The Canadian Journal of Statistics
Vol. 33, No. ?, 2005, Pages ???–???
La revue canadienne de statistique

Local likelihood density estimation for interval censored data

John BRAUN, Thierry DUCHESNE and James E. STAFFORD

Key words and phrases: EM algorithm; HIV; kernel density estimation; likelihood cross-validation; nonparametric maximum likelihood; self-consistency; smoothed histograms; symbolic computation.
MSC 2000: Primary 62G07; secondary 62N02.

Abstract: The authors propose a class of procedures for local likelihood estimation from data that are either interval-censored or that have been aggregated into bins. One such procedure relies on algorithm that generalizes existing self-consistency algorithms by introducing kernel smoothing at each step of the iteration. The entire class of procedures yields estimates that are obtained as solutions of fixed point equations. By discretizing and applying numerical integration, the authors use fixed point theory to study convergence of algorithms for the class. Rapid convergence is effected by the implementation of a local EM algorithm as a global Newton iteration. The latter requires an explicit solution of the local likelihood equations which can be found by using symbolic Newton–Raphson, if necessary.

Estimation de la densité par vraisemblance locale à partir de données censurées par intervalle

Résumé : Les auteurs proposent une classe de procédures pour l'estimation de la densité par vraisemblance locale lorsque les données sont censurées par intervalle ou qu'elles ont été regroupées en classes. L'une de ces procédures s'appuie sur un algorithme qui, en faisant appel à un noyau lissant à chaque itération, généralise les algorithmes auto-convergents déjà existants. Les estimations auxquelles la classe conduit sont des points fixes de certaines équations. En s'appuyant sur des techniques de discrétisation et d'intégration numérique, les auteurs se servent de la théorie des points fixes pour étudier la convergence des algorithmes de la classe. La convergence est accélérée par l'emploi d'un algorithme EM local dans l'itération globale de la méthode de Newton. Cette dernière fait intervenir une solution d'équations de vraisemblance locale qui, au besoin, peut être trouvée au moyen d'un algorithme de Newton–Raphson symbolique.

1. INTRODUCTION

Kernel density estimation is a simple and flexible method whose popularity is grounded in its interpretive appeal. Central to its use are kernel weights which depend on the proximity of an observation to the point of estimation, lending the estimator a local interpretation. In the context of interval censored data, an observation is only known to lie within some interval and it seems natural to define the weight as the conditional expectation of the kernel over that interval. Doing so not only yields an estimator that retains the interpretive appeal of a kernel density estimate, but also leads to some innovative techniques. For example, when the conditional expectation is computed with respect to the density estimate itself, a fixed point equation arises. Solving the equation iteratively leads to a generalization of the classical self-consistency algorithms of Efron (1967), Turnbull (1976) and Li, Watkins & Yu (1997). In addition, the estimator avoids some arbitrary aspects associated with the standard technique of directly smoothing the nonparametric

maximum likelihood estimator (NPMLE) of the cumulative distribution function. These details are discussed in Section 2 of this paper.

In Section 3, the kernel density estimator proposed in Section 2 is embedded in a broader class of local likelihood density estimators for interval censored data. The class generalizes the methods of Loader (1996) and Hjort & Jones (1996) by addressing the interval censoring through the use of an EM-type strategy. As in Loader (1996), we focus on local polynomial approximations of the log density. The coefficients of this polynomial form the basis of local linear and local quadratic estimators that have the potential to reduce bias.

Estimates of the coefficients of the local polynomial are computed using an EM-type algorithm and numerical integration. Here convergence to a unique solution is assured under certain conditions. By implementing the local EM algorithm as a global Newton iteration, very rapid convergence can be achieved. While such an implementation offers dramatic improvements in computational efficiency, it requires an explicit expression for the solution of the local likelihood equations, or rather, the M-step. When this is not directly available, the methods of symbolic computation given in Andrews & Stafford (2000) can be used. Here symbolic Newton–Raphson is shown to extend the exact results of Hjort & Jones (1996) for a Gaussian kernel but it also provides an explicit expression for any kernel. Issues concerning implementation and convergence are treated in Section 4 with an accompanying empirical study.

In Section 5, a direct analogy with likelihood cross-validation for completely observed data leads to a method for choosing a suitable value for the window size of the kernel. Concluding remarks are given in Section 6. Throughout this paper, applications are given for HIV data and data aggregated into a histogram. The latter demonstrates that our proposal is an effective alternative to the methods of Jones (1989) and Bellhouse & Stafford (1999).

2. A KERNEL DENSITY ESTIMATE FOR INTERVAL CENSORED DATA

Assume independent random variables $X_1, \ldots, X_n$ are drawn from a distribution with an unknown continuous univariate density f. These X's might not be directly observed. For each X, we assume a partition $\mathcal{T} = \{\tau_1, \tau_2, \ldots\}$ of the real line which is independent of X; the τ's might be monitoring times, for example. The observed data are of the form $I = (L, R)$, where

$$R = \inf\{\tau_j : \tau_j \geq X\}, \quad L = \sup\{\tau_j : \tau_j \leq X\}.$$

Thus, the observed data are a sequence of independent intervals $I_1, \ldots, I_n$. Such data may arise from panels in a study, bins in a histogram, or from visit times where an individual is tested for HIV. If X is observed we have $L = R$ and if $R = \infty$ ($L = -\infty$) then X is said to be right (left) censored.

2.1. Extending the usual kernel density estimate.

When we observe the complete sample $X_1, \ldots, X_n$, an appealing estimate of f is the usual kernel density estimate

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x\right), \tag{1}$$

with kernel function $K_h(u) = h^{-1} K(u/h)$. The appeal of $\hat{f}$ lies in the interpretation of each kernel weight, $K_h(X_i - x)$, in terms of the proximity of an observation X_i to x, the location of the kernel. When the data are interval censored, a natural extension of (1) is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left\{ K_h\left(X_i - x\right) | \, I_i \right\}. \tag{2}$$

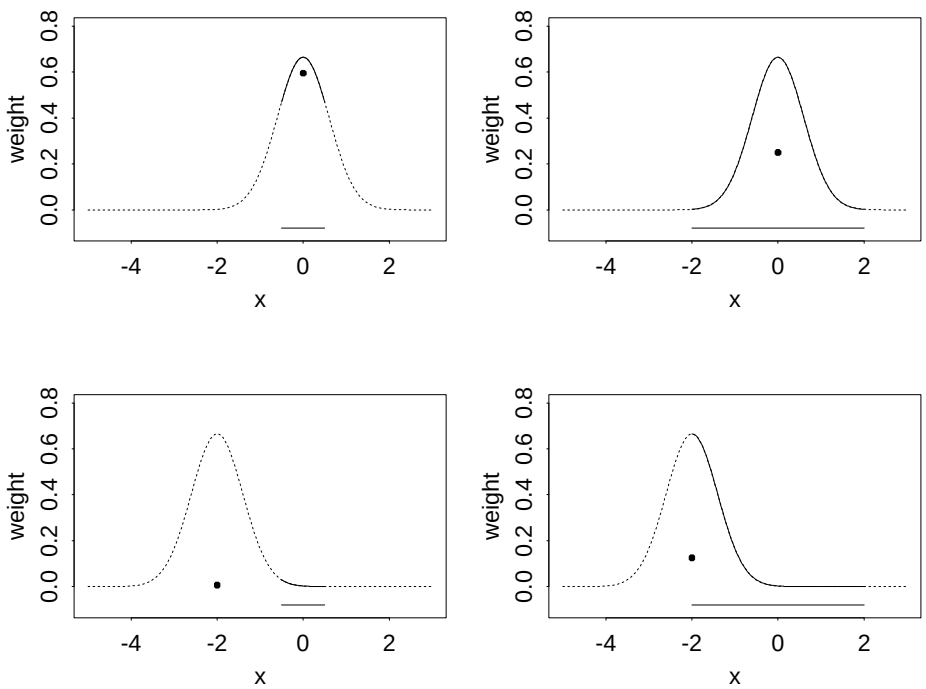

Figure 1: The kernel weight proposed in Section 2.1 is given for two intervals (solid horizontal lines), both centered at 0 but with different lengths. The columns of the display refer to a distinct interval; the rows to a distinct position of the kernel. When the kernel is also centered at 0, the method rewards precision by giving the shorter interval a greater weight (upper left panel). When the location of the kernel is shifted to "-2" it now assigns a larger weight to the longer interval (lower right panel). The latter is due to the longer interval being more "local" to "-2" than the shorter interval.

This retains an interpretation similar to (1) because the kernel weight for an observed interval is the average height of the kernel over that interval. Figure 1 depicts how this weight depends on the length of the interval and its proximity to the center of the kernel. Goutis (1997) proposed an estimator somewhat similar to (2) but in the context of the nonparametric estimation of a mixing density not applicable here.

In (2) the ith expectation is conditional on the random interval $I_i = [L_i, R_i]$ which arises from the partition $\mathcal{T}_i$, and which contains X_i by definition. A common approach is to assume the conditional distribution over the interval I_i is uniform, but because of the independence of X_i and the partition $\mathcal{T}_i$, the conditional distribution is in fact

$$F_{X_i|I_i}(x) = \frac{F(x) - F(L_i)}{F(R_i) - F(L_i)}$$

where $F(x) = \int^x f(t)dt$. Here the conditional distribution is itself unknown and must be estimated. One choice involves the kernel density estimate itself, and this results in a fixed point equation for $\hat{f}$,

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{\hat{f}} \left\{ K_h \left(X_i - x \right) \middle| I_i \right\}. \tag{3}$$

We propose solving (3) by an iterative algorithm where at the jth step

$$\hat{f}_j(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{f}_{j-1}} \left\{ K_h\left(X_i - x\right) \middle| I_i \right\}, \tag{4}$$

and $\hat{f}_0$ denotes the initial value. The conditional density over the ith interval at the jth step is

$$\hat{f}_{j-1;i}(t) = 1(t \in I_i)\frac{\hat{f}_{j-1}(t)}{c_{j-1;i}}$$

where

$$c_{j-1;i} = \int_{I_i} \hat{f}_{j-1}(t)dt.$$

We devote the remainder of this section, as well as Section 3, to placing the above estimator in a broadening context. Computational details are deferred to Sections 4, 5 where, for example, we prove that the fixed point of our implementation of (4) does not depend on $\hat{f}_0$. Finally, throughout the rest of the paper use of the subscript j^* will denote the fixed point of an algorithm, for example $\hat{f}_{j^*}$ denotes the fixed point of (4).

2.2. Kernel smoothing the NPMLE.

An alternative to solving (3) involves directly smoothing the non-parametric maximum likelihood estimator (NPMLE) of the cumulative distribution function, F, as is typically done for the empirical distribution function, F_n, and the Kaplan–Meier product limit estimator, $\hat{F}_k$. For example, (1) may be written as $\hat{f}(x) = E_{F_n}\{K_h(X - x)\}$; see Section 6.2.3 of Wand & Jones (1995) for details concerning smoothing $\hat{F}_k$.

When data are interval censored, Turnbull (1976) showed that the NPMLE, $\hat{F}_t$, is only defined up to an equivalence class of distributions over gaps called innermost intervals. Associated with each innermost interval is a probability mass whose location is left unspecified by the equivalence class. As a result it is not clear how to directly smooth $\hat{F}_t$. One possibility is to deal with the probability masses in a somewhat arbitrary fashion. For example, Pan (2000) suggests assigning them to the right-hand points of the innermost intervals and using

$$\hat{f}_t(x) = E_{\hat{F}_t}\left\{K_h\left(X - x\right)\right\}$$

to estimate f. However, $\hat{f}_t$ as defined is not unique, as we may prefer to place the probability masses somewhere else, such as at the midpoint of the innermost interval or at the left-hand point and so on. This complication does not arise when using (4), because the algorithm smooths the data directly at every step of the iteration rather than smoothing $\hat{F}_t$ once. Innermost intervals never explicitly enter into the calculation resulting in the advantage that (4) fills in the gaps of $\hat{F}_t$ in a data driven way.

Figure 2 compares $\hat{f}_{j^*}$ with $\hat{f}_t$, where both are applied to a group of hemophiliacs whose time of infection with the HIV virus was interval censored (De Gruttola & Lagakos 1989). The upper plot gives the original data ordered by the left end point. Time is measured in six month intervals and right censored observations are denoted by dotted lines. The lower plot gives $\hat{f}_{j^*}$ and $\hat{f}_t$; the same window size is used throughout. Evidently $\hat{f}_t$ is not uniquely determined, while the uniqueness of $\hat{f}_{j^*}$ will be demonstrated in Section 4. In addition, $\hat{f}_{j^*}$ does a better job of smoothing what may be a sampling anomaly on the left side of the plot without eroding the peak on the right. Both of these improvements can be attributed to introducing kernel smoothing at each step of a self-consistent algorithm (Section 2.3) rather than smoothing after such an algorithm has converged, as in the case of $\hat{f}_t$. Finally, Figure 3 gives the result of (4) for the first four iterations and for a variety of

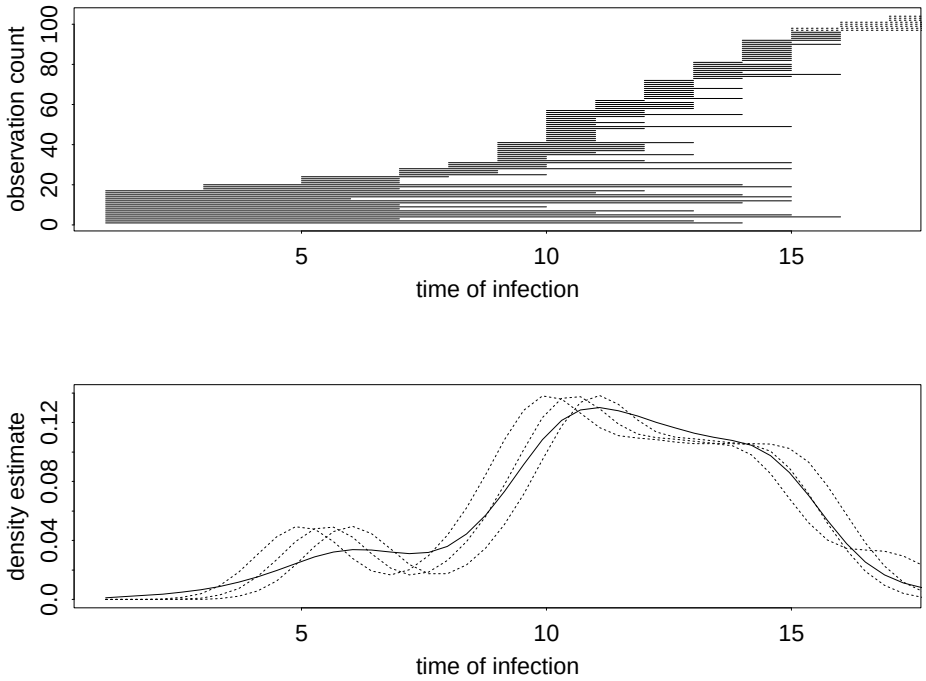

Figure 2: The top panel shows the original data by a line joining the left and right endpoints of each observed interval. The lower panel shows four kernel density estimates; the solid line indicates $\hat{f}_{j^*}$ and the three dotted lines indicate various kernel smoothed versions of Turnbull's estimator, $\hat{f}_t$. The differences in the three dotted lines are due to placing probability masses at the left-hand, right-hand, and mid-point of the innermost intervals.

initial values from a location-scale beta family. It is suggestive that the algorithm will converge to a unique solution that is independent of the initial value $\hat{f}_0$. See Section 4 for more details on convergence.

Note 1. Data that have been grouped into a histogram are interval censored. Here the NPMLE $\hat{F}_t$ has innermost intervals and probabilities that correspond to histogram bins and weights respectively. When smoothing a histogram, as in the manner of Jones (1989) say, it is immediate that placing weights in the center of bins is as arbitrary as Pan's suggestion given above. In Section 3 we use the methods of this paper as an alternative to Jones (1989) and show the effectiveness of polynomial adjustments.

2.3. Relationship to self-consistency.

The idea of filling in the gaps of Turnbull's $\hat{F}_t$ is not new. Li, Watkins & Yu (1997) propose an EM algorithm designed specifically for this purpose. In this section we show that for a vanishing window size, h, our algorithm coincides with that of Li, Watkins & Yu (1997) and hence with those of Efron (1967) and Turnbull (1976) as well.

Efron (1967) proposed an algorithm for approximating the cumulative distribution function

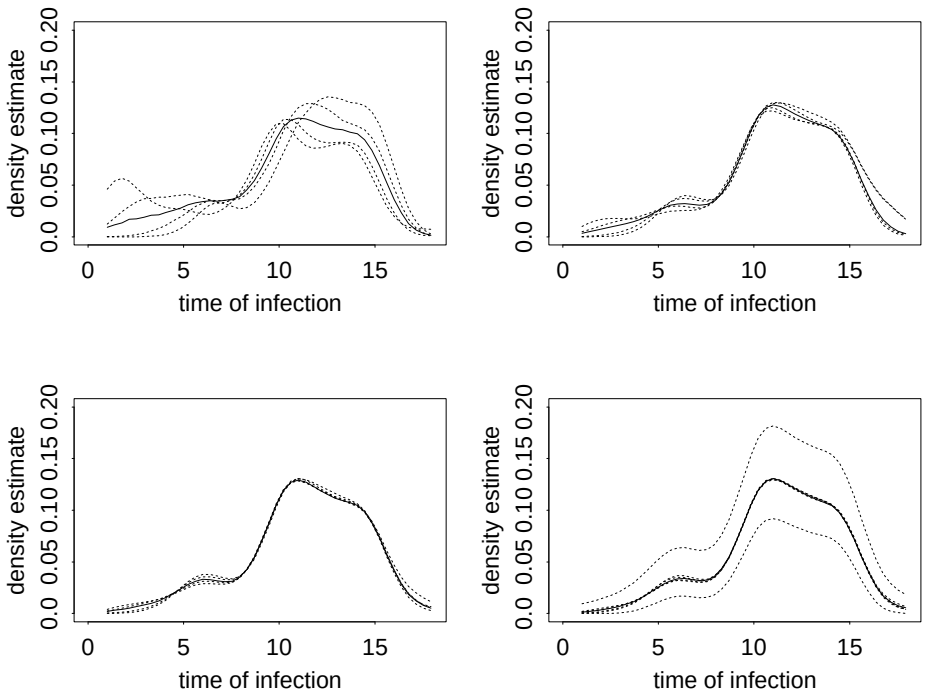

Figure 3: The figure gives from left to right then top to bottom $\hat{f}_j$ for $j = 1, \ldots, 4$. Each plot has several estimates of the density that correspond to different initial values of the algorithm. We appear to have convergence to a unique solution that does not depend on the initial value. The final plot also gives a simultaneous confidence band based on a bootstrap scheme (Davison & Hinkley 1997, p. 418) which is effective for all density estimates proposed.

when data are potentially right censored,

$$n\tilde{F}_j(x) = N(x) - \sum_{\substack{L_i < x \\ \delta_i = 0}} \frac{1 - \tilde{F}_{j-1}(x)}{1 - \tilde{F}_{j-1}(L_i)}.$$

Here $N(x) = \#\{X_i \leq x\}$, $\delta_i = 1$ if X_i is observed exactly, and $\delta_i = 0$ if X_i is right-censored. Efron showed that $\tilde{F}_j$ converges to a fixed point that coincides with the Kaplan–Meier product limit estimator and called this fixed point a self-consistent estimate. Turnbull (1976) then generalized this self-consistency algorithm to obtain the NPMLE $\hat{F}_t$ under general censoring and truncation schemes. The non-uniqueness of $\hat{F}_t$ over innermost intervals prompted Li, Watkins & Yu (1997) to propose an EM algorithm which coincides with $\hat{F}_t$ where it is uniquely defined, but converges over the innermost intervals to a value that depends on the starting point of the algorithm. The algorithm involves computing the conditional expectation of F_n at each step,

$$\check{F}_j(x) = E_{j-1}\left\{ F_n(x) | \, I_i \quad \forall i \right\}. \tag{5}$$

The following theorem shows that (5) can be obtained as a limit of (4) as the window width, h, of the kernel shrinks to zero at every step. In other words our algorithm modifies the usual self-consistency algorithms by introducing kernel smoothing at each step of the iteration.

THEOREM 1. *Let* $\hat{F}_j(x) = \int_{-\infty}^{x} \hat{f}_j(t)dt$ *be the estimate of the cumulative distribution function corresponding to the* j*th iterate of* (4)*, then*

$$\lim_{h \downarrow 0} \hat{F}_j(x) = \check{F}_j(x), \ \forall x, \ j = 1, \ 2, \ \ldots$$

Proof: $\check{F}_j$ may be rewritten as

$$\begin{aligned}
\check{F}_j(x) &= \mathrm{E}_{j-1}\left\{ F_n(x) \middle| I_i \ \ \forall i \right\} \\
&= \mathrm{E}_{j-1}\left\{ \frac{1}{n}\sum_{i=1}^{n} I(_i \leq x) \middle| I_1, \ldots, I_n \right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left\{ \frac{\check{F}_{j-1}(x) - \check{F}_{j-1}(L_i)}{\check{F}_{j-1}(R_i) - \check{F}_{j-1}(L_i)} \right\} 1_i(x) + 1(x \geq R_i) \right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{j-1}\left\{ 1(X_i \leq x) \middle| I_i \right\}, \quad j = 1, 2, \ldots
\end{aligned}$$

Note Li, Watkins & Yu (1997) use the third expression for computation. Defining $K^*(u) = \int_{-\infty}^{u} K(y)dy$ and using Tonelli's theorem to interchange expectation and integration we may similarly write

$$\begin{aligned}
\hat{F}_j(x) &= \int_{-\infty}^{x} \hat{f}_j(u)\, du = \int_{-\infty}^{x} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{j-1}\left\{ K_h\left(X_i - u\right) \middle| I_i \right\} \, du \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{j-1}\left\{ K_h^*\left(X_i - x\right) \middle| I_i \right\}.
\end{aligned}$$

Since $K_h^*(X_i - x) \leq 1$ for all h, we can bring the limit inside the expectation. The result obtains since $K_h^*(u - v) \to I(u \leq v)$ when $h \downarrow 0$. •

In the case of right-censored data, the following corollary is an immediate consequence of Theorem 1.

COROLLARY 1. *If* $R_i = \infty$ *for all interval censored data points, and* $L_i = R_i = X_i$ *otherwise, then*

$$\lim_{h \downarrow 0} \hat{F}_j(x) = \tilde{F}_j(x), \ \forall x, \ j = 1, 2, \ldots,$$

Proof. The results of Efron (1967), Turnbull (1976) and Li, Watkins & Yu (1997) imply that under right censoring, $\tilde{F}_j(x) = \check{F}_j(x)$; hence the result. •

Figure 4 displays five estimates of the cumulative distribution function for the hemophiliac data. These include $\hat{F}_{j^*}$, Turnbull's $\hat{F}_t$ and three separate estimates $\check{F}_{j^*}$ based on the algorithm Li, Watkins & Yu (1997) with different initial values. Here the innermost intervals are $\{(5,6),(7,8),(8,9),(9,10),(10,11),(11,12),(12,13),(13,14),(14,15),(15,16)\}$ leaving Turnbull's estimator undefined nearly everywhere in regions of interest (except on the interval (6,7) and at the points $\{8,9,10,11,12,13,14,15,16\}$. Furthermore, the estimate based on Li, Watkins & Yu (1997) is not uniquely determined and interpolates Turnbull's estimator, resulting in a rather implausible shape. However, $\hat{F}_{j^*}$ is defined in regions of interest, is uniquely determined and has a familiar appealing shape.

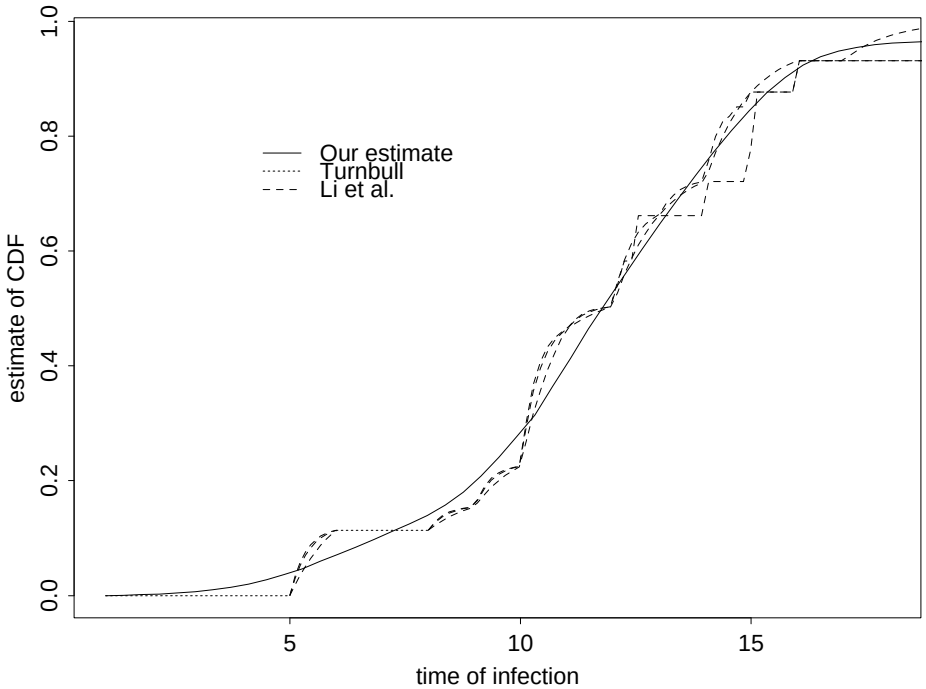

Figure 4: Five estimates of the cumulative distribution function for the hemophiliac data. These include $\hat{F}_{j^*}$, Turnbull's $\hat{F}_t$ and three separate estimates $\check{F}_{j^*}$ based on the algorithm of Li, Watkins and Yu with different initial values. While $\hat{F}_t$ is mostly undefined, and $\check{F}_{j^*}$ ill determined, $\hat{F}_{j^*}$ looks like a continuous cumulative distribution function.

2.4. An example: Gentleman & Geyer (1994).

The above theorem shows that the proposed estimator broadens the class of self-consistent algorithms by introducing kernel smoothing at each step of the iteration. We conjecture that the addition of kernel smoothing at each step of a convergent self-consistent algorithm is unlikely to introduce convergence problems. In fact, the following example suggests that smoothing improves convergence behaviour; see also Section 4.

Gentleman & Geyer (1994) consider an artificial data set where Turnbull's NPMLE $\hat{F}_t$ exists, but there are two fixed points of Turnbull's self-consistency algorithm. The data consist of six intervals $(0,1),(0,2),(0,2),(1,3),(1,3),(2,3)$ with three innermost intervals $(0,1),(1,2),(2,3)$. In this case, $\hat{F}_t$ can be shown to have weights 1/3, 1/3, 1/3. That is, 1/3, 1/3, 1/3 is a fixed point of Turnbull's algorithm, but so is 1/2, 0, 1/2. This highlights another difficulty with smoothing after a self-consistency algorithm has converged: the wrong fixed point might be smoothed to obtain the density estimate.

Interestingly, for reasonable values of h the estimator $\hat{f}_{j^*}$ seems to always result in a smoothed version of the NPMLE, even if we begin with an initial value that favours the fixed point 1/2, 0, 1/2. Consider the reduced data $(0,1),(0,2),(1,3),(2,3)$ that again has innermost intervals $(0,1)$, $(1,2)$, $(2,3)$ but where the NPMLE now has weights $1/2, 0, 1/2$. Figure 5 gives the result of our algorithm for the reduced and full data sets using the same window size in both cases. The density estimate for the reduced data has two modes because the NPMLE in this case has weights 1/2, 0, 1/2. However, the density estimate for the entire data set, where the NPMLE has weights 1/3,

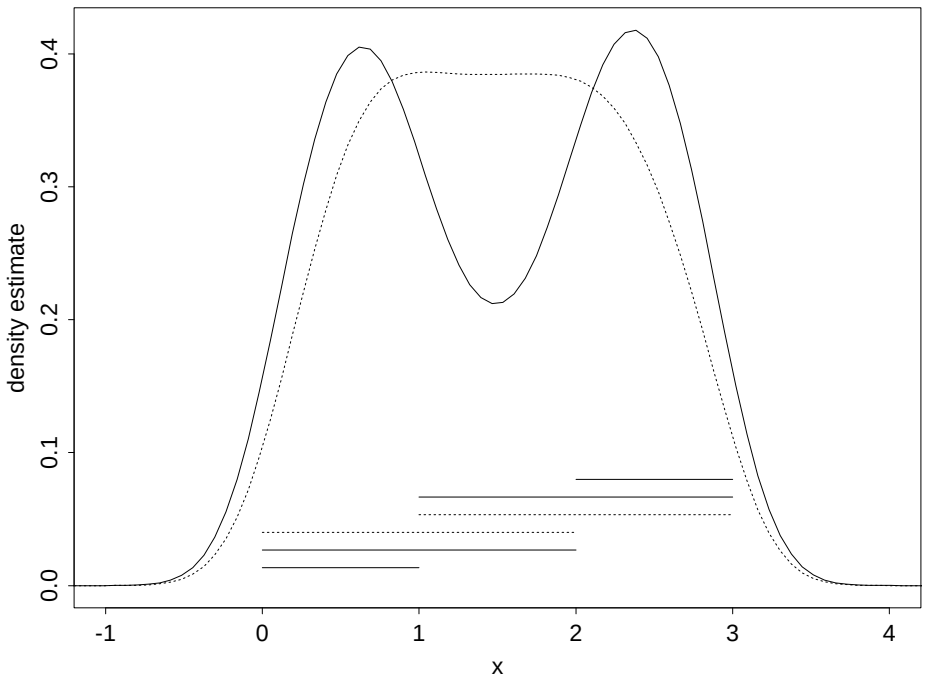

Figure 5: The estimate $\hat{f}_{j^*}$ for the Gentleman & Geyer (1994) data set and the reduced data set. The solid horizontal lines indicate the reduced data and the dotted horizontal lines represent the complement. The two curves are the density estimates for the reduced data (solid) and for the entire data set (dotted).

1/3, 1/3, has only one mode. This is true even when the density estimate for the reduced data is used as the initial value of our algorithm. That is, even though we begin with an initial value that favours the fixed point 1/2, 0, 1/2, our algorithm *still* converges to an estimate that smooths the weights for the NPMLE, namely 1/3, 1/3, 1/3.

What is going on is clear if one contrasts Turnbull's algorithm, where no smoothing occurs, with ours. The key interpretation is evident from Figure 1: smoothing permits all the data to influence the estimate at any location. Hence probability massed on the intervals $(0, 1)$ and $(2, 3)$ is smoothed repeatedly over the entire interval, where the extent to which this occurs depends on the window size h. The difficulty with Turnbull's algorithm, and subsequently Li, Watkins and Yu's, is that $h = 0$ and no smoothing takes place. Turnbull's algorithm can get stuck at local solutions, while smoothing permits our algorithm to move away from these regions. Finally there is a caveat. Theorem 1 implies that convergence of our algorithm to a unique fixed point will depend critically on h, and that as $h \downarrow 0$ our algorithm can exhibit the same difficulties as Turnbull's algorithm. These issues are addressed formally in Section 4.

3. A CLASS OF LOCAL LIKELIHOOD DENSITY ESTIMATES

In this section, we embed the fixed point of (3) in a class of local likelihood density estimates for interval censored data. Algorithm (4) is seen to be a member of a class of local EM algorithms associated with this likelihood.

Local likelihood techniques for density estimation were pioneered by Loader (1996) and Hjort & Jones (1996) in the context of completely observed data. Both propose estimating the unknown density locally at x by maximizing the criterion

$$\mathcal{L}(f,x) = \sum_{i=1}^{n} K_h(X_i - x)\log\{f(X_i)\} - n\int_{\Re} K_h(u-x)f(u)du \tag{6}$$

over some suitable class of functions. Loader (1996) suggests approximating the log density near x using polynomials, i.e.,

$$\log\{f(u)\} \quad = \quad \sum_{j=0}^{p} a_j(u-x)^j. \tag{7}$$

Upon substitution of the polynomial expansion into (6) one may estimate the coefficients $\{a_0,\ldots,a_p\}$ by maximizing to get $\{\hat{a}_0,\ldots,\hat{a}_p\}$. The density estimate is then

$$\hat{f}(x) = \exp(\hat{a}_0).$$

Loader (1996) and Hjort & Jones (1996) show that using $p = 0$ yields the usual kernel density estimator. The use of $p = 1$ and $p = 2$ yield linear and quadratic approximations that can reduce bias at the boundaries of the data, at peaks, points of inflection, and so on.

When the data are interval-censored, we propose replacing (6) with

$$\mathcal{L}(f,x) = \sum_{i=1}^{n} E[K_h(X_i - x)\log\{f(X_i)\}|I_i] - n\int_{\Re} K_h(u-x)f(u)du. \tag{8}$$

This device allows for essential use of the above machinery when we adopt an EM-type strategy. Use of the polynomial expansion renders (8) as

$$\mathcal{L}_p(f,x) = \sum_{i=1}^{n} \mathrm{E}\left\{K_h(X_i - x)\sum a_j(X_i - x)^j|I_i\right\} - n\int_{\Re} K_h(u-x)\exp\left\{\sum a_j(u-x)^j\right\}du \tag{9}$$

which is accompanied by a system of local likelihood equations for the coefficients of (7)

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{K_h(X_i - x)\frac{(X_i - x)^r}{h^r}|I_i\right\} = \int_{\Re} K_h(u-x)\frac{(X-x)^r}{h^r}\exp\left\{\sum a_j(u-x)^j\right\}du \tag{10}$$

for $r = 0,\ldots,p$. These equations retain the moment matching interpretation of Loader (1996), although here the sample moments on the left-hand side are not directly computable; the conditional expectations in (10) must be estimated by an iterative scheme. A local EM approach seems natural. It cycles through two steps at each iteration:

`E-step`: compute the relevant expectations using the current estimate of f restricted to the observed intervals;
`M-step`: solve the equations (10) to get updated estimates of $\{a_0,\ldots,a_p\}$ and f.

The algorithm differs from the typical EM algorithm, because, while expectation at the E-step is computed with respect to an estimate of the global parameter f, the equations (10) must be solved locally at each x. As such the typical arguments concerning convergence of the EM algorithm can not be brought to bear. The initial examples below show however that the local EM algorithm can lead to fixed point equations similar to (3). Convergence and efficient implementation of these are discussed in Section 4.

Example: A kernel density estimate for interval censored data. In the locally constant case, the polynomial at (7) is truncated at the leading term a_0. Solving the equations (10) at the jth step of the iteration of the local EM algorithm yields

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\hat{f}_{j-1}}\{K_h(X_i - x)|I_i\} \quad = \quad \int_{\Re} K_h(u-x)\exp(\hat{a}_0)du = \exp(\hat{a}_0) = \hat{f}_j(x).$$

Thus, the fixed point algorithm (4) is a special case of the local EM algorithm. Here the operator E has been subscripted by $\hat{f}_{j-1}$ to explicitly indicate that expectation is computed with respect to the density estimate from the previous iterate.

Example: Linear and quadratic adjustments. In the special case of a Gaussian kernel, truncating (7) at two or three terms results in explicit fixed point equations that provide a density estimate with linear and quadratic adjustments:

$$\hat{f}_{j,l}(x) = \hat{f}_j(x)\exp\left[-\frac{h^2}{2}\left\{\dot{\hat{f}}_j(x)/\hat{f}_j(x)\right\}^2\right], \tag{11}$$

$$\hat{f}_{j,q}(x) = \hat{f}_j(x)\hat{R}\exp\left[-\frac{h^2\hat{R}^2}{2}\left\{\dot{\hat{f}}_j(x)/\hat{f}_j(x)\right\}^2\right], \tag{12}$$

where $\hat{R} = (1 + h^2\hat{D})^{-1/2}$ with $\hat{D} = \ddot{\hat{f}}_j(x)/\hat{f}_j(x) - \{\dot{\hat{f}}_j(x)/\hat{f}_j(x)\}^2$. The equations resemble the exact results of Hjort & Jones (1996) except that now

$$\dot{\hat{f}}_j(x) = \frac{\partial}{\partial x}\hat{f}_j(x) = n^{-1}\sum \mathrm{E}_{\hat{f}_{j-1,r}}\left[\frac{\partial}{\partial x}\{K_h(X-x)\}|I_i\right], \tag{13}$$

$$\ddot{\hat{f}}_j(x) = \frac{\partial^2}{\partial x^2}\hat{f}_j(x) = n^{-1}\sum \mathrm{E}_{\hat{f}_{j-1,r}}\left[\frac{\partial^2}{\partial x^2}\{K_h(X-x)\}|I_i\right]. \tag{14}$$

The subscript $\hat{f}_{j-1,r}$ on the expected value operator E has $r = \ell, q$ depending on whether the linear or quadratic adjustments are used. This also applies to the expected value operator in the definition of $\hat{f}_j(x)$ for (11), (12). Establishing these results is analogous to the developments in Section 5 of Hjort & Jones (1996) and hence not given. Alternatively, they may be obtained through the use of symbolic computation where the advantage is that other explicit equations may be given for any arbitrary kernel. See Section 4.1 for details.

Example: The hemophiliac data. When applied to the hemophiliac data in Figure 6, the local linear and local quadratic estimators agree quite closely at the boundaries of the data, but differ considerably from the original, locally constant, estimator. In addition, the two adjustments themselves differ quite dramatically particularly at the peak of the estimate. Window sizes for all the estimates were determined using the cross-validation technique of Section 5. It could be argued that the quadratic adjustment is inappropriate here given that the amount of interval censoring renders the identification of such fine structure unlikely. In the next example the local quadratic estimator is viewed as more appropriate.

Example: Smoothing histogram data and the Ontario Health Survey. Histogram data consists of a set of bins B_r, with midpoints m_r, and weights p_r, $r = 1, \ldots, k$. Jones (1989) suggests smoothing histograms by computing

$$\hat{f}(x) = \sum_{r=1}^{k} \hat{p}_r K_h\left(\frac{m_r - x}{h}\right),$$

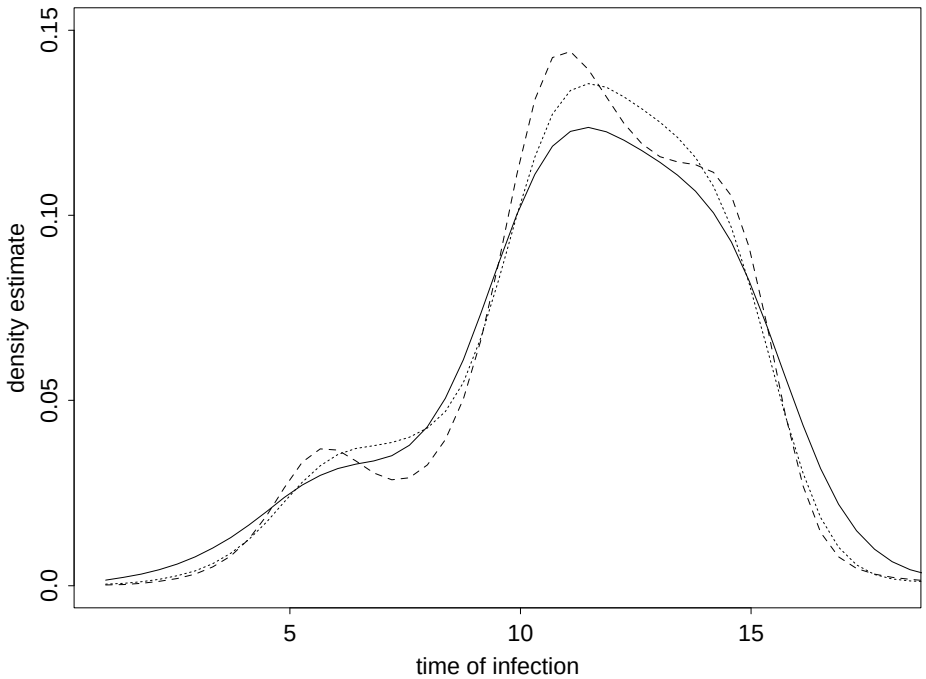

Figure 6: Density estimate of Section 2 (solid) with the local linear (dotted) and quadratic (dashed) estimators. All are members of the local likelihood class introduced in Section 3.

where, as in many statistical methods for histograms, the weights p_r are placed arbitrarily at the midpoints m_r. In addition, Jones (1989) notes that histograms are themselves binned kernel density estimates. As such, the above estimator is susceptible to multiple sources of bias since the original data have been smoothed, binned and then smoothed again.

Data that have been summarized as a histogram are interval censored since all we know about any observation X_i is to which bin, $I_i \in \{B_r; r = 1, \ldots, k\}$, it belongs. By construction, such data has an abundance of *ties*, i.e., $I_i = I_j$ whenever X_i and X_j belong to the same bin. Here the NPMLE $\hat{F}_t$ for the data $I_1, \ldots, I_n$ reproduces the histogram where the innermost intervals are the bins B_r and the probability masses are p_r. This may seem like a redundant observation, but, it should impress upon the reader that the methods of this paper may be applied directly to $I_1, \ldots, I_n$ as an alternative to Jones(1989). Not only will this remove the arbitrary use of midpoints, but concerns of bias may be addressed by local linear and local quadratic estimators.

In Figure 7 a histogram of body mass index based on the Ontario health survey is smoothed using Jones' $\hat{f}$ and various local likelihood estimators. The effect of the large bin $[33.5, 45.5]$ results in an unreasonable peak for $\hat{f}$ that does not appear in the other estimators. In addition, the linear and particularly the quadratic adjustments appear to correct bias at the peak of the smoothed histogram.

4. IMPLEMENTATION AND CONVERGENCE

We propose to implement the above local EM algorithms by computing conditional expectations using numerical integration. This approach, outlined in Section 4.1, leads to practical numerical

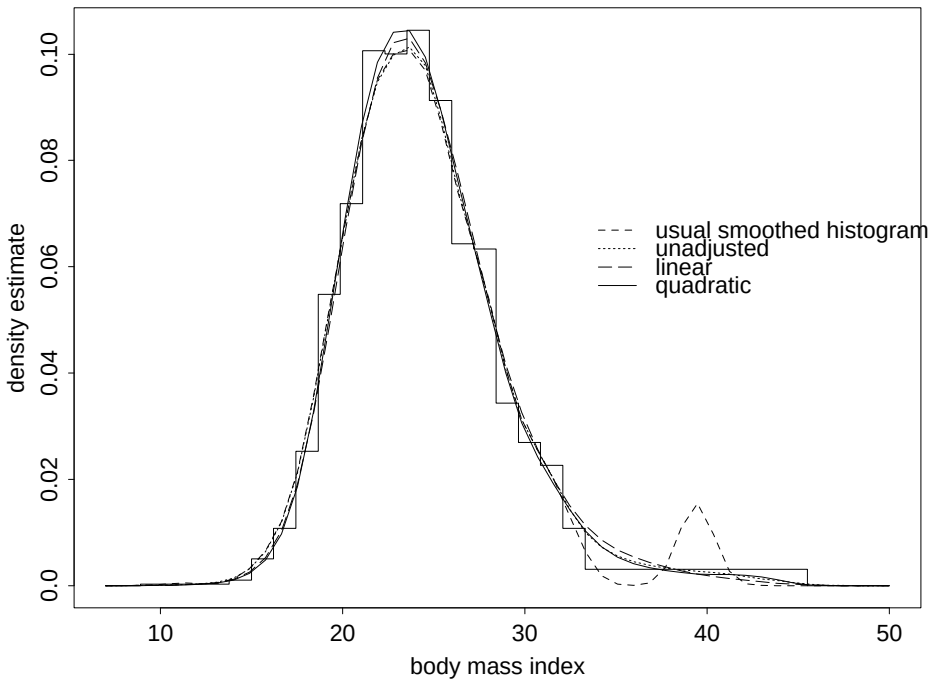

Figure 7: Histogram for body mass index, smoothed in various ways. Smoothing the histogram in the manner of Jones (1989) results in a second mode that is clearly an artifact of a large histogram bin. The local likelihood estimators eliminate this artifact and the local linear and quadratic estimators provide bias correction at the mode.

solutions, and it allows us to use fixed point theory to gain insight into convergence issues, including a proof of convergence to a unique estimate in the local constant case. These results set the stage for the development in Section 4.2 of an alternative Newton iteration which assures rapid convergence. It should be noted that establishing convergence of competing implementation strategies, like MCEM or multiple imputation, has proved to be very difficult. Furthermore, due to the quadratic convergence of Newton's method, our estimators can be calculated rapidly while the EM algorithm, for example, is notoriously slow.

Central to the Newton implementation is the need for an explicit solution for the local likelihood equations at the M-step. In cases where the local likelihood equations have a closed form solution this is straightforward; otherwise, an explicit expression for that solution can be found by symbolic Newton–Raphson (Andrews & Stafford 2000).

We conclude this section with a discussion of the use of an informal diagnostic tool which can be used to check whether a given iteration will converge. We demonstrate the usefulness of the tool with some numerical examples which also give an idea of how rapidly the iterations can converge.

4.1. Numerical integration and fixed point results.

Consider the estimator introduced in Section 2. Suppose we set out an equal-spaced mesh $\mathcal{M} = \{x^k\}_{k=1}^M$, with $\Delta = x^k - x^{k-1}$, and define $f^k = f(x^k)$ with $\mathbf{f} = (f^1, \ldots, f^M)^T$. If a trapezoidal quadrature rule is used for integration (ignoring correction at interval endpoints) then the fixed

point equation (3) becomes

$$f^k = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{\ell:x^\ell\in I_i} K_h(x^\ell - x^k) f^\ell \Delta}{\sum_{\ell:x^\ell\in I_i} f^\ell \Delta}\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{\ell:x^\ell\in I_i} K_h(x^\ell - x^k) f^\ell}{\sum_{\ell:x^\ell\in I_i} f^\ell}\right\}, \tag{15}$$

for $k = 1, \ldots, M$.

Viewing (15) as a prototype example, we may regard any local EM algorithm considered in Section 3 as an attempt to solve

$$\mathbf{f} \quad = \quad \mathcal{G}(\mathbf{f}) \tag{16}$$

by iterating the rule

$$\mathbf{f}_{j+1} = \mathcal{G}(\mathbf{f}_j) \tag{17}$$

until convergence. Here $\mathcal{G}$ depends on the choice of kernel, the bandwidth, the number of terms in the polynomial approximation and the form of integration, whether it's the trapezoidal rule or something more sophisticated. Thus, use of numerical integration gives us a general framework for dealing with the entire local likelihood class.

For iterations like (17), we may exploit a well-known fixed point theorem (Ortega 1972) that says if the image of a compact convex set D under a continuous mapping $\mathcal{G}$ lies in D, then $\mathcal{G}$ has a fixed point in D. Furthermore, a way of proving convergence of (17) to a unique fixed point is via the contraction mapping theorem (Ortega 1972, p. 152). Here $\mathcal{G}(\mathbf{f})$ is assumed to be continuously differentiable with $M \times M$ Jacobian $\nabla\mathcal{G}(\mathbf{f})$, and D is a closed, convex set. If $\mathcal{G}(\mathbf{f}) \in D$ and $0 \leq ||\nabla\mathcal{G}(\mathbf{f})|| \leq \alpha < 1$ whenever $\mathbf{f} \in D$, then $\mathcal{G}$ has a unique fixed point $\mathbf{f}_* \in D$, and for any $\mathbf{f}_0 \in D$, the iterates (17) converge to $\mathbf{f}_*$. The symbol $||.||$ denotes a vector-induced matrix norm (Ortega 1972, p. 20); we will use the infinity-norm (i.e. the maximum row-sum of the absolute values of the matrix entries).

4.1.1. Convergence of the locally constant iteration.

Let $\mathcal{G}_c$ denote the mapping whose kth component is given by the right-hand side of (15). The contraction mapping theorem allows us to prove convergence of the fixed point iteration based on $\mathcal{G}_c$ for sufficiently large bandwidths h. Thus, the theoretical behaviour of the density estimate is in agreement with what was observed in Section 2.4.

THEOREM 2 *Suppose $K(u)$ is a Hölder continuous symmetric probability density function with support in $[-1, 1]$, and such that $K(0) > 0$. Let*

$$D_h = \left\{(f^1, \ldots, f^M) : 0 \leq f^k \leq \sup_u \frac{K(u/h)}{h} \text{ and } \sum_{k:x^k\in I_i} f^k \geq \frac{K(0)}{hn} \text{ for } i = 1, \ldots, n\right\}.$$

There exists an $H > 0$ such that, for all $h > H$, $\mathcal{G}_c(\mathbf{f})$ has a unique fixed point $\mathbf{f}_*$ in D_h, and for any $\mathbf{f}_0 \in D_h$, the corresponding fixed point iteration converges to $\mathbf{f}_*$.

Proof. A routine calculation shows that, for any fixed $h > 0$, D_h is a closed and convex subset of R^M. Next, suppose $\mathbf{f} \in D_h$. Because of the nonnegativity of the kernel and of $\mathbf{f}$, it follows from the definition of $\mathcal{G}_c$ that all components of $\mathcal{G}_c(\mathbf{f})$ are nonnegative, and that

$$\mathcal{G}_c^k(\mathbf{f}) \leq \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\ell:x^\ell\in I_i}\sup_u K_h(u) f_\ell}{\sum_{\ell:x^\ell\in I_i} f_\ell} \leq \sup_u \frac{K(u/h)}{h}.$$

We may then deduce that $\mathcal{G}_c(\mathbf{f}) \in D_h$, since

$$\sum_{k:x^k \in I_i} \mathcal{G}_c^k(\mathbf{f}) \geq \frac{\frac{1}{n}\sum_{\ell:x^\ell \in I_i}\sum_{k:x^k \in I_i} f^\ell K_h(x^\ell - x^k)}{\sum_{\ell:x^\ell \in I_i} f^\ell} \geq \frac{K_h(0)}{n}.$$

We next check the Jacobian condition. Differentiating gives

$$\frac{\partial \mathcal{G}_c^k(\mathbf{f})}{\partial f^j} = \frac{1}{n}\sum_{i=1}^{n} 1(x^j \in I_i)\left[\frac{\sum_{\ell:x^\ell \in I_i} f^\ell \left\{K_h(x^j - x^k) - K_h(x^\ell - x^k)\right\}}{\left(\sum_{\ell:x^\ell \in I_i} f^\ell\right)^2}\right]. \tag{18}$$

For $\mathbf{f} \in D_h$, a crude upper bound for the absolute value of this is

$$\frac{(M+1)^\gamma \Delta^\gamma L}{h^{\gamma+1} n}\sum_{i=1}^{n} 1(x^j \in I_i)\frac{1}{\sum_{\ell:x^\ell \in I_i} f^\ell} \leq \frac{(M+1)^\gamma \Delta^\gamma L n}{h^\gamma K(0)},$$

where L and $\gamma \in (0,1)$ arise from the Hölder continuity condition on K:

$$|K(x) - K(y)| \leq L|x-y|^\gamma.$$

The upper bound can be made arbitrarily small by taking h large enough. Thus, there exists a bandwidth H such that $\alpha = ||\nabla \mathcal{G}_c(\mathbf{f})|| < 1$, for all $\mathbf{f} \in D_H$. Taking h to be any bandwidth h exceeding H, we can thus assure convergence of a fixed point iteration based on $\mathcal{G}_c$ to a unique vector in D_h. •

Note 2. The proof can be extended to kernels with noncompact support such as the Gaussian kernel.

Note 3. The facts that $\mathcal{G}_c(\mathbf{f}) \in D_h$ when $\mathbf{f} \in D_h$ for any $h > 0$, and that $\mathcal{G}_c$ is continuous are sufficient to assure existence of (though not uniqueness of or convergence to) a fixed point of $\mathcal{G}_c$. This follows from the fixed point theorem of Ortega (1972) quoted earlier.

Note 4. A practitioner might be concerned that a sufficiently large bandwidth for convergence may be larger than a bandwidth that meets a particular optimality requirement such as that given by cross-validation; our numerical work suggests that such optimal bandwidths are usually large enough to guarantee convergence.

Note 5. When the iteration converges, the result is a vector which approximates a probability density as the following argument demonstrates. In particular, we will show that if there is a solution to (15) in D_h, then that solution approximates a probability density, when the kernel has a bounded derivative. Nonnegativity of f^k is immediate. Properties of Riemann integration and the kernel imply that

$$\begin{aligned}\sum_{k=1}^{M} f^k \Delta &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\ell:x^\ell \in I_i} f^\ell \sum_{k=1}^{M} K_h(x^\ell - x^k)\Delta}{\sum_{\ell:x^\ell \in I_i} f^\ell} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\ell:x^\ell \in I_i} f^\ell \int K_h(x^\ell - y)dy}{\sum_{\ell:x^\ell \in I_i} f^\ell} + O(M^{-1}) = 1 + O(M^{-1}).\end{aligned}$$

4.1.2. The linear and quadratic adjustments.

Moving away from the locally constant case complicates the details considerably. Concerning ourselves only with the case of a Gaussian kernel where closed form solutions exist, we give some indication of these complications and how one might proceed. Extension to arbitrary kernels through the use of symbolic computation are considered in Section 4.3.

Based on a trapezoidal quadrature approximation to the linear adjustment at (11), we define the continuous mapping $\mathcal{G}_l$ as

$$\mathcal{G}_l(\mathbf{f}) = \mathcal{G}_c(\mathbf{f}) \exp\left[-\frac{h^2}{2}\left\{\dot{\mathcal{G}}_c(\mathbf{f})/\mathcal{G}_c(\mathbf{f})\right\}^2\right]$$

where the operations multiplication, exponentiation, and so on, are componentwise. Occasionally, one or more of the components of $\mathcal{G}_c(\mathbf{f})$ are 0, in which case the corresponding component of $\mathcal{G}_l(\mathbf{f})$ will be defined as 0. Here, $\dot{\mathcal{G}}_c$ is defined via a trapezoidal version of the right-hand side of (13). Now if we define

$$D = \left\{(f^1, f^2, \ldots, f^M) : 0 \leq f^k \leq \sup_u \frac{K(u/h)}{h}\right\},$$

and note that

$$\mathcal{G}_l^k(\mathbf{f}) \leq \mathcal{G}_c^k(\mathbf{f}), \tag{19}$$

then, given all components of $\mathcal{G}_l$ are nonnegative, we have $\mathcal{G}_l(\mathbf{f}) \in D$; hence $\mathcal{G}_l$ satisfies the conditions for the existence of a fixed point in D. Note that because of (19), the linear adjustment will require normalization, upon convergence, in order to approximate a proper probability density.

Unfortunately, it is not possible to verify the derivative condition for convergence to a unique fixed point of $\mathcal{G}_l$ using either D or the set D_h defined in Theorem 2. While it seems likely that the condition may be verified for an appropriate domain, the matter is not pursued here other than to conjecture that the requirement on h for such convergence will be more stringent (i.e., larger bandwidths will be required) than in the locally constant case. We may also define a mapping $\mathcal{G}_q$ for the local quadratic case, using trapezoidal versions of (12), (13) and (14). However, it is not possible to prove that $\mathcal{G}_q$ has a fixed point, even in D, since the $\hat{R}$ factor involved in its definition may take on negative values. Thus, it is possible for $\mathcal{G}_q$ to map elements of D to vectors having negative components.

In the absence of a proof of global convergence, local convergence should be considered. An iteration is said to converge locally to a fixed point $\mathbf{f}_*$, if such convergence occurs whenever the initial guess is close enough to $\mathbf{f}_*$. Ostrowski's Theorem (Ortega 1972) states that if the iteration function is differentiable in a neighbourhood of a fixed point $\mathbf{f}_*$, and the spectral radius (maximum absolute eigenvalue) of the Jacobian, $\rho(\nabla\mathcal{G}(\mathbf{f}))$ is less than 1 in a neighbourhood of $\mathbf{f}_*$, then the fixed point iteration converges locally to $\mathbf{f}_*$.

Local convergence for the iterations based on $\mathcal{G}_l$ and $\mathcal{G}_q$, for large enough h, follows from a verification of the spectral radius condition. Because $\rho(\nabla\mathcal{G}(\mathbf{f})) \leq ||\nabla\mathcal{G}(\mathbf{f})||$, for any vector-induced matrix norm, we could proceed by similar arguments to those used to prove global convergence in the locally constant case. However, it is well known that any fixed point iteration, with a smooth enough iteration function, can be rendered locally convergent upon conversion to a Newton iteration. Since the iteration functions under consideration are smooth, we will see, in the next subsection, that conversion to a Newton iteration is a practical way of assuring local convergence. It should be noted that a Newton implementation assures local convergence for any $h > 0$ in the constant and linear cases.

4.2. Conversion to a Newton iteration.

We can improve upon any fixed point iteration proposed in Section 3 by re-expressing it as a Newton iteration. For example, in the locally constant case the result is a stable, quadratically

convergent algorithm provided that the initial guess is close enough to the solution. In addition, theoretical and numerical considerations suggest that a Newton iteration also works well in the local linear and quadratic cases, provided minor adjustments are made. In particular, conversion of the fixed point scheme for the linear adjustment to a Newton iteration assures local convergence, since we have shown that a fixed point of $\mathcal{G}_l$ exists in a convex domain.

Letting

$$\mathcal{U}(\mathbf{f}) \quad = \quad \mathbf{f} - \mathcal{G}(\mathbf{f}) \tag{20}$$

the goal is to compute the fixed point of (16) by solving $\mathcal{U}(\widehat{\mathbf{f}}) = 0$. Differentiating (20) with respect to $\mathbf{f}$ gives

$$\nabla\mathcal{U}(\mathbf{f}) = I - \nabla\mathcal{G}(\mathbf{f}),$$

and a Newton iteration can be constructed from

$$\mathbf{f}_{j+1} = \mathbf{f}_j - [\nabla\mathcal{U}(\mathbf{f}_j)]^{-1}\,\mathcal{U}(\mathbf{f}_j). \tag{21}$$

For example, for $\mathcal{G}_c$ we may facilitate the Newton iteration without difficulty through use of the gradient $\nabla\mathcal{G}$ with (j,k) component being simply

$$(\nabla\mathcal{G})_{jk} = \frac{\partial}{\partial f^k} f^j = \frac{1}{n} \sum_{\{i:x^k\in I_i, 1\le i\le n\}} \left\{ \frac{K_h(x^j - x^k)}{\sum_{\ell:x^\ell\in I_i} f^\ell} - \frac{\sum_{\ell=1}^m K_h(x^j - x^\ell) f^\ell}{\left(\sum_{\ell:x^\ell\in I_i} f^\ell\right)^2} \right\}.$$

Similar expressions can be found for $\mathcal{G}_l$ and $\mathcal{G}_q$.

4.3. Newton iteration through symbolic computation.

Use of Newton iteration depends on the computation of the gradient $\nabla\mathcal{G}$, which in turn requires an explicit solution of the local likelihood equations. Immediate examples are given by $\mathcal{G}_c, \mathcal{G}_l$ and $\mathcal{G}_q$, but the method may be applied to the entire local likelihood class through the use of symbolic computation (Andrews & Stafford 2000). Symbolic Newton–Raphson permits the solution $\hat{a}_0$, and hence $\exp(\hat{a}_0)$ to be written explicitly in terms of the expected value operator $\mathrm{E}_{\hat{f}}$ without ever having to evaluate $\mathrm{E}_{\hat{f}}$. This expression then becomes the basis for the Newton iteration. In other words, for any local EM algorithm we only have to solve the local likelihood equations at the M-step once throughout, and render iteration on the remaining E-step only. The result is a general fixed point equation similar to $\mathcal{G}_c, \mathcal{G}_l, \mathcal{G}_q$ that is suitable for Newton implementation. Full details are given in a *Mathematica* notebook which is available upon request. However, the result of the linear case is summarized here. Letting μ_r denote the rth moment of the kernel K, then in general the fixed point iteration for the local linear estimator can be expressed explicitly as

$$\hat{f}_{j,l}(x) = \hat{f}_j(x) \exp\left\{ -\frac{1}{2} \left[\left\{ \frac{\hat{\dot{f}}_j}{\hat{f}_j(x)} \right\}^2 h^2\mu_2 + \left\{ \frac{\hat{\dot{f}}_j}{4\hat{f}_j(x)} \right\}^4 h^4(3\mu_2 - \mu_4) \right] \right\}$$

where

$$\hat{\dot{f}}_j = \frac{1}{n} \sum_{i=1}^n \mathrm{E}_{\hat{f}_{j,\ell}} \left\{ K_h(X - x)(X - x) | I_i \right\}.$$

Note for the case of a Gaussian kernel $\hat{\dot{f}}_j = \dot{\hat{f}}_j$, $3\mu_2 - \mu_4 = 0$ and the above expression simplifies to that given in Section 3. Also, in general, if one ignores the second term in the exponent, the linear adjustment has a remarkably simple form that greatly resembles the closed form solution in the Gaussian case. In addition, the second term in the exponent may be interpreted in terms of a

class of implicit kernels that share cumulants with known kernels but for which there is a closed form solution at the `M-step`. Issues of this type are deferred for a more general treatment of our Newton iteration.

4.4. A convergence diagnostic based on the spectral radius.

In cases where global convergence has not been established, as in the local linear or local quadratic cases, the spectral radius of the Jacobian of the iteration function provides an informal way of checking whether an iteration will converge. In Section 4.1.2, we noted that Ostrowski's theorem assured local convergence of an iterative scheme based on $\mathcal{G}$ if $\rho\{\nabla\mathcal{G}(\mathbf{f}_*)\} < 1$ at or near a fixed point $\mathbf{f}_*$. Since the iteration functions $\mathcal{G}$ that we are dealing with are smooth, it is not unreasonable to conjecture that if $\rho\{\nabla\mathcal{G}(\mathbf{f}_0)\} < 1$, then convergence will ensue. Moreover, if this spectral radius condition holds for a succession of iterates, our confidence that the iteration will converge should increase.

This convergence diagnostic is useful in additional ways. First, if $\rho\{\nabla\mathcal{G}_c(\mathbf{f}_0)\} < 1$, then we have some assurance that the bandwidth has been chosen large enough for the theory of Section 4.1.1 to apply, yielding global convergence. Second, if an iterative scheme based on $\mathcal{G}$ has been converted to a Newton iteration, the condition $\rho(\nabla\mathcal{G}\{\mathbf{f}_0)\} < 1$ provides a check as to whether the initial guess is close enough to the solution for convergence of the Newton scheme to occur.

This last observation suggests a general strategy for obtaining an iterative scheme which will converge rapidly. One applies a convergent fixed point iteration of the form (16) until its spectral radius is less than one and uses the resulting estimate as the initial value for the Newton iteration. This is analogous to the use of so-called "Spacer Steps" in nonlinear programming (see Luenberger 1973), where a Newton–Raphson iteration is interspersed with occasional steps of another method, which is known to converge, though more slowly. The theory of Section 4.1 ensures that this will work for locally constant density estimation provided a sufficiently large bandwidth is used.

The rest of this subsection is devoted to a description of empirical results that we have obtained for specific data sets.

4.4.1. Applying the diagnostic to the original iteration.

We first applied the iteration based on $\mathcal{G}_c$ to the hemophiliac data (excluding right-censored observations), starting with a uniform initial density, $m = 200$ and various values of h. The first five columns of Table 1 list the bandwidths, the spectral radii at the 1st, 2nd and final steps of the original iteration, together with the respective numbers of steps required for the iteration to converge. Convergence was declared when successive approximations differed by less than 10^{-7} in the Euclidean norm. Note that such convergence occurs for all values of h considered, and the number of required iteration steps decreases with increasing h. When $h = .3$ and $h = .4$, the initial spectral radius exceeds 1, but convergence occurs anyway. We also note that if h is very small, then the spectral radius remains above 1 for all steps of the iteration.

Applying the spectral radius diagnostic to the linear adjustments $\mathcal{G}_l$ gives similar results. As can be seen in Table 2, decreasing the bandwidth is associated with increased initial and terminal spectral radii and with decreasing speed of convergence. Again, we encountered no convergence failures when the spectral radius remains below 1, i.e., when the bandwidth is reasonably large. This leads us to conjecture that an analogue of Theorem 2 holds for the linear adjustment.

We have noted already that it is possible for the iteration based on $\mathcal{G}_q$ to hit negative values of the estimate, causing convergence difficulties, especially for small values of h. This kind of problem can be eliminated by imposing a simple constraint at each step of the iteration. Nonnegativity is enforced by setting to 0 any components of $\mathbf{f}_j$ that may have fallen below 0 at the most recent iterate. With this simple adjustment, we observe similar convergence and diagnostic behaviour in Table 3 for $\mathcal{G}_q$ as we saw earlier for $\mathcal{G}_l$ and $\mathcal{G}_c$.

h	ρ_1	ρ_2	ρ_∞	Original	Newton	Hybrid
.01	7.74	1.26	1.0	125	fails	fails
.3	1.475	.795	.800	53	fails	4
.4	1.180	.699	.716	41	50	3
.5	.981	.614	.633	30	3	3
.6	.878	.537	.588	26	3	3
.7	.777	.481	.551	24	3	3
1	.564	.364	.433	19	3	3
1.2	.488	.325	.368	17	3	3

Table 1: Spectral radius values, ρ_1, ρ_2, ρ_∞, for the iteration based on $\mathcal{G}_c$ applied to the hemophiliac data (without right-censored data) at the first, second and final steps. The last three columns list the number of steps to convergence to within a tolerance of 10^{-7} for each method. (The hybrid method consists of a single step of the original iteration before application of the Newton iteration.)

h	ρ_1	ρ_2	ρ_∞	Original	Newton	Hybrid
.3	1.502	.781	.858	53	fails	13
.4	1.242	.700	.796	51	fails	8
.5	1.044	.629	.756	47	4	4
.6	.926	.560	.730	47	fails	4
.7	.809	.506	.713	43	4	4
.8	.701	.443	.692	38	4	4
1	.580	.388	.659	31	4	4
1.2	.501	.348	.607	28	4	4

Table 2: Spectral radius values, ρ_1, ρ_2, ρ_∞, for the iteration based on $\mathcal{G}_l$ applied to the hemophiliac data (without right-censored data) at the first, second and final steps. The last three columns list the number of steps to convergence to within a tolerance of 10^{-7} for each method. For $h = 0.3$ and $h = 0.6$, the positivity constraint had to be enforced, even in the hybrid case.

h	ρ_1	ρ_2	ρ_∞	Original	Newton	Hybrid
.3	2.280	1.070	.942	229	fails	95
.4	1.970	.943	.891	114	18	14
.5	1.660	.885	.857	110	fails	6
.6	1.500	.815	.832	84	fails	6
.7	1.375	.752	.819	77	fails	6
.8	1.239	.687	.808	72	fails	6
.9	1.162	.661	.773	62	6	6
1.0	1.084	.554	.747	56	6	6
1.1	.975	.500	.736	53	fails	5
1.2	.921	.465	.732	50	7	5
1.3	.868	.443	.732	47	8	5
1.4	.851	.419	.734	46	5	5

Table 3: Spectral radius values for, ρ_1, ρ_2, ρ_∞, the iteration based on $\mathcal{G}_q$ applied to the hemophiliac data (without right-censored data) at the first, second and final steps. (Nonnegativity was enforced in all cases). The last three columns list the number of steps to convergence to within a tolerance of 10^{-7} for each method. For $h = 0.3$, only 38 steps were required for convergence if the Newton iteration was applied after two steps of the original iteration.

When applied to the Gentleman & Geyer (1994) data considered in Section 2.4, the original iteration converges to a smoothed version of the nonparametric maximum likelihood estimate for all values of h tried (i.e., $h = 0.2$, 0.3, 0.4, 0.5). The spectral radius diagnostic is less than 1 for all iterates in all cases.

We have tested the spectral radius diagnostic on several other artificial data sets, and we have found that the spectral radius is usually less than 1 at reasonable values of the bandwidth parameter h. The initial and terminal values of the spectral radius both tend to increase with decreasing h. The speed of convergence tends to decrease as the spectral radius increases. These observations are consonant with the theory set out in Section 4.1 as well as the assertions made in Section 2.4.

We have encountered several examples where the spectral radius is initially above 1 before dropping below 1, where it remains until convergence is apparently achieved. These examples typically involved very small bandwidths. We encountered no examples of non-convergence when the spectral radius is eventually less than 1.

4.4.2. Applying the diagnostic to Newton's iteration.

The final two columns of Tables 1, 2 and 3 give the number of steps required for convergence of the Newton iteration applied respectively to $\mathcal{G}_c$, $\mathcal{G}_l$ and $\mathcal{G}_q$. When the initial spectral radius of $\nabla\mathcal{G}_c$ is less than 1, the associated Newton iteration usually converges. When the initial spectral radius exceeds 1, convergence behaviour is less predictable. Sometimes there is reasonably rapid convergence (i.e., 3 or 4 steps); otherwise, the Newton iteration fails to converge. However, in the latter cases, it is usually sufficient to apply a single step of the original scheme before applying the Newton iteration. Rapid convergence ensues almost every time.

In addition, we note that when Newton's method is applied to $\mathcal{G}_c$ with $h = .01$, it apparently converges as well, but to a slightly different solution. The spectral radius at this other solution was 1.42.

Non-uniqueness has arisen, because the bandwidth is not large enough, and the Newton iteration is converging locally, but to the wrong fixed point. Because $||\nabla\mathcal{G}(\mathbf{f})|| > \rho\{\nabla\mathcal{G}(\mathbf{f})\}$, the conditions of Theorem 2 are not met. The failures of the Newton iteration at low values of h point out the potential usefulness of our diagnostic, since in all such cases, the diagnostic exceeds 1 at the first step.

Similar results can be obtained for the linear adjustment $\mathcal{G}_l$ (with normal kernel) as for the constant case. The only difficulty to note here is that when the initial spectral radius is above 1, it is possible for the Newton iteration to hit negative values of the estimate, causing convergence difficulties. Again, these difficulties can be alleviated by imposing the simple nonnegativity constraint at each step of the iteration.

Applying Newton's method to the quadratic adjustment $\mathcal{G}_q$ requires the imposition of the nonnegativity constraint even when the initial spectral radius of the original iteration is less than 1. Without this constraint, Newton's method almost always fails to converge. This perhaps explains why our theory fails to provide a sufficient condition for convergence in this case. The imposition of nonnegativity (as described in Section 4.4.1) tends to fix the convergence problem, so that the convergence behaviour is once again similar to that for the constant case.

When applied to the Gentleman & Geyer (1994) data, the Newton iteration also converges in all cases tried; however, for $h = 0.2$ and $h = 0.3$, it converges to the incorrect self-consistent estimate. (As usual, we are starting with a uniform initial guess.) In both of these cases, the initial values of the spectral radius exceed 1.0. Again, the conditions of Theorem 2 for uniqueness of the fixed point are not met. Interestingly, the original scheme converged properly, suggesting that the original scheme is more stable than our theory suggests.

In the cases of convergence of the Newton iteration to the correct solution, the spectral radius is always less than 1.0. If started with 1 step of the original iteration, the Newton iteration converges to the correct solution in 2 steps.

5. CHOICE OF SMOOTHING PARAMETER

A central component of kernel density estimation is the choice of the window size h or the smoothing parameter. We propose a method based on likelihood cross-validation and analogous to the case where the data are completely observed (Silverman 1986). For the latter, cross-validation aims to maximize

$$CV(h) = \prod_{i=1}^{n} \hat{f}_h^{(-i)}(X_i) \tag{22}$$

with respect to h. It acts as a surrogate for the Kullback–Leibler distance between f and $\hat{f}_h$ on the basis of its expectation

$$\mathrm{E}\{CV(h)\} \approx -\int f(t)\log\{f(t)/\hat{f}(t)\}dt + \int f(t)\log\{f(t)\}dt.$$

The superscript $(-i)$ in (22) indicates that $\hat{f}_h^{(-i)}(X_i)$ is obtained by eliminating a point of support, X_i, from the NPMLE, F_n, and using only the remaining data. Here the estimator is explicitly subscripted by h and the notation $\hat{f}_h$ can represent any density estimate proposed in this paper.

An analogy in the case of interval censored data leads to eliminating the points of support for $\hat{F}_t$, namely the innermost intervals. Denoting an innermost interval as J_r, $r = 1, \ldots, m$, where we suppose there are m such intervals given by $\hat{F}_t$, we define the cross-validated likelihood as

$$CV(h) = \prod_{r=1}^{m} \int_{J_r} \hat{f}_h^{(-r)}(t)dt,$$

where $\int_{J_r} \hat{f}_h^{(-r)}(t)dt$ is obtained by dropping the innermost interval J_r when estimating the density. Dropping an innermost interval is accomplished by removing all intervals in the original sample that contribute to its presence but not to the presence of any other innermost interval. Often this is not possible and the elimination of one interval leads to the elimination of others so that ultimately the method resembles a form of k-fold cross-validation. However, the method conveniently addresses the question of tied observations which are common for interval censored data (Figures 2 and 6). For example, the hemophiliac data contains only 40 distinct intervals in a sample of size 105. In addition it also handles two observed intervals that are not tied but have a high degree of overlap. If they both contain the same innermost interval then they are both eliminated from the cross-validation process when that innermost interval is dropped.

The scheme worked well in a limited simulation study using 40 samples of size 20 with event times from a Weibull distribution and interval censoring determined by an independent homogeneous Poisson process. Table 4 compares average values of the Kullback–Leibler distance with our method of likelihood cross-validation where the latter picks a value of the smoothing parameter that is close to the value which optimizes the Kullback-Leibler distance. We have used this method of cross-validation in the paper without comparing it to alternatives like k-fold cross-validation. A more thorough investigation of the scheme is needed, but beyond the scope of this paper.

6. DISCUSSION

There is a rich literature on smooth inference methods for interval-censored data. Examples include Rosenberg (1995), Joly & Commenges (1999), Kooperberg & Stone (1992), Tanner & Wong (1987) and papers by Betensky and co-authors. A survey of statistical methods for interval-censored data is given by Lindsay & Ryan (1998). Often data augmentation algorithms (EM and multiple imputation, respectively) are used for smooth inference methods for interval-censored data and little about convergence has been formally developed. The approach taken in this paper, that of

h	0.25	0.42	0.50	0.58	0.67	0.75	0.83	0.92	1.08
KL	0.1769	0.1309	0.1234	0.1149	**0.1093**	0.1131	0.1146	0.1190	0.1358
CV	−3.845	−3.807	−3.789	−3.779	−3.770	−3.765	**−3.764**	−3.768	−3.779

Table 4: Kullback–Leibler distance and the logarithm of the proposed cross-validated likelihood for the unadjusted density estimate. The cross-validated likelihood is maximized at a value of the window size that is reasonably close to the value that minimizes the Kullback–Leibler distance.

recasting a local EM algorithm as a Newton iteration, has permitted some formal developments concerning convergence. Its applicability to other local likelihood methods will be pursued in subsequent work.

ACKNOWLEDGEMENTS

We are grateful to David Andrews, Jerry Lawless, Derick Peterson and Rob Tibshirani for useful discussions. We are particularly grateful to Anthony Davison for very carefully proof reading a revision of this paper. We would also like to thank Richard Lockhart and the Natural Sciences and Engineering Research Council of Canada for supporting this research through individual operating grants.

REFERENCES

D. F. Andrews & J. E. Stafford (2000). *Symbolic Computation for Statistical Inference.* Oxford University Press, Oxford, United Kingdom.

D. R. Bellhouse & J. E. Stafford (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407–424.

R. A. Betensky, J. C. Lindsey, L. M. Ryan & W. P. Wand (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics*, 55, 238–245.

A. C. Davison & D. V. Hinkley (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge, United Kingdom.

V. De Gruttola & S. W. Lagakos (1989). Analysis of doubly-censored survival data, with applications to AIDS. *Biometrics*, 45, 1–11.

B. Efron (1967). The two sample problem with censored data. In *Fourth Berkeley Symposium on Mathematical Statistics*, University of California Press, Berkeley, CA, pp. 831–853.

R. Gentleman & C. J. Geyer (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika*, 81, 618–623.

C. Goutis (1997). Nonparametric estimation of a mixing density via the kernel method. *Journal of the American Statistical Association*, 92, 1445–1450.

N. L. Hjort & M. C. Jones (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24, 1619–1647.

P. Joly & D. Commenges (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. *Biometrics*, 55, 887–890.

M. C. Jones (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84, 733–741.

C. Kooperberg & C. J. Stone (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1, 301–328.

L. Li, T. Watkins & Q. Yu (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, 24, 531–542.

J. C. Lindsey & L. M. Ryan (1998). Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine*, 17, 219–238.

C. R. Loader (1996). Local likelihood density estimation. *The Annals of Statistics*, 24, 1602–1618.

D. G. Luenberger (1973). *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, Reading, MA.

J. M. Ortega (1972). *Numerical Analysis: A Second Course.* Academic Press, New York.

W. Pan (2000). Smooth estimation of the survival function for interval censored data. *Statistics in Medicine*, 19, 2611–2624.

P. S. Rosenberg (1995). Hazard function estimation using B-splines. *Biometrika*, 51, 874–887.

B. Silverman (1986). *Density Estimation for Statistics and Data Analysis*, London, Chapman-Hall.

M. A. Tanner & W. H. Wong (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, 29, 23–32.

B. M. Turnbull (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, *Series B*, 38, 290–295.

M. P. Wand & M. C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

Received 29 June 2003
Accepted 31 July 2004

W. John BRAUN: braun@stats.uwo.ca
Department of Statistical and Actuarial Sciences
University of Western Ontario, London, Ontario, Canada N6A 5B7

Thierry DUCHESNE: duchesne@mat.ulaval.ca
Département de mathématiques et de statistique, Université Laval
Québec, Canada G1K 7P4

James E. STAFFORD: stafford@utstat.toronto.edu
Department of Public Health Sciences, University of Toronto
Toronto, Ontario, Canada M5S 1A8