Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314013489

Nonparametric kernel methods for curve estimation and measurement errors

Aurore Delaigle

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

Abstract. We consider the problem of estimating an unknown density or regression curve from data. In the parametric setting, the curve to estimate is modelled by a function which is known up to the value of a finite number of parameters. We consider the nonparametric setting, where the curve is not modelled a priori. We focus on kernel methods, which are popular nonparametric techniques that can be used for both density and regression estimation. While these methods are appropriate when the data are observed accurately, they cannot be directly applied to astronomical data, which are often measured with a certain degree of error. It is well known in the statistics literature that when the observations are measured with errors, nonparametric procedures become biased, and need to be adjusted for the errors. Correction techniques have been developed, and are often referred to as deconvolution methods. We introduce those methods, in both the homoscedastic and heteroscedastic error cases, and discuss their practical implementation.

Keywords. methods: data analysis, methods: statistical, stars: statistics, galaxies: statistics.

1. Nonparametric curve estimation

1.1. *Regression estimation*

In the regression problem, we are interested in modelling the unknown relationship between two random variables X and Y. Specifically, we wish to estimate the unknown regression curve $m(x) = E(Y|X = x)$ from a sample of independent and identically distributed (i.i.d.) data $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfying

$$Y_i = m(X_i) + \epsilon_i, \tag{1.1}$$

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

There are many examples in astronomy where one is interested in modelling the relationship between two variables X and Y. For example, in the Hubble diagram, (X, Y)= (redshift, distance modulus). In Fig. 1, we show a scatterplot of the (X_i, Y_i)'s for $n = 557$ SNe from the Union2 compilation; see Amanullah *et al.* (2010).

In the parametric estimation setting, we assume that we know the shape of m up to the value of a finite number, d say, of parameters. Then, estimating the regression curve m reduces to the estimation of these unknown parameters. For example, if we assume that the regression curve is a quadratic curve, then $m(x) = \theta_0 + \theta_1 x + \theta_2 x^2$, where θ_0, θ_1 and θ_2 are unknown parameters that need to be estimated from the data. Several approaches are possible for computing these estimators, such as the least squares and maximum likelihood procedures.

Parametric estimators can have excellent properties, such as fast convergence rates, but this is only true when the parametric model (i.e. the assumed shape for m) is correct, at least approximately. In particular, if we do not have sufficient information about m, and

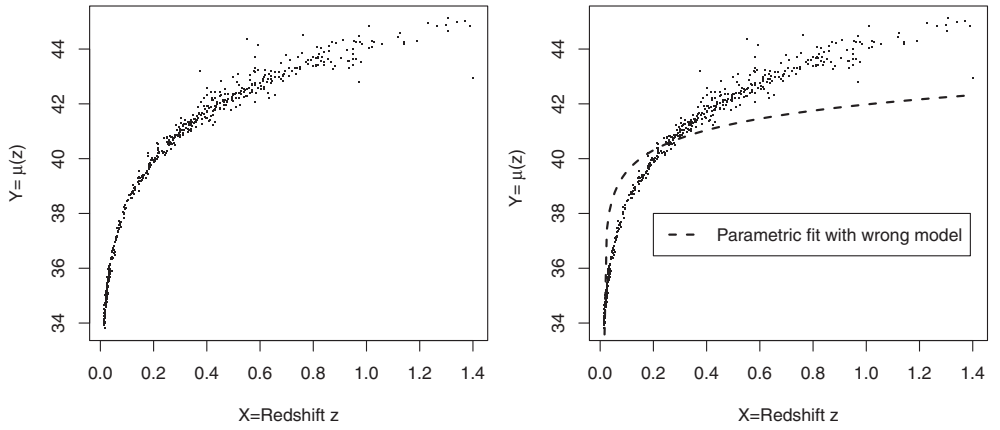

Figure 1. Left: distance moduls versus redshift for 557 SNe from the Union2 compilation. Right: parametric estiimator of $E(Y|X = x)$ using a wrong parametric model.

use a parametric model that is far from the truth, then parametric estimators produce biased, inconsistent estimators of m. See for example Fig. 1, where we depict a parametric estimator of m obtained when assuming that $m(x) = \log(\theta_0 + \theta_1\, x)$.

One way to overcome this difficulty is to estimate m using a nonparametric estimator, that is, an estimator that does not require to formulate a parametric model. For excellent introductions to techniques of nonparametric estimation of regression curves, see for example Wand & Jones (1994) and Fan & Gijbels (1996). To understand how such an estimator may be constructed, recall that the regression curve $m(x)$ is the expectation of Y, conditional on $X = x$. Motivated by this, we could think of constructing an estimator of m by taking $\hat{m}_0(x) = \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$, the empirical mean of the Y_i's. Clearly, this estimator is too naive since it estimates m by a constant. A more sophisticated approach for estimating m at a point x could be to take the average of only the Y_i's whose X_i is relatively close to x. Letting I denote the indicator function, with this approach we would estimate m by

$$\hat{m}_1(x) = \sum_{i=1}^{n} Y_i \cdot I(X_i \text{ close to } x) \Big/ \sum_{i=1}^{n} I(X_i \text{ close to } x).$$

As long as we define "close to x" properly, this estimator produces a reasonable estimator of m, but is not sufficiently smooth to be attractive. See Fig. 2 below for an illustration of the non smoothness of the estimator in the closely related density estimation problem introduced in section 1.2.

A more sophisticated approach consists in using all the data (X_i, Y_i), assigning to each pair (X_i, Y_i) a weight $w(X_i)$ which is small if X_i is far from x, and large if X_i is close to x. This leads to the estimator

$$\hat{m}_2(x) = \sum_{i=1}^{n} Y_i\, w(X_i) \Big/ \sum_{i=1}^{n} w(X_i).$$

It remains to define the weights $w(X_i)$. In Statistics, a very popular way to choose the weights is to take $w(X_i) = K\{(x - X_i)/h\}$, where $h > 0$ is a smoothing parameter called the bandwidth and K is a weight function called the kernel, wich is usually smooth and symmetric. Computed with these particular weights, the estimator $\hat{m}_2$ of m is called the

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

Nadaraya-Watson estimator. It is defined by

$$\hat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^{n} Y_i\, K\{(X_i - x)/h\}}{\sum_{i=1}^{n} K\{(X_i - x)/h\}}. \tag{1.2}$$

The Nadaraya-Watson estimator is a particular case of a more general class of nonparametric estimators of m called local polynomial estimators. Local polynomial estimators are constructed in a very intuitive way, as follows: while many curves can not be expressed as a polynomial, locally around each point x, if they are smooth, they can be well approximated by a polynomial (one way to understand this is through Taylor's expansion). Motivated by this, the local polynomial estimator of $m(x)$, of order p, is obtained by fitting a polynomial of order p locally around x (that is, using mostly the data (X_i, Y_i) for which X_i is close to x). Formally, at each x, approximate the function $m(u)$ by a pth order polynomial $m_{\mathrm{pol},p}(u) = \beta_{0,x} + \beta_{1,x}(u-x) + \ldots + \beta_{p,x}(u-x)^p$, and estimate the parameters $\beta_{k,x}$ by $\hat{\beta}_{k,x}$, obtained by minimising a local least squares sum which puts more weight on observations whose X_i is close to x:

$$(\hat{\beta}_{0,x}, \ldots, \hat{\beta}_{p,x}) = \mathrm{argmin}_{(\beta_{0,x},\ldots,\beta_{p,x})} \sum_{i=1}^{n} \{Y_i - m_{\mathrm{pol},p}(X_i)\}^2 K\{(X_i - x)/h\}.$$

Then, estimate $m(x)$ by $\hat{m}(x) = \hat{m}_{\mathrm{pol},p}(x) = \hat{\beta}_{0,x} + \hat{\beta}_{1,x}(x-x) + \ldots + \hat{\beta}_{p,x}(x-x)^p = \hat{\beta}_{0,x}$.

It can be proved that the Nadaraya-Watson estimator is equal to the local polynomial estimator of order $p = 0$, which is also called the local constant estimator. In both theory and practice, the Nadaraya-Watson is known to suffer from boundary effects when the density f_X of the X_i's is compactly supported and is not continuous at the endpoints of its support. Specifically, the bias of the Nadaraya-Watson estimator near such endpoints tends to be larger (see below for an illustration). The local polynomial estimator of order $p = 1$, which is also called the local linear estimator, is less affected by this boundary effect, and tends to perform better than the local constant estimator. It is one of the most popular nonparametric regression estimators. It can be written as

$$\hat{m}_{\mathrm{LL}}(x) = \frac{S_2(x)T_0(x) - S_1(x)T_1(x)}{S_2(x)S_0(x) - S_1^2(x)},$$

where, for $k = 0, 1$, $T_k(x) = n^{-1}h^{-k-1}\sum_{i=1}^{n} Y_i\,(X_i - x)^k\, K\{(X_i - x)/h\}$ and for $k = 0, 1, 2$, $S_k(x) = n^{-1}h^{-k-1}\sum_{i=1}^{n}(X_i - x)^k K\{(X_i - x)/h\}$. The right panel of Fig. 3 compares the local constant and the local linear estimators computed from the data in Fig. 1. The boundary problem of the local constant estimator is apparent for x small.

1.2. *Density estimation*

Similar ideas can be used to estimate a density f_X from i.i.d. data $X_1, \ldots, X_n \sim f_X$. For excellent introductions to the problem, see Silverman (1986) and Wand & Jones (1994). There too, we can construct a nonparametric estimator of $f_X(x)$ using mainly the observations X_i that are close to x. To define this estimator, recall that $f_X(x) = F_X'(x)$, where $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$ is the cumulative distribution function. This implies that

$$f_X(x) = \lim_{h\to 0} \frac{F_X(x+h) - F_X(x-h)}{2h} = \lim_{h\to 0} \frac{P(x-h \leqslant X \leqslant x+h)}{2h},$$

which can be estimated by replacing the unknown probability by a proportion computed from the data:

$$\hat{f}_X(x) = \frac{\sum_{i=1}^{n} I(x-h \leqslant X_i \leqslant x+h)}{2nh}, \tag{1.3}$$

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

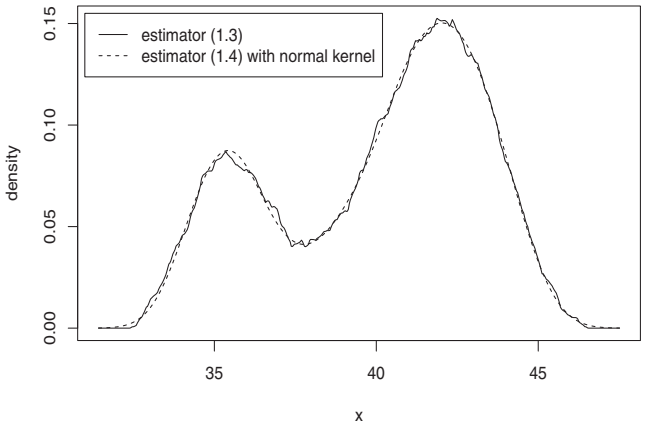

Figure 2. Estimator of the density of distance modulus constructed from the data shown in Fig. 1, using the non smooth estimator at (1.3) or the kernel density estimator at (1.4) with the standard normal kernel.

where h is a small positive number. As in the regression case, while this estimator is reasonable, its lack of smoothness make is relatively unattractive. See for example Fig. 2, where we show this estimator in the case where the X_i's are the distance moduli from $n = 557$ SNe from the Union2 compilation and $h = 0.8$.

A more sophisticated version of this naive estimator is the kernel density estimator defined by

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

where the kernel K is a smooth and symmetric density and $h > 0$ is a bandwidth. In Fig. 2, we depict this estimator in the case where the X_i's are the distance moduli from 557 SNe from the Union2 compilation, taking K to be the density of a standard normal random variable, and $h = 0.8$. It is clear from this example that this estimator is very similar to one at (1.3). However, the fact that it is nice and smooth makes it more attractive. Such density estimators are useful to understand properties of a population. For example in this case the two modes of the density suggest two groups or clusters. See Sun *et al.* (2002) for interesting aspects of the detection of bumps using nonparametric density estimators in astronomy.

1.3. *Choosing the bandwidth h and the kernel K*

As long as it is smooth, the choice of the kernel K is not very important and does not play a major role in the quality of the estimator. It is usually chosen to be a smooth and symmetric density, such as the density of a standard normal random variable.

The role of the bandwidth h is much more important. It dictates the closeness of an observation X_i to x. For example, in the regression case, if h is too small, most observations will be deemed far from x, and the estimator $\hat{m}_{\mathrm{NW}}(x)$ will essentially be based on the few observations (X_i, Y_i) for which X_i is very close to x. As a result it will tend be too wiggly. On the other hand, if h is too large, most observations will be considered to be close to x, and the estimator $\hat{m}_{\mathrm{NW}}(x)$ will be quite similar to the naive estimator $\hat{m}_0(x)$ introduced above. The bandwidth plays the same role for the more general local polynomial estimators. For example in Fig. 3 we show the local linear estimator of $m(x) = E(Y|X = x)$ computed from the data plotted in Fig. 1, using a standard normal kernel and three different bandwidths: a too small bandwidth, which

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

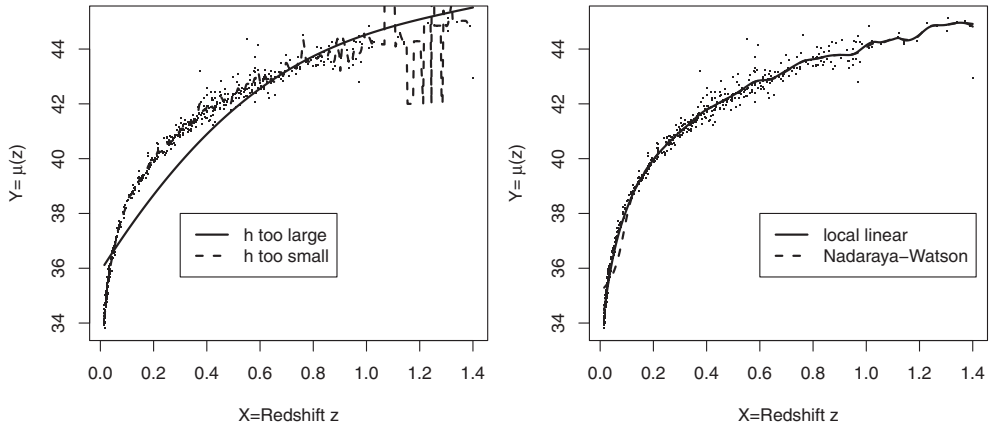

Figure 3. Local linear estimator constructed from the data shown in Fig. 1, using three bandwidths: a too large or too small bandwidth (left graph) or a good bandwidth (right graph). The right graph depicts the local linear and the Nadaraya-Watson estimators, both computed with a standard normal kernel and a bandwidth chosen by an automatic procedure.

produces an estimator which almost interpolates the data and causes numerical difficulty for x large, a too large bandwidth, which oversmoothes the data, and a good bandwidth, computed using one of the automatic procedures described below.

In practice, h should preferably be chosen by a fully automatic data-driven procedure and not by eye (the user may find that the estimator with a given bandwidth looks more attractive than one with a different bandwidth, but in the absence of detailed information about the true curve, the user's impression does not necessarily reflect the reality). Let $\hat{m}$ denote one of the regression estimators introduced above, which depend on a bandwidth h. Ideally, if we knew the curve m, we would choose h to minimise the error committed by estimating m by $\hat{m}$. This error is not unique. It could be the L_2 distance between m and $\hat{m}$, the L_1 distance, or any other sensible criterion. In nonparametric regression, we often employ an L_2 distance called the conditional mean integrated squared error, defined by $\mathrm{MISE}(h) = \mathrm{E}\left[\int\{\hat{m}(x) - m(x)\}^2\,dx \middle| X_1, \ldots, X_n\right]$. With the latter, the ideal bandwidth is defined by $h_{\mathrm{opt}} = \mathrm{argmin}_h\,\mathrm{MISE}(h)$.

Of course we cannot compute this bandwidth in practice, since it depends on the unknown m. However, a large statistics literature has been devoted to developing estimators of the MISE which can be computed from the data. Once a good estimator of the MISE has been computed in this way, the bandwidth can be chosen by minimising this MISE estimator. Perhaps the most popular data-driven bandwidth is the so-called plug-in bandwidth of Ruppert *et al.* (1995), which is obtained in such a way. There, in a first step the MISE is approximated by its asymptotic dominating part (i.e. the part that dominates the MISE as the sample size n increases) denoted by AMISE (asymptotic mean integrated squared error). In a second step, the unknown quantities of the AMISE are replaced by estimators computed from the data, producing an estimator $\widehat{\mathrm{AMISE}}$ of AMISE. Finally, the plug-in bandwidth h_{PI} is defined by $h_{\mathrm{PI}} = \mathrm{argmin}_h\,\widehat{\mathrm{AMISE}}(h)$, or, more commonly, by $h_{\mathrm{PI}} = \mathrm{argmin}_h\,\widehat{\mathrm{AMISE}}_w(h)$, where $\widehat{\mathrm{AMISE}}_w$ denotes a weighted version of the $\widehat{\mathrm{AMISE}}$. See section 4 of Fan & Gijbels (1996) for a more detailed description. We refer to the R package `KernSmooth` of Wand R port by Ripley (2011) for R codes for computing this bandwidth for the local linear estimator (see `dpill`).

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

Another popular data-driven bandwidth for computing regression estimators is the cross-validation bandwidth h_{CV}. It is defined by

$$h_{\mathrm{CV}} = \mathrm{argmin}_h \sum_{i=1}^{n} \{Y_i - \hat{m}^{(-i)}(X_i)\}^2, \tag{1.5}$$

where $\hat{m}^{(-i)}$ denotes the version of the estimator $\hat{m}$ computed without using the ith observation. For example, in the case of the Nadaraya-Watson estimator,

$$\hat{m}_{\mathrm{NW}}^{(-i)}(x) = \frac{\sum_{j\neq i}^{n} Y_j\, K\{(X_j - x)/h\}}{\sum_{j\neq i}^{n} K\{(X_j - x)/h\}}.$$

This bandwidth is simple to define but it tends to be too small. Moreover, it is not always unique (the sum on the right hand side of (1.5) does not always have a unique minimum).

In the density case, typically the bandwidth is chosen to minimise an estimator of $\mathrm{MISE} = \mathrm{E}\int\{\hat{f}_X(x) - f(x)\}^2\,dx$ (e.g. the plug-in bandwidth), or of $\mathrm{ISE} = \int\{\hat{f}_X(x) - f_X(x)\}^2\,dx$ (e.g. the cross-validation bandwidth). The plug-in bandwidth (Sheather & Jones (1991)) is constructed using ideas similar to those explained above in the regression case. Like there, it usually performs very well in practice. We refer to the R function `bw.SJ` for computing this bandwidth in the case where K is the standard normal kernel. The cross-validation bandwidth, h_{CV}, suffers from the same difficulties as those mentioned in the regression case above, but it is simpler to define than the plug-in bandwidth. Specifically,

$$h_{\mathrm{CV}} = \mathrm{argmin}_h \Big\{ \int \hat{f}_X^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_X^{(-i)}(X_i) \Big\},$$

where $\hat{f}_X^{(-i)}(x) = \{(n-1)h\}^{-1}\sum_{j\neq i}^{n} K\{(X_j - x)/h\}$.

We conclude this section with three remarks on the bandwidth. First, it is important to note that different kernels usually require different bandwidths. This is easy to understand since $\hat{m}$ and $\hat{f}_X$ both depend on K. In particular, the MISE and thus the optimal bandwidth depend on K. Likewise, h_{PI} and h_{CV} both depend on K. What this means in practice is that if we have computed h_{PI} or h_{CV} for a given kernel, those bandwidths are generally not appropriate for kernel estimators computed with another kernel. Another important remark is that a good bandwidth should depend on the sample size n. Specifically, as n increases the optimal bandwidth decreases. What this means in practice is that a bandwidth computed for a sample of a given size n is generally not appropriate for a kernel estimator computed from a sample of a different size. Finally, good bandwidths are usually different for density and for regression estimators. For example, a bandwidth computed by `dpill` should not be used to compute a kernel density estimator.

2. Measurement errors

2.1. *Introduction*

In astronomy, data are rarely measured with perfect accuracy and the quantities we observe are often approximated versions of those we are interested in. When computed with data contaminated by measurement errors, the nonparametric procedures introduced in the previous section are not valid and need to be corrected for the measurement errors. In this section we consider the classical measurement error problem, where, instead of observing the variable X of interest, we only manage to observe $W = X + U$, where U represents a measurement error. Importantly, X and U are independent and $E(U) = 0$. See Carroll *et al.* (2006) for an introduction to measurement errors.

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

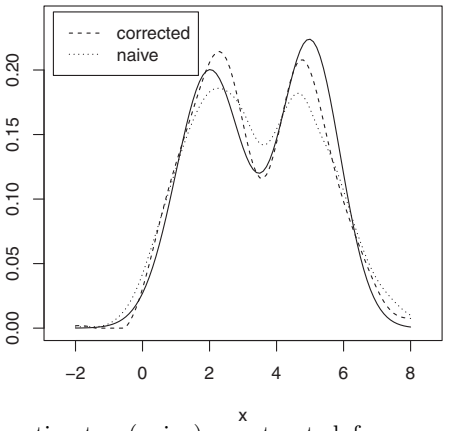

Figure 4. Kernel density estimator (naive) constructed from a sample $W_1, \ldots, W_n$ of size $n = 1000$ contaminated with normal errors U_i such that $\text{var}(U_i) = 0.2\ \text{var}(X_i)$, and modified kernel estimator (corrected) that takes measurement errors into account. The true density f_X is depicted by a continuous line.

Clearly, if we compute the kernel density estimator at (1.4) using data $W_1, \ldots, W_n$ having the distribution of W, instead of data $X_1, \ldots, X_n$ having the distribution of X, then instead of obtaining a consistent estimator of the density f_X, we will obtain a consistent estimator of the density f_W of W. For example, in Fig. 4, we show the kernel density estimator computed from a sample $W_1, \ldots, W_n$ of size $n = 1000$, where, for each i, $W_i = X_i + U_i$, the X_i's have a bimodal density shown in Fig. 4, and the U_i's are normally distributed, with $\text{var}(U_i) = 0.2\ \text{var}(X_i)$. This estimator, which ignores the presence of measurement errors, is often referred to as a naive estimator. It is a consistent estimator of f_W, but a non consistent, biased, estimator of f_X. In this example the bias is noticeable from the fact that the peaks and the valleys of the estimator are attenuated compared to those of f_X. In Fig. 4 we also show a corrected estimator that takes the measurement errors into account. It is a consistent estimator of f_X, and, for example, is able to better estimate the peaks and the valleys of f_X.

Likewise, if we compute one of the regression estimators of $m(x) = E(Y|X = x)$ defined in section 1.1 from data $(W_1, Y_1), \ldots, (W_n, Y_n)$ instead of data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $W_i = X_i + U_i$ with X_i and U_i independent as above, then instead of a consistent estimator of $m(x)$, we will obtain a consistent estimator of $E(Y|W = x)$.

While the measurement errors U_i are not observed, often in astronomy we can compute the distribution of U_i. Exploiting this fact, we shall assume throughout section 2 that the distribution of the U_i's in known, and under this assumption we shall see how to adapt the kernel density and regression estimators to this errors-in-variables context. In particular we shall see how to transform them into consistent estimators of f_X and m.

2.2. *Deconvolution kernel density estimator*

Suppose we observe i.i.d. data $W_1, \ldots, W_n$, where, for $i = 1, \ldots, n$, $W_i = X_i + U_i$ with $X_i \sim f_X$ and $U_i \sim f_U$ independent. The error density f_U is known, $E(U_i) = 0$ and the goal is to estimate the density f_X from the W_i's. For $V = X$, U and W, let $\phi_V(t) = \int e^{itx} f_V(x)\, dx$ denote the characteristic function of V. By the Fourier inversion theorem, we can write

$$f_X(x) = \frac{1}{2\pi} \int e^{-itx} \phi_X(t)\, dt.$$

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

Moreover, the independence of X_i and U_i implies that $\phi_W(t) = \phi_X(t)\,\phi_U(t)$. Therefore, assuming that $\phi_U(t) \neq 0$, we have $\phi_X(t) = \phi_W(t)/\phi_U(t)$. Since f_U is known, then ϕ_U is known too, and since we observe data $W_1, \ldots, W_n$, then we can estimate $\phi_W(t)$ by the empirical characteristic function $\hat{\phi}_W(t) = n^{-1}\sum_{j=1}^{n} e^{itW_j}$. Therefore, we can estimate $\phi_X(t)$ by $\hat{\phi}_X(t) = \hat{\phi}_W(t)/\phi_U(t)$. From there, it is tempting to define an estimator of $f_X(x)$ by $\hat{f}_X(x) = (2\pi)^{-1}\int e^{-itx}\hat{\phi}_X(t)\,dt$. However, $\hat{\phi}_X(t)$ is a very poor estimator of $\phi_X(t)$ for $|t|$ large, which makes the estimator $\hat{f}_X(x)$ inappropriate.

To overcome this difficulty, we need to modify $\hat{\phi}_X(t)$ so as to put less emphasis on it for $|t|$ large. One way to do this is to replace $\hat{\phi}_W(t)$ by the Fourier transform of the kernel estimator $\hat{f}_W(w) = (nh)^{-1}\sum_{j=1}^{n} K\{(W_i - w)/h\}$ of f_W. Indeed, it can be proved that the Fourier transform of $\hat{f}_W$ is given by $\tilde{\phi}_W(t) = \hat{\phi}_W(t)\phi_K(ht)$, where ϕ_K denotes the Fourier transform of the kernel K. Based on this, we can define a new estimator of $\phi_X(t)$ by $\tilde{\phi}_X(t) = \tilde{\phi}_W(t)/\phi_U(t) = \hat{\phi}_X(t)\phi_K(ht)$. Now the factor $\phi_K(ht)$ is small when $|t|$ is large, which reduces the impact of $\hat{\phi}_X(t)$ when the latter is a poor estimate of $\phi_X(t)$. Motivated by this, the deconvolution kernel estimator of Carroll & Hall (1988) and Stefanski & Carroll (1990) is defined by

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

It can be rewritten as

$$\hat{f}_X(x) = (nh)^{-1}\sum_{j=1}^{n} K_U\{(W_j - x)/h\}, \tag{2.1}$$

where

$$K_U(x) = (2\pi)^{-1}\int e^{-itx}\phi_K(t)/\phi_U(t/h)\,dt. \tag{2.2}$$

It is interesting to note that $\inf_x K_U(x) < 0$, even if $\inf_x K(x) \geqslant 0$. As a result, in finite samples, while $\hat{f}_X$ integrates to 1, it is often not positive everywhere, although in general, $\hat{f}_X(x)$ vanishes only at points x where $f_X(x)$ is rather small. Since a density is always positive, it is convenient to replace $\hat{f}_X(x)$ by $\tilde{f}_X(x) = \max\{0, \hat{f}_X(x)\}$. This is the estimator we used to construct the corrected estimator shown in Fig. 4. If needed, $\tilde{f}_X$ can also be rescaled so that it integrates to 1. See Hall & Murison (1993).

While the choice of the kernel K is usually not important in kernel estimation procedures from data measured without errors, one has to be more careful in this case since the kernel needs to be such that the integral in (2.2) exists. This is not trivially the case. For example if K is the standard normal kernel and $U \sim N(0, \sigma^2)$, then this integral only exists for sufficiently large values of h. However, when the sample size n is large, we should use a sufficiently small bandwidth (as already noticed earlier, h should decrease to zero as n increases, and this is true both in theory and in practice). To ensure that the integral at (2.2) exists, in the deconvolution literature it is standard to take a kernel whose characteristic function is compactly supported. Two such kernels are usually employed: the sinc kernel, denoted here by K_1, whose Fourier transform is defined by $\phi_{K_1}(t) = I(|t| \leqslant 1)$, and the kernel which we shall denote here by K_2, whose Fourier transform is defined by $\phi_{K_2}(t) = (1 - t^2)^3 \cdot I(|t| \leqslant 1)$. See, for example, Fan (1991).

2.3. *Errors-in-variables regression estimator*

In the errors-in-variables regression context, we observe i.i.d. data $(W_1, Y_1), \ldots, (W_n, Y_n)$, where, for $i = 1, \ldots, n$, $W_i = X_i + U_i$ with $X_i \sim f_X$ and $U_i \sim f_U$. Moreover, $E(U_i) = 0$

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

and $Y_i = m(X_i) + \epsilon_i$, where $E(\epsilon_i|X_i) = 0$, $\text{var}(\epsilon_i|X_i) < \infty$, and the U_i's are independent of the ϵ_i's, the Y_i's and the X_i's. Finally m and f_X are unknown but f_U is known, and the goal is to estimate m from the (W_i, Y_i)'s.

To construct a consistent estimator of m in this context, we start by comparing the standard kernel density estimator with the deconvolution kernel estimator introduced in the previous section. In particular, comparing (1.4) and (2.1), we can see that the deconvolution kernel density estimator takes the same form as the standard kernel density estimator, except that K is replaced by K_U and the X_i's are replaced by the W_i's. This motivates us to modify the Nadaraya-Watson estimator at (1.2) in a similar manner, and define a kernel estimator that takes measurement errors into account by

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i \, K_U\{(W_i - x)/h\}}{\sum_{i=1}^n K_U\{(W_i - x)/h\}}. \tag{2.3}$$

This estimator was introduced by Fan & Truong (1993). Under appropriate regularity conditions, including the one that $|\phi_U(t)| > 0$ for all t, it can be proved that it is a consistent estimator of m.

It is also possible to define a version of local polynomial estimators which takes the measurement errors into account. These estimators are less easy to define, and we refer to Delaigle *et al.* (2009) for details. See also Delaigle (2014) for a more general description of how to construct consistent nonparametric estimators in errors-in-variables problems. In practice, unlike the error-free case, in the errors-in-variables context the local constant estimator, which corresponds to the estimator defined at (2.3), tends to perform better than the local linear estimator.

As indicated in section 2.2, the function K_U is not positive everywhere and, in practice, the denominator of the right hand side of (2.3) can vanish (or be very close to zero) at some points x, which creates numerical problems. The latter can be avoided by preventing the denominator from getting too small. One way to do this is to replace (2.3) by

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i \, K_U\{(W_i - x)/h\}}{\max\left[\sum_{i=1}^n K_U\{(W_i - x)/h\}, \rho\right]},$$

where ρ is a positive number, sometimes referred to as a ridge parameter.

2.4. *Heteroscedastic errors*

As highlighted by Feigelson & Babu (2012), in astronomy, the measurement errors are often heteroscedastic. There, the independent contaminated data $W_1, \ldots, W_n$ are such that $W_i = X_i + U_i$, with X_i and U_i independent, $X_i \sim f_X$ and $U_i \sim f_{U_i}$. In particular, each observation may have its own error density f_{U_i}, which we assume to be known. As before, $E(U_i) = 0$. In this case, the observations W_i that are contaminated by "a lot of noise" contain less information than the observations that are contaminated by "less noise". Intuitively, when constructing estimators, the least contaminated observations should be given more emphasis than the most contaminated observations. It remains to define the notions of "a lot of noise", "less noise", etc, and to seek a way of giving more importance to more reliable observations.

First, contrary to what may be thought, "a lot of noise" does not always mean "a large error variance". In nonparametric errors-in-variables problems, the effect of measurement errors is also measured by the speed at which the characteristic function of the noise, $\phi_U(t)$, tends to zero as $|t|$ tends to infinity. Specifically, the faster $\phi_U(t)$ tends to zero as $|t|$ increases, the more the measurement errors affect the quality of the estimators. For example, if the errors U_i are normally distributed, then nonparametric estimators of f_X and m converge at a logarithmic rate, i.e. like $(\log n)^{-\alpha}$ for some $\alpha > 0$. By contrast, if

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

the errors U_i have a Laplace distribution, then this rate is rather $n^{-\alpha}$ for some $\alpha > 0$. Of course the variance of the U_i's also plays a role in the quality of nonparametric estimators: the larger that variance, the more difficult it is to estimate f_X and m.

These considerations indicate that combining observations that are contaminated by errors which are not identically distributed is a rather subtle problem. For example, a naive construction could be as follows. Let ϕ_{U_i} and ϕ_{W_i} denote the characteristic functions of U_i and W_i, and assume that $|\phi_{U_i}(t)| > 0$ for all t. Then we have

$$\phi_X(t) = \phi_{W_j}(t)/\phi_{U_j}(t) = n^{-1}\sum_{j=1}^{n} \phi_{W_j}(t)/\phi_{U_j}(t).$$

Since $E(e^{itW_j}) = \phi_{W_j}(t)$, using the same ideas as in section 2.2, and in particular using the Fourier inversion theorem, we could define an estimator of $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{2\pi n}\sum_{j=1}^{n} \int e^{-itx}\frac{e^{itW_j}}{\phi_{U_j}(t)}\phi_K(ht)\,dt.$$

As highlighted by Delaigle & Meister (2008), this would be a consistent estimator, but with rather poor properties. For example, suppose that half of the observations were observed with Laplace error, and the other half with normal errors. Then it can be proved that the converge rate of this estimator would be logarithmic. In other words, the estimator would inherit from the convergence rate induced by the least favourable errors U_i. In this example, we would do worse by using all the observations than by using only the observations contaminated by Laplace errors (the latter would lead to a convergence rate of order $n^{-\alpha}$ for some $\alpha > 0$). This indicates that the observations W_j were not combined in an adequate way since our estimator should improve as we use more data.

Delaigle & Meister (2008) proposed an estimator which does not suffer from this problem. They proceed as follows. To understand their estimator, note that $\phi_X(t) = \phi_{W_j}(t)/\phi_{U_j}(t)$ implies that $\phi_X(t) = \phi_{W_j}(t)\bar{\phi}_{U_j}(t)/|\phi_{U_j}(t)|^2$ or again that $\phi_X(t)|\phi_{U_j}(t)|^2 = \phi_{W_j}(t)\bar{\phi}_{U_j}(t)$ (here $\bar{a}$ denotes the conjugate of a complex number a). In turn this implies that $\phi_X(t)\sum_{j=1}^{n}|\phi_{U_j}(t)|^2 = \sum_{j=1}^{n}\phi_{W_j}(t)\bar{\phi}_{U_j}(t)$, so that

$$\phi_X(t) = \sum_{j=1}^{n}\phi_{W_j}(t)\bar{\phi}_{U_j}(t)\Big/\sum_{k=1}^{n}|\phi_{U_k}(t)|^2.$$

Motivated by the fact that $E(e^{itW_j}) = \phi_{W_j}(t)$, using arguments similar to those used above, Delaigle & Meister (2008) propose to estimate $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{2\pi n}\sum_{j=1}^{n} \int e^{-itx}\frac{e^{itW_j}\,\bar{\phi}_{U_j}(t)}{\sum_{k=1}^{n}|\phi_{U_k}(t)|^2}\,\phi_K(ht)\,dt. \tag{2.4}$$

See Delaigle & Meister (2008) for properties of this estimator.

To illustrate this estimator, we used the data described by De Blok *et al.* (2001) and used by Wang & Wang (2011). They concern the velocity of $n = 318$ stars from 26 low surface brightness galaxies, and for these data the variance of each U_i is available. In Fig. 5 we show the estimator $\hat{f}_X(x)$ at (2.4) computed from these data, assuming that the U_i's are normally distributed with the known error variances, or Laplace distributed with those error variances. In this example the two estimators are so close that they can hardly be distinguished on the graph (this is not always the case!).

In the regression case, where we observe independent data $(W_1, Y_1), \ldots, (W_n, Y_n)$ with the W_i's as above and where $Y_i = m(X_i) + \epsilon_i$ as in section 2.3, Delaigle & Meister (2007)

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

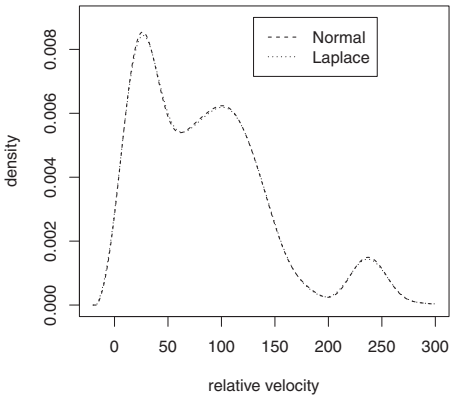

Figure 5. Deconvolution kernel estimator of the density of relative velocity for 318 stars from 26 low surface brightness galaxies, using the estimator of Delaigle & Meister (2008) for heteroscedastic errors and assuming that the errors have a Laplace distribution or a normal distribution.

suggest the following regression estimator:

$$\hat{m}(x) = \frac{\sum_{j=1}^n Y_j \int e^{-itx} \left\{\bar{\phi}_{U_j}(t) \big/ \sum_{k=1}^n |\phi_{U_k}(t)|^2\right\} \phi_K(ht)\,dt}{\sum_{j=1}^n \int e^{-it(x-W_j)} \left\{\bar{\phi}_{U_j}(t) \big/ \sum_{k=1}^n |\phi_{U_k}(t)|^2\right\} \phi_K(ht)\,dt}.$$

2.5. *Bandwidth choice and code for computing the estimators*

As in the case where the data are observed without measurement errors, in order for the estimators introduced above to work well, the bandwidth h needs to be chosen with a lot of care. In the density case, the plug-in techniques of Sheather & Jones (1991) can be adapted to the measurement error context. See Delaigle & Gijbels (2002) and Delaigle & Gijbels (2004). A cross-validation bandwidth can also be constructed; see Stefanski & Carroll (1990).

The situation is much more complex in the regression case. For example, there the plug-in techniques depend on many more unknown functions than in the standard error-free context, which makes them particularly unattractive. Moreover, in the measurement error context it is not possible to compute the cross-validation bandwidth defined at (1.5): even though we can compute the estimator $\hat{m}^{(-i)}$, we need to compute it at X_i, which we cannot do since we only observe the W_i's. Thus it does not seem that standard bandwidth selection techniques can be used in this context. Delaigle & Hall (2008) suggested a procedure based on Simulation Extrapolation (SIMEX) which can be applied to select the bandwidth of a variety of errors-in-variables problems, including the one of regression estimation.

Matlab codes for computing all the estimators described in section 2 are available on the author's webpage at `www.ms.unimelb.edu.au/~aurored`. On that webpage, code for computing the plug-in, cross-validation and SIMEX bandwidths in the errors-in-variables density and regression estimation problems are also available. Some limited R codes written by Achilleas Achilleos, and which focus on density estimation with a local bandwidth as described in Achilleos & Delaigle (2012), are also available there. An R package called `decon` written by Wang & Wang (2011) also exists. However, at the time of writing this paper, this package did not seem to compute the bandwidths in an appropriate manner. See Delaigle (2014) for a detailed description of the problems with this R package, and Delaigle & Gijbels (2007) for a description of numerical issues that can be encountered when computing deconvolution kernel estimators.

https://doi.org/10.1017/S1743921314013489 Published online by Cambridge University Press

Acknowledgements

This work was supported by the Australian Research Council. The author thanks Véronique Delouille for useful discussion.

References

Amanullah, R., *et al.* (2010). Spectra and Hubble space telescope light curves of six type Ia supernovae at $0.511 < z < 1.12$ and the Union2 compilation. *ApJ*, **716**, 712.

Achilleos, A. & Delaigle, A. (2012). Local bandwidth selectors for deconvolution kernel density estimation. *Statistics and Computing*, **22**, 563–577.

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

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

De Blok, W., McGaugh, S. & Rubin, V. (2001). High-resolution rotation curves of low surface brightness galaxies: Mass Models. *Astr J.*, **122**.

Delaigle, A. (2014). Kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. *Australian and New Zealand J. Statist.*, **56**, 105–124.

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

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

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

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

Delaigle, A. & Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.*, **102**, 1416–1426.

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

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

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

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

Feigelson, E. D. & Babu, G. J. (2012). Modern Statistical Methods for Astronomy. With R Applications. *Cambridge University Press.*

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

Matt Wand R port by Brian Ripley (2011). KernSmooth: Functions for kernel smoothing for Wand & Jones (1995). R package version 2.23-6.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257–1270.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* Series B, **53**, 683–690.

Silverman, B., W. (1986). Density Estimation for Statistics and Data Analysis. *CRC Press.*

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

Sun, J., Morrisson, H., Harding, P., & Woodroofe (2002). Mixtures and bumps: errors in measurement. *Technical report.* Case Western Reserve University.

Wand, M. P., & Jones, M. C. (1994). Kernel smoothing. *CRC Press.*

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