SiZer for Length Biased, Censored Density and Hazard Estimation

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

Jacobo de Uña - Álvarez
Department of Statistics and O. R.
University of Vigo
Campus Lagoas - Marcosende
36200 Vigo
Spain

February 18, 2001

Abstract

Length biased data require appropriate modification of smooth estimates of the density and hazard rate. The SiZer map is a visual device for summarizing a range of useful statistical inferences about the shape of the density and the hazard rate curves. In this paper, SiZer is adapted to correctly account for the length biased adjustment. Both the censored and uncensored cases are treated.

Footnote: J. S. Marron was supported by NSF Grant DMS-9971649, Jacobo de Uña - Álvarez was supported by DGES grant PB98-0182-C02-02 and the Xunta de Galicia Grant PGIDT00PXI20704PN

1 Introduction

When analyzing sample of times, a number of problems related to "loss of information", often emerge. These problems include length biasing and various types of censoring and truncation.

In a length biased sampling process the data appear in the sample with probabilities that are proportional to their lengths. Important examples of this situation arise naturally in sampling from stochastic processes such as queues, telephone networks and renewal processes. Other fields of potential application

are health sciences (e.g. analysis of survival times) and micro-economics (e.g. the study of unemployment times). As noted by Woodroofe (1985), the length bias phenomenon is closely related to that of truncation, which was been the object of much investigation in survival analysis. Some important references on curve estimation for length biased sampling (and more general selection bias models) include Vardi (1982, 1985), Horváth (1985), Bhattacharyya, Franklin, and Richardson (1988), Jones (1991), Ahmad (1995), El Barmi and Simonoff (2000) and Cristóbal and Alcalá (2000).

The presence of right censored data is one of the most common features of survival data. Right censoring is caused by the occurrence of a risk that precedes the end of the time period of analysis. Since the initial paper by Kaplan and Meier (1958), in which the time honored product-limit estimate of a survival function was first introduced, a large number of methods for addressing censored information in various scenarios have been proposed and studied. Recent contributions on nonparametric density and hazard estimation under right censoring include Tanner and Wong (1983), Marron and Padgett (1987), Lo, Mack and Wang (1989), Patil (1993), Müller and Wang (1994), González-Manteiga, Cao, and Marron (1996), Hess, Serachitopol and Brown (1999) and Jiang and Doksum (2000).

The main contribution of this paper, is the adaptation of the SiZer method, Chaudhuri and Marron (1999, 2000), to length biased density and hazard estimation, including the right censored case.

The SiZer method for length biased, censored hazard estimation is illustrated in Figure 1. The data are durations of unemployment (in years), for 700 married women, gathered from 1987 to 1997 in Galicia, Spain. These times are shown as black dots in Figure 1a. A quantity of interest is the "instantaneous chance of leaving unemployment", i.e. the hazard rate, $\lambda(x)$ (defined in Section 2.2 below). People leave unemployment when they are no longer actively seeking a job, either through finding a job, or for other reasons.

Analysis of these data is complicated by the two factors of length-bias and censoring discussed above. The times are right censored because some of the subjects were still unemployed at the end of their follow-up period (no woman was followed for more than 1.5 years). Censored times are shown as green dots in Figure 1a. Both the black (uncensored) and green (censored) dots are "jittered", i.e. a random height has been added for easy visual separation. The times are length biased, because the women are not followed continuously over time. Instead the surveys, from which these data came, are such that the chance of being in the sample is proportional to the length of time on unemployment.

The blue curves in Figure 1a are a family of smooths which are kernel hazard estimates (defined in (4) below), for a wide range of bandwidths. The small bandwidths result in a very wiggly hazard estimate, while the large bandwidths give a smooth result. The family of smooths shows "features at different levels of resolution". The thick black curve in Figure 1a is a member of this family, that shows the features expected in these data. There is an early period of increasing hazard rate, followed by a long period of decrease (those who have been job searching for some time, are less likely to leave unemployment soon).

But the family of smooths makes it clear that this impression is strongly driven by the chosen window width. In particular, a smaller bandwidth gives many probably spurious bumps, and a larger bandwidth smears out the peak, while moving it to the right.

SiZer addresses this confounding of the bandwidth with the needed statistical inference (which features in the smooths are "really there"?). Classical approaches to this problem involve a first step of choosing the "right" bandwidth. Unfortunately there has never been a consensus as to the "best" method of doing this, and few of them focus on the problem of statistical significance of features. In SiZer this problem is avoided by taking a "scale space view", which means considering the full range of bandwidths, as shown in the family in Figure 1a. A simple characterization of important features in the data, is given in terms of where the smooth increases, and where it decreases. For example at a bump, the smooth first increases, then decreases. SiZer combines this concept with statistical inference. In particular, a confidence interval for the derivative is computed, at each location, and for each bandwidth. If this confidence interval lies entirely above the origin, then the derivative is "statistically significantly increasing", and the color blue is used. If the confidence interval lies entirely below the origin, then the derivative is "statistically significantly decreasing", and the color red is used. When the confidence interval contains the origin, the sign of derivative is not clear, and the intermediate color of purple is used. The aggregation of these colors, over both locations and bandwidth (on the log scale), is the SiZer map, shown in Figure 1b.

The largest bandwidths are represented at the top of Figure 1b, and in the grossly oversmoothed members of the family in Figure 1a. The blue - purple - red pattern shows that these curves first increase, and then decrease, and that is important underlying structure at this level of resolution of the data. The smallest bandwidths appear at the bottom of Figure 1b, and in the very spiky family members in Figure 1a. The predominant SiZer color of purple shows that these spikes are not important underlying structure, and suggests that they are simply artifacts of the sampling process. The thick black curve in Figure 1a, corresponds to the white line in Figure 1b. The SiZer inference at that scale, shows that the initial increase is statistically significant (blue), followed by a purple region (near the peak, where the slope is not very different from 0), followed by a significant region of decrease (red), followed by an unclear region (purple). Note that in the second purple region (for about 3.5 to 4.5 years), the black curve increases slightly. Since the SiZer map does not become blue, it is suggesting that this increase is not strong enough to be different from increases explainable by the sampling variability.

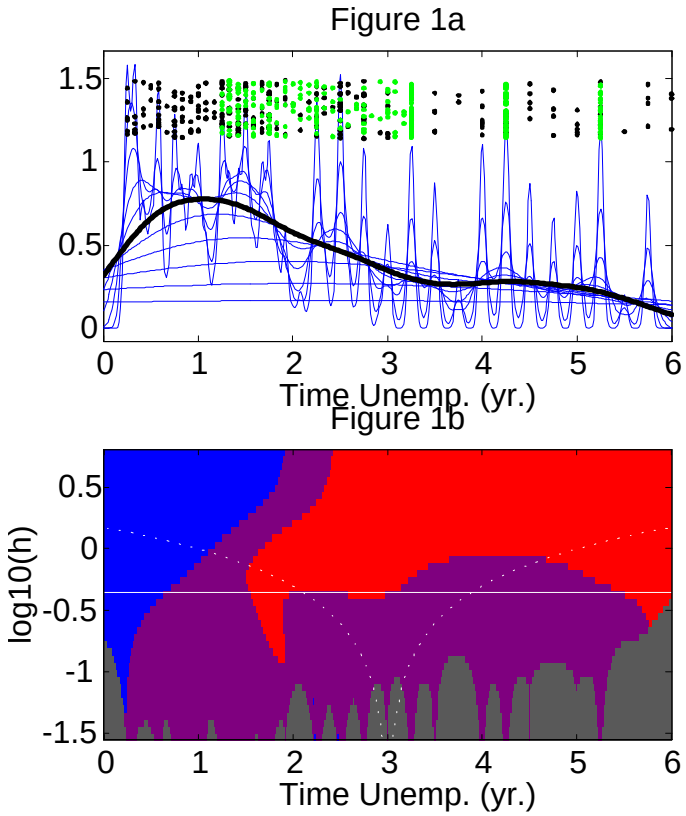

FIGURE 1: *Family of length-biased, censored hazard estimates (blue curves in part a) for the Galician Unemployment data, with corresponding SiZer map in part b.*

Note that for small bandwidths, sometimes the SiZer color of gray is used. This points out regions where there is not enough data present for reliable inference, as discussed in Section 2.4.2.

Both the shape of the smooths in Figure 1a, and the inferences given by the SiZer map in Figure 1b have been carefully adjusted to account for both the right censoring, and the length biased aspect, of these data. In Section 3, it is seen that when either of these issues is ignored, serious distortion occurs.

Because the SiZer map is a simultaneous display of the results of a large number of statistical inferences, it is very important to resolve the multiple comparison problem. This is done, using an "Effective Sample Size" approach, as in Chaudhuri and Marron (1999). The present modification of the Effective Sample Size is given in Section 2.4.2. Utility of SiZer is substantially enhanced by use of a fast binned implementation described in Section 2.4.3.

2 Mathematical development

This Section contains the details of the calculations underlying the new versions of SiZer proposed in this paper.

2.1 Length biased density estimation

Length biased density estimation is appropriate when a full data set is unavailable, and the chance of an observation appearing in the sample is proportional to its value. As noted by Cox (1969) and by Bhattacharyya, Franklin and Richardson (1988), this is equivalent to the estimation of a probability density $f(x)$, based on a random sample, $X_1^*, ..., X_n^*$, from the probability density $f^*(x) = \mu^{-1} x f(x)$, where $\mu = \int_0^\infty x f(x) dx$. Note that length bias only makes sense when the variable of interest X has support on $(0, \infty)$. A straightforward application of the conventional kernel density estimation method

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

(where $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$) estimates $f^*(x)$ not $f(x)$. A simple adjustment, proposed by Jones (1991), which yields an estimate of $f(x)$ is

$$\widehat{f}_h^0(x) = \frac{\mu_n^0}{n} \sum_{i=1}^{n} \frac{K_h(x - X_i^*)}{X_i^*} \tag{1}$$

where $\mu_n^0 = \left(n^{-1} \sum_{i=1}^n 1/X_i^*\right)^{-1}$, the harmonic mean of the data.

2.2 Length biased hazard estimation

The setting for length-biased hazard estimation is similar, except that now the goal is estimation of the hazard rate $\lambda(x) = f(x)/\bar{F}(x)$, where $\bar{F}(x) = 1 - F(x)$ is the complementary cumulative distribution function of $f(x)$. A direct length biased estimate of $\lambda(x)$ is obtained by convolution of the kernel function and an appropriate estimate of the cumulative hazard rate

$$\hat{\lambda}_h^0(x) = \int_0^\infty K_h(x - y) \frac{dF_n^0(y)}{\bar{F}_n^0(y)} = \frac{\mu_n^0}{n} \sum_{i=1}^{n} \frac{K_h(x - X_i^*)}{X_i^* \bar{F}_n^0(X_i^*)}, \tag{2}$$

where $\bar{F}_n^0(x) = 1 - F_n^0(x)$, the complement of the maximum likelihood estimator of $F(x)$, under length biased sampling, proposed by Vardi (1982), given by

$$F_n^0(x) = \frac{\mu_n^0}{n} \sum_{i=1}^{n} \frac{1_{\{X_i^* \le x\}}}{X_i^*}. \tag{3}$$

2.3 Censored versions

Censored estimation is needed when the data are not completely observed. I.e. the observed X_i are of the form $X_i = \min(X_i^*, C_i)$, where the X_i^* are length biased data as defined in Section 2.1, and the C_i are right censoring times, assumed to be independent of the X_i^*. In addition to X_i, it is also known whether the observation was censored or not, via the censoring indicator $\delta_i = 1_{\{X_i^* \leq C_i\}}$.

Natural density and hazard estimates, result again from convolutions of the kernel function, with appropriately modified cumulatives:

$$\hat{f}_h(x) = \int_0^\infty K_h(x-y)\, dF_n(y) = \frac{\mu_n}{n} \sum_{i=1}^n \frac{\delta_i K_h(x - X_i)}{X_i \bar{G}_n(X_i)},$$

$$\hat{\lambda}_h(x) = \int_0^\infty K_h(x-y)\, \frac{dF_n(y)}{\bar{F}_n(y)} = \frac{\mu_n}{n} \sum_{i=1}^n \frac{\delta_i K_h(x - X_i)}{X_i \bar{F}_n(X_i) \bar{G}_n(X_i)}, \tag{4}$$

where

$$\mu_n = \left(n^{-1} \sum_{i=1}^n \frac{\delta_i}{X_i \bar{G}_n(X_i)} \right)^{-1},$$

where F_n is the natural extension of Vardi's maximum likelihood estimate under right censoring, given by

$$F_n(x) = \frac{\mu_n}{n} \sum_{i=1}^n \frac{\delta_i 1_{\{X_i \leq x\}}}{X_i \bar{G}_n(X_i)},$$

where $\bar{F}_n(x) = 1 - F_n(x)$, and where $\bar{G}_n$ is the product limit estimator of the survival function of the censoring variable $\bar{G}$.

Note that in the case of no censoring, $F_n(x)$ reduces to $F_n^0(x)$ defined in (3) above, and consequently $\hat{f}_h(x)$ and $\hat{\lambda}_h(x)$ reduce to $\hat{f}_h^0(x)$ and $\hat{\lambda}_h^0(x)$, defined at (1) and (2), respectively.

The estimate F_n was proposed by de Uña-Álvarez (2000). That paper also provides asymptotic analysis including consistency and asymptotic normality for general empirical integrals.

2.4 SiZer specifics

This section contains the derivations that form the computational core of SiZer. We treat the most general case of censored, length biased density and hazard estimation, because other variations follow as consequences of this.

The statistical inference that underlies SiZer is based on derivatives. Following Jiang and Marron (2000), derivatives are naturally estimated by differentiation. The density derivative, $f'(x)$, is estimated by

$$\hat{f}'_h(x) = \frac{\mu_n}{n} \sum_{i=1}^{n} \frac{\delta_i K'_h(x - X_i)}{X_i \bar{G}_n(X_i)}$$

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

$$\hat{\lambda}'_h(x) = \frac{\mu_n}{n} \sum_{i=1}^{n} \frac{\delta_i K'_h(x - X_i)}{X_i \bar{F}_n(X_i) \bar{G}_n(X_i)}$$

where

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

2.4.1 Variance calculation

The variance of the density derivative estimate is

$$\begin{aligned} var\left(\hat{f}'_h(x)\right) &\simeq var\left(\frac{\mu}{n} \sum_{i=1}^{n} \frac{\delta_i K'_h(x - X_i)}{X_i \bar{G}(X_i)}\right) = \\ &= \frac{\mu^2}{n} var\left(\frac{\delta_i K'_h(x - X_i)}{X_i \bar{G}(X_i)}\right), \end{aligned}$$

where the approximations $\bar{G}_n \simeq \bar{G}$ and $\mu_n \simeq \mu$ are relevant and useful because their asymptotic rate of convergence is faster than that of $\hat{f}'_h(x)$. Except for the fact that $\bar{G}$ is unknown, the variance factor could be estimated by the sample variance

$$s^2\left(K'_{\bar{G},1}, ..., K'_{\bar{G},n}\right) = n^{-1} \sum_{i=1}^{n} \left(K'_{\bar{G},i}\right)^2 - \left(n^{-1} \sum_{i=1}^{n} K'_{\bar{G},i}\right)^2,$$

where, for any function H, the dependence on x and h is suppressed in the notation

$$K'_{H,i} = \frac{\delta_i K'_h(x - X_i)}{X_i H(X_i)}.$$

Finally we use the approximation

$$\frac{\mu^2}{n} s^2\left(K'_{\bar{G},1}, ..., K'_{\bar{G},n}\right) \simeq \frac{\mu_n^2}{n} s^2\left(K'_{\bar{G}_n,1}, ..., K'_{\bar{G}_n,n}\right).$$

The variance of the hazard rate estimate is

$$\begin{aligned} var\left(\hat{\lambda}_h^{'}(x)\right) &\simeq var\left(\frac{\mu}{n}\sum_{i=1}^{n}\frac{\delta_i K_h^{'}(x-X_i)}{X_i\bar{F}(X_i)\bar{G}(X_i)}\right) = \\ &= \frac{\mu^2}{n}var\left(\frac{\delta_i K_h^{'}(x-X_i)}{X_i\bar{F}(X_i)\bar{G}(X_i)}\right) \simeq \\ &\simeq \frac{\mu_n^2}{n}s^2\left(K^{'}_{\bar{F}_n\bar{G}_n,1},...,K^{'}_{\bar{F}_n\bar{G}_n,n}\right), \end{aligned}$$

using a sequence of approximations parallel to the above.

2.4.2 Effective Sample Size

The important simultaneous inference part of SiZer is based on the notion of Effective Sample Size. In conventional SiZer, this is essentially the "number of data points in each kernel window". Jiang and Marron (2000) showed that in the censored density and hazard estimation cases, careful definition of this Effective Sample Size is a pivotal issue.

Following their approach, we define

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

Note that when K is the uniform kernel, then this is the number of (uncensored) observations in the data window. An important point is that the counts are not given the usual Kaplan-Meier weights. Such weights are usually used to reflect what is known about the censored observations. It was seen in Jiang and Marron (2000) that SiZer with such weighting in ESS found spurious structure, that was not representative of important underlying structure. For the same reason, no adjustment factor of X_i^{-1}, which allows length biased weighting of observations, appears in ESS.

The ESS is also used to find "SiZer regions where there is not enough data present for inference", shaded gray in Figure 1b. Here the simple Binomial rule "$np > 5$" is used.

2.4.3 Fast computation

Because SiZer relies on a large number of smooths, effective data analysis is facilitated by use a fast computational method. The linear binning described in Fan and Marron (1994) is especially well suited to SiZer. Details of the binned implementation of $\hat{f}_h^{'}(x)$ and $\hat{\lambda}_h^{'}(x)$, are similar to those given in Chaudhuri and Marron (1999), except that the kernels now are divided by appropriate cumulative distribution functions and by the length of datum, X_i.

In particular, for the equally spaced grid of points $\{x_j : j = 1, ..., g\}$ define

the bin counts

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

$$c_{\bar{G}_n,j} = \sum_{i=1}^{n} \frac{\delta_i w_{ij}}{X_i \bar{G}_n(X_i)}$$

and

$$c_{\bar{F}_n \bar{G}_n,j} = \sum_{i=1}^{n} \frac{\delta_i w_{ij}}{X_i \bar{F}_n(X_i) \bar{G}_n(X_i)},$$

where $w_{ij} = \left(1 - \frac{|X_i - x_j|}{\Delta}\right)_+$, for $i = 1, ..., n$ and $j = 1, ..., g$, with Δ denoting the bin width, i.e. the grid spacing. Then for density SiZer

$$\hat{f}'_h(x_j) \simeq \frac{\mu_n}{n} \sum_{j'=1}^{g} \kappa'_{j-j'} c_{\bar{G}_n,j'},$$

where $\kappa'_{j-j'} = K'_h(x_j - x_{j'})$. The approximated standard deviation of $\hat{f}'_h(x_j)$ becomes

$$\widehat{sd}_f(x_j) = \mu_n n^{-1/2} \sqrt{n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 \left(c_{\bar{G}_n,j'}\right)^2 / c^0_{1,j'} - \left(n^{-1} \sum_{j'=1}^{g} \kappa'_{j-j'} c_{\bar{G}_n,j'}\right)^2},$$

where the frist term under the square root comes from the approximations

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \left(\frac{\delta_i K'_h(x_j - X_i)}{X_i \bar{G}_n(X_i)}\right)^2 &\simeq n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 \sum_{i=1}^{n} \frac{\delta_i w_{ij'}}{X_i^2 \bar{G}_n(X_i)^2} \simeq \\ &\simeq n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 \frac{c^0_{1,j'}}{(x_{j'})^2 \bar{G}_n(x_{j'})^2} \simeq \\ &\simeq n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 \left(c_{\bar{G}_n,j'}\right)^2 / c^0_{1,j'}. \end{aligned}$$

Similarly for hazard rate SiZer,

$$\hat{\lambda}'_h(x_j) \simeq \frac{\mu_n}{n} \sum_{j'=1}^{g} \kappa'_{j-j'} c_{\bar{F}_n \bar{G}_n,j'},$$

and the approximated standard deviation of $\hat{\lambda}_h^{'}(x_j)$ is

$$\widehat{sd}_\lambda(x_j) = \mu_n n^{-1/2} \sqrt{n^{-1} \sum_{j'=1}^{g} \left(\kappa'_{j-j'}\right)^2 \left(c_{\bar{F}_n \bar{G}_n, j'}\right)^2 / c^0_{1,j'} - \left(n^{-1} \sum_{j'=1}^{g} \kappa'_{j-j'} c_{\bar{F}_n \bar{G}_n, j'} \right)^2}.$$

Finally the binned version of the Effective Sample Size is

$$\overline{ESS}_h(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} c^0_{1,j'} / K_h(0),$$

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

3 Application

In this section, we revisit the Galician Unemployment data, analyzed in Figure 1. In particular it is seen that both the censoring and the length biased adjustments are critical.

Figure 2 shows the result of a naive estimate of the hazard rate, λ, when information about both the censoring and the length bias is ignored. The family of smooths have very large spikes at a few time points. The reason is that many observations take on these values. Careful comparison with Figure 1a reveals that almost all of these data points are censored (shown in green there). This means that they should be "at some larger time", and not bunched up. The Kaplan-Meier weights provide a sensible redistribution of these points, over the large durations. The censored versions of smoothers given in Section 2.3 correctly incorporate these weights to appropriately adjust for this, which is why the smooths in Figure 1a do not exhibit these massive spikes.

The lack of adjustment for censoring also has a major impact on the SiZer map in Figure 2b. In particular, on the lower right (more than about 3 years, and small bandwidth), the strong blue - red patterns suggest that these spikes are important features of the data. This inference would be correct if these were independent identically distributed data from a probability density, as assumed by this version of SiZer. Under that assumption, the large number of data piled up in these locations is indeed strong evidence of major spikes in the density. But, with the additional information that these data have been censored, it is clear that careful adjustment is needed to learn more about the data.

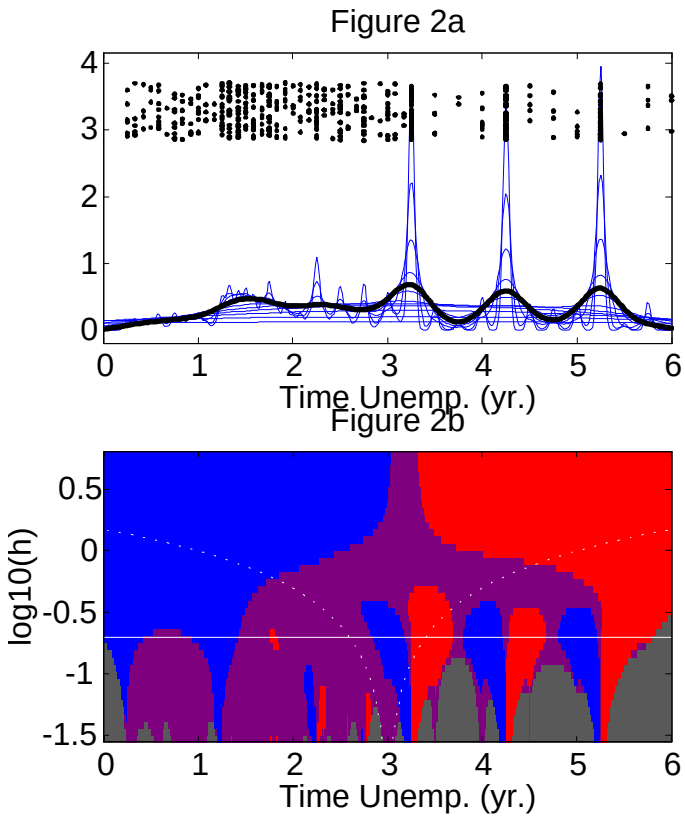

FIGURE 2: *Direct hazard estimation (part a), plus SiZer, for the Galician Unemployment data, with no adjustment for length bias or censoring.*

Figure 2 shows that the censoring adjustment dominates the analysis of these data. But the length bias adjustment is also important, as shown in Figure 3, where this is not done (although correction is done for censoring). This time the data in the jitter plot in Figure 3a are again colored green to reflect censoring (same as is in Figure 1a).

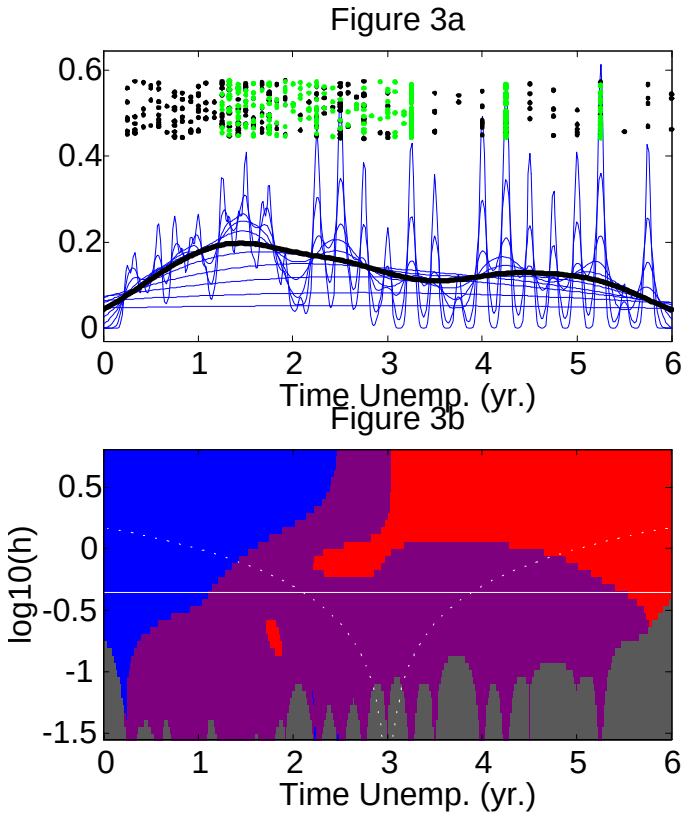

FIGURE 3: *Hazard estimation (part a), plus SiZer, for the Galician Unemployment data, with censoring adjustment, but no length bias adjustment.*

A careful look at the vertical axes in Figure 1a and Figure 3a, shows that the estimated hazard rates in Figure 3a (maxima around 0.2 - 0.4) tend to be much less than in Figure 1a (maxima around 0.8 - 1.2). Also the peak has been shifted to the right. The white bars in Figures 1b and 3b, and the black curves in Figures 1a and 3a, all represent the same bandwidth (i.e. level of smoothing), $\log_{10}(h) = -0.35$. Note that here also the hazard rate is severely underestimated without the length biased adjustment, and also the SiZer inference is substantially decreased. In particular, at this level of smoothing, the SiZer map is only red (significant decrease) on the far right.

These examples show that the issues of both length bias, and also censoring are important, and that careful adjustment is worthwhile.

Easy exploration of these issues and more, can be done using the movie versions of Figures 1, 2 and 3. These are available in the web directory: http://www.unc.edu/depts/statistics/postscript/papers/marron/SZLenBias/.

4 Acknowledgments

We are very grateful to María Soledad Otero Giráldez and Gema Álvarez Llorente, of the University of Vigo, for providing the Galician Unemployment data.

References

[1] Ahmad, I. A. (1995) On multivariate kernel estimation for samples from weighted distributions, *Statistics and Probability Letters*, 22, 121-129.

[2] Bhattacharyya, B. B., Franklin, L. A. and Richardson, G. D. (1988) A comparison of nonparametric unweighted and length-biased density estimation of fibres, *Communications in Statistics, Series A*, 17, 3629-3644.

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

[4] Chaudhuri, P. and Marron, J. S. (2000) Scale view of curve estimation, *Annals of Statistics*, 28, 408-428.

[5] Cox, D. R. (1969) Some sampling problems in technology, in *New Developments in Survey Sampling*, ed. N. L. Johnson and H. Smith, Wiley, New York, 506-527.

[6] Cristóbal, J. A. and Alcalá, J. T. (2000) Nonparametric regression estimators for length biased data. *Journal of Statistical Planning and Inference* 89, 145-168.

[7] El Barmi, H. and Simonoff, J. S. (2000) Transformation-based density estimation for weighted distributions, *Journal of Nonparametric Statistics*, 12, 861-878.

[8] Fan. J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, Journal of Computational and Graphical Statistics, 3, 35-56.

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

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

[11] Horváth, L. (1985) Estimation from a length-biased distribution, *Statistics and Decisions*, 3, 91-113.

[12] Jones, M. C. (1991) Kernel density estimation for length biased data, *Biometrika*, 78, 511-520.

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

[14] Jiang, J. C. and Doksum, K. (2000) On local polynomial estimation of hazard functions and their derivatives under random censoring, unpublished manuscript.

[15] Jiang, J. and Marron, J. S. (2000) SiZer for censored density and hazard estimation, unpublished manuscript.

[16] 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.

[17] 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.

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

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

[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] de Uña Álvarez, J. (2000) Product limit estimation for length-biased censored data, unpublished manuscript.

[22] Vardi, Y. (1982) Nonparametric estimation in the presence of length bias, *Annals of Statistics*, 10, 616-620.

[23] Vardi, Y. (1985) Empirical distributions in selection bias models, with discussion, *Annals of Statistics*, 13, 178-205.

[24] Woodroofe, M. (1985) Estimating a distribution function with truncated data, *Annals of Statistics*, 13, 163-177.