.*, **77**, 748-753.

Nakamura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*, **77**, 127–137.

Pensky, M. (2002). Density deconvolution based on wavelets with bounded supports . *Statist. Prob. Lett.*, **56**, 2033-2053.

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

Press, W.H. , Flannery, B.P. , Teukolsky, S.A., Vetterling, W.T. (1992). Numerical Recipes in C, The Art of Scientific Computing (Second Edition). Cambridge University Press.

R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* Series B, **53**, 683690

Staudenmayer, J., and Ruppert, D. (2004). Local Polynomial regression and SIMEX, *J. Roy. Statist. Soc.* Series B, **66**, 17–30.

Stefanski, L.A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Commun. Statist. Theory*, **18**, 4335–4358.

Stefanski, L.A. and Carroll, R.J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**, 703–716.

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

Wang, X.F. and Wang, B. (2011). Deconvolution estimation in measurement error models: The R package decon. *J. Statist. Soft.*, **39**, 1–24.