

## Local likelihood density estimation based on smooth truncation

BY PEDRO DELICADO

*Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya,  
08034 Barcelona, Spain  
pedro.delicado@upc.edu*

### SUMMARY

Two existing density estimators based on local likelihood have properties that are comparable to those of local likelihood regression but they are much less used than their counterparts in regression. We consider truncation as a natural way of localising parametric density estimation. Based on this idea, a third local likelihood density estimator is introduced. Our main result establishes that the three estimators coincide when a free multiplicative constant is used as an extra local parameter.

*Some key words:* Local polynomial regression; Nonparametric estimation; Truncated density.

### 1. INTRODUCTION

Consider the nonparametric regression model  $Y = m(X) + \varepsilon$ , where the error term  $\varepsilon$  has an absolutely continuous distribution independent of  $X$ , with zero mean and finite variance. Let  $(x_1, y_1), \dots, (x_n, y_n)$  be a random sample of  $(X, Y)$ . The unknown function  $m(t) = E(Y|X = t)$  is the regression function. Local polynomial regression is a standard nonparametric approach for estimating  $m(t)$  (Wand & Jones, 1995, p. 114; Simonoff, 1996, p. 138; Fan & Gijbels, 1996, p. 18; Bowman & Azzalini, 1997, p. 50). The method can be interpreted as the result of locally maximising the loglikelihood of a polynomial regression model with normal errors, with density function denoted by  $f(y|X = x; \theta)$ , with  $\theta \in \mathbb{R}^k$  being the local polynomial coefficients. The estimator of  $m(t) = E(Y|X = t)$  is  $\hat{m}(t) = E\{Y|X = t, \hat{\theta}(t)\}$ , where  $\hat{\theta}(t)$  is chosen as the maximiser of

$$\sum_{i=1}^n w(x_i - t) \log f(y_i|X = x_i, \theta),$$

and where the expectation is taken with respect to the parametric model. A common choice for the weight function is  $w(u) = K(u/h)/h$ , where  $K$  is a symmetric unimodal density function, called the kernel function. The extension to other types of conditional dependence, such as binary or counting response, is straightforward, and consists of modifying the parametric likelihood appropriately. Generalised linear models are flexible enough to be used as local parametric models (Loader, 1999, p. 59; Wand & Jones, 1995, § 6.5; Fan & Gijbels, 1996, § 5.4; Bowman & Azzalini, 1997, § 3.4). Local likelihood nonparametric fitting combines a clear justification, appropriate theoretical properties, flexibility to fit a wide range of datasets and automatic boundary adaptation.

Consider now the nonparametric density estimation problem. Let  $X$  be a random variable with density function  $f$ , and let  $t \in \mathbb{R}$ . Let  $x_1, \dots, x_n$  be  $n$  independent observations of  $X$ . The goal is to estimate  $f(t)$ . The main difficulty in using local likelihood ideas in density estimation is that there is no explanatory variable on which to condition. The only way to make conditional inference is by conditioning on  $X = t$ . By analogy with regression problems, the naive version of the localised

loglikelihood function for density estimation is

$$\sum_{i=1}^n w(x_i - t) \log f(x_i, \theta),$$

or any multiple thereof, where  $f(x; \theta)$  belongs to a class  $\mathcal{F}$  of local parametric models. Nevertheless, density estimation of  $X$  at  $t$  must depend not only on what happens in a neighbourhood of  $t$  but also on the remaining observations: the estimated density might vary if the proportion of observations outside the neighbourhood of  $t$  varies. As a result the naive localised loglikelihood function does not work. As pointed out by Copas (1995), the corresponding score function has nonzero expectation, leading to invalid inference.

To avoid this, Copas (1995), Loader (1996) and Hjort & Jones (1996) proposed corrections to the naive approach that provide consistent density function estimators.

The local likelihood problem formulation proposed by Copas (1995) is

$$\max_{\theta} \sum_{i=1}^n \left[ \bar{w}(x_i - t) \log f(x_i; \theta) + \{1 - \bar{w}(x_i - t)\} \log \left\{ 1 - \int_{\mathbb{R}} \bar{w}(u - t) f(u; \theta) du \right\} \right], \quad (1)$$

where  $\bar{w}(u) = hw(u) = K(u/h)$ . The resulting score function now has zero mean. Copas (1995) establishes a parallelism between local estimation and censoring models: the weight of  $x_i$  in local estimation is compared to the probability of observing  $x_i$  in censoring models. In fact he requires that  $\bar{w}(0) = 1$  and that  $\bar{w}(u)$  decrease with  $|u|$ . Standard techniques in censored data analysis lead the author to the problem in (1). Copas (1995) admits that the censoring process is artificial in this context. An alternative approach, based on local maximum likelihood, is possible when  $\bar{w}(u)$  is the indicator function of  $B(t, h) = [t - h, t + h]$ . Assume that the model  $f(x; \theta)$  is appropriate in  $B(t, h)$ . The contribution to the local likelihood function of  $x_i \in B(t, h)$  is  $f(x_i; \theta)$  while the information that  $x_i \notin B(t, h)$  provides about  $\theta$  is that  $x_i$  belongs to a set having probability  $1 - P_{\theta}\{B(t, h)\}$ . Then the loglikelihood function coincides with the objective function in (1).

Loader (1996) and Hjort & Jones (1996) formulate the local likelihood problem as

$$\max_{\theta} \sum_{i=1}^n w(x_i - t) \log f(x_i; \theta) - n \int_{\mathbb{R}} w(u - t) f(u; \theta) du. \quad (2)$$

These papers comment on the good performance of the resulting estimator in the presence of edge effects, as well as the parallelism between its large-sample properties and those of local polynomial regression. Loader claims that the usual loglikelihood function should be written as

$$\sum_{i=1}^n \log f(x_i; \theta) - n \left\{ \int_{\mathbb{R}} f(u; \theta) du - 1 \right\},$$

and that (2) is its natural localised version. Hjort & Jones (1996, § 2.3) provide five additional arguments for the use of (2). Of those, from our point of view, the most convincing argument for (2) is that it leads to the nearest local parametric approximation to the true density, in terms of local Kullback–Leibler distance (Hjort & Jones, 1996, § 2.1).

Observe that the Copas and Loader–Hjort–Jones proposals are general ways of defining local likelihood density estimators. A local parametric family  $\mathcal{F}$  must be specified to define an estimator completely. Two different families,  $\mathcal{F}^I$  and  $\mathcal{F}^{II}$  say, will produce two different Copas or Loader–Hjort–Jones estimators. The theoretical properties of the Copas and Loader–Hjort–Jones estimators have been studied for a generic family  $\mathcal{F}$ , and when a particular family is used these properties have to be discussed in detail. Other papers studying local likelihood density estimation are Eguchi & Copas (1998), Kim et al. (2001), Park et al. (2002) and Hall & Tao (2002).

Local likelihood density estimation has not become as popular as local regression in spite of the theoretical and practical similarities. A possible explanation for that could be that arguments leading to local likelihood regression are much more direct than in the density case. In this paper we consider truncation, using a uniform kernel, as the most natural way to localise parametric density estimation.

## 2. A NEW PROPOSAL BASED ON TRUNCATION

Let  $B(t, h) = [t - h, t + h]$  for  $h > 0$ . If  $\text{pr}\{X \in B(t, h)\} > 0$  then

$$f(t) = f\{t|X \in B(t, h)\} \text{pr}\{X \in B(t, h)\}. \quad (3)$$

The first factor is the density of  $X$  truncated to  $B(t, h)$ . The second factor can be estimated by the sampling proportion

$$F_n\{B(t, h)\} = \frac{n_{th}}{n}, \quad (4)$$

where  $n_{th} = \#\{x_i \in B(t, h)\} = \sum_{i=1}^n I_{B(t, h)}(x_i)$ . For small  $h$  the unknown function  $f$  can be approximated by a parametric model on  $B(t, h)$ ,

$$\mathcal{F}_0 = \{f_0(x; \theta) : \theta \in \Theta \subseteq \mathbb{R}^k\}$$

say, where  $f_0$  is a known nonnegative function with  $\int_{\mathbb{R}} f_0(x, \theta) dx = 1$  for all  $\theta$ . For instance, the class  $\mathcal{F}_0^1$  may contain the  $N(\mu, \sigma^2)$  density functions, with  $\theta = (\mu, \sigma)$ . Then we have that

$$f_0\{x; \theta|X \in B(t, h)\} = \frac{f_0(x; \theta)}{\int_{B(t, h)} f_0(u; \theta) du}$$

is a reliable estimator of  $f\{x|X \in B(t, h)\}$ , provided that  $\theta$  is adequately chosen. We propose to select  $\theta$  by maximum truncated likelihood, solving the optimisation problem

$$\max_{\theta} \sum_{x_i \in B(t, h)} \log \frac{f_0(x_i; \theta)}{\int_{B(t, h)} f_0(u; \theta) du}. \quad (5)$$

Let  $\hat{\theta}_h(t)$  be the solution to this problem. The estimator of  $f(t)$  is then

$$\hat{f}_h(t) = \frac{f_0\{t; \hat{\theta}_h(t)\}}{\int_{B(t, h)} f_0\{u; \hat{\theta}_h(t)\} du} F_n\{B(t, h)\}. \quad (6)$$

Uniform kernels are often replaced by smoother kernel functions to obtain smoother non-parametric estimators (Silverman, 1986, p. 13). We define the uniform kernel by  $K^U(u) = \frac{1}{2}I_{[-1, 1]}(u)$ . For any function  $K(u)$  and any  $h > 0$  we write  $K_h(u) = (1/h)K(u/h)$ . Thus the objective function of the problem in (5) can be written as

$$\sum_{i=1}^n 2hK_h^U(x_i - t) \log \frac{f_0(x_i; \theta)}{\int_{\mathbb{R}} 2hK_h^U(u - t)f_0(u; \theta) du}.$$

We now replace the uniform kernel by a generic weight function  $w(u - t) = K_h(u - t)$ . The smoothed version of the problem in (5) is then obtained as

$$\max_{\theta} \sum_{i=1}^n w(x_i - t) \log \frac{f_0(x_i; \theta)}{\int_{\mathbb{R}} w(u - t)f_0(u; \theta) du}; \quad (7)$$

constants not affecting the optimisation have been removed. Let  $\hat{\theta}^{\text{ST}}(t)$  be the maximum; ST stands for smooth truncation.

The smoothed version of the term  $F_n\{B(t, h)\} = (1/n) \sum_{i=1}^n I_{B(t, h)}(x_i)$  is  $2h\hat{f}_w(t)$ , where  $\hat{f}_w(t)$  is the usual kernel estimator of  $f(t)$ , defined by  $\hat{f}_w(t) = \sum_{i=1}^n w(x_i - t)/n$ .

Taking into account the truncation rationale that led from the problem in (5) to the estimator  $\hat{f}_h(t)$  defined in (6), our proposal for local likelihood density estimation is

$$\hat{f}^{\text{ST}}(t) = \frac{f_0\{t; \theta^{\text{ST}}(t)\}}{\int_{\mathbb{R}} w(u - t)f_0\{u; \hat{\theta}^{\text{ST}}(t)\} du} \hat{f}_w(t). \quad (8)$$

Observe that  $f_0(\cdot; \theta)$  need not be a density function: it is enough that  $f_0$  is nonnegative and has finite integrals over finite intervals. For instance, the class  $\mathcal{F}_0^{\text{II}}$  may contain the functions  $f_0(x; \theta) = \exp(ax + bx^2)$ , with  $\theta = (a, b)$ .

We define a more flexible parametric model, including a free multiplicative constant:

$$\mathcal{F}_1 = \{f_1(x; c, \theta) = cf_0(x; \theta) : c > 0, \theta \in \Theta \subseteq \mathbb{R}^k\}. \tag{9}$$

For the examples  $\mathcal{F}_0^{\text{I}}$  and  $\mathcal{F}_0^{\text{II}}$ , the corresponding  $\mathcal{F}_1$  families coincide and can be expressed as  $\mathcal{F}_1^{\text{I}} = \mathcal{F}_1^{\text{II}} = \{f_1(x; \alpha, \beta, \gamma) = \exp(\alpha + \beta x + \gamma x^2)\}$ .

Observe that  $\mathcal{F}_1$  is closed over multiplication by positive constants: for all  $g(x) \in \mathcal{F}_1$  and all  $\delta > 0$ ,  $\delta g(x) \in \mathcal{F}_1$ . Moreover, if  $\mathcal{F}_0$  has this property then  $\mathcal{F}_1$  coincides with  $\mathcal{F}_0$ . Two important parametric classes of functions having this property are the polynomial parametric model considered by Hjort & Jones (1996) and the log-polynomial parametric model introduced by Loader (1996), where  $\mathcal{F}_1^{\text{I}} = \mathcal{F}_1^{\text{II}}$  is obtained for second-degree polynomials.

We now establish the numerical equivalence of our proposal and the problems corresponding to (1) (Copas, 1995) and (2) (Loader, 1996; Hjort & Jones, 1996) when the working parametric family is  $\mathcal{F}_1$ , that is a family closed over multiplication by positive constants. We first consider the problem in (1) when the parametric family is  $\mathcal{F}_1$ , and the maximisation is carried out over  $(c, \theta)$ . Some care has to be taken in order to have positive arguments for the log function in the second term of the objective function: in fact the appropriate extended parametric model for the problem in (1) is

$$\mathcal{F}_1^{\text{C}} = \left\{ f_1(x; c, \theta) = cf_0(x; \theta) : c > 0, \theta \in \Theta \subseteq \mathbb{R}^k, c \int_{\mathbb{R}} \bar{w}(u-t)f_0(u; \theta) < 1 \right\}.$$

Let  $(\hat{c}^{\text{C}}(t), \hat{\theta}^{\text{C}}(t))$  be the solution and let  $\hat{f}^{\text{C}}(t) = f_1\{t; \hat{c}^{\text{C}}(t), \hat{\theta}^{\text{C}}(t)\}$  be the corresponding estimator of  $f(t)$ .

Now we consider the problem in (2), when the parametric family is  $\mathcal{F}_1$  and the maximisation is carried out over  $(c, \theta)$ . Let  $(\hat{c}^{\text{L}}(t), \hat{\theta}^{\text{L}}(t))$  be the solution. The resulting estimator of  $f(t)$  is  $\hat{f}^{\text{L}}(t) = f_1\{t; \hat{c}^{\text{L}}(t), \hat{\theta}^{\text{L}}(t)\}$ .

Observe that in (7) it is enough to take  $\mathcal{F}_0$  as the local parametric model.

**THEOREM 1.** *In the previous context,  $\hat{\theta}^{\text{C}}(t) = \hat{\theta}^{\text{L}}(t) = \hat{\theta}^{\text{ST}}(t)$ . Moreover*

$$\hat{c}^{\text{C}}(t) = \hat{c}^{\text{L}}(t) = \frac{\hat{f}_w(t)}{\int_{\mathbb{R}} w(u-t)f_0\{u; \hat{\theta}^{\text{ST}}(t)\}du}.$$

Finally  $\hat{f}^{\text{C}}(t) = \hat{f}^{\text{L}}(t) = \hat{f}^{\text{ST}}(t)$ .

The proof is deferred to the Appendix.

Under the same assumptions, the three local likelihood problems are equivalent to maximising the naive localised likelihood function, subject to the kernel estimator of  $f(t)$  being equal to its expected value under the parametric model; see the note following the proof of Theorem 1.

*Remark 1: Asymptotic properties of  $\hat{f}^{\text{ST}}(t)$ .* The numerical equivalence established in Theorem 1 implies that the asymptotic properties of the new estimator coincide with those of the previous ones. Hjort & Jones (1996) study small-bandwidth asymptotics for the Loader–Hjort–Jones local likelihood density estimator when a generic parametric family, with  $p$  parameters, is used. They prove that the asymptotic bias and variance of the estimator depend only on the number  $p$  of local parameters fitted. Their results apply in particular to the local parametric model  $\mathcal{F}_1$  for  $p = k + 1$ . Therefore, following §§ 3 and 4 in Hjort & Jones (1996), we can say that, when we use a second-order kernel  $K$  and  $k$  parameters in  $\mathcal{F}_0$ , the bias of  $\hat{f}^{\text{ST}}(t)$ , as well as those of  $\hat{f}^{\text{C}}(t)$  and  $\hat{f}^{\text{L}}(t)$ , is  $O(h^2)$  for  $k = 0$  or  $k = 1$ , and is  $O(h^4)$  for  $k = 2$  or  $k = 3$ . The variance is always  $O\{(nh)^{-1}\}$ .

Large-bandwidth asymptotics were discussed in Eguchi & Copas (1998) for a wider class of local likelihood methods including both the Copas and the Loader–Hjort–Jones estimators; for this

class, Park et al. (2002) find small-bandwidth asymptotic results similar to those of Hjort & Jones (1996). Theorem 1 guarantees that the asymptotic results established in Eguchi & Copas (1998) also apply to  $\hat{f}^{\text{ST}}(t)$ , given the required assumption that the true density function  $f(t)$  is in the semiparametric band

$$\bigcup_{c, \theta} \{g: D(g(\cdot), cf_0(\cdot; \theta)) = O(n^{-(1+\alpha)})\},$$

where  $n$  is the sample size,  $\alpha > 0$ , and  $D(g, f)$  is the Kullback–Leibler distance between  $g$  and  $f$ .

*Remark 2: Computational considerations for  $\hat{f}^{\text{ST}}(t)$ .* From Remark 1 it follows that the three estimators differ just in terms of computational considerations. We will compute only  $\hat{f}^{\text{L}}(t)$  and  $\hat{f}^{\text{ST}}(t)$ , for two reasons. First, the direct numerical computation of  $\hat{f}^{\text{C}}(t)$  requires us to solve a constrained optimisation problem, since  $(c, \theta)$  must be such that  $cf_0(t; \theta)$  is in  $\mathcal{F}_1^{\text{C}}$ , that is much more expensive than computing  $\hat{f}^{\text{L}}(t)$  or  $\hat{f}^{\text{ST}}(t)$ . Secondly, the proof of Theorem 1 in the Appendix gives the closed-form expression for the optimal value  $c$  for a fixed  $\theta$ , and shows that including this formula in the implementation of  $\hat{f}^{\text{C}}(t)$  leads to solving the same problem as in  $\hat{f}^{\text{ST}}(t)$ .

The main difference between solving the problem in (2) to obtain  $\hat{f}^{\text{L}}(t)$  and that in (7) to obtain  $\hat{f}^{\text{ST}}(t)$  is that in the former the optimisation variable has dimensional  $k + 1$ , and in the latter the dimension is  $k$ . This favours  $\hat{f}^{\text{ST}}(t)$ . Nevertheless, the theoretical levels of computational complexity coincide: in both cases the evaluation of the objective function, its gradient and its Hessian matrix requires numerical integrations and sums of  $O(nh)$  terms. The evaluation of  $\hat{f}^{\text{ST}}(t)$  involves the kernel density estimator  $\hat{f}_w(t)$ , but it can be determined at no extra cost from previously computed quantities. On the other hand the objective function in (2) is simpler than that in (7), and this favours  $\hat{f}^{\text{L}}(t)$ .

In order to evaluate the practical performance of both estimators we have designed the following computer experiment. As local parametric model we have considered the log-polynomial model

$$\mathcal{F}_0 = \left\{ \exp\left(\sum_{j=1}^k \theta_j x^j\right) : \theta_j \in \mathbb{R} \right\}, \quad \mathcal{F}_1 = \left\{ \exp\left(\sum_{j=0}^k \theta_j x^j\right) : \theta_j \in \mathbb{R} \right\}.$$

A Newton–Raphson algorithm was implemented in R to solve the problems in (2) and (7) numerically. For each problem we derive the formulae for the gradient and the Hessian matrix; numerical integration is required.

We have considered samples from a mixture of normal distributions, with theoretical density function  $f(x) = \frac{3}{4}\phi_N(x; \mu = 0, \sigma = 1) + \frac{1}{4}\phi_N(x; \mu = \frac{3}{2}, \sigma = \frac{1}{3})$ , where  $\phi_N(x; \mu, \sigma)$  is the density function of a  $N(\mu, \sigma^2)$ . Thus the interval  $[-3, 3]$  has probability almost equal to 1. This density function appears as an example in the second chapter of Wand & Jones (1995). We use a sample size of  $n = 100$  and the degree of local polynomials is  $k = 2$ , equivalent to locally fitting a normal density, multiplied by a constant. We use an Epanechnikov kernel and a window width of  $h = 1.25$ .

Figure 1 shows the results of four iterations of the Newton–Raphson algorithm. The initial values of the parameters were such that the corresponding density was uniform in  $[-3, 3]$ . After four iterations both estimators are almost identical, as predicted by the theory. The main differences are in regions where the density is close to 0. In one iteration the estimator  $\hat{f}^{\text{ST}}(t)$  gives good results, and two iterations are enough to arrive at the final solution. The convergence of  $\hat{f}^{\text{L}}(t)$  is slower; the four iterations are needed. In terms of computer time,  $\hat{f}^{\text{ST}}(t)$  requires about 5% more time than  $\hat{f}^{\text{L}}(t)$  to complete the four iterations. The process of fitting  $\hat{f}^{\text{ST}}(t)$  presents sporadic numerical stability problems when the objective function in (7) is evaluated at parameters  $\theta$  such that  $f_0(t, \theta)$  is close to the machine precision, because the log argument is close to 0/0. In this sense  $\hat{f}^{\text{L}}(t)$  is more robust.

We have also experimented with other sample sizes  $n$ , other degrees  $k$  of polynomials, other smoothing parameters  $h$  and other types of kernel. We also have used a theoretical model with density  $f(x) = (2 - x)I_{[0,1]}(x)$ . The results were qualitatively similar.

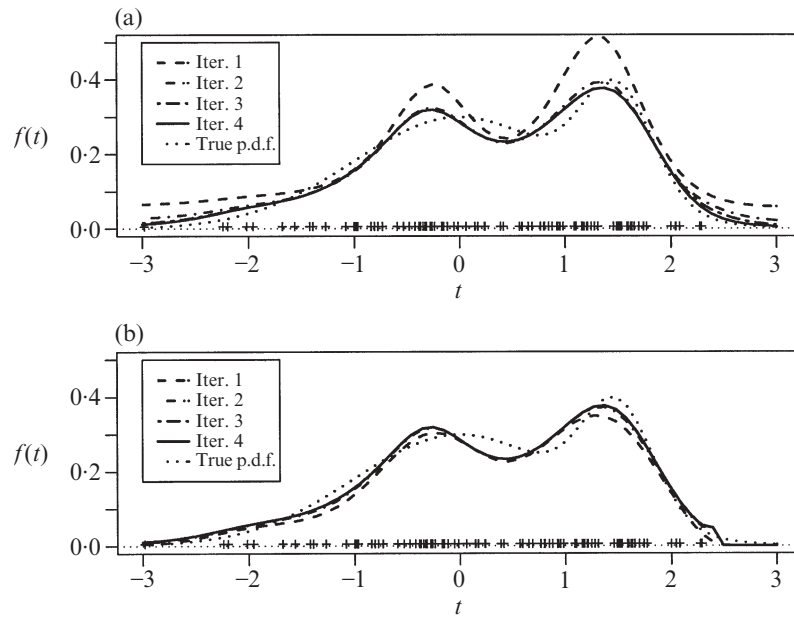


Fig. 1: Results from the computing experiment. (a) Loader-Hjort-Jones estimator. (b) Smooth-truncation estimator. Four iterations for each estimator are shown, and the dotted line corresponds to the true density; p.d.f., probability density function.

We conclude that  $\hat{h}^{\text{ST}}(t)$  requires fewer iterations than  $\hat{h}^{\text{L}}(t)$  to reach the solution, but the optimisation problem leading to  $\hat{h}^{\text{L}}(t)$  is simpler and numerically more stable.

*Remark 3:*  $\hat{\theta}^{\text{ST}}(t)$  and the  $T$ -version of Eguchi & Copas (1998). It is appropriate to compare the estimator (8) with the  $T$ -version, where  $T$  stands for truncation, of the local likelihood estimator proposed by Eguchi & Copas (1998). It is easy to see that, apart from a constant, the corresponding local likelihood function coincides with the objective function in (7). Thus the estimated parameter in the  $T$ -version is equal to  $\hat{\theta}^{\text{ST}}(t)$ . The difference stems from the definition of the density estimator: the  $T$ -version estimator is  $\hat{f}^T(t) = f_0\{t; \hat{\theta}^{\text{ST}}(t)\}$ , while estimator (8) includes a normalising factor, i.e. the denominator, and a weight factor,  $\hat{f}_w(t)$ . We have seen that the natural estimator for  $f(t)$  derived from truncation is  $\hat{f}^{\text{ST}}(t)$  and not  $\hat{f}^T(t)$ . It is therefore not surprising that the  $T$ -version has undesirable properties for small  $h$ , as reported by Eguchi & Copas (1998) and by Park et al. (2002).

*Remark 4.* Two problems affect the three estimators we have dealt with in this paper, namely how to choose the bandwidth  $h$  and the fact that the estimated density is not a bona fide function, in that it does not integrate to one. The bandwidth choice is considered in § 8.3 of Hjort & Jones (1996). They suggest using a plug-in rule or least-squares crossvalidation, the latter being less reliable. For the second difficulty, ideas in Gajek (1986) and Hall & Murison (1993) can be applied in our context to create bona fide densities without changing asymptotic properties.

#### ACKNOWLEDGEMENT

Research partially supported by the Spanish Ministry of Education and Science and Fondo Europeo de Desarrollo Regional, and by the E.U. Pattern Analysis, Statistical Modelling and Computational Learning Network of Excellence. Positive interchange with J. de Uña and J. Ginebra is gratefully acknowledged, as well as the help of G. Silverstone Tarr as copy editor. The comments of the editor and an anonymous reviewer substantially improved this paper.

APPENDIX

*Proof of Theorem 1*

We start by proving that (5) is equivalent to the version of (2) corresponding to the uniform kernel. The equivalence between  $\hat{\theta}^L(t)$  and  $\hat{\theta}^{ST}(t)$ , and the value of  $\hat{c}^L(t)$  can be established in a completely parallel way.

The problem in (5) is equivalent to the constrained optimisation problem

$$\max_{c, \theta} \sum_{x_i \in B(t, h)} \log f_1(x_i; c, \theta), \tag{A1}$$

subject to

$$\int_{B(t, h)} f_1(u; c, \theta) du = 1,$$

in the sense that  $(c, \theta)$  is a solution of (A1) if and only if  $\theta$  is a solution of (5) and

$$c = \left\{ \int_{B(t, h)} f_0(u; \theta) du \right\}^{-1}.$$

The estimator of  $f(t)$  can be written as  $\hat{f}_h(t) = f_1\{t; \hat{c}_h(t), \hat{\theta}_h(t)\} F_n\{B(t, h)\}$ , where  $(\hat{c}_h(t), \hat{\theta}_h(t))$  is the optimiser of (A1). Remember that  $\mathcal{F}_1$  is closed for products by positive constants. Let  $\hat{c}_h^*(t) = \hat{c}_h(t) F_n\{B(t, h)\}$ . Then

$$f_1\{x; \hat{c}_h(t), \hat{\theta}_h(t)\} F_n\{B(t, h)\} = f_1\{x; \hat{c}_h^*(t), \hat{\theta}_h(t)\}$$

for all  $x$ , which verifies that  $\int_{B(t, h)} f_1\{u; \hat{c}_h^*(t), \hat{\theta}_h(t)\} du = F_n\{B(t, h)\}$ . We conclude that problem (A1) is equivalent to

$$\max_{c, \theta} \sum_{x_i \in B(t, h)} \log f_1(x_i; c, \theta), \tag{A2}$$

subject to

$$\int_{B(t, h)} f_1(u; c, \theta) du = F_n\{B(t, h)\},$$

because  $(c^*, \theta)$  is the solution of (A2) if and only if  $(c, \theta)$  is the solution of (A1) and  $c^* = c F_n\{B(t, h)\}$ . Thus, the estimator of  $f(t)$  is obtained directly as  $\hat{f}_h(t) = f_1\{t; \hat{c}_h^*(t), \hat{\theta}_h(t)\}$ , where  $(\hat{c}_h^*(t), \hat{\theta}_h(t))$  is the optimiser of (A2). The Lagrangian function associated with (A2) is

$$l_{th}^0(c, \theta, \lambda) = \sum_{x_i \in B(t, h)} \log f_1(x_i; c, \theta) - \lambda \left[ \int_{B(t, h)} f_1(u; c, \theta) du - F_n\{B(t, h)\} \right].$$

If we set the partial derivative of  $l_{th}^0$  with respect to  $c$  equal to zero, it follows that  $\lambda = n$ . We then define  $l_{th}(c, \theta) = l_{th}^0(c, \theta, n)$ , so that

$$l_{th}(c, \theta) = \sum_{x_i \in B(t, h)} \log f_1(x_i; c, \theta) - n \left[ \int_{B(t, h)} f_1(u; c, \theta) du - F_n\{B(t, h)\} \right],$$

and conclude that (A2) is equivalent to

$$\max_{c, \theta} l_{th}(c, \theta). \tag{A3}$$

Note that the last term in  $l_{th}(c, \theta)$  does not depend on  $(c, \theta)$  and thus can be deleted. Therefore, (A3) is the version of (2) corresponding to the uniform kernel, and the first part of the proof is complete.

We now study the equivalence between the problems in (1) and (7). The local likelihood function maximised in (1) is, after addition and subtraction of  $\sum_i \bar{w}(x_i - t) \log \int_{\mathbb{R}} \bar{w}(u - t) c f_0(u; \theta) du$ ,

$$\begin{aligned} l(c, \theta) &= \sum_{i=1}^n \bar{w}(x_i - t) \log \frac{f_0(x_i; \theta)}{\int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du} \\ &\quad + \left\{ \sum_{i=1}^n \bar{w}(x_i - t) \right\} \log \left\{ c \int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du \right\} \\ &\quad + \left\{ n - \sum_{i=1}^n \bar{w}(x_i - t) \right\} \log \left\{ 1 - c \int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du \right\}. \end{aligned}$$

The first term does not depend on  $c$ . Let  $n_w = \sum_{i=1}^n \bar{w}(x_i - t) = nh \hat{f}_w(t)$  and

$$P_w(\theta) = \int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du.$$

The sum of the second and third terms is

$$g_\theta(c) = n_w \log \{c P_w(\theta)\} + (n - n_w) \log \{1 - c P_w(\theta)\}.$$

The maximiser of  $g_\theta(c)$  is

$$\hat{c}(t; \theta) = \frac{n_w}{n P_w(\theta)} = \frac{\hat{f}_w(t)}{\int_{\mathbb{R}} w(u - t) f_0(u; \theta) du}$$

and  $g_\theta\{\hat{c}(t; \theta)\}$  is constant in  $\theta$ . The optimum value of  $\theta$  is therefore the solution of

$$\begin{aligned} \max_{\theta} \sum_{i=1}^n \bar{w}(x_i - t) \log \frac{f_0(x_i; \theta)}{\int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du} \\ = h \left\{ \sum_{i=1}^n w(x_i - t) \log \frac{f_0(x_i; \theta)}{\int_{\mathbb{R}} w(u - t) f_0(u; \theta) du} \right\} - h \sum_{i=1}^n w(x_i - t) \log h, \end{aligned}$$

which is equivalent to the problem in (7). Therefore,  $\hat{\theta}^C = \hat{\theta}^{ST}$  and  $\hat{c}^C(t) = \hat{c}(t; \hat{\theta}^C)$  is as stated in the Theorem.

Finally, observe that for any  $\theta$  the pair  $(\theta, \hat{c}(t; \theta))$  satisfies

$$\hat{c}(t; \theta) \int_{\mathbb{R}} \bar{w}(u - t) f_0(u; \theta) du = \frac{n_w}{n} = \frac{\sum_{i=1}^n \bar{w}(x_i - t)}{n} \leq \frac{\sum_{i=1}^n \bar{w}(0)}{n} = 1,$$

taking into account that  $\bar{w}$  has its maximum at 0 and that  $\bar{w}(0) = 1$ . It follows that the optimisation was done within the parametric model  $\mathcal{F}_1^C$ .  $\square$

*Remark A1.* Suppose that in problem (A2) we replace the uniform kernel by a generic one. Then the equality constraint becomes  $\int_{\mathbb{R}} w(u - t) f_1(u; c, \theta) du = n^{-1} \sum_{i=1}^n w(x_i - t)$  and problem (A2) becomes

$$\max_{c, \theta} \sum_{i=1}^n w(x_i - t) \log f_1(x_i; c, \theta),$$

subject to

$$\{f_1(\cdot; c, \theta) * w\}(t) = \hat{f}_w(t),$$

thus proving the claim following Theorem 1.

## REFERENCES

- BOWMAN, A. W. & AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.
- COPAS, J. B. (1995). Local likelihood based on kernel censoring. *J. R. Statist. Soc. B* **57**, 221–35.
- EGUCHI, S. & COPAS, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. R. Statist. Soc. B* **60**, 709–24.
- FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.
- GAJEK, L. (1986). On improving density estimators which are not bona fida functions. *Ann. Statist.* **14**, 1612–8.
- HALL, P. & MURISON, R. D. (1993). Correcting the negativity of high-order kernel density estimators. *J. Mult. Anal.* **47**, 103–22.
- HALL, P. & TAO, T. (2002). Relative efficiencies of kernel and local likelihood density estimators. *J. R. Statist. Soc. B* **64**, 537–47.
- HJORT, N. L. & JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–47.
- KIM, W. C., PARK, B. U. & KIM, Y. G. (2001). On Copas' local likelihood density estimator. *J. Korean Statist. Soc.* **30**, 77–87.
- LOADER, C. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–18.
- LOADER, C. (1999). *Local Regression and Likelihood*. New York: Springer.
- PARK, B. U., KIM, W. C. & JONES, M. C. (2002). On local likelihood density estimation. *Ann. Statist.* **30**, 1480–95.
- SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer.
- WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

[Received May 2004. Revised November 2005]