SiZer for Censored Density and Hazard Estimation

Jiancheng Jiang
Department of Probability and Statistics
Peking University
Beijing 100871, People's Republic of China

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
USA

February 13, 2003

Abstract

The SiZer method is extended to nonparametric hazard estimation and also to censored density and hazard estimation. The new method allows quick, visual statistical inference about the important issue of statistically significant increases and decreases in the smooth curve estimate. Instead of being straightforward, this extension has required the opening of a new avenue of research on the interface between statistical inference and scale space.

Footnote: Jiancheng Jiang was supported by Chinese NSF Grant 39930160. J. S. Marron was supported by NSF Grant DMS-9971649.

1 Introduction

Nonparametric hazard rate estimation is a standard tool in survival analysis, dating back at least to Watson and Leadbetter (1964a,b) and Rice and Rosenblatt (1976).

For practical use, a critical issue is understanding where the hazard rate curve increases and where it decreases. A confounding issue is the bandwidth, i.e. the window width or smoothing parameter. SiZer addresses both of these problems, in the context of nonparametric density and regression estimation, by combining a scale space approach to smoothing with a useful visualization of simultaneous statistical inference.

An illustrative example is given in Section 1.1. The extension of SiZer developed in this paper is not straightforward, as shown in Section 1.2. The obvious idea of simply plugging reweighted data into SiZer gives invalid statistical inference. Hence, a non-obvious statistical accounting for the reweighting is developed. Some real data examples are shown in Section 1.3. Precise mathematical development is given in Section 2. Important computational issues are discussed in Section 3.

It is straightforward to simultaneously extend these ideas to both censored density estimation, and also to censored hazard rate estimation. This is because all three of these cases fit very simply into a general form of estimator, using an elegant common notation, perhaps first published by Patil (1990, 1993). Hence all three cases are treated simultaneously here. For reasons of presentation, various aspects of this paper are usually illustrated by focusing on just one of the three cases first.

Some other important related references include Tanner and Wong (1983), Marron and Padgett (1987), Lo, Mack and Wang (1989), Sarda and Vieu (1991), Müller and Wang (1994), González-Manteiga, Marron, and Cao (1996), Kousassi and Singh (1997), Stute (1999), Hess, Serachitopol and Brown (1999), and Jiang and Doksum (2003).

1.1 An Illustrative Example

In this paper, the SiZer ideas are extended to censored density, to hazard rate, and to censored hazard rate estimation. The SiZer method is illustrated in the context of censored density estimation in Figure 1. Also shown there is the importance of the correct use of censoring reweightings, when censoring is present.

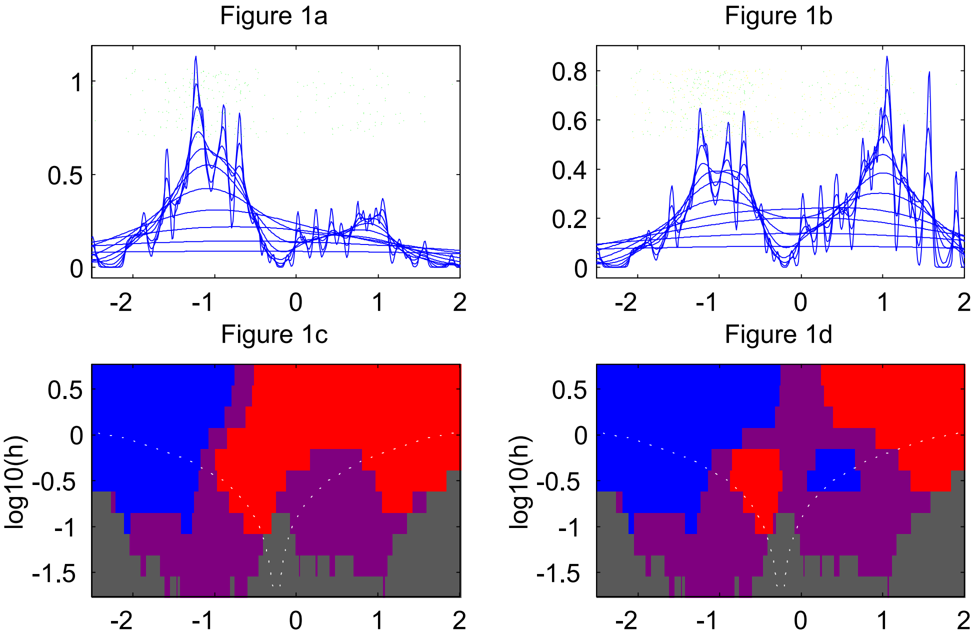

FIGURE 1: *Simulated censored density estimation, $n = 500$ data points (uncensored values shown as jittered green dots, censored as yellow) from a symmetric bimodal Gaussian mixture, with 50% censoring from the same distribution. Figures 1a and b are family plots, Figures 1c and d are the corresponding SiZer maps. In Figures 1a and c, only the uncensored data are used, while Figures 1b and d properly use the information in the censored data, showing that censored data can not be simply ignored.*

The data in Figure 1 are a random sample of $n = 500$ simulated observations from the bimodal Gaussian mixture density with 50% of the data being $N\left(-1, \frac{1}{4}\right)$ and the other 50% $N\left(1, \frac{1}{4}\right)$. The data are censored by an independently sampled data set from the same distribution. I.e. when each original data point is smaller than the corresponding censoring point, it is kept. When the censoring value is smaller it is kept instead. Full mathematical details of censoring are given in Section 2.2. Figure 1a illustrates a naive approach to censoring: ignore the censored data and use standard methods, in this case kernel density estimation. The raw data are shown as jitter plots, i.e. a random height is used to separate them for convenient visualization, with green for uncensored observations, and yellow for censored values. In keeping with scale space ideas, a family of kernel density estimates (indexed by the bandwidth, i.e. level of smoothing) is shown as overlaid blue curves, based only on the uncensored green data. The peak on the left is much higher than the peak on the right, because there are more uncensored observations in that region.

The result of naively plugging these data in to the standard SiZer algorithm, as developed in Chaudhuri and Marron (1999) is shown in Figure 1c. The SiZer map is indexed by location in the horizontal direction, and bandwidth (amount of smoothing) in the vertical direction, thus covering the full "scale space", i.e. the family of smooths shown in blue in Figure 1a. At each point, the color blue is used to indicate that the corresponding smooth is significantly increasing, and red is used for significant decrease. When there is no statistically significant change, the intermediate color of purple is used. For the smaller bandwidths, i.e. finer levels of resolution, the blue curves in Figure 1a have many small peaks and valleys. The SiZer map shows that these are spurious sampling artifacts, because the color is either purple indicating that the features are not significant, or gray indicating that there is not enough data in the smoothing window for reliable statistical inference. For larger bandwidths, near the top of the map, the first peak is flagged as statistically significant by the blue and red regions on each side. However for the peak on the right, only the decrease is flagged as statistically significant, but not the increase in the region $x \in [0, 1]$. That is to say, this important feature of the underlying distribution has been obscured by the censoring.

Figures 1a and 1c are an example of the gross bias that is usually present if censoring effects are ignored. Effective adjustment for this bias is done via reweighting of the data using the Kaplan Meier, i.e. product limit, weight function. A precise definition of this weight function is given in Section 2.2. The effect of this reweighting, for the same simulated data, is shown in Figure 1b. Note that the yellow censoring data points are now included among the green uncensored data. Using the Kaplan Meier weights to upweight the green points on the right side, the blue curves in Figure 1b now properly reflect the symmetric bimodal Gaussian mixture density from which the data were drawn. A careful look at the small bandwidth peaks on the right side shows that the smooths have simply been "magnified" to achieve this effect. In particular, the smallest scale bumps follow the same pattern in Figures 1a and 1b, because their structure is driven by the same set of green dots.

While this vertical rescaling is well understood for estimation, its impact on the statistical inference is much less obvious. Indeed our first attempt at inference, which was to simply feed reweighted data into standard SiZer, failed in the direction of flagging spurious sampling artifacts as being statistically significant (sometimes on the basis of a single data point). The reason for this failure is revealed in Section 1.2. These reasons are then used to develop a statistically valid version of SiZer in the reweighting context in Section 2.

1.2 Reweighted SiZer

As made clear in Patil (1990, 1993), essentially the same "reweighting" technique underlies standard estimators in each of the three cases of censored density, hazard rate, and censored hazard estimation. While these are similar to each other, none allows a straightforward extension of the usual SiZer ideas. In particular, simply plugging reweighted data into standard SiZer fails as noted in the last section. A central contribution of this paper is the development of a correct adaptation of the statistical significance implicit in SiZer, in the presence of the “reweighting” that lies at the core of these estimation methods.

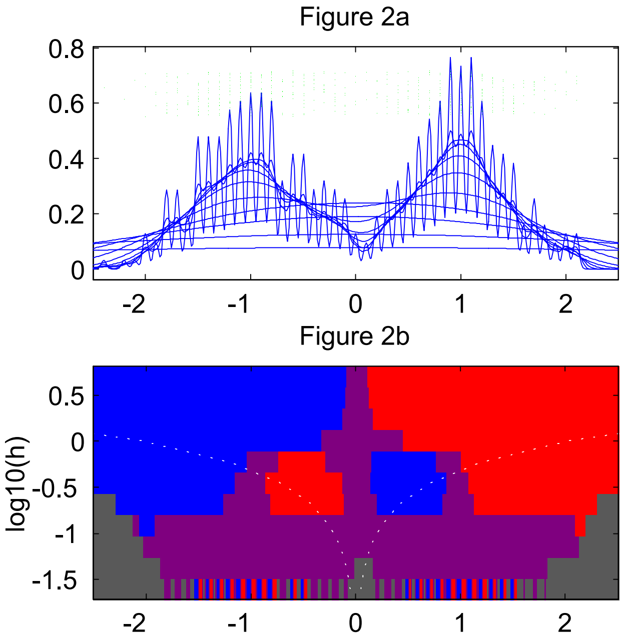

FIGURE 2: *Simulated density estimation, using $n =$ 500 data points from a symmetric bimodal Gaussian mixture, rounded to the nearest 0.1. This shows how rounding creates fine scale significant features in the SiZe map.*

The central problem underlying the statistical inference is illustrated in Figure 2, which again shows a simulated example based on $n = 500$ data points from the same symmetric bimodal Gaussian mixture distribution used in Figure 1. However, this time there is no censoring, but the data have been rounded to the first to the nearest 0.1, note the equal spacing of the green dots. The family plot in Figure 2a shows that at coarse scales, i.e. large bandwidths, the data rounding has no important effect, since it is negligible with respect to the amount of smoothing being done. However, at finer scales, i.e. smaller bandwidths, the rounding has a very strong effect creating many equally spaced spikes in the undersmoothed density estimate.

The rounding affects the SiZer map in Figure 2b in a similar way. At the largest scales, near the top, the SiZer map shows unimodality, which is the true behavior at those levels of resolution. For medium scales, the resolution is such that bimodality is present, and SiZer shows that structure is statistically significant. For most of the finer scales, i.e. smaller bandwidths, the SiZer map is purple, indicating no significant structure. But at the smallest scale, SiZer flags many small peaks as significant, one for each rounded point. The reason for this is that SiZer finds increasing and decreasing points in an assumed continuous underlying density. At very small scale, the replications in the data created by the rounding are viewed as “very tight clusters”, which are indeed statistically significant. This same phenomenon was illustrated in a real data set that had heavy rounding in Figure 5 of Chaudhuri and Marron (1999).

While SiZer *should* flag this type of feature, at these small scales, in the case of rounding, it should *not* flag them in the cases of censored density and hazard estimation that are the focus of this paper. This is why simply plugging weighted data into conventional SiZer gives invalid statistical inference: the weighted points get treated as "tight clusters" at small scales. With enough weight, even a single data point can create an apparently significant spike, even at quite large scales. Thus for these settings, the statistical inference done by SiZer needs to be appropriately adapted to the weighting scheme being used. This is developed in Section 2. Figure 1d is one sensible application of our corrected version of SiZer. More examples are shown in the next section.

1.3 More examples

Figure 3 shows a censored SiZer analysis of the Stanford Heart Transplant Data, from Kalbfleisch and Prentice (1980). The data, originally from Crowley and Hu (1977) are the survival times (in days) of potential heart transplant recipients from their date of acceptance into the transplant program. There is censoring since some patients were lost to follow up before they died and since some patients were still alive on the closing date of the study.

Analysis from the point of view of density estimation is shown in Figures 3a and 3c. This shows that are many deaths very soon after transplantation, and a long decreasing tail. Because of the relatively poor way in which the kernel density estimator handles boundaries, see e.g. Figure 2.16 of Wand and Jones (1995), at larger scales the estimates first increase at the left edge. SiZer shows that both the overall decrease (the large red region) is statistically significant, and so are the boundary effects (the thin blue region right at the edge).

For these data there is more interest in analysis of the hazard rate, as done in Figures 3b and 3d. The hazard rate is carefully defined in Section 2, but the intuitive idea is the instantaneous rate at which patients die. The estimate is a reweighting of the kernel density estimate, as can be seen from the fact that the small scale spikes in Figure 3b are simply magnifications of those in Figure 3a, but the scale is more appropriate for survival considerations. A central question is: when is the hazard rate increasing, and when is it decreasing? The color scheme of SiZer is well suited to address this issue. Furthermore, this question is much more directly answered by the SiZer approach, than by more conventional confidence intervals. The red in the middle left of the SiZer map in Figure 3d shows that the hazard rate significantly decreases during that time period, i.e. as transplants "settle in", chances of survival increase. The red near the top on the right shows that there is also a longer term improvement of the chances of survival after one has survived for a substantial period.

These findings are consistent with those of Jiang and Doksum (2003). An inconsistency is the blue region at the left end. As noted above this is due to poor boundary behavior of the kernel density estimator that underlies this inference. The local polynomial estimator developed by Jiang and Doksum (2003) avoids this problem, which is why their hazard estimate is mostly monotonically decreasing. An interesting open problem is to adapt SiZer ideas to the Jiang and Doksum local polynomial hazard estimator.

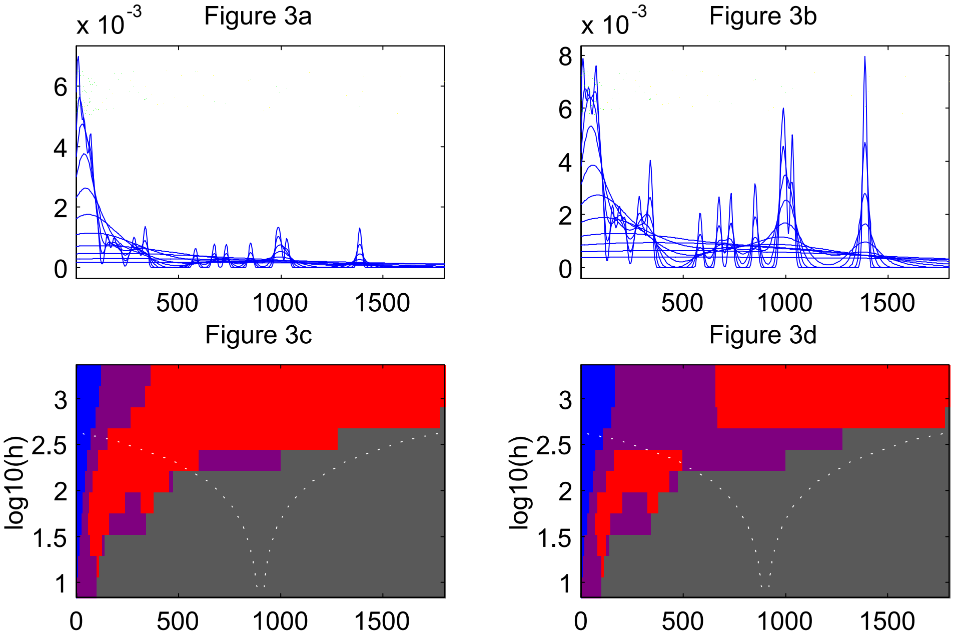

FIGURE 3: *Days to death after heart transplant. Jitter plot, and family of censored density estimates in Figure 3a, with corresponding SiZer map in Figure 3c. Family of censored hazard estimates in Figure 3b, with corresponding SiZer map in Figure 3d.*

Figure 4 shows a SiZer analysis of the device lifetime data of Aarset(1987). These data are uncensored.

The density estimates in Figure 4a suggest a “U-shape” density. However the SiZer map in Figure 4c flags only the right hand peak as statistically significant. This is likely due to the same inefficiency of the kernel density estimator near the boundary.

Of more interest for these data is the hazard rate analysis shown in Figures 4b and 4d. The dominant color in the SiZer map is blue which shows that the hazard rate generally increases over time, which is consistent with the expected wearing over time of mechanical components. A disappointing feature of the family of hazard estimates in Figure 4b, is that there is a spike only on the right side, while other analyses, including Aarset(1987) and Mudholkar, Srivastava and Kollia (1996) find a “bathtub” shape, that includes a spike on the left as well. This again is because of the poor boundary behavior of the kernel density estimator. This problem could also be addressed by a version of hazard estimation SiZer that is based on the local polynomial method of Jiang and Doksum (2003).

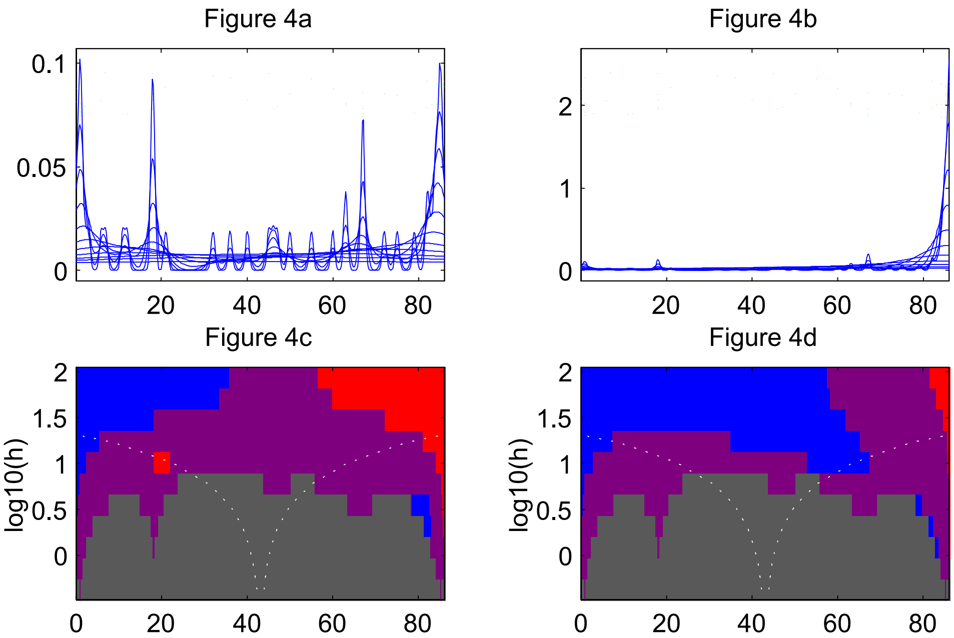

FIGURE 4: *Device lifetime data. Family of censored density estimates in Figure 4a, with corresponding SiZer map in Figure 4c. Family of censored hazard estimates in Figure 4b, with corresponding SiZer map in Figure 4d.*

Both of these examples illustrate an important property of SiZer: it provides a generally good big picture assessment for initial exploratory purposes. However, for addressing any specific problem, e.g. the boundary questions brought up in Figures 3 and 4, it may not be as effective as a method that specifically targets that issue (although we do not know of a currently implemented method that gives better statistical inference of this type at the boundary). Hence we propose SiZer as a broad based method for initially finding structure in data (and for the perhaps more important task of quickly understanding what structures are mere sample artifacts). After one has an idea about what to look for, then other methods can provide deeper insights. Often the next useful step is modelling, e.g. as done by Mudholkar, Srivastava and Kollia (1996) for the device lifetime data.

2 Mathematical Development

Our extension of SiZer is most transparently explained in the context of hazard rate estimation. Hence this is developed in Section 2.1. Then the extension to censored density and censored hazard estimation is done in Section 2.2.

2.1 Hazard Rate Mathematics

For data $X_1, ..., X_n$ independent, identically distributed with cumulative distribution function $F(x)$, and probability density $f(x) = F'(x)$, the maximum likelihood estimate of F is the empirical cumulative distribution function

$$F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i),$$

where

$$1_{(-\infty,x]}(u) = \begin{cases} 1 & \text{if } u \in (-\infty, x] \\ 0 & \text{if } u \notin (-\infty, x] \end{cases}.$$

The kernel density estimate of f is

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, for the kernel function K and the bandwidth h. For f supported on $(0, \infty)$ the hazard rate is

$$\lambda(x) = \frac{f(x)}{1 - F(x)},$$

and its cumulative is

$$\Lambda(x) = \int_0^x \lambda(u)\, du.$$

Watson and Leadbetter (1964a) showed, but see for example Proposition 1 of Shorack and Wellner (1986) for a much different way to arrive at the same conclusion, that a natural estimate of the hazard rate is

$$\begin{aligned} \widehat{\lambda}_h(x) &= n^{-1} \sum_{i=1}^{n} \frac{K_h(x - X_i)}{1 - F_n(X_i)} \qquad (1) \\ &= n^{-1} \sum_{i=1}^{n} \frac{K_h\left(x - X_{(i)}\right)}{1 - F_n\left(X_{(i)}\right)} \\ &= n^{-1} \sum_{i=1}^{n} \frac{K_h\left(x - X_{(i)}\right)}{1 - i/n} \\ &= \sum_{i=1}^{n} \frac{K_h\left(x - X_{(i)}\right)}{n - i}, \end{aligned}$$

where the $X_{(i)}$ are the order statistics, with $X_{(1)} \leq \cdots \leq X_{(n)}$.

Derivatives are estimated by differentiation. The density derivative, $f'(x)$, is estimated by

$$\widehat{f}'_h(x) = n^{-1} \sum_{i=1}^{n} K'_h(x - X_i), \qquad (2)$$

and the hazard rate derivative, $\lambda'(x)$, is estimated by

$$\widehat{\lambda}_h'(x) = n^{-1}\sum_{i=1}^{n}\frac{K_h'(x - X_i)}{1 - F_n(X_i)} = \sum_{i=1}^{n}\frac{K_h'\left(x - X_{(i)}\right)}{n - i},$$

where

$$K_h'(x) = \frac{\partial}{\partial x}K_h(x) = \frac{1}{h^2}K'\left(\frac{x}{h}\right).$$

The variance of the density derivative estimate is:

$$var\left(\widehat{f}_h'(x)\right) = var\left(n^{-1}\sum_{i=1}^{n}K_h'(x - X_i)\right) = n^{-1}var\left(K_h'(x - X_i)\right),$$

the variance factor of which is estimated by the sample variance

$$\begin{aligned} s^2\left(K_{0,1}', ..., K_{0,n}'\right) &= n^{-1}\sum_{i=1}^{n}\left(K_{0,i}'\right)^2 - \left(n^{-1}\sum_{i=1}^{n}K_{0,i}'\right)^2 && (3) \\ &= n^{-1}\sum_{i=1}^{n}\left(K_{0,i}'\right)^2 - \left(\widehat{f}_h'(x)\right)^2 && (4) \end{aligned}$$

where

$$K_{0,i}' = K_h'(x - X_i).$$

Using the approximation

$$F_n(x) \approx F(x), \tag{5}$$

the variance of the derivative hazard rate is approximated by

$$\begin{aligned} var\left(\widehat{\lambda}_h'(x)\right) &= var\left(n^{-1}\sum_{i=1}^{n}\frac{K_h'(x - X_i)}{1 - F_n(X_i)}\right) \\ &\approx var\left(n^{-1}\sum_{i=1}^{n}\frac{K_h'(x - X_i)}{1 - F(X_i)}\right) \\ &= n^{-1}var\left(\frac{K_h'(x - X_i)}{1 - F(X_i)}\right). \end{aligned}$$

Except for the fact that F is unknown the variance factor here could be estimated by the sample variance

$$s^2\left(K_{F,1}', ..., K_{F,n}'\right),$$

where for any cumulative distribution function $H(x)$, dependence on x and h is suppressed in the notation

$$K_{H,i}' = \frac{K_h'(x - X_i)}{1 - H(X_i)}. \tag{6}$$

Another application of (5) gives the approximation

$$s^2\left(K'_{F_n,1},...,K'_{F_n,n}\right). \tag{7}$$

This is an important point where there is a critical difference between this development, and simply using the reweighted data in ordinary SiZer. In particular, the variance factor (7), now appropriately uses the weights. Thus, an isolated point with a heavy weight is no longer flagged as significant, because the variance estimate also increases when the weights are heavier.

SiZer gets its "simultaneous inference" properties (i.e. it addresses the multiple comparison problem) using a "number of independent blocks" calculation done in Section 3 of Chaudhuri and Marron (1999). The basis of this is the Effective Sample Size:

$$ESS_h(x)=\frac{\sum_{i=1}^n K_h(x-X_i)}{K_h(0)}, \tag{8}$$

which measures the "number of points in each kernel window" (this is exactly true if K is the uniform density window). Correct adaptation to the hazard context requires yet another careful twist. Naive reweighting would suggest that denominators of $1-F(X_i)$ should be inserted. But the independent blocks calculation is based on the number of *independent* pieces of information, so instead the formula (8) should be retained in the same form.

Thus a hazard rate version of SiZer comes from modifying the density estimation version, replacing the terms

$$K'_h(x-X_i)$$

in (2) by

$$\frac{K'_h(x-X_i)}{1-F_n(X_i)},$$

and replacing the variables

$$K'_{0,i}$$

in (3) by

$$K'_{F_n,i}.$$

2.2 Censored Estimation Mathematics

The basic structure of a censored observation starts with an unobserved survival time T_i, and an unobserved censoring variable C_i. The observed information only includes the value of the one that happens first, together with an indicator of whether that is the survival time, or the censoring time. In particular, censored data comes in the form $(X_1,\delta_1),...,(X_n,\delta_n)$, where $X_i=\min(T_i,C_i)$ and $\delta_i=1(T_i\le C_i)$, where

$$1(T_i\le C_i)=\begin{cases}1 & \text{if } T_i\le C_i\\ 0 & \text{if } T_i>C_i\end{cases}.$$

The standard “random sample” assumption is that $T_1, ..., T_n$ are independent, identically distributed with cumulative distribution function $F(x)$ and that $C_1, ..., C_n$ are independent (and independent of the T_i), identically distributed with cumulative distribution function $G(x)$. Note that the cumulative distribution function of X_i, is $L(x)$, where the corresponding cumulative survival function is $\overline{L}(x) = \overline{F}(x)\,\overline{G}(x)$, using the notation $\overline{H}(x) = 1 - H(x)$, for any cumulative distribution function $H(x)$.

The goal is estimation of the survival probability density $f(x) = F'(x)$ and the corresponding hazard rate

$$\lambda(x) = \frac{f(x)}{\overline{F}(x)},$$

with cumulative

$$\Lambda(x) = \int_0^x \lambda(u)\,du.$$

The cumulative distribution functions F and G can be estimated by the Kaplan Meier (1958), i.e. Product Limit, estimators (also the maximum likelihood estimators) given by

$$\begin{aligned}\overline{F}_n &= \begin{cases} 1 - \prod\limits_{X_{(i)} \le x} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}}, & \text{if } x \le X_{(n)} \\ 0, & \text{if } x > X_{(n)}. \end{cases}, \\ \overline{G}_n &= \begin{cases} 1 - \prod\limits_{X_{(i)} \le x} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}}, & \text{if } x \le X_{(n)} \\ 0, & \text{if } x > X_{(n)}. \end{cases},\end{aligned}$$

where the $\left(X_{(i)}, \delta_{(i)}\right)$ are the order statistics version of the data with $X_{(1)} \le \cdots \le X_{(n)}$.

A natural kernel density estimate of f is

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} \frac{\delta_i K_h(x - X_i)}{\overline{G}_n(X_i)}.$$

For f supported on $(0,\infty)$ the corresponding estimate of the hazard rate is

$$\begin{aligned}
\widehat{\lambda}_h(x) &= n^{-1}\sum_{i=1}^{n}\frac{\delta_i K_h(x-X_i)}{\overline{G}_n(X_i)\,\overline{F}_n(X_i)} \\
&= n^{-1}\sum_{i=1}^{n}\frac{\delta_i K_h(x-X_i)}{\overline{L}_n(X_i)} \\
&= n^{-1}\sum_{i=1}^{n}\frac{\delta_{(i)} K_h\left(x-X_{(i)}\right)}{\overline{L}_n\left(X_{(i)}\right)} \\
&= n^{-1}\sum_{i=1}^{n}\frac{\delta_{(i)} K_h\left(x-X_{(i)}\right)}{1-i/n} \\
&= \sum_{i=1}^{n}\frac{\delta_{(i)} K_h\left(x-X_{(i)}\right)}{n-i}.
\end{aligned}$$

Note that these have a structure very similar to the hazard rate estimator (1), which is why it is straight forward to extend SiZer to these cases as well.

Derivatives are again estimated by differentiation. The density derivative, $f'(x)$, is estimated by

$$\widehat{f}'_h(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i K'_h(x-X_i)}{\overline{G}_n(X_i)}. \tag{9}$$

The hazard rate derivative, $\lambda'(x)$, is estimated by

$$\widehat{\lambda}'_h(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i K'_h(x-X_i)}{\overline{L}_n(X_i)}.$$

The variance of the density derivative estimate is:

$$var\left(\widehat{f}'_h(x)\right) = var\left(n^{-1}\sum_{i=1}^{n}\frac{\delta_i K'_h(x-X_i)}{\overline{G}_n(X_i)}\right) = n^{-1}var\left(\frac{\delta_i K'_h(x-X_i)}{\overline{G}_n(X_i)}\right),$$

and for the hazard rate

$$var\left(\widehat{\lambda}'_h(x)\right) = var\left(n^{-1}\sum_{i=1}^{n}\frac{\delta_i K'_h(x-X_i)}{\overline{L}_n(X_i)}\right) = n^{-1}var\left(\frac{\delta_i K'_h(x-X_i)}{\overline{L}_n(X_i)}\right).$$

Using the approximation methods leading to (7), these variance factors are estimated by

$$s^2\left(K'_{G_n,1},...,K'_{G_n,n}\right),$$

using again the notation (6), and by

$$s^2\left(K'_{L_n,1},...,K'_{L_n,n}\right)$$

respectively.

The Effective Sample Size follows in a similar spirit. Again the basis is the number of independent pieces of uncensored data, resulting in the formula

$$ESS_h(x) = \frac{\sum_{i=1}^{n} \delta_i K_h(x - X_i)}{K_h(0)}.$$

Thus the censored density and censored hazard rate version of SiZer come from modifying the density estimation version, replacing the terms

$$K_h'(x - X_i)$$

in (2) by

$$\frac{\delta_i K_h'(x - X_i)}{\overline{G}_n(X_i)}$$

and

$$\frac{\delta_i K_h'(x - X_i)}{\overline{L}_n(X_i)}$$

respectively, and by replacing the variables $K_{0,i}'$ in (3) by $K_{G_n,i}'$ and $K_{L_n,i}'$ respectively.

3 Fast Computation

Because SiZer relies on a large number of smooths, it is important to use a fast computational method. Several such are discussed by Fan and Marron (1994). The binned approach is especially well suited to SiZer.

Details of the binned implementation of $\widehat{f}_h'(x)$ are similar to those given in Chaudhuri and Marron (1998), which are based on those of Fan and Marron (1994), except that the kernels are now divided by appropriate cumulative distribution functions. In particular, for the equally spaced grid of points $\{x_j : j = 1, ..., g\}$, let the corresponding bin counts (computed by some method, we have always used the "linear binning" described in Fan and Marron (1994)) be $\{c_{0,j} : j = 1, ..., g\}$. Then for density SiZer

$$\widehat{f}_h'(x_j) \approx n^{-1}\overline{S}_0'(x_j),$$

where

$$\overline{S}_0'(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'}' c_{0,j'}$$

and

$$\kappa_{j-j'}' = K_h'(x_j - x_{j'}).$$

The approximated standard deviation of $\widehat{f}_h'(x_j)$, is

$$\widehat{sd}(x_j) = n^{-1/2}\sqrt{n^{-1}\sum_{j'=1}^{g}\left(\kappa_{j-j'}'\right)^2 c_{0,j'} - \left(n^{-1}\sum_{j'=1}^{g}\kappa_{j-j'}' c_{0,j'}\right)^2}.$$

The censored and hazard versions of SiZer require reconsideration of the linear binning algorithm. When a data point X_i is between grid points x_j and x_{j+1}, linear binning assigns weight

$$w_{i,j} = \frac{X_i - x_j}{x_{j+1} - x_j}$$

to the bin centered at x_j, and weight

$$w_{i,j+1} = \frac{x_{j+1} - X_i}{x_{j+1} - x_j}$$

to the bin centered at x_{j+1}, and weight 0 to all other bins. These result in bin counts

$$c_{0,j} = \sum_{i=1}^{n} w_{i,j}.$$

For a generic estimated cumulative distribution function H_n, these bin counts are replaced by

$$c_{H_n,j} = \sum_{i=1}^{n} \frac{w_{i,j}\delta_i}{\overline{H}_n\left(X_i\right)}.$$

This results in the binned approximation to the generic estimator:

$$n^{-1} \sum_{i=1}^{n} \frac{\delta_i K_h'\left(x - X_i\right)}{\overline{H}_n\left(X_i\right)} \approx n^{-1} \overline{S}'_{H_n}(x_j),$$

where

$$\overline{S}'_{H_n}(x_j) = \sum_{j'=1}^{g} \kappa'_{j-j'} c_{H_n,j'},$$

To similarly approximate $\widehat{sd}$, use

$$\widehat{sd}(x_j) = n^{-1/2} \sqrt{n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 c_{H_n,j'} \left(\frac{c_{H_n,j'}}{\widetilde{c}_{0,j'}}\right) - \left(n^{-1} \sum_{j'=1}^{g} \kappa'_{j-j'} c_{H_n,j'}\right)^2},$$

where the factor of $\left(\frac{c_{H_n,j'}}{c_{0,j'}}\right)$ in the second moment term gives the second factor of $\frac{1}{\overline{H}_n}$ that appears in the second moment. Finally, the binned version of the Effective Sample Size needs to be based on the *unadjusted* bin counts of the uncensored data

$$\overline{ESS} = \frac{\sum_{j'=1}^{g} \kappa_{j-j'} \widetilde{c}_{0,j'}}{K_h\left(0\right)},$$

where

$$\widetilde{c}_{0,j} = \sum_{i=1}^{n} w_{i,j} \delta_i$$

and

$$\kappa_{j-j'} = K_h(x_j - x_{j'}).$$

References

[1] Aarset, M. V. (1987) How to identify a bathtub hazard rate, *IEEE Transactions on Reliability*, R-36, 106-108.

[2] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807-823.

[3] Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant Survival data, *Journal of the American Statistical Association*, 72, $27 - 36$.

[4] González-Manteiga, W., Marron, J. S. and Cao, R. (1996) Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation, *Journal of the American Statistical Association*, 91, 1130-1140.

[5] Hess, K. R., Serachitopol, D. M. and Brown, B. W. (1999) Hazard function estimators: a simulation study, *Statistics in Medicine*, 18, 3075-3088.

[6] Jiang, J. C. and Doksum, K. (2003) On local polynomial estimation of hazard functions and their derivatives under random censoring, *Constance van Eeden Volume*, IMS.

[7] Kalbfleisch, J. D. and Prentice, R. I. (1980) *The statistical analysis of failure time data*, Wiley, New York.

[8] Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457-481.

[9] Kousassi, D. A. and Singh, J. (1997) A semiparametric approach to hazard estimation with randomly censored observations, *Journal of the American Statistical Association*, 92, 1351-1356.

[10] Lo, S. H., Mack, Y. P. and Wang, J. L. (1989) Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator, *Probability Theory and Related Fields*, 74, 461-473.

[11] Marron, J. S. and Padgett, W. J. (1987) Asymptotically optimal bandwidth selection for kernel density estimators from randomly right censored samples, *Annals of Statistics*, 15, 1520-1535.

[12] Mudholkar, G. S., Srivastava, D. K. and Kollia, G. D. (1996) A generalization of the Weibull distribution with application to the analysis of survival data, *Journal of the American Statistical Association*, 91, 1575-158.

[13] Müller, H. G. and Wang, J. L. (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics*, 50, 61-76.

[14] Patil, P. N. (1990) Automatic smoothing parameter selection in hazard rate estimation, PhD Dissertation, University of North Carolina, Institute of Statistics, Mimeo Series #2033.

[15] Patil, P. N. (1993) On the least squares cross-validation bandwidth in hazard rate estimation, *Annals of Statistics*, 21, 1792-1810.

[16] Rice, J. and Rosenblatt, M. (1976) Estimation of the log survivor function and hazard function, *Sankhya, A*, 38, 60-78.

[17] Sarda, P. and Vieu, P. (1991) Smothing parameter selection in hazard rate estimation, *Statistics and Probability Letters*, 11, 429-434.

[18] Stute, W. (1999) Nonlinear censored regression, *Statistica Sinica*, 9, 1089-1102.

[19] Shorack, G. R. and Wellner, J. A. (1986) *Empirical processes with applications*, Wiley, New York.

[20] Tanner, M. A. and Wong, W. H. (1983) The estimation of the hazard function from randomly censored data by the kernel method, *Annals of Statistics*, 11, 989-993.

[21] Watson, G. S. and Leadbetter, M. R. (1964a) Hazard Analysis, I, *Biometrika*, 51, 175-184.

[22] Watson, G. S. and Leadbetter, M. R. (1964b) Hazard Analysis, II, *Sankhya A*, 26, 110-116.