Estimating survival from partially observed data

Xun Zhang
Division of Clinical Epidemiology, Montreal General Hospital
Masoud Asgharian
Department of Mathematics and Statistics, McGill University, Montreal
David B. Wolfson
Department of Mathematics and Statistics, McGill University, Montreal

Working draft, May 15, 2002

Abstract:

Often, in cross-sectional-follow-up studies, survival data are obtained from prevalent cases only. This sampling mechanism introduces lenght-bias. An added difficulty is that in some cases the times of the onset cannot be ascertained or are recorded with great uncertainty. Such was the situation in the Canadian Study of Health and Aging (CSHA), an ongoing nation-wide study of dementia conducted by Health Canada. This paper proposes methods to estimate the survival function nonparametrically, when the data are length-biased and only partially observed. By using the "forward recurrence times" only, we suggest how one can overcome the difficulty caused by missing onset times, while by using the "backward recurrence times" only, one can avoid the cost and effort of follow-up. We illustrate our methods through an application to data derived from the CSHA.

◀1 : ▶

1 Introduction

The Canadian Study of Health and Aging (CSHA) (CSHA 1994) is one of the largest epidemiological studies of dementia conducted so far. In the study, over 10,000 subjects were screened for cognitive impairment. A total of 1,132 people were identified with dementia, including Alzheimer's disease, as CSHA-1, the initial data collection process, which took place between February 1991 and May 1992. During this initial phase, dates of onset of symptoms were ascertained from caregivers. Those subjects with dementia were then followed for a further five-year period. CSHA-2 data collection started in January 1996 and ended by May 1997.

One might term the CSHA a cross-sectional-follow-up study. The observed survival times do not constitute a random sample from the true survival distribution. Rather, they are *length-biased* (Cox 1969); only those who have survived to the sampling time point (CSHA-1 in our example) have a chance of being selected into the study.

Consider a cohort of subjects who experience two major events: the initiating event and the terminating event. Let X denote the time interval between these two events, which is of primary interest. Suppose that the random variable X is independent of the ◀1|2▶initiating event. This is the case in many practical simulations. For example, in the CSHA, the initiating event corresponds to the onset of dementia, the terminating event to the death of the individual, and X to the survival time of an individual from onset. For the majority of dementias, survival times have remained relatively independent of dates of onset. One focus of CSHA was to estimate the survival distibution of subjects with an umbrella diagnosis of dementia, as well as the survival distributions of the subgroups of those diagnosed with possible Alzheimer's disease, probable Alzheimer's disease, and vascular dementia.

Wolfson et al (2001) analyzed the CSHA data in the presence of right censoring, and estimated the survival function of patients with dementia adjusting for length-bias. There are several difficulties that often arise when cross-sectional data are used fro survival analysis, some of which arose in the analysis of the CSHA data.

1. The survival times and the censoring times are not independent; in fact censoring is informative, which means that most of the ◀2|3▶standard results of surviavl analysis, established under the assumption of non-informative censoring, are not automatically valid. This drawback can be avoided when estimation of the survival function is of interest, by using a so-called "conditional" estimator (Wang 1991). An added benefit of conditioning is that the estimator obtained is robust against non-stationarity of the onset times (Wang 1991). Alternatively, under stationarity (our situation) Asgharian et al (2001) derived the nonparametric maximum likelihood estimator and its asymptotic.

2. There may either be complete or partial ignorance of the initiation date. In fact, among the 1,132 CSHA subjects, 185 of them, i.e., more than 10% of the data, have missing recorded dates of onset. If the full (possibly censored) observed survival times are to be used in the analysis, these 185 subjects must be excluded. Apart from a fairly large loss of information, their exclusion could be related to their survival. This would lead to an aditional bias, apart from length-bias. Further, since the onset of dementia is insidious, the dates of onset recorded cannot be precise.

◀3|4▶

We shall show how the difficulties laid out in 1 and 2 above may be overcome when basing our analysis on partially observed prevalence data. Recently, Helmer et al (2000, 2001) have proposed a method that accounts for onset date uncertainty, in incident data.

The main objective in this paper is to estimate the survival function, non-parametrically, in studies involving cross-sectional sampling, and in which part of the data could be missing or unreliable. We shall consider two ways in which our data may be incomplete: Type I data: The initiation dates are unkown or only known with great uncertainty. Type II data: Alternatively, assuming the initiation dates are known it may be desired to estimate the survival function based only on the current survival times collected as part of a cross-sectional study without follow-up. The latter type of study, of course, avoids the cost and effort of follow-up, and its analysis is discussed here because of its similarity with the analysis of Type I data.

For data of type I we are required to estimate the survival function using only the (possibly censored) "forward recurrence times", while for data of type II we must use only the "backward recurrence times". It may sometimes be assumed in length-biased sampling that the (unseen) initiating events follow a stationary Poisson process on a certain interval (Blumenthal 1967, Cox 1969). We shall refer to analyses based on this assumption as "unconditional methods"; this paper focuses on "unconditional methods", in contrast to "conditional methods," for fully observed survival times, that allow the initiating points to be arbitrary (Wang 1991). We shall exploit the key relationship between the forward recurrence time density and the unbiased survival function,

$$f_{FT}(x) = \frac{1 - F(x)}{\mu} = \frac{S(x)}{\mu} \tag{1}$$

where $F(x)$ is the distribution function of X and $S(x) = 1 - F(x)$ is the survival function. Equation (1(1.2)) is also well-known in the theory of renewal processes as the stationary ◀4|5▶forward and backward recurrence time densities (Resnick 1992, chapter 3). It is convenient that the "backward recurrence times" have the same distribution as the "forward recurrence times", owing to the stationary assumption (Cox 1969), which means that (1(1.2)) holds when we replace $f_{FT}(x)$ by $f_{BT}(x)$, the backward recurrence time density.

It is often reasonable to assume that we have random right censoring of the forward recurrence times. Thus the Kaplan-Meier estimator of the forward recurrence time survivor function retains its usual properties when based on the randomly right-censored [forward recurrence] times that arise from type I data.

A kernel estimator that averages this Kaplan-Meier estimator may be used to estimate $f_{FT}(x)$ and hence, finally, we may define

$$\hat{S}(x) = \frac{\hat{f}_{FT}(x)}{\hat{f}_{FT}(0)}, \tag{2}$$

using (1(1.2)). In similar fashion one may estimate $S(x)$ by using only the uncensored backward recurrence times and a kernel estimator of $f_{BT}(x)$ based on the empirical distribution function of the backward recurrence times.

In section 2 we introduce the terminology of this paper. Section 3 considers kernel density estimators. It is a well-known phenomenon that if a probability density function has bounded support then kernel density estimators are often biased at and near the boundary points of of the support. Since we need to estimate $f_{FT}(0)$, where 0 is the boundary point we must address this issue. A method of boundary correction for kernel density estimation, the reflection method (Zhang et al 1999), will be discussed in section 3. Section 4 outlines some algorithms that we use for estimation of the survival function, ◀5|6▶$S(x)$ as well as its mean and median. A bootstrap scheme is also given to estimate the variances of these estimators. A simulation study is performed in section 5, which demonstrates the plausibility of our method. Finally in section 6 we apply our methods to the CSHA data to estimate the survival function of patients with dementia, and compare our results with those found by using the full survival times after excluding those with missing dates of onset.

While much of the ensuing discussion is motivated by the problem of estimating survival with dementia from a cross-sectional follow-up, the methods will be seen to have much broader applicability.

◀7▶

2 Terminology and a proposed estimator

Let U and V be the times of meaningful initiating and terminating events, respectively, and let

$$X = V - U \tag{3}$$

denote the "lifetime". Suppose that the random variable X has the cumulative distribution function F with the probability density function f. Let μ be the (finite) mean of X, which is unknown.

In a cross-sectional study, suppose that for a random variable T, only an individual whose time of initiation $U' \leq T$ and whose failure time V' satisfies $V' > T$, will be observed. Let

$$X' = U' - V'. \tag{4}$$

◀7|8▶The random variable X' is left truncated, or length-biased, and T is called the left truncation time of X.

Let

$$T_F = V' - T \tag{5}$$

and let

$$T_B = T - U'. \tag{6}$$

Then T_F and T_B are called, respectively, the *forward* and *backward recurrence times* in analogy with their counterparts in renewal theory (Resnick 1992). Notice that

$$X' = T_B + T_F. \tag{7}$$

In cross-sectional sampling, under the stationary assumption (Wang, 1991) it is well-known that X had the p.d.f., X',

$$f_{LB}(x) = \frac{xf(x)}{\mu}, \tag{8}$$

where f is the density of X (see (3(2.1))) (Cox 1969).

◀8|9▶

Let C' be the censoring variable which measures the time interval from initiation U' to the time of censoring or of potential censoring, and let

$$Y' = \min(X', C'). \tag{9}$$

Further, write

$$C' = T_B + D', \tag{10}$$

where D' is the forward recurrence censoring time measured from the recruitment date. Since X' (see (7(2.5))) and C' have T_B in common, they are not, in general, independent. The censoring mechanism of X' is, therefore, informative, and, consequently, the Meier estimator is not the non-parametric maximum likelihood estimator for the survival function of the length-biased survival time (Vardi 1985).

Under the assumption that the initiating events follow a stationary Poisson process, which we shall term "the stationarity assumption", we shall exploit the relationship, (1(1.2)), between the forward recurrence time density and the unbiased survival function to estimate latter.

◀9|10▶

Since

$$S(z) = \mu f_{FT}(z), \qquad z \leq 0, \tag{11}$$

where $S(0) = 1$,

$$\mu = 1/f_{FT}(0). \tag{12}$$

It follows that $f_{FT}(z)$ is a decreasing function.

◀10|11▶

Let

$$Z = \min(T_F, D'),$$

the censored forward recurrence time, and suppose that D' is independent of T_F, so called random censoring. Let δ be the censoring indicator, which takes the value 1 if $T_F \leq D'$, and 0 otherwise.

Suppose that $T_{F1}, T_{F2}, \ldots T_{Fn}$ are n i.i.d. of forward recurrence times with p.d.f. (1(1.2)) and with corresponding censoring times $D'_1, D'_2, \ldots, D'_n$. We actually observe

$$\{(Z_i, \delta_i); i = 1, 2, \ldots, n\}, \tag{13}$$

where $Z_i = \min(T_{Fi}, D_i')$ and δ_i are the censoring indicators. From (11(2.10)) and (12(2.11)) a natural estimator for $S(z)$ is then

$$\hat{S}(z) = \hat{f}_{FT}(z)/\hat{f}_{FT}(0), \tag{14}$$

although it should be pointed out that this is not the nonparametric maximum likelihood estimator of $S(z)$. Our task is, therefore, to estimate the density function $f_{FT}(z)$ from the observed data, $\{(Z_i, \delta_i); i = 1, 2, \ldots, n\}$, for all $z \leq 0$, noting that

$$\hat{\mu} = 1/\hat{f}_{FT}(0) \tag{15}$$

would be an estimator of the mean μ of the unbiased lifetime X.

In section 3 we discuss point estimation of f_{FT} and later address the issue of how to estimate the variance of $\hat{S}(z)$.

◀11|12▶

3 Kernel density estimator

◀12|13|14▶

It $\tilde{G}$ is the empirical distribution function of sample $\{X_i; i = 1, 2, \ldots, n\}$, which is the nonparametric maximum likelihood estimator of G (NPLME), then the kernel density estimator of the p.d.f. g, of G, is given by,

$$\begin{aligned} \tilde{g}_n(x) &= K_{h_n} * d\tilde{G}(x) \\ &= \frac{1}{h_n} \int K(\frac{x-y}{h_n}) d\tilde{G}(y), \end{aligned} \tag{16}$$

for given kernel K (Devroye, 1984).

◀14|15▶

In the presence of right censoring, however, the expression (16(3.4)) suggests the use of the Kaplan-Meier estimator of G [,] to replace the empirical distribution function [,] $\tilde{G}$. Let $\tilde{G}$ denote the Kaplan-Meier estimator of the distribution function G from the censored data $\{(Y_i, \delta_i); i = 1, 2, \ldots, n\}$. That is, $\tilde{G} = 1 - \tilde{S}$, where $\tilde{S}$ [is] the Kaplan-Meier estimator of the survival function.

Indeed, it can be shown that the modified kernel estimator

$$\begin{aligned} \tilde{g}_n(x) &= K_{h_n} * d\tilde{G}(x) \\ &= \frac{1}{h_n} \int K(\frac{x-y}{h_n}) d\tilde{G}(y) \end{aligned} \tag{17}$$

is consistent for g (Padgett and McNichols 1984).

Equation (17(3.5)) is the key in the sequel.

The density of the forward recurrence times f_{FT}, defined by (1.2), has support $[0, \infty]$ with 0 as the only boundary point. Further, in order to use (14(2.13)) to estimate $S(t)$, estimation of $\mu = f_{FT}(0)$ is crucial. We shall, without loss of generality, therefore, focus on this special case, assuming that g has support $[0, \infty]$.

When the support of the p.d.f., g, has boundary points, the kernel estimator $\tilde{g}_n$ defined by (16(3.4)) is often seriously biased at and near the boundary points (Schuster 1985). ◀15|16▶Various methods have been developed to adjust for this bias at the expense of increasing the variance of the estimator. Among these, the most notable one, perhaps, is the reflection method (Schuster 1985, Zhang et al 1999): this consists of reflecting the density g to the other side of the boundary point 0 and modifying the estimator (16(3.4)) as

$$\tilde{g}_n^R = \frac{1}{h_n} \int \left(K(\frac{x-y}{h_n}) + K(\frac{x+y}{h_n}) \right) d\tilde{G}(y). \tag{18}$$

This estimator works well when $g'(0) = 0$, as an estimator of g. Zhang et al (1999) proposed an improved estimator which minimizes the Mean Square Error (MSE) at and near the boundary point. Their estimator is

$$\tilde{g}_n^t = \frac{1}{h_n} \int \left(K(\frac{x-y}{h_n}) + K(\frac{x+t(y)}{h_n}) \right) d\tilde{G}(y). \tag{19}$$

where t is the transform

$$t(y) = y + dy^2 + Ad^2y^3 \tag{20}$$

with $d = g'(0)/g(0)$, provided that g is differentiable at $x = 0$, that $g(0) > 0$, and $A > \frac{1}{3}$. Unfortunately, the choice of A requires the knowledge of $g''(0)$, the second order derivative of g at zero. If we have no information about $g''(0)$, we could take $A = 1$ to avoid serious bias. Zhang et al (1999) suggested that although d is unknown it can be estimated by

$$d_n = \frac{\log \tilde{g}_n(h_n) - \log \tilde{g}_n^0(0)}{h_n}, \tag{21}$$

◀16|17▶where $\tilde{g}_n(h_n)$ is the usual kernel estimator of g at h_n, defined by (16(3.4)), and

$$\tilde{g}_n^0(0) = \frac{1}{nbh_0} \sum_{i=1}^n K_{(0)}(\frac{-X_i}{bh_n}). \tag{22}$$

Here $K_{(0)}$ is a so-called endpoint kernel supported on $[-1, 0]$ satisfying

$$\int_{-1}^0 K_{(0)}(t)\,dt = 1, \quad \int_{-1}^0 tK_{(0)}(t)\,dt = 0, \quad \text{and} \quad \int_{-1}^0 t^2K_{(0)}(t)\,dt \neq 0,$$

and

$$b = \left(\frac{\left(\int t^2K(t)\,dt\right)^2 \int (K_{(0)}(t))^2\,dt}{\left(\int t^2K_{(0)}(t)\,dt\right)^2 \int (K(t))^2\,dt} \right)^{1/5}. \tag{23}$$

The function $\tilde{g}_n^0(0)$ is an estimator of g proposed by Zhang and Karunamuni (1998).

Thus, in the presence of censoring, a proposed kernel density estimator is

$$\hat{g}_n^t(x) = \frac{1}{h_n} \int \left(K(\frac{x-y}{h_n}) + K(\frac{x+t_n(y)}{h_n}) \right) d\hat{G}(y), \tag{24}$$

where

$$t_n(y) = y + d_n y^2 + A d_n^2 y^3$$

with the problematic A an unresolved difficulty.

$$d_n = \frac{\log \hat{g}_n(h_n) - \log \hat{g}_n^0(0)}{h_n}. \tag{25}$$

In particular, if $g'(0)$ is known to be 0, we propose,

$$\hat{g}_n^t(x) = \frac{1}{h_n} \int \left(K(\frac{x-y}{h_n}) + K(\frac{x+y}{h_n}) \right) d\hat{G}(y). \tag{26}$$

◀17|18▶

4 An algorithm and a bootstrap scheme

In this section we present an algorithm for estimating the survival function, $S(x)$, of the unbiased survival time X, using the estimator $\hat{S}(x)$ defined by (14(2.13)). We assume that the derivative $f'_{FT}(0)$ exists.

4.1 Algorithm

1. Given the observed data (13(2.12)) of censored forward recurrence times

$$\{(Z_i, \delta_i); i = 1, 2, \ldots, n\},$$

where $Z_i = \min(T_{Fi}, D_i)$ and δ_i are the censoring indicators, calculate the Kaplan-Meier estimator $\hat{G}$ of the distribution function of the forward recurrence times T_F. For convenience we assume $Z_1 \leq Z_2 \leq \ldots \leq Z_n$. Then

$$\hat{G}(x) = 1 - \prod_{i=1}^{k_x} (\frac{n-i}{n-i+1})^{\delta_i}.$$

Here k_x is the value of k such that $x \in [Z_k, Z_{k+1}]$.

2. Calculate the variance $\hat{\sigma}^2$ of $Z_1, Z_2, \ldots, Z_n$, and let the bandwidth

$$h_n = \hat{\sigma}(\frac{15e\sqrt{2\pi}}{8n})^{\frac{1}{5}} = 1.6644\hat{\sigma}n^{-\frac{1}{5}},$$

(see (Devroy 1984) for a discussion of bandwidth choice). ◀18|19▶Although our choice of bandwidths is probably not "optimal" because of the censoring, our experience with large data sets is that the optimality of bandwidths is not important.

3. Choosing an appropriate kernel K and a suitable endpoint kernel $K_{(0)}$, calculate the crude kernel density estimator (17(3.5))

$$\tilde{g}_n(x) = \frac{1}{h_n} \int K(\frac{x-y}{h_n}) d\tilde{G}(y)$$

for $x = h_n$ and the endpoint kernel density estimate (22(3.10)) at $x = 0$:

$$\tilde{g}_n^0(0) = \frac{1}{nbh_0} \sum_{i=1}^{n} K_{(0)}(\frac{-X_i}{bh_n}).$$

4. Estimate the derivative of $\log(f_{FT}(x))$ at $x = 0$, i.e., $f_{FT}'(0)/f_{FT}(0)$, by (25(3.13)):

$$d_n = \frac{\log \hat{g}_n(h_n) - \log \hat{g}_n^0(0)}{h_n}.$$

5. Estimate the density function for $x \leq 0$ by (24(3.12)):

$$\hat{f}_{FT}(x) = \frac{1}{h_n} \int \left(K(\frac{x-y}{h_n}) + K(\frac{x+t_n(y)}{h_n}) \right) d\hat{G}(y),$$

where $t_n(y) = y + d_n y^2 + d_n^2 y^3$. Let g_i denote the jump of $\hat{G}$ at Z_i. Since $g_i = 0$ if T_{Fi} is censored (i.e., when $T_{Fi} \neq Z_i$),

$$\begin{aligned} \hat{f}_{FT}(x) &= \frac{1}{h_n} \sum_{i=1}^{n} \left(K(\frac{x-T_{Fi}}{h_n}) + K(\frac{x+t_n(T_{Fi})}{h_n}) \right) g_i \\ &= \frac{1}{h_n} \sum_{i=1}^{n} \left(K(\frac{x-Z_i}{h_n}) + K(\frac{x+t_n(Z_i)}{h_n}) \right) g_i. \end{aligned}$$

◀19|20▶

6. Estimate the mean μ of the lifetime X by (14(2.13)):

$$\hat{\mu} = 1/\hat{f}_{FT}(0).$$

7. Combine steps 5 & 6 to finally obtain an estimator of the survival function $\hat{S}(x)$ by (15(2.14)):

$$\hat{S}(x) = \hat{\mu}\hat{f}_{FT}(x) = \hat{f}_{FT}/\hat{f}_{FT}(0), \qquad x \leq 0.$$

Remark: $d = f_{FT}'(0)/f_{FT}(0) = S'(0)$. In cases where $S'(0)$ is known, or can be approximated from prior knowledge of the distribution X, steps 3 & 4 can be omitted.

4.2 Bootstrap scheme

In order to find confidence intervals of the estimators derived above, we use the bootstrap (Davidson and Hinkley 1997). There are several bootstrap resampling schemes with regard to survival data. The most straightforward method is simply to resample from the observed pairs $\{(Z_1, \delta_i); i = 1, 2, \ldots, n\}$ (Efron 1981). The procedure is as follows: let $\hat{H}$ be the empirical distribution function on $\mathbb{R} \times \{0, 1\}$, of the observed n pairs; the distribution puts mass $1/n$ at each pair (Z_i, δ_i).

1. Draw a bootstrap sample $\{(Z_i^*, \delta_i^*); i = 1, 2, \ldots, n\}$ by independently sampling n times with replacement from the set $\{(Z_1, \delta_i); i = 1, 2, \ldots, n\}$. This is equivalent to drawing a random sample from $\hat{H}$.

◀20|21▶

2. Applying steps 1–7 listed in the previous section to these artificial bootstrapped data, calculate accordingly $\hat{\mu}^*$, $\hat{S}^*(x)$, and $\hat{m}^*$, the estimated mean, survival function, and median, respectively.

3. Repeat independently steps 1 & 2 N times, obtaining $\{\hat{\mu}_k; k = 1, 2, \ldots, N\}$, $\{\hat{S}_k^*(x); k = 1, 2, \ldots, N\}$, and $\{\hat{m}_k^*; k = 1, 2, \ldots, N\}$.

4. Calculate the bootstrap variances of $\hat{\mu}$, $\hat{S}(x)$, and $\hat{m}$, i.e., the sample variances respectively from those data obtained in step 3. Confidence intervals can be constructed based on normal theory.

5. Various other types of bootstrap confidence intervals, such as percentile, basic bootstrap, and studentized confidence intervals, can also be constructed based on the data obtained in setp 3 (Davidson and Hinkley 1997).

◀21|22▶

5 Simulations

Consider the one-parameter Gamma family

$$f_a(x) = \frac{x^{a-1}e^{-x}}{\Gamma(\alpha)}, \quad x \geq 0 \tag{27}$$

with parameter $\alpha > 0$. Let X be a random variable with density (5.1), mean

$$\mu = \mathrm{mean}(X) = \alpha \tag{28}$$

and survival function

$$S_\alpha(x) = \frac{1}{\Gamma(\alpha)} \int_x^\infty t^{a-1} e^{-t}\, dt, \quad x \geq 0, \tag{29}$$

which cannot, in general, be expressed in closed form. The density of the forward recurrence times induced by length-biased sampling, is, by (1(1.2)) and (28(5.2)),

$$g_a(x) = \frac{S_a(x)}{\mu} = \frac{S_a(x)}{\alpha}, \quad x \leq 0. \tag{30}$$

The goal of the following simulation study was to evaluate the performance of the nonparametric estimator, $\hat{S}(x)$, computed from the forward recurrence times only. If $\alpha < 1$, the density g has no derivative at 0, and so our method does not apply; if $\alpha = 1$, then (27(5.1)), (29(5.3)), and (30(5.4)) are all identical to the p.d.f. of the exponential distribution. Hence, we assumed $\alpha > 1$. In this case $g'(0) = 0$, and formula (26(3.14)) can be used for the estimation of (30(5.4)).

We began by examining the behaviour of the estimator $\hat{S}_\alpha(x)$ when there is no censoring.

Let $\{Y_i; i = 1, 2, \ldots, n\}$ be a random sample from probability density (30(5.4)), let

$$h_n = \hat{\sigma}(\frac{15e\sqrt{2\pi}}{8n})^{\frac{1}{5}} = 1.6644\hat{\sigma}n^{-\frac{1}{5}},$$

◀22|23▶the bandwidth, where $\hat{\sigma}_n^2$ is the sample variance of $\{Y_i; i = 1, 2, \ldots, n\}$, and let K be a kernel. Without censoring, the kernel density estimate of (30(5.4)) is

$$\hat{g}_\alpha(x) = \frac{1}{h_n}\sum_{i=1}^{n}\left(K(\frac{x - Y_i}{h_n}) + K(\frac{x + Y_i}{h_n})\right), \quad x \geq 0. \tag{31}$$

Our simulations consisted of three parts. In all simulations, samples of size 500 were used. To generate samples from density (30(5.4)), Von Neumann's rejection method (Devroye 1985, chapter 8) was employed. Three typical kernels were examined. They are the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},$$

the Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2), \quad -1 \leq x \leq 1,$$

and the box kernel

$$K(x) = \frac{1}{2} \quad -1 \leq x \leq 1.$$

As pointed out earlier, since the estimation of the density function at the boundary point $x = 0$, is crucial, the first part of the simulation was to investigate the estimation of the mean μ for different values of α:

$$\hat{\mu}_\alpha = \frac{1}{\hat{g}_\alpha(0)},$$

which also estimates α since $\alpha = \mu$. The summary results are reported in Tables 5.1–3 for the three different kernels given above. The value of $\overline{\hat{\mu}_\alpha}$ represents the average over 400 ◀23|24▶repetitions. The bias and the variance of $\hat{\mu}_\alpha$ were also estimated based on the 400 repetitions.

Table 5.1. Estimation of $\mu = \alpha$ with the Gaussian kernel

α	$\overline{\hat{\mu}_\alpha}$	Bias	Variance
1.5	1.5399	0.0399	0.0239
2	2.0374	0.0374	0.0510
3	3.0543	0.0543	0.1704
5	5.1752	0.1752	0.8811
10	10.6078	0.6078	9.8478

Table 5.2. Estimation of $\mu = \alpha$ with the Epanechnikov kernel

α	$\overline{\hat{\mu}_\alpha}$	Bias	Variance
1.5	1.5499	0.0499	0.0503
2	2.0573	0.0573	0.1153
3	3.1134	0.1134	0.4265
5	5.4065	0.4065	2.9459
10	10.9903	0.9903	21.4525

Table 5.3. Estimation of $\mu = \alpha$ with the box kernel

α	$\overline{\hat{\mu}_\alpha}$	Bias	Variance
1.5	1.5567	0.0567	0.0407
2	2.0466	0.0466	0.0999
3	3.1176	0.1176	0.4067
5	5.3701	0.3701	2.0475
10	10.9541	0.9541	20.0773

◀24|25▶

The second part of the simulation examined the behavior of the nonparametric estimator for the survival function $S_\alpha(x)$:

$$\hat{S}_\alpha(x) = \frac{\hat{g}_\alpha(x)}{\hat{g}_\alpha(0)}, \quad x \geq 0.$$

We plot 10 typical realizations of the estimator in Figures 5.1–3 with the three given kernels.

◀25|26▶

Figure 5.1. Estimates of $S_\alpha(x)$ with the Gaussian kernel

◀26|27▶

Figure 5.2. Estimates of $S_\alpha(x)$ with the Epanechnikov kernel

◀27|28▶

Figure 5.3. Estimates of $S_\alpha(x)$ with the box kernel

◀28|29▶

Remarks:

1. The modified kernel density estimator works very well for densities with substantial mass near the boundary point (Zhang et al 1999). This is the case for the densities g_α defined by (30(5.4)) with small values of α, as illustrated through Tables 5.1–3 and Figures 5.1–3. For larger α, there is much less mass

near $x = 0$. We demonstrate this by plotting the densities g_α for $\alpha = 2$ and $\alpha = 5$ in Figure 5.4. The mean and the variance of g_α can be easily calculated: $\mathrm{E}[g_\alpha(X)] = (\alpha+1)/2$ and $\mathrm{Var}[g_\alpha(X)] = (\alpha+1)(\alpha+5)/12$. With the value of α increasing, the mean of the sample is further away from 0 and the variance becomes larger. As a result, the bias and the variability of our estimators increase as α increases.

Figure 5.4. Densities g_α for $\alpha = 2$ & 5

2. Among the three kernels, the Gaussian kernel performs best. There is little difference between the Epanechnikov kernel and the box kernel.

◀29|30▶

3. For large values of α, our estimators still perform reasonably well when the sample size is large. This is illustrated in Table 5.4 and Figure 5.5, in which samples of size 1000 and the Gaussian kernel was used. The values in Table 5.4 were calculated over 400 repetitions.

Table 5.4. Estimation of $\mu = \alpha$ for large α

α	$\overline{\hat{\mu}_\alpha}$	Bias	Variance
10	10.4434	0.4434	3.5558
15	15.6831	0.6831	13.0304

Figure 5.5. Estimates of $S_{10}(x)$ from samples of size 1000

◀30|31▶

In the last part of the simulation, we investigated the eefect on the estimator due to censoring. Samples from the density g_α (see (30(5.4))), were randomly censored with an exponential random variable with mean λ and with censoring proportion set at about 15%. For instance, with $\alpha = 3$ and $\lambda = 12$, the censoring probability is about 0.15. In this case, formula (31(5.5)) is no longer valid. Instead, to estimate g_α, the density in (30(5.4)), we use the kernel density estimator

$$\hat{g}_\alpha(x) = \frac{1}{h_n n} \sum_{i=1}^{n} \left(K(\frac{x-z_i}{h_n}) + K(\frac{x+z_i}{h_n}) \right) g_i, \quad x \le 0,$$

where $Z_i, i = 1, 2, \ldots, n$, are the censored data, and g_i is the jump of the Kaplan-Meier estimator at Z_i (see (26(3.14))).

Again, samples of size 500 were taken, and estimation was based on 400 repetitions. Summary results of the estimator $\hat{\mu}_\alpha$ are reported in Table 5.5, and 10 realizations of the estimator $\hat{S}_\alpha$ are plotted in Figure 5.6, using the Gaussian kernel. We did this for moderate values of α. With large α, the sample size needs to be increased in order to achieve reasonble accuracy, as noted in Remark 1 above.

Table 5.5 Estimation of $\mu = \alpha$ in the presence of censoring

α	$\overline{\hat{\mu}_\alpha}$	Bias	Variance
3	3.0601	0.0601	0.0346
5	5.0863	0.0863	0.0946

◀31|32▶
Figure 5.6 Estimates of $S_\alpha(x)$ in the presence of censoring
◀32|33▶

6 Application to survival with dementia

In the Canadian study of Health and Aging (see Introduction), there were questions raised about the reliability of the recorded dates of onset of patients, and concerns over a substantial proportion of missing onset times. These concerns can be avoided by using the methods described in Sections 2 and 3. In the CSHA, the forward recurrence times are the time intervals from CSHA-1, the first phase of the study during the year of 1991, to failure; when there is censoring, they are the time intervals from CSHA-1 to loss of follow-up or to CSHA-2, when the study ended in 1996. Among the 1,332 patients diagnosed with dementia, a total number of 901 patients were classified into the categories of "probable Alzheimer's disease", "possible Alzheimer's disease", and "vascular dementia", with 433, 277, and 201 subjects in these categories, respectively. One of the goals of the study was the estimation of the survival distribution from onset, of individuals with dementia of these three types.

We first apply the methods described in Sections 2 and 3 by assuming that the derivative of the survival function at time zero is zero. This is essentially equivalent to the assumption that subjects cannot fail in some very small interval after onset, which seems reasonable by the nature of the disease.

We computed the estimator, $\hat{S}(t)$, of the unbiased survival distribution, $S(t)$, from these forward recurrence times, following the procedure proposed in 4.1, with the ◀33|34▶Gaussian kernel $K = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Then the bootstrap scheme described in 4.2, repeated 1,000 times, was used to calculate a 95% pointwise confidence band for $S(t)$. The results are summarized in Figure 6.1.

We further computed the estimated survival distributions for the three different diagnostic categories, respectively. They are plotted in Figure 6.2. There is little difference between the survival distributions of the groups "possible Alzheimer's disease" and "vascular dementia". Although the estimated survival probability of the group "probable Alzheimer's disease" is greater than those of the two other groups, the difference is not statistically significant.

Table 6.1 displays the estimated mean and median survival times of these groups, along with their respective 95% confidence intervals, which were computed over 1,000 bootstrap replications. Notice that all the confidence intervals of the three groups overlap.

Table 6.1. Estimated Mean and Median Survival Times by Diagnosis Category

Diagnosis	Total group	Probable Alzheimer's	Possible Alzheimer's	Vascular Dementia
Median survival time (months)	53	56	51	47
95% Confidence interval	$(47, 59)$	$(49, 63)$	$(41, 61)$	$(37, 57)$
Mean survival time (months)	61	67	62	56
95% Confidence interval	$(55, 67)$	$(58, 76)$	$(52, 72)$	$(46, 66)$

◀34|35▶

Figure 6.1. Estimated Survival Curve with 95% Confidence Band

Figure 6.2. Comparison of Survival Curves — By Diagnostic Categories

◀35|36▶

Although the survival probability at any time between 1 to 5 years post onset of dementia can be roughly approximated from Figures 6.1 and 6.2, we list the estimated survival probabilities at 1–5 years post onset, with their 95% confidence intervals, in Table 6.2.

Table 6.2. Estimated Survival Probabilities at 1–5 Years Post Onset of Dementia

Diagnosis	1 Year	2 Year	3 Year	4 Year	5 Year
Total group	0.95 (0.90, 1.00)	0.84 (0.73, 0.94)	0.70 (0.60, 0.80)	0.58 (0.48, 0.68)	0.39 (0.31, 0.48)
Probable Alzheimer's	0.99 (0.93, 1.00)	0.94 (0.78, 1.00)	0.81 (0.63, 0.98)	0.65 (0.50, 0.81)	0.43 (0.31, 0.56)
Possible Alzheimer's	0.94 (0.89, 0.99)	0.79 (0.66, 0.91)	0.65 (0.49, 0.81)	0.54 (0.39, 0.69)	0.38 (0.24, 0.51)
Vascular Dementia	0.92 (0.87, 0.97)	0.77 (0.63, 0.90)	0.64 (0.48, 0.80)	0.49 (0.34, 0.64)	0.30 (0.18, 0.42)

One of the drawbacks of using the forward recurrence times is that we are unable to estimate the distribution beyond the follow-up period. In our example, we cannot estimate the survival distribution beyond the 5-year period, the follow-up time period of the CSHA. We do not have this obstacle while working with the backward recurrence times.

To illustrate our method, we also used the backward recurrence times to estimate the survival distribution of patients with dementia. As mentioned in the introduction, among the subjects diagnosed in the CSHA, about 10% of them had missing dates of onset. Out of the total number of 901 subjects in the three categories of "probable Alzheimer's disease", "possible Alzheimer's disease", and "vascular dementia", 81 had no dates of ◀36|37▶onset recorded. Thus there are 820 backwards recurrence times, with 396, 251, and 173 in the three categories, respectively. Table 6.3 gives the summary results of the mean and median survival times, Figure 6.3 plots the survival distribution for the total group, along with its pointwise 95% confidence band, and Figure 6.4 plots the survival distributions for the three categories.

Table 6.3. Estimated Median Survival Times by Diagnosis Category

— Using Backwards Recurrence Times

Diagnosis	Total group	Probable Alzheimer's	Possible Alzheimer's	Vascular Dementia
Median survival time (months)	55	58	68	57
95% Confidence interval	$(50, 60)$	$(51, 65)$	$(59, 77)$	$(48, 66)$
Mean survival time (months)	66	68	82	67
95% Confidence interval	$(61, 71)$	$(61, 75)$	$(73, 91)$	$(57, 77)$

◀37|38▶

Figure 6.3. Estimated Survival Curve with 95% Confidence Band
— Using Backward Recurrence Times

Figure 6.4. Comparison of Estimated Survival Curves
— By Diagnostic Categories

◀38|39▶

There is considerable discrepancy between Figure 6.2 and Figure 6.4. From Figure 6.2 we see that patients in the category "probable Alzheimer's disease" have better survival than that of the patients in the other two categories; while from Figure 6.4 we find that the patients in the category of "possible Alzheimer's disease" have better survival. It is important to point out that it is not the methodology that causes the discrepancy but the two figures are plotted based on *different* data sets. This shows that the missing dates of onset were not *randomly* missing and we conclude that patients with missing onset may be different from the others. It is, therefore, dangerous to simply exclude those patients with missing onset in the study. We further illustrate this by overlaying Figure 6.1 and Figure 6.3 together in Figure 6.5, the two survival curves estimated based on forward and backward recurrence times, respectively, of the two different data sets. They are not comparable methodologically, but the comparison of the survival of two slightly different groups of patients.

If we wish to compare the results of the two methods, we should use the same set of subjects. Figure 6.6 overlays the two estimated survival curves by using forward and backward recurrence times, respectively, based on the same group of patients in the CSHA — a total of 820 subjects in the three categories "probable Alzheimer's disease", "possible Alzheimer's disease", and "vascular dementia", excluding those who had missing dates of onset. The close correspondence of the two curves shown in Figure 6.6 also supports our "stationary assumption" imposed at the very beginning, the asumption that the initiating events follow a stationary Poisson process. Under this assumption, the distributions of the forward and backward recurrence times are very similar.

◀39|40▶

Figure 6.5. Comparison of Survival Curves
— Forward and Backward Recurrence Times

Figure 6.6. Comparison of Survival Curves
— Forward and Backward Recurrence Times, Same Subjects

◀40|41▶

Notice that in Figure 6.6 there is a moderate separation of the two estimated survival curves beyond 30 months. Censoring of the forward recurrence times might account for the discrepancy.

Comparing our results, especially Table 6.1, with those found by Wolfson et al (2000) by using the full survival times after excluding those with missing dates of onset, our estimated median survival times are much longer, though all the corresponding 95% confidence intervals overlap. Their estimated median survival time for the total group was 40 months with a 95% confidence interval of (32, 48), and their estimated medians for the three categories, "probable Alzheimer's disease" "possible Alzheimer's disease", and "vascular dementia", were 38 (18, 58) months, 42 (28, 54) months, and 40 (28, 52) months respectively.

One possible explanation for this discrepancy could be that the asumption we made at the beginning of this section, that the derivative of the survival function at time zero is zero, is not accurate. To avoid this assumption, we have to estimate the derivative of the logarithm of the density of the forward recurrence times at zero, $f'_{FT}(0)/f_{FT}(0)$, following steps 3 and 4 in 4.1, In doing so we used the Gaussian kernel and the following endpoint kernel (recall (22(3.10))) proposed by Zhang and Karunamuni (1998):

$$K_{(0)}(t) = 12(1+t)(0.5+t)I_{[-1,0]}(t),$$

where $I_{[-1,0]}(t)$ denotes the indicator function of the interval $[-1, 0]$.

Summary results are shown in Table 6.4, which are very close to those estimates calculated by Wolfson et al (2000), although our methods are different from theirs since they used the full survival data. We also plot the estimated survival curve for the total ◀41|42▶group along with a 95% confidence band in Figure 6.7, as well as the estimated survival curves for the three categories, in Figure 6.8.

Table 6.4. Estimated Mean and Survival Times by Diagnosis Category

Diagnosis	Total group	Probable Alzheimer's	Possible Alzheimer's	Vascular Dementia
Median survival time (months)	44	51	41	41
95% Confidence interval	$(36, 52)$	$(39, 63)$	$(29, 53)$	$(31, 51)$
Mean survival time (months)	50	55	50	48
95% Confidence interval	$(43, 57)$	$(37, 73)$	$(40, 60)$	$(40, 56)$

Figure 6.7. Estimated Survival Curve with 95% Confidence Band

◀42|43▶

Figure 6.8. Comparison of Survival Curves
— By Diagnostic Categories

Observe that in Figue 6.8, the survival of patients in the category of "probable Alzheimer's disease" is much better than the other two groups, compared to what we have seen in Figure 6.2 and 6.4. There is no contradiction here since we get much wider 95% confidence intervals this time This is mainly due to the bias introduced by the estimation of $f'_{FT}(0)/f_{FT}(0)$ and the lack of knowledge of $f''_{FT}(0)$ (see (20(3.8))). This is the price we pay for not using the full information (full survival times); sometimes we do not have a choice. Notice that all the confidence intervals calculated in Table 6.4 overlap the corresponding confidence intervals presented in Table 6.1. This also suggests that the assumption, that the derivative of the survival function at time zero is zero, we made at the beginning, may not be unreasonable.

◀43|44▶

The same procedures can be also applied to the backward recurrence times. The estimated survival curves are plotted in Figures 6.9. Confidence intervals can be obtained by bootstrapping.

Figure 6.9. Estimated Survival Curves
— By Diagnostic Categories

◀44|45▶

7 Conclusion

The methods proposed in this thesis represent a first attempt to estimate a full survival distribution nonparametrically based only on forward or backward recurrence times. They allow us to estimate the full survival distribution without following up subjects. This would surely save cost and time. Moreover, backward recurrence times cannot be censored. Alternatively, by using only the forward recurrence times our methods allow us to estimate the survival distribution without knowing the onset times. For many situations this is the case. Finally our methods may be useful even if we had full data. For we could estimate the survival function of those with missing onset times and compare it with the survival function of those with observed onset times. Differences between these two estimated survival functions may indicate that those with missing onset times are not missing at random.

Our methods work very well with large data sets with substantial mass near the origin. The estimator (14(2.13)), $\hat{S}(t) = \hat{f}_{FT}(t)/\hat{f}_{FT}(0)$ (or $\hat{S}(t) = \hat{f}_{BT}(t)/\hat{f}_{BT}(0)$), is very sensitive to the estimation of the value of the density of the forward or backward recurrence times at the origin. It is always difficult to estimate a density function at a boundary point. When there are few observations near the origin, serious bias could occur. In particular, without the knowledge of the derivative of the survival function at time zero, we have to, in addition, estimate $f'_{FT}(0)/f_{FT}(0)$ (or its counterpart with the backward recurrence times). This could ◀45|46▶introduce further serious bias. Therefore, our methods should be used with caution when there are not enough data points near the origin.

◀46|47▶

References

References

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2000). Length-biased sampling with right censoring: an unconditional approach. To appear.

Blumenthal, S. (1967). Proportional sampling in life length studies. *Technometrics* **9**: 205–218.

Colett, D. (1994). *Modelling Survival Data in Medical Research.* Chapman & Hall.

Cox, D. R. (1969). Some Sampling Problems in Technology. In *New Developments in Survey Sampling.* Edited by Johnson & Smith. John Wiley & Sons.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* John Wiley & Sons.

CSHA (1994). Canadian Study of Health and Aging: Study methods and prevalence of dementia. *Canadian Medical Association Journals* 1994; **150**: 899–913.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications.* Cambridge University Press.

Devroye, L. and Gyrgi, L. (1985). *Nonparametric Density Estimation.* John Wiley & Sons.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**: 312–319.

Folland, G. B. (1984). *Real Analysis.* John Wiley & Sons.

◀47|48▶

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley and Sons.

McFadden, J. A. (1962). On the lengths of intervals in a stationary point process. *Journal of the Royal Statistical Society. Series B* **24**: 364–382.

Padgett, W. T. and McNichols, D. T. (1984). Nonparametric density estimation from censored data. *Communications in Statistics: Theory and Methods* **13**: 1581–1661.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**: 1065–1076.

Resnick, S. I. (1992). *Adventures in Stochastic Processes.* Birkhauser.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**: 832–837.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics: Theory and Methods* **14**: 1123–1136.

Stein, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press.

Vardi, Y. (1985). Empirical distributions in selection bias model. *Annals of Statistics* **13**: 178–205.

Wang M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86**: 130–143.

Wolfson, C., Wolfson, D. B., et al (2000). A reevaluation of the duration of survival after the onset of dementia. *The New England Journal of Medicine* **344**: 1111-1116.

Zhang, S., Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *Journal of Statistical Planning and Inference* **70**: 301–316.

◀48|49▶

Zhang, S., Karunamuni, R. J., and Jones, M. C. (1999). An improved estimator of the density function at the boundary. *Journal of the American Statistical Association* **94**: 1231–1241.