A kernel density estimate for interval censored data

Thierry Duchesne* and James E. Stafford†*

Abstract

In this paper we propose a kernel density estimate for interval-censored data. It retains the simplicity and intuitive appeal of the usual kernel density estimate and is easy to compute. The estimate results from an algorithm where conditional expectations of a kernel are computed at each iteration. These conditional expectations are computed with respect to the density estimate from the previous iteration, allowing the estimator to extract more information from the data at each step. The estimator is applied to HIV data where interval censoring is common.

In terms of the cumulative distribution function the algorithm is shown to coincide with those of Efron (1967), Turnbull (1976), and Li et al. (1997), as the window size of the kernel shrinks to zero. Viewing the iterative scheme as a generalized EM algorithm permits a natural interpretation of the estimator as being close to the ideal kernel density estimate where the data is not censored in any way. Simulation results support the conjecture that kernel smoothing at every iteration does not effect convergence. In addition, comparison to the standard kernel density estimate, based on smoothing Turnbull's estimator, reflect favourably on the estimator for all criteria considered. Use of the estimator for scatterplot smoothing is considered in a final example.

Keywords: Cross-validation, EM algorithm, HIV, importance sampling, interval censoring, kernel smoothing, Kullbeck-Leibler, mean squared error, Monte Carlo integration, nonparametric maximum likelihood, scatterplot smoothing, self-consistency.

1 Introduction

We propose a kernel density estimate to be used in the presence of interval censored data, i.e. data that are observed to lie within an interval but whose exact value is unknown. The

*Department of Statistics, University of Toronto

†Department of Public Health Sciences, University of Toronto

estimate results from a recursive scheme that generalizes the algorithms of Efron (1967), Turnbull (1976) and Li et al. (1997) by kernel smoothing the data at each iteration

$$\hat{f}_j(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{j-1}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\middle| X \in I_i\right]. \tag{1}$$

Here expectation is with respect to the previous iterate conditional on the observed interval. Convergence of the algorithm implies that $\hat{f}_j$ approaches some density for which the application of (1) has no effect. Efron (1967) called such a fixed point a self-consistent estimator.

The estimator retains the simplicity and intuitive appeal of a kernel density estimate. In fact, this simplicity avoids some of the awkward aspects associated with kernel smoothing Turnbull's estimator, F_t, of the cumulative distribution function (cdf)

$$\hat{f}_t(x) = \int_{\Re} \frac{1}{h}K\left(\frac{x-u}{h}\right) dF_t(u)$$

which is a standard technique. Turnbull's F_t is a non-parametric maximum likelihood estimator (NPMLE) that is not uniquely defined over the whole real line but only up to an equivalence class of distributions that may differ over gaps called "innermost" intervals. Associated with these gaps are probability masses whose distribution over the gap is left unspecified and that proves to be troublesome when computing $\hat{f}_t$. Pan (2000) suggests *arbitrarily* assuming that jumps occur at the right-hand points of the gaps which may be appropriate if the censoring proportion and the length of the censoring intervals are small. However, if most observations are interval censored with interval lengths that can be large, as is often the case with HIV/AIDS data, then assuming that the jumps occur at the right-hand point of the interval may cause considerable bias in the estimator. This complication never arises when computing (1) because we smooth the data directly at every iteration rather that smoothing a NPMLE once. Moreover, this smoothing process distributes probability mass over each observed interval using a conditional density determined by the previous iterate. This process is data driven rather than *arbitrary.*

Figure 1 depicts use of the estimator as applied to a group of heavily treated hemophiliacs (De Gruttola and Lagakos, 1989) whose time of infection with the HIV virus was interval censored. 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 $\hat{f}_t$ and our estimator $\hat{f}_4$ based on four iterations of the algorithm. The choice of $j = 4$ is based on both simulations and visual inspection of the estimator for several values of $j > 4$. The latter can be made common practice as successive iterates are based on an importance sampling scheme where the time to compute an iterate does not increase with the number of iterations. The estimator $\hat{f}_t$ was computed assuming jumps occur at the center of innermost intervals rather than the right-hand point, which causes the estimate to be shifted to the right. Window sizes were chosen using a method of cross-validation discussed in §5. What is evident from the plot is that $\hat{f}_4$ does a better job of smoothing what appears to be a sampling anomaly on the left side of the plot without eroding the peak on the right. It eliminates sampling artifacts in the estimate without degrading the estimate itself, and to some extent overcomes the fact that smoothing the NPMLE does not recover the information lost by the non-parametric estimation (Pan, 2000). By smoothing at every iteration it does a better job of borrowing information from neighbouring data points in the smoothing process. This is borne out in simulations of mean squared error. This example is used throughout the paper to illustrate other aspects of the estimator and a separate example concerning HIV infection and infant mortality is given in §7.

Innermost intervals, whose concept is not entirely straightforward, never explicitly enter into the calculation resulting in the advantage that our estimator fills in the gaps of Turnbull's F_t. This idea of filling in the gaps is not new as Li *et al.* (1997) embed Turnbull's NPMLE in an EM algorithm designed specifically for this purpose. They obtain an estimator that will converge to the NPMLE where the NPMLE is uniquely defined, and to some cdf that depends on the starting point of the algorithm where the NPMLE has gaps. In §3 we show that as the window size, h, shrinks to zero our algorithm coincides with that of Li et al. (1997) and hence with the algorithms of Efron (1967) and Turnbull (1976) as well.

The remainder of this paper is organized as follows. In §2 the estimator is proposed as a natural extension of the usual kernel density estimate in the complete data case (no censoring). It is formally defined through a generalized EM algorithm where the "M" step is characterised by optimizing an "MSE" criterion. This criterion is quite natural as it involves the complete data kernel density estimate, $\hat{f}_c$, allowing the estimator to be interpreted as minimizing the distance between itself and the ideal estimate $\hat{f}_c$. Numerical implementation of the method is discussed in §4 and the choice of the smoothing parameter is considered in §5. The question of convergence of the algorithm is considered in §3. Although the developments are not rigorous, the conjecture is that use of kernel smoothing at every iteration does not perturb algorithms, that are known converge, to such an extent that they no longer converge. The argument is supported by simulation results in §6. Finally, in §7 the method is used to provide kernel weights for scatterplot smoothing. Throughout the paper analogies with the complete data case make developments transparent.

2 Definition of the estimator

In the presence of complete data $X_1, \ldots, X_n$ the standard kernel density estimate,

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

may be written as an expectation with respect to the empirical distribution, F_n, of the sample

$$\hat{f}_c(x) = E_{F_n}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right].$$

When the data are interval censored, so that $X_i \in I_i \; \forall i$ and only $I_i = (L_i, R_i)$ is observed, it seems natural to express the kernel density estimate in terms of iterated expectation

$$\hat{f}(x) = E_{F_n}\left[E\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\middle| X \in I\right]\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\middle| X \in I_i\right].$$

Here conditional expectation is computed with respect to the distribution for the true value of X_i over the interval I_i. Goutis (1997) uses such a strategy for the nonparametric estimation of a mixing density.

This conditional distribution is itself unknown and must be estimated. A natural choice is data driven and involves using the kernel density estimate itself to approximate each conditional distribution. This results in an iterative algorithm with the following smooth estimate of the density at the jth step:

$$\hat{f}_j(x) \;\; = \;\; \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{j-1}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\middle| X \in I_i\right]$$

where

$$\mathrm{E}_k\left[g(X) | \, X \in I_i\right] = \begin{cases} \int_{L_i}^{R_i} g(t) \hat{f}_{k;i}(t) dt & L_i \neq R_i \\ g(X_i) & L_i = R_i = X_i. \end{cases}$$

The conditional density $\hat{f}_{k;i}(\cdot)$ over the interval I_i is defined as

$$\hat{f}_{k;i}(t) = 1_i(t) \hat{f}_k(t) \Big/ c_{k;i}$$

where $1_i(\cdot)$ is the indicator function for the interval I_i and $c_{k;i}$ is its unconditional expectation under $\hat{f}_k$. At the $(k+1)^{st}$ iterate it is the conditional density $\hat{f}_{k;i}(x)$ that is used to smoothly distribute a probability mass of $1/n$ over the interval I_i. Note how this differs from, for example, the product limit estimator which distributes the mass associated with a right censored observation X_i to only those uncensored observations that exceed X_i and not to the entire interval $[X_i, \infty)$.

Given the estimator weights a data point by computing the average height of the kernel over the observed interval consider Figure 2 which depicts how the weight depends on the length and proximity of the interval to the location of the kernel. In the figure the weights for two intervals, centered at 0 but with different lengths, are shown for different positions of the kernel. When the kernel is also centered at 0, the method rewards precision by giving the shorter interval a greater weight. However, when the location of the kernel is shifted so it overlaps predominantly with the longer interval, it assigns a larger weight to this interval even though it is less precise. This is due to the longer interval being more "local" than the shorter interval to the point of estimation, or the center of the kernel. The longer interval is local because there is non-zero probability that the true observation is in a region close to "-2" while this is not the case for the shorter interval.

While the above derivation has intuitive appeal the estimator may be formally defined as minimizing an integrated squared distance between some arbitrary function and the ideal estimator $\hat{f}_c$. We first present the following result.

Theorem 2.1 *Let $\mathcal{F}$ be the set of absolutely continuous density functions in L_2. Suppose that X is distributed with density $f^* \in \mathcal{F}$. Let $\mathcal{C}(f) = \int_{-\infty}^{\infty} \left\{\hat{f}_c(x) - f(x)\right\}^2 dx$, where $\hat{f}_c(x) = (nh)^{-1} \sum_{i=1}^{n} K((x - X_i)/h)$, and assume that $h^{-1}K((\cdot - u)/h) \in \mathcal{F}$ for any fixed $h > 0$ and $u \in I\!R$. Then $\hat{f} = E_{f^*}[(nh)^{-1} \sum_{i=1}^{n} K((x - X_i)/h) | X_i \in I_i, \forall i]$ solves*

$$\hat{f} = \arg\min_{f \in \mathcal{F}} E_{f^*} \left[\mathcal{C}(f) \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right].$$

Proof: Let

$$\begin{aligned} \hat{f} &= \arg\min_{f \in \mathcal{F}} E_{f^*} \left[\mathcal{C}(f) \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right] \\ &= \arg\min_{f \in \mathcal{F}} E_{f^*} \left[\int_{-\infty}^{\infty} \left\{\hat{f}_c(x) - f(x)\right\}^2 dx \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right]. \end{aligned}$$

Under the assumptions about f^* and K, this expectation is finite and hence

$$E_{f^*} \left[\int_{-\infty}^{\infty} \left\{\hat{f}_c(x) - f(x)\right\}^2 dx \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right] = \int_{-\infty}^{\infty} E_{f^*} \left[\left\{\hat{f}_c(x) - f(x)\right\}^2 \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right] dx.$$

By minimizing the positive integrand for every fixed x, we minimize the integral. Thus for a fixed value of x the definition of conditional expectation implies that

$$E_{f^*} \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) - f(x) \right\}^2 \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right]$$

is minimized with respect to $f(x)$ at

$$\hat{f}(x) = E_{f^*} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \middle| \begin{matrix} X_i \in I_i, \\ i = 1, \ldots, n \end{matrix} \right] = \frac{1}{n} \sum_{i=1}^{n} E_{f^*} \left[\frac{1}{h} K\left(\frac{x - X_i}{h}\right) \middle| X_i \in I_i \right]. \square$$

Note the criterion is quite restrictive. It explicitly involves $\hat{f}_c$ and hence the form of the optimal estimator is not surprising. Nevertheless, the result is useful due to the interpretation it lends the estimator. In terms of squared distance the estimator gets as close as possible to the ideal kernel density estimate. In addition, since we do not know the true density f^*, we replace it with any current guess for f^*, say $\hat{f}_{j-1}$. Hence the estimator may be regarded as resulting from a generalized EM algorithm with:

E-step: $\forall i$ define $\hat{f}_{j-1;i}(x)$ and compute $w_i(x) = E_{j-1}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\middle| X \in I_i\right]$
M-step: Compute $\hat{f}_j(x) = \bar{w}(x) = \frac{1}{n}\sum_{i=1}^n w_i(x)$.

Computational issues concerning the E-step are considered in §4.

Figure 3 gives the result of the first four iterations of the algorithm for the hemophiliac data. In this data set, patients who were infected at the time of entry were assigned a left-hand point of $L_i = 1$ which resulted in a number of lengthy intervals commencing from the beginning of the study. The common practice in the HIV literature of assuming a uniform distribution over each interval is clearly inappropriate from an inspection of the data. For the estimates in Figure 3 we used a uniform distribution as our starting point (right censored observations were given a weight of 0). As one expects, differences between the first two iterates are quite large as the initial assumption of a uniform distribution is adjusted by the density estimate itself which places more weight on the later period of the study. Convergence is achieved after four iterations.

3 Properties of the estimator

When the complete data kernel density estimate, $\hat{f}_c$, is used to estimate the cdf as

$$\hat{F}_c(x) = \int_{-\infty}^{x} \hat{f}_c(u)du,$$

the estimate $\hat{F}_c$ reduces to the NPMLE as $h \downarrow 0$. Here the NPMLE is the empirical distribution function F_n. An analogous development holds for the estimator $\hat{f}_j$ as well. In this section we show the algorithm (1) reduces to that of Efron (1967), Turnbull (1976) and Li et al. (1997) as $h \downarrow 0$. Since each of these converges under broad conditions we conjecture that the use of kernel smoothing at each iteration does not perturb the algorithm to such an extent as to effect convergence. The simulation results of §6 support this.

Efron (1967) proposed an iterative scheme for approximating the survivor function at a point x, $S(x) = P[X \geq x]$:

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

where $N(x) = \#X_i \geq x$, $\delta_i = 1$ if X_i is observed exactly and $\delta_i = 0$ if X_i is right-censored ($R_i = \infty$). Efron shows $\tilde{S}_j$ converges to a fixed point that coincides with the Kaplan-Meier product limit estimator, that is the NPMLE. Turnbull (1976) generalized this algorithm to obtain a NPMLE of the distribution function under general censoring and truncation schemes. Li et al. (1997) proposed an estimator that is the fixed point of an EM algorithm. Their estimator coincides with Turnbull's estimator where Turnbull's estimator is uniquely defined, and converges to a value that depends on the starting point of the iterative scheme where Turnbull's estimator is not uniquely defined. The iterative scheme proposed by Li et al. (1997) involves computing the conditional expectation of F_n at each step

$$\check{F}_j(x) = E_{j-1}\left[F_n(x) \middle| X_i \in I_i \quad \forall i\right].$$

The following theorem shows that Li et al.'s estimator can be obtained as a limit of our estimator when we let the window width of the kernel shrink to zero at every step.

Theorem 3.1 *Let $\hat{F}_j(x)$ be the estimate of the cdf corresponding to the density estimate (1). Assuming both algorithms have the same initial value, 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

$$\begin{aligned}
\check{F}_j(x) &= E_{j-1}\left[F_n(x) \middle| X_i \in I_i \quad \forall i\right] \\
&= E_{j-1}\left[\frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x] \middle| X_1 \in I_1, \ldots, X_n \in I_n\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{\check{F}_{j-1}(x) - \check{F}_{j-1}(L_i)}{\check{F}_{j-1}(x) - \check{F}_{j-1}(L_i)}\right) 1_i(x) + I[x \geq R_i]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{j-1}\left[I[X_i \leq x] \middle| X_i \in I_i\right], \quad j = 1, 2, \ldots
\end{aligned}$$

Note Li et al. (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

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

Since $K^*((x - X_i)/h) \leq 1$ for all h, we can bring the limit inside the expectation. The result obtains since $\lim_{h\downarrow 0} K^*((u - v)/h) = I[v \leq u]$, $\forall u, v \in \Re$. □

In the case of right-censored data the algorithm (1) reduces to that of Efron (1967) as an immediate consequence of Theorem 3.1.

Corollary 3.1.1 *If $R_i = \infty$ for all interval censored data points, then*

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

Proof: Under right censoring, $1 - \tilde{S}_j(x) = \check{F}_j(x)$ and hence the result. □

The above developments naturally lead to the consideration of convergence of the algorithm to a fixed point. Series expansions, (Silverman 1986), similar to those for the complete data kernel density estimate, $\hat{f}_c$, show that $\hat{F}_j$ is equivalent to $\check{F}_j$ to second order. The following theorem does not prove convergence but it does show that convergence of the algorithm is linked to the convergence of $\check{F}_j$ and may be inherited from $\check{F}_j$. In other words, the convergence of $\check{F}_j$ is a necessary condition for the convergence of $\hat{F}_j$. Li et al. (1997) show that $\check{F}_j$ converges when $\check{F}_0$ is a strictly increasing distribution function.

Theorem 3.2 *Assume, $\int_\Re K(u)du = 1$, $\int_\Re uK(u)du = 0$ and $\int_\Re u^2K(u)du = \sigma_K^2 < \infty$, then, assuming both algorithms have same initial value we have*

$$\hat{F}_j(x) = \check{F}_j(x) + O(h^2), \quad \forall x \ j = 1, 2, \ldots$$

The proof of this Theorem can be found in the Appendix.

The assumptions of the theorem are typical of most popular kernel functions, including the Gaussian kernel. The effect of the $O(h^2)$ term depends on both the properties of the kernel as well as the size of h. In the simulations of §6 the $O(h^2)$ term does not disturb convergence.

4 Implementation through importance sampling

Computing an iterate in the recursive scheme (1) requires the computation of a conditional expectation for each interval censored observation. For an interval I this conditional expectation has the form

$$\mu_I = E_j\left[\frac{1}{h}K\left(\frac{x - X}{h}\right)\middle| X \in I\right] = \int_L^R \frac{1}{h}K\left(\frac{x - X}{h}\right)\hat{f}_{j;I}(X)dX$$

which, except in special cases, will not be computable in closed form. Rather than numerically approximating the integral involved in the expectation, we estimate it by a sample mean in a Monte Carlo scheme that is fast and easy to implement. Thus the iterative algorithm involves a sampling process which iterates until we are confident that we are sampling from the fixed point of (1).

Two sampling schemes are considered where the second is an approximation of the first. The first method involves sampling exactly from $\hat{f}_{j;I}$ using an acceptance/rejection method where candidate values Y are generated from the distribution with density $\hat{f}_j$ and accepted if $Y \in I$

```
1. generate $Y \sim \hat{f}_j$
2. if $Y \in I$ set $X \longleftarrow Y$ otherwise goto 1.
```

The first step is accomplished by the following recursive scheme:

```
1. sample with replacement from $\{I_1, \ldots, I_n\}$ to get $I^\star$
2. sample $\mu$ from $f_{j-1;I^\star}$
3. sample $Y$ from $\frac{1}{h}K(\frac{x-\mu}{h})$
```

where the recursion occurs at step 2. Once a sample $X_1, \ldots, X_B$ is obtained, $\hat{\mu}_I$ is computed as

$$\hat{\mu}_I = \frac{1}{B}\sum_{k=1}^{B}\frac{1}{h}K\left(\frac{x - X_k}{h}\right)$$

and since the sampling is exact

$$\hat{\mu}_I \overset{B\to\infty}{\longrightarrow} E_j\left[\frac{1}{h}K\left(\frac{x - X}{h}\right)\middle| X \in I\right]$$

Thus we can limit the effect of Monte Carlo error by choosing B to be as large as we want.

The difficulty with this exact sampling method is that it punishes precision in the data. When an interval is narrow the acceptance/rejection step will largely reject proposals. Thus obtaining a large sample may take a long time and given the scheme is recursive the impact can be substantial. To offset this we use an importance sampling scheme where the time to compute an iterate does not increase with the number of iterations as in the exact sampling scheme above. Based on

$$E_j\left[\frac{1}{h}K\left(\frac{x - X)}{h}\right)\middle| X \in I\right] = E_g\left[\frac{1}{h}K\left(\frac{x - X}{h}\right)w(X)\right],$$

where g is some distribution over the interval I that is easy to sample from, and $w(X) = \frac{\hat{f}_{j;I}(X)}{g(X)}$ is the importance sampling weight, $\hat{\mu}_I$ becomes

$$\hat{\mu}_I = \sum_{k=1}^{B}\frac{1}{h}K\left(\frac{x - X_k}{h}\right)w_k^\star,$$

where $w_k^\star = w(X_k)/\sum_{l=1}^{B} w(X_l)$ and the above acceptance/rejection scheme is replaced by

1. Generate $X_k \sim g, k = 1, \ldots, B$
2. Compute $\hat{\mu}_I$.

The only additional complication is to compute the sampling weights $w_k^\star$ which, upon inspection, simplify in a convenient way. Ultimately they involve the height of the unconditional kernel density estimate $\hat{f}_j$ thus avoiding computation of the constants $c_{j;I}$

$$w_k^\star \; = \; \frac{\hat{f}_{j;I}(X_k)}{g(X_k)} \Big/ \sum_{l=1}^{B} \frac{\hat{f}_{j;I}(X_l)}{g(X_l)} = \frac{\hat{f}_j(X_k)}{c_{j;I}g(X_k)} \Big/ \sum_{l=1}^{B} \frac{\hat{f}_j(X_l)}{c_{j;I}g(X_l)} = \frac{\hat{f}_j(X_k)}{g(X_k)} \Big/ \sum_{l=1}^{B} \frac{\hat{f}_j(X_l)}{g(X_l)}.$$

Finally, using the values of the kernel density estimate $\hat{f}_j$ at a sufficiently fine grid, $\hat{f}_{j-1}(X_k)$ can be accurately computed by interpolation.

Using the hemophiliac data figure 4 gives the result of a simulation study for values of $B = 10$ & 100 respectively. For each plot the kernel density estimate, based on 4 iterations of the algorithm, was computed 100 times for a fixed value of B. The plot depicts the resulting pointwise mean and 99 % percentile interval. The method works quite well for samples of size 100.

5 Choice of the smoothing parameter

A central component of kernel density estimation is the choice of the smoothing parameter. We propose an automatic method for this purpose based on likelihood cross-validation that is is analogous to the complete data case (Silverman, 1986). In the presence of complete data $X_1, \ldots, X_n$ likelihood cross-validation aims to maximize

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

with respect to the smoothing parameter h. The superscript indicates X_i is left out when the estimate $\hat{f}_h^{(-i)}(X_i)$ is computed and the method works because $E[CV(h)]$ involves the Kullbeck-Leibler distance between f and $\hat{f}$:

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

In the case of interval censored data it is natural to mimic the above strategy through analogy. In the above, $\hat{f}_h^{(-i)}(X_i)$ is obtained by eliminating a point of support, X_i from the NPMLE, namely F_n. By eliminating X_i the contribution to CV at that point of support uses only the remaining data. In our case, the support of the NPMLE, F_t, are the innermost intervals defined as $J_r = (p_r, q_r), r = 1, \ldots, m$ where $p_r \in \{L_i, i = 1, \ldots, n\}$, $q_r \in \{R_i, i = 1, \ldots, n\}$ and $J_r \cap I_i$ equals J_r or $\emptyset, \forall r, i$ (see Turnbull 1976, or Li et al., 1997 for a more detailed discussion of innermost sets). For simplicity of exposition, and without loss of generality, assume all data are interval censored. In this case, the cross-validated likelihood is defined as

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

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. This conveniently addresses the question of tied observations which are common for interval censored data. For example, the hemophiliac data contains only 40 distinct intervals in a sample of size 105. In addition it also addresses the issue of how to handle two observed intervals that are not tied but have a high degree of overlap. If they both overlap completely with the eliminated innermost interval then they are both eliminated when estimating the contribution to the cross-validation process for that interval.

While the scheme is admittedly adhoc, it worked well in a limited simulation study using 40 samples. A description of how data was generated is given in §6. Table 1 compares average values of our cross-validated likelihood with the Kullbeck-Leibler distance for both $\hat{f}_j$ and $\hat{f}_t$. In both cases our cross-validated likelihood is quite accurate when compared to the Kullbeck-Leibler distance. It obtains its maximum at, or near, the value of the window size that minimizes the Kullbeck-Leibler distance. In addition, the ideal window size is smaller for the proposed estimator, $\hat{f}_j$, indicating it uses more information in the data. Note the method contradicts our aim of simplicity as knowledge of the innermost intervals is required for computation. Ultimately, a method that is independent on innermost intervals, like k-fold cross-validation as considered in Pan (2000), may be preferred.

6 A simulation study

For the estimator $\hat{f}_j$ two patterns of behaviour are evident in the following simulation study: convergence and improvement over the standard kernel density estimate, $\hat{f}_t$. Five criteria are considered of which three are useful for comparing $\hat{f}_j$ and $\hat{f}_t$. The remaining two assess the dependence of the estimator on the initial value, f_0, of the algorithm. All criteria assess convergence. Define the squared distance, ρ, between two functions, u and v, as

$$\rho(u, v) = \sum_{x \in \mathcal{X}} \{u(x) - v(x)\}^2$$

where $\mathcal{X}$ is a fixed grid of equally spaced points spanning the range of the data. This distance is central to all convergence criteria with the exception of one involving the Kullbeck-Leibler distance.

If the algorithm converges to a fixed point, $\hat{f}$, that is independent of the initial value, then it is said to be a contraction mapping, Υ, if for some suitably defined space of densities $\mathcal{F}$, Υ is such that

$$\begin{aligned}\Upsilon \ : \quad & \mathcal{F} \longrightarrow \mathcal{F} \\ & \hat{f}_j = \Upsilon(\hat{f}_{j-1}) \\ & \hat{f} = \Upsilon(\hat{f}) \\ & \rho(\hat{f}_j, \hat{g}_j) < \rho(\hat{f}_{j-1}, \hat{g}_{j-1}), \ j > 1\end{aligned}$$

where $\hat{f}_j$ and $\hat{g}_j$ are the density estimates at the j^{th} step for two arbitrary but different starting points $f_0, \ g_0 \in \mathcal{F}$. The first two columns of table assess the behaviour of the estimator as a contraction mapping. The "squared distance" column gives the average value of $\rho(\hat{f}_j, \hat{g}_j), \ j = 1, \ldots, 10$ based on 100 samples

$$\bar{\rho}(\hat{f}_j, \hat{g}_j) = \frac{1}{100} \sum_{i=1}^{100} \rho_i(\hat{f}_j, \hat{g}_j)$$

where $\rho_i(\cdot, \cdot)$ denotes the value of $\rho(\cdot, \cdot)$ for the i^{th} sample. The "contraction" column give the proportion of samples that satisfy the condition

$$\rho_i(\hat{f}_j, \hat{g}_j) < \rho_i(\hat{f}_{j-1}, \hat{g}_{j-1}), \ j > 1.$$

For each of the 100 random samples we generated 20 failure times from a Weibull distribution with shape parameter 1.75 and scale parameter 3. Independently, we then generated "visit times" using a homogeneous Poisson process. Each failure time was interval-censored by the visit times that bracketed it. For each sample, we computed the iterative scheme (1) using a Gaussian kernel with $h = 1$. We used a sample size of $B = 100$ for the importance sampling scheme described in §4. The initial values of the density for the iterative scheme are various scale and location shifted beta distributions. Here f_0 and g_0 are based on beta(5, 2) and beta(2, 5) distributions respectively.

The criteria "MSE$_r$", $r = 1, 2$ use $\bar{\rho}(u, v)$ to assess the expected value of the squared distance $\rho(u, v)$ under the Weibull(1.75,3) distribution. When $r = 1$, the function u is set to be the true density, f, and $v = \hat{f}_j$. Thus MSE$_1$ estimates the actual mean squared error of the estimator. For MSE$_2$ the true density is replaced by the ideal estimator, $\hat{f}_c$, and the criterion assesses the closeness of $\hat{f}_j$ to the ideal estimator as discussed in §2. The final column assesses an estimate of the Kullbeck-Leibler distance

$$\kappa(f, \hat{f}_j) = E\left[\log\left\{f(X) \Big/ \hat{f}_j(X)\right\}\right].$$

For each of the 100 samples an estimate,

$$\hat{\kappa}_i(f, \hat{f}_j) = \overline{f(X) \Big/ \hat{f}_j(X)},$$

is itself based on a sample $X \sim \text{Weibull}(1.75, 3)$. As in §4, the values of the density estimate $\hat{f}_j(X)$ are found by interpolation using the values of $\hat{f}_j$ computed at the grid $\mathcal{X}$. The entry given in the table is

$$\bar{\kappa}(f, \hat{f}_j) = \frac{1}{100} \sum_{i=1}^{100} \hat{\kappa}_i(f, \hat{f}_j).$$

Note the MSE and Kullbeck-Leibler criteria are evaluated for Turnbull's estimator as well where $\hat{f}_j$ is simply replaced by $\hat{f}_t$. All columns in the table give standard errors in brackets.

The results reported in Table 2 show the algorithm tends to reach convergence after 3-6 iterations. The contraction criterion improves until the 6^{th} iteration after which its behaviour is consistent with what would be expected if the Monte Carlo scheme of §4 involved sampling from the fixed point (once convergence is reached we expect the condition $\rho_i(\hat{f}_j, \hat{g}_j) < \rho_i(\hat{f}_{j-1}, \hat{g}_{j-1})$ to hold % 50 of the time). The squared distance criterion shows the distance between estimators with different initial values gets very small indeed by the 6^{th} iteration. The MSE and Kullbeck-Leibler criteria reach their minimum after only 3 iterations of the algorithm after which they remain fairly constant. The large improvement between the first and second iterates, and the smaller improvement between the second and third iterates, show how the estimator continues to extract more information out of the data after the first iteration. It is these improvements that result in the estimator having better properties than $\hat{f}_t$ for these criteria. Finally, as an example, Figure 5 shows four successive iterates for the hemophiliac data based on various initial values of the algorithm.

7 Use as a scatterplot smoother

Kernel weights are useful in regression as well as density estimation. In the regression context we consider the use of kernel weights where a covariate is interval censored. The techniques described here are understood to be applicable to multiple regression problems where an additive or generalized additive model are used. This and the context where the response in a regression is also interval censored are deferred. The purpose of the following example is to exhibit the flexibility of the methods of the paper rather than to perform the ideal data analysis. The data used is a subset of a larger dataset concerning HIV infection and infant mortality (Hughes and Richardson, 2001). Here we only consider infants with no interval censoring in the response (i.e. infants that died) and that were infected with HIV.

Consider the model

$$E[Y] = g(x)$$

where the covariate x may be interval censored. In terms of a scatterplot, interval censoring in a covariate means only the y coordinate is known. Any smoothing process that uses kernel weights whose size is determined by some nearest neighbourhood technique needs modification, as such neighbourhoods are determined by the covariate. Suppose, for example, that a running mean smoother is used where

$$\hat{y} = \hat{g}(x) = \sum_{y_j \in \mathcal{N}_x} v_j y_j.$$

where $\mathcal{N}_x$ is the nearest neighbourhood for x. Typically, the weights v_j are given as $v_j = \frac{1}{h} K\left(\frac{x - X_j}{h}\right)$ however the X_j are not observed. In keeping with the spirit of this paper v_j is replaced by

$$\mu_{I_j} = E_{\hat{f}_x}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right) \middle| X \in I_j\right]$$

where expectation is computed with respect to the fixed point $\hat{f}_x$ of (1) restricted to the interval I_j. Note the estimate $\hat{f}_x$ is the density estimate for the covariate X. As in §4 expectation is approximated by an importance sampling algorithm and so the recipe for computing $\hat{g}$ is

1. Generate a sample $X_1, \ldots, X_B$ from the chosen importance sampling distribution for the interval I_j
2. Using $\hat{f}_x$ compute the sampling weights $w_k^\star$ as in §4
3. approximate μ_{I_j} by $\hat{\mu}_{I_j} = \sum_{k=1}^{B} \left\{\frac{1}{h} K\left(\frac{x - X_k}{h}\right) w_k^\star\right\}$
4. Compute $\hat{g}(x) = \sum_{y_j \in \mathcal{N}_x} \hat{\mu}_{I_j} y_j$

Figure 6 gives four plots for the infant data. The first of these is a scatterplot of the times of death for the sixty infants used in the fitting process. The covariate is the time of infection with the HIV virus. It is interval censored and hence a scatterplot "point" is actually a line obtained by joining the right and left endpoints of each interval. The second plot gives the fitted density estimate for the covariate based on four iterations of our algorithm. The cross validation technique of §5 was used to pick the "ideal" window size of 3.75. The data were originally collected in a study of the effect of breast feeding on infection. However, for the infants used here the primary source of infection seems to be invetro. The time point 0 indicates the time of birth of the child. Note intervals for the covariate extend to -1 indicating infection may have taken place before the birth of the child. The remaining two plots use this density estimate to kernel smooth the scatterplot using a running mean, although any linear smoother could be used. Several window sizes were arbitrarily chosen. The plots raise many interesting issues worthy of further study. For example, the smooths are dominated in an unreasonable fashion by three points on the right suggesting that either the techniques be made robust or the choice of smoothing parameter be made adaptive, or both! These and o ther issues, like how to handle interval censored response as well, will be considered in future work.

Acknowledgements

The authors are grateful to Jeff Rosenthal, David Andrews, Rob Tibshirani and Jerry Lawless for inspiring discussions. We particularly appreciate many useful conversations with Paul Corey. They also wish to thank the Natural Sciences and Engineering Research Council of Canada for supporting this research through individual operating grants.

Appendix

Proof of Theorem 3.2: The proof is given for the case where all data are interval censored. Recall,

$$\begin{aligned}\tilde{F}_j(x) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{\tilde{F}_{j-1}(x)-\tilde{F}_{j-1}(L_i)}{\tilde{F}_{j-1}(R_i)-\tilde{F}_{j-1}(L_i)}\right)1_i(x)+I[x\geq R_i]\right\}\\ &= \int^x \tilde{f}_j(t)dt\end{aligned}$$

where

$$\begin{aligned}\tilde{f}_j(x) &= \frac{d}{dx}\tilde{F}_j(x)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{f_{j-1}(x)}{\tilde{F}_{j-1}(R_i)-\tilde{F}_{j-1}(L_i)}1_i(x)\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}\tilde{f}_{j-1}(x)1_i(x)\end{aligned}$$

Now standard calculations like those of Silverman (1984) allow expansion about $h=0$ of each conditional expectation and thus expansion of $\hat{f}_j$ $\forall j$

$$\begin{aligned}\hat{f}_j(x) &= \frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{R_i}h^{-1}K\left(\frac{x-t}{h}\right)f_{j-1;i}(t)dt\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}h^{-1}\int_{\Re}K\left(\frac{x-t}{h}\right)f_{j-1}(t)1_i(t)dt\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}h^{-1}\int_{\Re}K\left(\frac{x-t}{h}\right)g(t)dt,\quad g(t)=f_{j-1}(t)1_i(t)\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}\int_{\Re}K(u)g(x-hu)du\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}\int_{\Re}K(u)\left\{g(x)-hu\dot{g}(x)+\frac{1}{2}h^2u^2\ddot{g}(x)+\cdots\right\}du\\ &= \frac{1}{n}\sum_{i=1}^{n}c_{i,j-1}^{-1}\left\{g(x)\int_{\Re}K(u)du-h\dot{g}(x)\int_{\Re}uK(u)du+\frac{1}{2}h^2\ddot{g}(x)\int_{\Re}u^2K(u)du+\cdots\right\}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} c_{i,j-1}^{-1} 1_i(x) f_{j-1}(x) \left\{1 + \frac{1}{2}h^2 \ddot{g}(x)\sigma_K^2 + \cdots\right\} \\
&= \tilde{f}_j(t) + O(h^2)
\end{aligned}
$$

and therefore

$$
\hat{F}_j(x) = \tilde{F}_j(x) + O(h^2), \quad j = 1, 2, \ldots, \ \Box.
$$

References

[1] Efron, B. (1967) "The two sample problem with censored data", *Fourth Berkeley Symposium on Mathematical Statistics*, University of California Press, 831-853.

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

[3] De Gruttola, V. and Lagakos, S. W. (1989). Analysis of Doubly-Censored Survival Data, with Applications to AIDS. *Biometrics.* **45**, 1-11.

[4] Hughes, J. P. and Richardson, R. (2001). Analysis of a Randomized Trial to Prevent Vertical Transmission of HIV-1. *J of Amer. Statist. Assoc.* In press.

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

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

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

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

[9] Wu, C.F. (1983) "On the convergence of the EM algorithm", *Annals of Statistics*, **11**, 95-103.

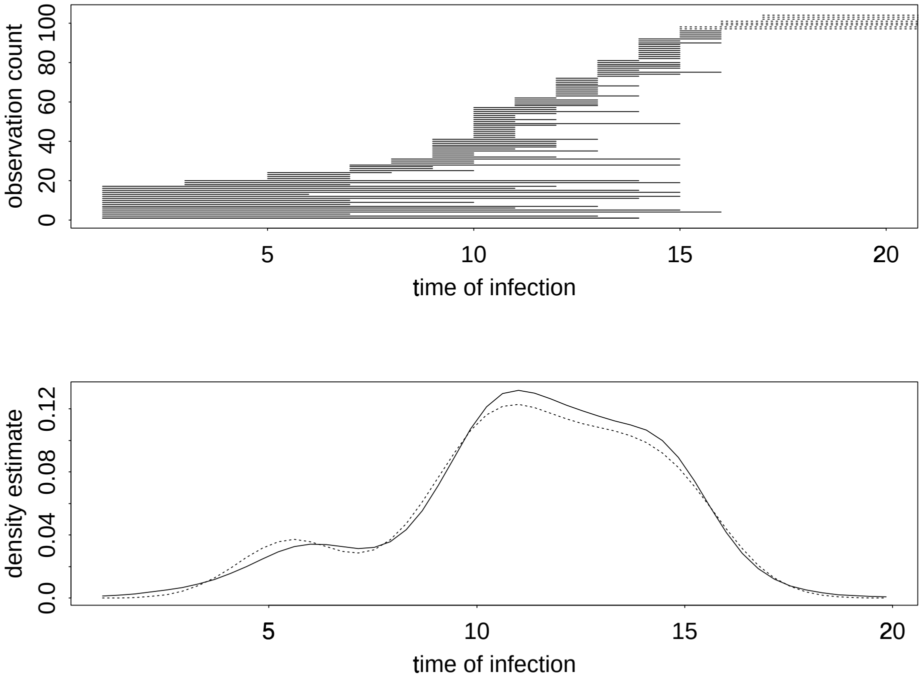

Figure 1: The first plot gives the original data with a line joining the left and right endpoints of each interval. The second plot gives two kernel density estimates with the solid line being that of the method proposed, $\hat{f}_4$, after 4 iterations and the dotted line a kernel smoothed version of Turnbull's estimator, $\hat{f}_t$.

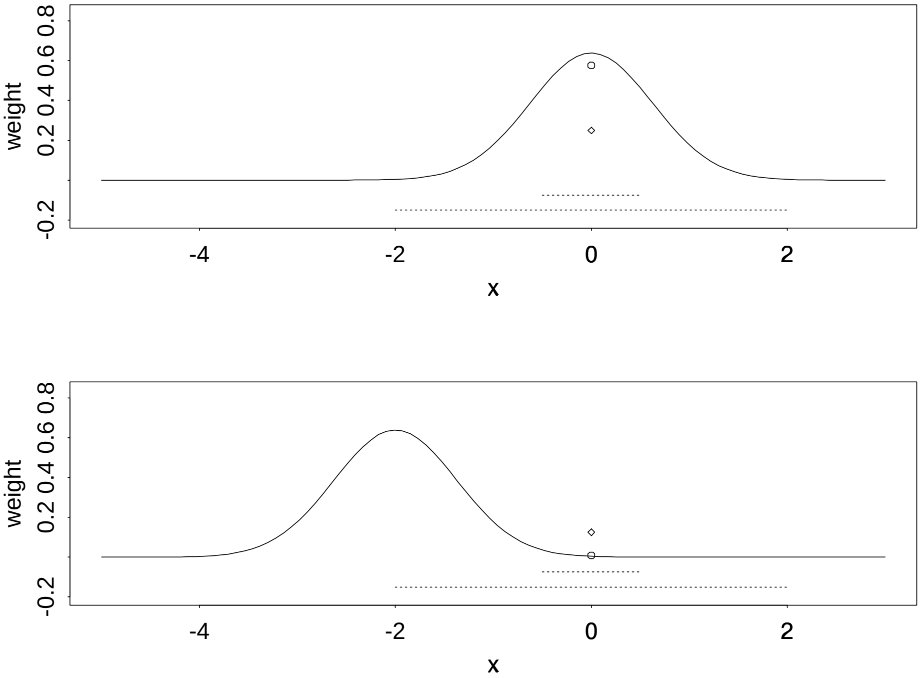

Figure 2: The plot depicts how the kernel density estimate works. The diamond shaped plotting character gives the weight for the longer interval.

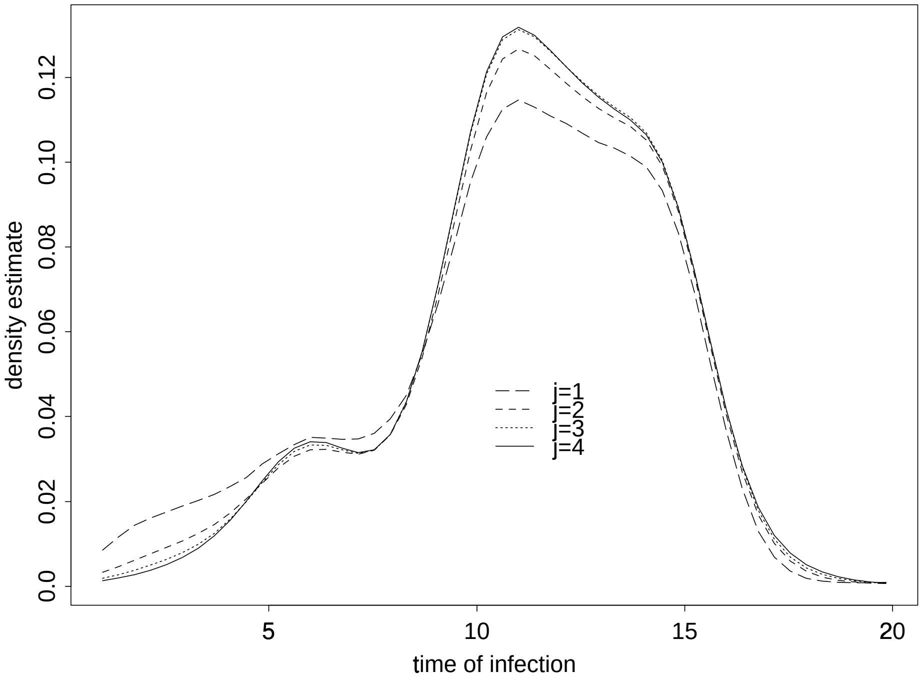

Figure 3: The plot gives the kernel density estimate $\hat{f}_j$ for each of the first four iterations of the algorithm. Convergence appears to have been reached by the third iteration.

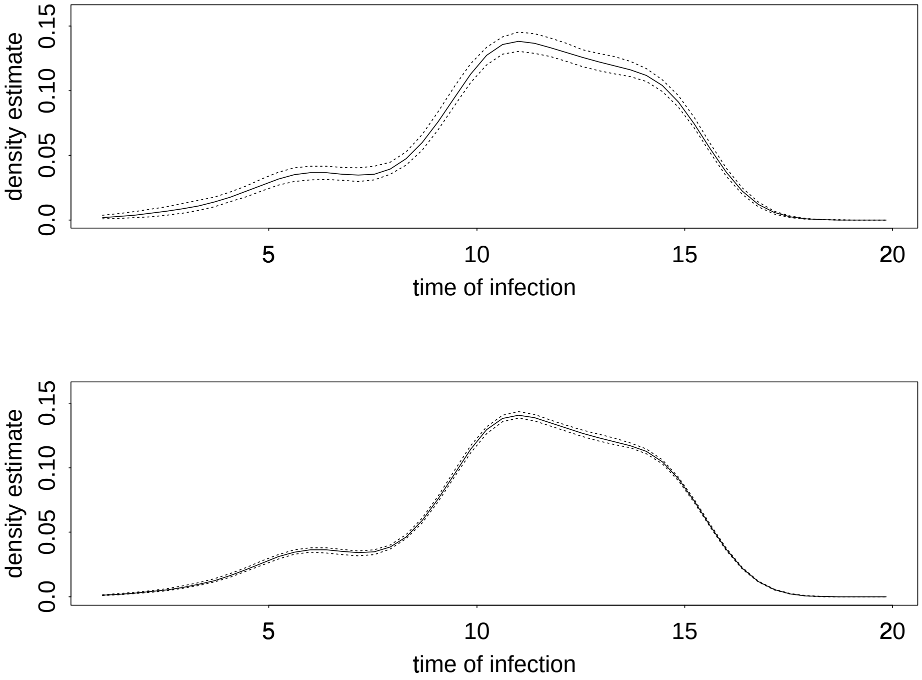

Figure 4: Plots of the pointwise mean and 99% percentile intervals for B with values of 10 and 100 respectively.

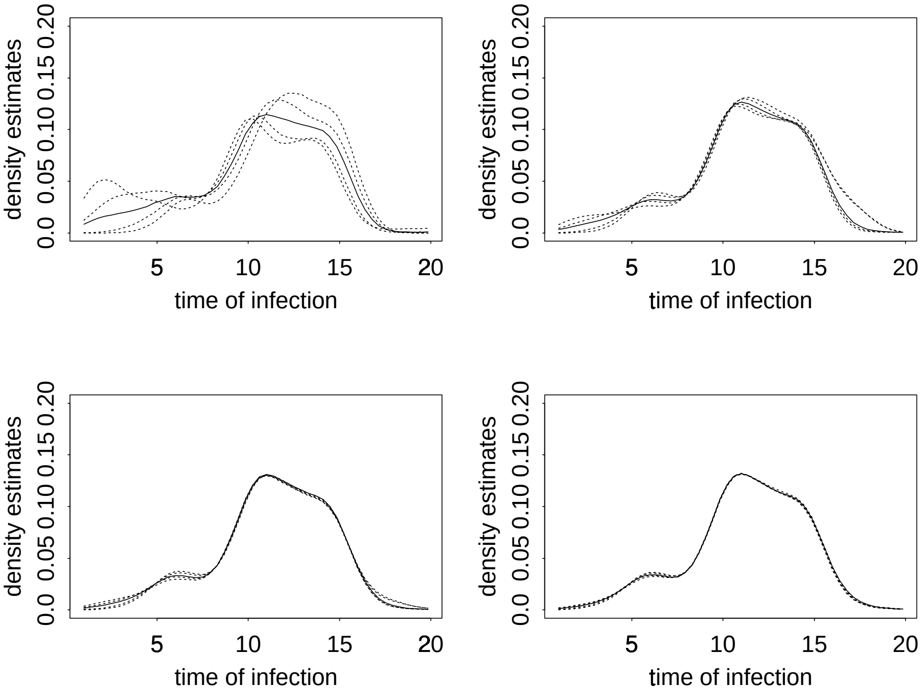

Figure 5: Plots of $\hat{f}_j, j = 1, \ldots, 4$ for different initial values of the algorithm.

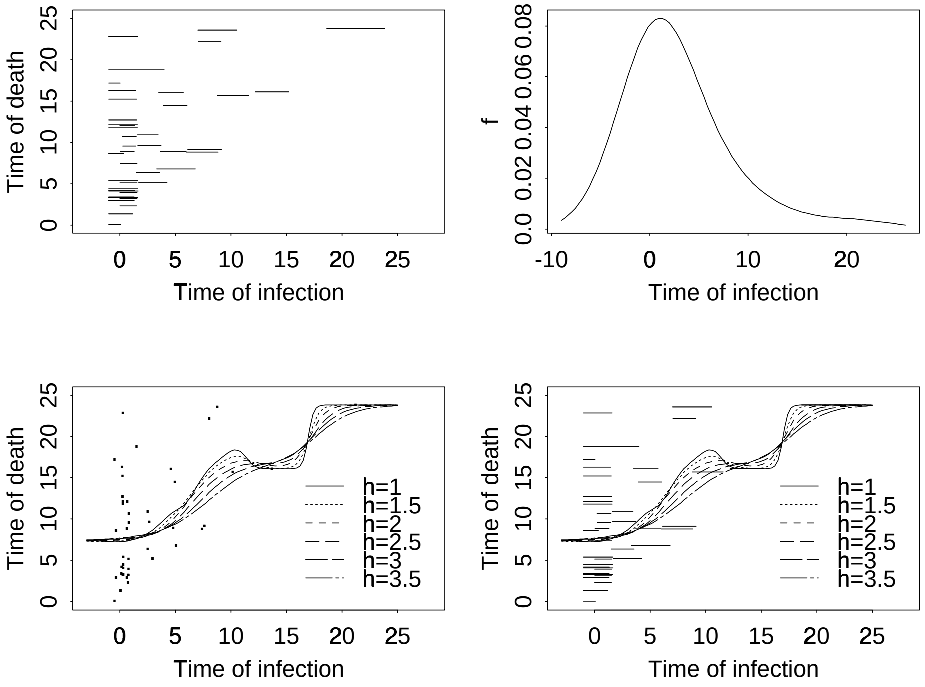

Figure 6: Clockwise from the upper left are plots of the data, a density estimate for the interval censored covariate and two plots of the fitted curve for various window sizes. The first of these includes the scatterplot data reported as the midpoint of the interval and the second gives the interval itself.

Table 1: Cross validation (CV) and Kullbeck-Leibler (KL) distances

h	CV ($\hat{f}_4$)	KL ($\hat{f}_4$)	CV ($\hat{f}_t$)	KL ($\hat{f}_t$)
0.25	-3.820	0.1769	$-\infty$	∞
0.33	-3.790	0.1489	-5.668	∞
0.42	-3.762	0.1309	-4.895	∞
0.50	-3.741	0.1234	-4.473	∞
0.58	-3.724	0.1149	-4.225	0.4805
0.67	-3.716	**0.1093**	-4.073	0.3706
0.75	-3.711	0.1131	-3.977	0.2782
0.83	**-3.709**	0.1146	-3.915	0.2336
0.92	-3.713	0.1190	-3.876	0.1948
1.00	-3.718	0.1277	-3.851	0.1743
1.08	-3.726	0.1358	-3.837	0.1673
1.17	-3.736	0.1477	-3.830	0.1574
1.25	-3.747	0.1589	**-3.829**	**0.1554**
1.33	-3.759	0.1721	-3.832	0.1612
1.42	-3.772	0.1854	-3.838	0.1684
1.50	-3.884	0.1991	-3.846	0.1742

Table 2: The behavior of various criteria

j	Squared distance	Contraction	MSE$_1$	MSE$_2$	Kullbeck-Leibler
1	1.2e-01 (3.7e-02)	·	0.0700 (0.00312)	0.0242 (0.00154)	0.179 (0.0059)
2	1.2e-02 (9.4e-03)	1.000(N/A)	0.0515 (0.00317)	0.0130 (0.000875)	0.131 (0.0049)
3	1.6e-03 (2.0e-03)	1.000(N/A)	0.0493 (0.00322)	0.0122 (0.000844)	0.125 (0.0047)
4	2.4e-04 (4.4e-04)	1.000(N/A)	0.0490 (0.00325)	0.0122 (0.000866)	0.126 (0.0050)
5	5.5e-05 (8.0e-05)	0.91(0.0286)	0.0492 (0.00328)	0.0123 (0.000869)	0.125 (0.0050)
6	2.7e-05 (2.6e-05)	0.69(0.0462)	0.0492 (0.00329)	0.0124 (0.000884)	0.126 (0.0050)
7	2.8e-05 (2.4e-05)	0.47(0.0499)	0.0491 (0.00328)	0.0123 (0.000881)	0.124 (0.0052)
8	2.6e-05 (2.7e-05)	0.53(0.0499)	0.0491 (0.00328)	0.0123 (0.000879)	0.125 (0.0049)
9	2.8e-05 (3.0e-05)	0.46(0.0498)	0.0493 (0.00331)	0.0123 (0.000877)	0.125 (0.0049)
10	2.5e-05 (2.5e-05)	0.53(0.0499)	0.0491 (0.00329)	0.0123 (0.000882)	0.126 (0.0052)
$\hat{f}_t$	·	·	0.0760 (0.00579)	0.0385 (0.00302)	0.149 (0.0067)