This is to certify that the

dissertation entitled

Estimation in Interval Censorship Models

presented by

Zhiming Wang

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

Joseph C. Gardiner
Major professor

Date October 28, 1993

ESTIMATION IN INTERVAL CENSORSHIP MODELS

By

Zhiming Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1993

ABSTRACT

ESTIMATION IN INTERVAL CENSORSHIP MODELS

By

Zhiming Wang

Interval-censored data arise in the analysis of survival times where there is only periodic assessment of subjects for outcomes of interest. We address the problem of estimating the survival distribution of the time of onset of some biologic event when its time of occurrence is not observed directly, but is known to have taken place in some time interval determined by the pattern of the examination times. This event time may also be right-censored if the event has not occurred at the time of the last assessment.

We introduce a statistical model incorporating the examination times and the event status at these times that indicates whether or not the outcome has already taken place. Two cases are considered, one in which the number of assessments that are made over the period of study is fixed, and in the other where this number is random. Identifiability of the distribution of the event time is proved.

A class of estimators of the survival function is proposed in an interval censorship model in which there is one examination time. These estimators are shown to satisfy certain self-consistency equations and this provides a means of their computation using the EM algorithm. The relationship between this class of estimators and the nonparametric maximum likelihood estimates is investigated. We establish the strong consistency of our estimators and derive the weak convergence of certain functionals.

A corresponding Bayesian estimator is constructed under the assumption of a Dirichlet process prior, and is shown that it does not necessarily converge to the nonparametric maximum likelihood estimator.

Simulation studies are presented under parametric assumption with two examination times. Comparisons are made between the nonparametric Turnbull estimate of the survival function and the maximum likelihood estimate under the parametric model.

To my wife Liqin and my daughter Cynthia

ACKNOWLEDGMENTS

I thank my advisor Professor Joseph C. Gardiner for his patient guidance and encouragement during the preparation of this dissertation.

I would also like to thank Professor R.V. Ramamoorthi for an introduction to some problems studied here and for fruitful discussions.

I extend my sincere thanks to Professor Habib Salehi and Shlomo Levental for having served on my guidance committee and for their comments.

Finally I would also wish to thank Mr. Alexander White for helping me to improve the language of this thesis.

TABLE OF CONTENTS

Chapter Page

1. INTRODUCTION AND SUMMARY 1

2. IDENTIFIABILITY 6
   - 2.1 Interval Censorship Model 6
   - 2.2 Model I 8
   - 2.3 Model II 12

3. STRONG CONSISTENCY AND WEAK CONVERGENCE 16
   - 3.1 Introduction 16
   - 3.2 Self-Consistency 18
   - 3.3 Strong Consistency 22
   - 3.4 Weak Convergence 27

4. BAYESIAN ESTIMATION 34
   - 4.1 Introduction 34
   - 4.2 Right Censoring 35
   - 4.3 Interval Censoring 38

5. SIMULATIONS 43

APPENDIX 49

REFERENCES 54

Chapter 1

INTRODUCTION AND SUMMARY

The past two decades have witnessed the rapid development of statistical methodology for the analysis of survival times from censored data. These data arise in several clinical, biomedical and epidemiological studies where outcomes of interest are response times, for example, time to tumor appearance, time to relapse, or time to death or failure. These endpoints, however, may not be observed in all subjects for various reasons. Subjects in the study may be lost to follow-up due to withdrawal or to the occurrence of an end point unrelated to the outcome of interest in the study. These data are turned right censored. The typical situation where right censorship occurs is when subjects enter a study at different (random) times and are followed until a specific endpoint is observed. The time of its occurrence T, measured from entry, will be right censored if by the time of termination of the study the event of interest has not taken place.

The situation where the event/outcome of interest may be subject to left or right censoring is generally referred to as double censoring. Leiderman (1973) describes a study of infant precocity in a group of Kenyan children in which subjects entered the study at different times and were tested periodically to ascertain whether they had acquired certain skills. The time of onset T of the development of these skills was of interest. We note that T is left censored if a child at the time of entry into the study has attained the desired level of development, and right censored if, by the last available assessment time, the child has not reached the goal.

Interval censorship arises where continuous monitoring of outcomes of interest is impractical, and assessments of the study subjects can be conducted only periodically. The precise time T at which the outcome occurred is not observed, but is known to have taken place within a specified time interval determined by the sequence of observation times $\{W_k : k \geq 1\}$. What is known is that the event occurred in some time interval $(W_k, W_{k+1}]$. Therefore, the event time T is said to be interval censored. For example, Peckham (1991) reported for the European Collaborative Study regarding the development of AIDS in 600 children born to mothers who tested positive for the human immune deficiency virus-1. These children were examined at birth and at 3 months intervals until 18 months of age, and again every 6 months thereafter until 4 years of age. Thus changes in their clinical status were not observed directly. They were seen only at the age at which disease was first detected and the age at the previous assessment and hence the time of onset of infection or disease is interval censored.

In each of the cases of censorship described here, a basic problem is the estimation of the distribution of T from a sample of observations under nonparametric circumstances. For right censored data, Kaplan & Meier (1958) introduced the product limit (PL) estimator, which has become the cornerstone for almost all analysis of right censored survival data. The properties of the PL estimator, its asymptotic theory and the statistical inferential procedures based on it, have been extensively investigated by several researchers. Turnbull (1974) introduced an algorithm for estimating survival curves from double censored data, and Chang & Yang (1987) described a statistical model under which identifiability of the distribution of T and the strong consistency of an estimator of it were established. The weak convergence of the estimator was proved by Chang (1990). Samuelson (1989) describes another approach to the problem of estimation using a formulation based on counting processes.

Turnbull (1976), improving on a method of Peto (1973), developed an algorithm for computing an estimate of the survival function from interval censored observations. His method of estimation is based on maximum likelihood considerations and yields a system of equations (self-consistent equations) that may be solved using the EM algorithm.

An investigation of the properties of the Turnbull estimator was made very recently by Groeneboom and Wellner (1992), who explored a connection in the construction of the nonparametric maximum likelihood estimator (NPMLE) and isotonic regression. They established the strong consistency and weak convergence of the NPMLE in two models of interval censorship.

The focus of this thesis is on three issues in interval censorship: Identifiability of the distribution of T in a general model of interval censorship, Consistency of a sequence of estimators of the distribution of T and a Bayesian approach to this estimation.

Chapter 2 introduces a general model of interval censorship that incorporates the entire pattern of potential examination times over the duration of a study. We consider two designs. The first permits only a fixed number of inspections of the units, while the other allows the number of assessments made during the follow-up period to be random. Within a nonparametric formulation and under some mild conditions, we prove the identifiability of the distribution of the event time and describe its relationship to other interval censorship models.

Chapter 3 studies the strong consistency and weak convergence of the estimate of survival function. The estimator is defined implicitly through two equations ((3.3) and (3.4)) and we demonstrate that their solution is equivalent to the solution of the self-consistency equation. Turnbull (1976) has suggested that a self-consistent estimator is a maximum likelihood estimator (MLE). However, we have found self-consistent estimators that depend on the initial mass assigned in the EM-algorithm and do not lead to maximum likelihood estimators. An example is given in section 3.2 showing we have different self-consistent estimates derived from one set of interval censored data. Strong consistency is proved for the class of self-consistent estimators under certain conditions which are also satisfied by the MLE. Therefore the MLE is also strongly consistent under these conditions.

In Section 3.4, we study the weak convergence of our estimator. Let $S(t)$ be the survival function of T and $S^{(n)}(t)$ be a self-consistent estimator of $S(t)$. For the double censorship model, Chang (1990) shows that $\sqrt{n}(S^{(n)}(t)-S(t))$ converges weakly to a Gaussian process. However, in our interval censoring model, we do not achieve such a property. Instead, we show that $\sqrt{n}\int_0^t (S^{(n)}(s)-S(s))ds$ converges weakly to a normal distribution.

In Chapter 4 we construct a Bayesian estimator of $S(t)$ under the assumption of a Dirichlet process prior. For right censored data, Susarla & van Ryzin (1976) demonstrated a nonparametric Bayesian solution to estimation under squared error loss. The resulting Bayesian estimator was shown to reduce to the PL estimator in the cases where $\alpha(R^+)\to 0$, where $\alpha(\cdot)$ is a finite measure on R^+ which serves as a parameter of the Dirichlet process prior. By following the process of construction of the Bayes estimate of $S(t)$ for right censored data, we find that, for right censored data, the estimator satisfies certain self-consistency equations. Several relationships between the Kaplan-Meier PL estimator and Bayes analog are obtained and an expression for this estimate under interval censorship is presented in section 4.3. It turns out that this Bayes solution does not necessarily reduce to the MLE as $\alpha(R^{+}) \to 0$.

Examples simulation studies are presented in Chapter 5. The underlying survival time T is assumed exponential and the inspection times are Gamma-distributed. We compare the Turnbull estimate of the survival function with that of the MLE obtained under the parametric assumption. For the former we employ our APL programs to carry out the task of solution of the self-consistency equations. The SAS LIFEREG procedure is used for the parametric model.

Chapter 2

IDENTIFIABILITY

2.1 Interval Censorship Model

Let T denote the time of occurrence of the event or response of interest, and let $\{W_k : k \geq 1\}$ be the increasingly ordered sequence of potential observation times, with all variables being measured from a common time origin. At the k-th examination time W_k, the event status is designated by the indicator $\delta_k = [\, T \leq W_k \,]$, that is $\delta_k = 1$ if the response has occurred by time W_k and $\delta_k = 0$ otherwise. We set $W_0 = 0$ and for each t, define

$$N(t) = \min \{\, k \geq 1: \ \delta_k = 1, W_k \leq t \,\}, \tag{2.1}$$

if this minimum exists and set $N(t) = +\infty$ otherwise. Also $M(t) = \max \{\, k \geq 1: \ W_k \leq \mathrm{t} \,\}$ is the last examination at or before t, which takes place at time W_M (see Figure 2.1). Since throughout this paper t is held fixed, we shall not exhibit the explicit dependence of the variables on t.

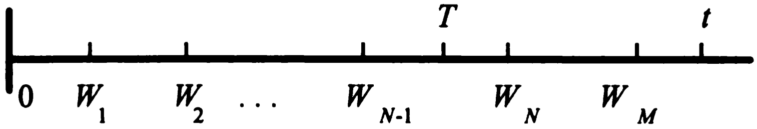

Figure 2.1

Under the assumption that $W_1 \leq t$ almost surely, $M(t)$ is finite, also N in (2.1) when finite, marks the first examination at which a positive diagnosis of the event occurring at the unobserved time T is made by time t. Then $L = W_{N-1}$ and $R = W_N$, define respectively the time of the last negative assessment and the time of the first positive assessment. In this case we have $T \in (L , R]$. On the other hand if $N = +\infty$, we only have the knowledge that $T > W_M$, which makes T right censored. Therefore the information available on T is that $T \in (W_{N-1}, W_N]$ when N is finite, and for $N = \infty$, $T > W_M$. Define $W_\infty = W_M$. The problem addressed in this chapter is the identifiability of the distribution of T on the basis of the datum (W_{N-1}, W_N, N, M). Specifically, if N' and M' are defined similarly for the nonnegative variable T' and the positive increasing sequence $\{W_k': k \geq 1\}$ then we wish to show that $\{T, W_k, k = 1, 2, \cdots\}$ and $\{T', W_k', k = 1, 2, \cdots\}$ have the same distribution on $[0, t]$, when the distributions of (W_{N-1}, W_N, N, M) and $(W_{N'-1}', W_{N'}', N', M')$ are the same.

Turnbull (1976) described a method of maximum likelihood estimation of the distribution of T based on data of the type $T \in (L , R]$, with R possibly infinite when T is right censored. Groeneboom and Wellner (1992) have developed estimation methods for two interval censoring models. In the simplest of these, called Case 1, there is a single examination time X, and the observable aspect of the model is the pair (X , δ) where $\delta = [T \leq X]$. This situation is obtained by taking $X = W_k$, for any examination time W_k. In the second case, called Case 2, there are two examination times X and Y, with $X < Y$, with the observable datum being $(X , Y , \gamma , \gamma ')$, where $\gamma = [T \leq X]$ and $\gamma ' = [T \in (X , Y]]$. In the context of periodic assessments Case 2 is also too wide, since the times X and Y can be obtained from the examination history in several ways. For instance, from any two examination times W_i, W_j with $W_i < W_j$, we get Case 2 by defining $X = W_i$ and $Y = W_j$. Our model, on the other hand, focuses on the entire examination pattern and defines the variables that are relevant.

For our identifiability problem, we consider two situations. In both the times of inspection are random. In the first, a fixed number of inspections is made in each subject, that is M is a constant. We refer to this case as 'Model I'. In the second situation, the follow-up period t is fixed and then the number of inspections M over the period $[0, t]$ is a random number. We refer to this case as 'Model II'.

Let $S_T(x) = P(T > x)$ denote the survival distribution of T, and $V_k = W_k - W_{k-1}$, $k \geq 1$. The basic assumptions that we make to establish our results are the following:

(A1) T is independent of $\{ W_k; k = 1, 2, \cdots \}$.

(A2) $\{V_k : k \geq 1\}$ are independent and nonnegative, with the density g_k of V_k being continuous and positive on $(0, \infty)$.

(A3) $0 < S_T(x) < 1$ for $x \in (0, \infty)$, and S_T is continuous.

The conditions (A1) and (A3) were used by Chang and Yang (1987) to prove identifiability in a model for double censorship. Identifiability of the distribution of T fails when the inspection times are degenerate; that is $g_k = 0$ a.s.. See Remark (2) at the end of the proofs for àn example. Therefore in this nonparametric framework we will not have identifiability if the inspections are made at fixed times. Only the probabilities assigned by the distribution of T to the fixed intervals $(W_{k-1}, W_k]$ can be identified.

2.2 Model I:

Random Inspection with Fixed Number of Inspections During Follow-up

Since M is assumed here to be fixed, let $M = m$ (≥ 1), where m is an integer. Also N is the first integer k for which $T \leq W_k$, or $N = \infty$ if no such k exists. We first define several sub-distributions that will be used in our proofs. All arguments x, x_1, $x_2 \in R^+ = [0, \infty)$.

Let

$$Q_1(x) = P(W_1 \le x, N=1),$$

$$Q_\infty(x) = P(W_m \le x, N=\infty),$$

and for $k \ge 2$,

$$Q_k(x_1, x_2) = P(W_{k-1} \le x_1, W_k \le x_2, N=k).$$

From our assumptions, W_k will have a density f_k which is continuous and positive on $[0, \infty)$. In fact $f_1 = g_1$ and $f_k(x) = \int_0^x f_{k-1}(u) g_k(x-u) du$, $k \ge 2$. Then Q_1, Q_∞ and Q_k for $k \ge 2$ can be expressed as follows:

$$Q_1(x) = \int_0^x \{1 - S_T(u)\} f_1(u) du, \tag{2.2}$$

$$Q_\infty(x) = \int_0^x S_T(u) f_m(u) du, \tag{2.3}$$

$$Q_k(x_1, x_2) = \int_0^{x_1} \int_u^{x_2} \{S_T(u) - S_T(v)\} f_{k-1}(u) g_k(v-u) dv du, \quad k \ge 2. \tag{2.4}$$

Our first theorem shows that T is identified from the datum (W_{N-1} , W_N, N).

Theorem 2.1

Under assumptions (A1)-(A3), S_T *is uniquely determined by* Q_1, Q_∞, *and* Q_k ($k = 2, \cdots, m$).

Proof: For the random variables T', W_k', $k \ge 1$, we define $V_k' = W_k' - W_{k-1}'$ ($W_0' = 0$) and N', M' as in (2.1). Let $S_1(x) = P(T' > x)$ and f_k', g_k' be the densities of W_k', V_k' respectively. Assume the conditions similar to (A1) - (A3), and suppose they produce the same sub-distributions Q_1, Q_∞, and Q_k ($k = 2, \cdots, m$) as in (2.2) - (2.4). We get

$$\{1 - S_T(x)\} f_1 (x) = \{1 - S_1(x)\} f_1'(x), \qquad (2.5)$$

$$S_T(x) f_m (x) = S_1(x) f_m' (x), \qquad (2.6)$$

$$\{S_T(x_1) - S_T(x_2)\} f_{k-1} (x_1)\, g_k(x_2 - x_1) = \{S_1(x_1) - S_1(x_2)\} f_{k-1}'(x_1)\, g_k'(x_2 - x_1), \qquad (2.7)$$

where $k = 2, 3, \cdots, m$. We will prove $S_1 \equiv S_T$.

If $m = 1$, we only have (2.5) and (2.6), and together these yield $f_1 (x) = f_1' (x)$. Hence $S_T = S_1$ since f_1 is assumed positive.

For $m \geq 2$, rewrite (2.5) - (2.7) in the form

$$\{1 - S_T(x)\}\ \{f_1 (x) - f_1'(x)\} = \{S_T(x) - S_1(x)\} f_1'(x), \qquad (2.8)$$

$$S_T(x)\{f_m' (x) - f_m(x)\} = \{S_T(x) - S_1(x)\} f_m'(x), \qquad (2.9)$$

$$\begin{aligned} &\{(S_T - S_1)(x_1) - (S_T - S_1)(x_2)\} f_{k-1} (x_1)\, g_k(x_2 - x_1) \\ &= \{S_1(x_1) - S_1(x_2)\}\ \{f_{k-1}'(x_1)\, g_k'(x_2 - x_1) - f_{k-1}(x_1)\, g_k(x_2 - x_1)\}, \end{aligned} \qquad (2.10)$$

where $k = 2, 3, \cdots, m$.

Let $h = S_T - S_1$. Then if $h \neq 0$ we must have one of the following two cases:
Case (I): $h(x) \geq 0$ (or symmetrically $h(x) \leq 0$), $x \in R^+$.
Case (II): There exist $x_1, x_2 \in R^+$, such that $h(x_1) > 0$ and $h(x_2) < 0$.
Now we prove both cases will lead to a contradiction.

In Case (I), since h is continuous and $h(0) = 0$, there exists $t_0 \in R^+$ such that $h(x) \geq 0$ for $x \in [0, t_0]$, and $h(x) < h(t_0)$ for $x \in [0, t_0)$. From (2.10) with $k = m$,

$$f_{m-1}'(x)\, g_m'(t_0 - x) < f_{m-1}(x)\, g_m(t_0 - x),\ \ x \in [0, t_0).$$

So, $$f_m'(t_0) = \int_0^{t_0} f_{m-1}'(x) g_m'(t_0 - x) dx$$

$$< \int_0^{t_0} f_{m-1}(x) g_m(t_0 - x) dx$$

$$= f_m(t_0). \tag{2.11}$$

Since $S_T(t_0) - S_1(t_0) > 0$, we get from (2.9)

$$f_m'(t_0) > f_m(t_0). \tag{2.12}$$

Hence (2.11) and (2.12) are contrary to each other. Symmetrically if $h \leq 0$, the proof is the same. Therefore Case (I) cannot hold.

In Case (II), since h is continuous and $h(0) = 0$, there exist t_0 , t_1 with $0 \leq t_0 < t_1$, such that $h(t_0) = 0$, $h(t_1) = 0$, and $h > 0$ on (t_0 , t_1) (or $h < 0$ on (t_0, t_1), in which case the proof is the same). Let us assume $t_0 > 0$, for otherwise we would essentially have Case (I) again.

By (2.8) and (A3) we have

$$f_1(t_0) = f_1'(t_0),\ f_1(t_1) = f_1'(t_1), \text{ and } f_1(x) > f_1'(x) \text{ for } x \in (t_0 , t_1). \tag{2.13}$$

Taking $x_1 = t_0$ and $x_2 \in (t_0 , t_1)$ in (2.10) with $k = 2$, we obtain

$$f_1'(x_1)\, g_2'(x_2 - x_1) < f_1(x_1)\, g_2(x_2 - x_1),$$

and in view of (2.13), also

$$g_2'(x) < g_2(x), \text{ for } x \in (0,\ t_1 - t_0). \tag{2.14}$$

Taking $x_1 \in (t_0, t_1)$and $x_2 = t_1$ in (2.10) with $k = 2$, and in view of (2.13), we obtain

$$f_1'(x_1)\, g_2'(x_2 - x_1) > f_1(x_1)\, g_2(x_2 - x_1)$$

$$> f_1'(x_1)\, g_2(x_2 - x_1).$$

So, $$g_2'(x) > g_2(x), \text{ for } x \in (0,\ t_1 - t_0). \tag{2.15}$$

Hence (2.14) and (2.15) are contrary to each other. Therefore Case (II) cannot hold.

Corollary 2.1

The identifiability above is total identifiability, that is $\{ f_k , g_k , k = 1, 2, \cdots , m \}$ *are also identified by* $Q_1 , Q_2 , \cdots , Q_m, Q_\infty$.

Proof: We need to prove $f_k \equiv f_k'$, $g_k \equiv g_k'$, $k = 1, 2, \cdots , m$.
We have already proved $S_T(x) = S_1(x)$ for $x \in R^+$. By (2.5) and (A3) we get $g_1 = f_1 = f_1' = g_1'$ on R^+. In (2.7) with $k = 2$, take $x_1 = 0, x_2 = x \in R^+$, then

$$\{1 - S_T(x)\} f_1 (0)\, g_2 (x) = \{1 - S_1(x)\} f_1'(0)\, g_2' (x),$$

and it follows that $g_2(x) = g_2'(x)$. From $f_2(x) = \int_0^x f_1(u) g_2(x-u)du$, we obtain $f_2 (x) = f_2' (x)$. By the same arguments and inductively we can establish $f_k (x) = f_k'(x)$ and $g_k (x) = g_k'(x)$ for $1 \le k \le m$.

2.3 Model II:

Random Inspection with Fixed Follow-up Period

In this case the number of inspections M made over the observational period $[0, t]$ is not fixed in advance. Then M becomes a random integer that is the total number of inspections made in this period. Note that N takes on values $1, 2, \cdots, M$ and $+\infty$. Let

$$Q (x_1 , x_2 , n , m) = P(W_{N-1} \le x_1, W_N \le x_2, N = n, M = m), \qquad (2.16)$$

where $0 \le x_1, x_2 \le t, 1 \le n \le m$ and $m \ge 1$; and

$$Q_\infty (x, m) = P (W_m \le x, N = \infty , M = m), \qquad (2.17)$$

where $0 \le x \le t$ and $m \ge 1$. The following theorem shows that T is identified from the datum (W_{N-1}, W_N, N, M).

Theorem 2.2

Under assumptions (A1) - (A3), *on* [0, *t*], S_T *is uniquely determined by the sub-distributions of* (2.16) *and* (2.17).

Proof: Let

$$Q_1(x) = P(W_1 \leq x, N = 1),$$

$$= \sum_{m=1}^{\infty} P(W_{N-1} \leq x_1, W_N \leq x, N = 1, M = m),$$

$$= \sum_{m=1}^{\infty} Q(x_1, x, 1, m),$$

and

$$Q_2(x_1, x_2) = P(W_1 \leq x_1, W_2 \leq x_2, N = 2),$$

$$= \sum_{m=1}^{\infty} Q(x_1, x_2, 2, m),$$

where $x, x_1, x_2 \in [0, t]$.

By the same argument leading to (2.2)-(2.4), Q_1 and Q_2 can be expressed as follows:

$$Q_1(x) = \int_0^x \{1 - S_T(u)\} f_1(u) du, \tag{2.18}$$

$$Q_2(x_1, x_2) = \int_0^{x_1} \int_u^{x_2} \{S_T(u) - S_T(v)\} f_1(u) g_2(v-u) dv du, \tag{2.19}$$

where $x, x_1, x_2 \in [0, t]$.

For $x \in [0, t]$, let $\overline{G}_2(x) = P(V_2 > x) = \int_x^{\infty} g_2(u) du$, then

$$Q_\infty(x, 1) = P(W_1 \leq x, T \geq W_1, M = 1),$$

$$= P(W_1 \leq x, T \geq W_1, W_1 + V_2 > t),$$

$$= \int_0^x P(T > u, V_2 > t-u) f_1(u) du,$$

$$= \int_0^x S_T(u) f_1(u) \overline{G}_2(t-u) du. \tag{2.20}$$

We will prove S_T is uniquely determined by Q_1, Q_2 and Q_∞.

As in the proof of theorem 2.1, for the random variables T', W_k', $k \geq 1$, we define $V_k' = W_k' - W_{k-1}'$ ($W_0' = 0$) and N', M' as in (2.1). Let $S_1(x) = P(T' > x)$ and f_k', g_k' be the density of W_k', V_k' respectively. Assume the conditions similar to (A1) - (A3), and suppose they produce the same sub-distributions Q and Q_∞ as in (2.16) and (2.17). From (2.18)-(2.20) we get

$$\{1 - S_T(x)\} f_1(x) = \{1 - S_1(x)\} f_1'(x), \tag{2.21}$$

$$\{S_T(x_1) - S_T(x_2)\} f_1(x_1) g_2(x_2 - x_1) = \{S_1(x_1) - S_1(x_2)\} f_1'(x_1) g_2'(x_2 - x_1), \tag{2.22}$$

$$S_T(x) f_1(x) \overline{G}_2(t - x) = S_1(x) f_1'(x) \overline{G}_2'(t - x), \tag{2.23}$$

where $x, x_1, x_2 \in [0, t]$, and $\overline{G}_2'(x) = P(V_2' > x) = \int_x^\infty g_2'(u)du$.

Since $\overline{G}_2$ and $\overline{G}_2'$ are survival distributions, taking $x = t$ in (2.23) yields

$$S_T(t) f_1(t) = S_1(t) f_1'(t). \tag{2.24}$$

Together with (2.21), this gives $f_1(t) = f_1'(t)$, and since f_1 and f_1' are positive, we also get $S_T(t) = S_1(t)$. Hence S_T and S_1 agree at zero and t. Suppose $S_T \neq S_1$ on $[0, t]$. There exist t_0 and t_1 with $0 \leq t_0 < t_1 \leq t$ such that $S_T(t_0) = S_1(t_0)$, $S_T(t_1) = S_1(t_1)$, and $S_T > S_1$ on (t_0, t_1) (or $S_T < S_1$ on (t_0, t_1)). The rest of the proof follows the same arguments as in Case (II) of section 2.3 in order to obtain a contradiction. Therefore $S_T \equiv S_1$ on $[0, t]$.

Corollary 2.2

The densities $\{f_k, g_k;\ k \geq 1\}$ *are also identified by the sub-distributions* Q *and* Q_∞.

Proof: The proof is the same as corollary 2.1, note we have the sub-distributions:

$$
\begin{aligned}
Q_k(x_1, x_2) &= P(W_{k\text{-}1} \le x_1, W_k \le x_2, N_t = k) \\
&= Q(x_1, x_2, k, k) + Q(x_1, x_2, k, k{+}1) + \cdots \\
&= P(W_{k\text{-}1} \le x_1, W_k \le x_2, W_{k\text{-}1} < T < W_k) \\
&= \int_0^{x_1} \int_u^{x_2} \{S_T(u) - S_T(v)\} f_{k-1}(u) g_k(v-u) dv du,
\end{aligned}
$$

where $x_1, x_2 \in [0, t]$, $k \ge 2$.

Remark:

(1). Theorem 2.1 and 2.2 will continue to hold if the assumption of continuity of the densities g_k in (A2) is replaced by either right or left continuity of the g_k.

(2). If $g_k = 0$ on some interval, then the identifiability fails. For example, in model I with a simple inspection ($m = 1$), suppose $f_1 = g_1 = 0$ on (1, 2) and S_T and S_1 coincide outside (1, 2), but differ on (1, 2). Then S_T and S_1 will produce the same sub-distributions Q_1 and Q_∞ in (2.2) and (2.3). A similar example for the two-inspection case ($m = 2$) is given by Chang & Yang (1987).

Chapter 3

STRONG CONSISTENCY AND WEAK CONVERGENCE

3.1 Introduction

In this chapter, we study the Case 1 interval censorship model, which is a special case of our Model I discussed in Section 2.2 with m=1. Let T be the failure time on $R^+ = [0, \infty)$ with survival function $S(t) = P(\ T > t\ )$, Y be the inspection time on $R^+ = [0, \infty)$ with survival function $S_Y(t) = P(\ Y > t\ )$, and $\delta=[T \leq Y]$. We observe n independent and identically distributed copies of (Y, δ), $\{(Y_i, \delta_i): i=1, 2, \cdots, n\}$.

Assume the following conditions to hold:

(B1) T and Y are independent.

(B2) S_Y is continuous, and $S_Y(s) - S_Y(t) > 0$, for $\forall\ 0 \leq s < t < \infty$.

(B3) S is continuous, and $S(s) - S(t) > 0$, for $\forall\ 0 \leq s < t < \infty$.

Define

$W_1(t)=P(Y>t, \delta=1)$,

$W_2(t)=P(Y>t, \delta=0)$, $t \in [0, \infty)$,

which can be written in terms of the survival functions as

$$W_1(t) = -\int_t^\infty (1-S(s))dS_Y(s), \tag{3.1}$$

$$W_2(t) = -\int_t^\infty S(s)dS_Y(s)\ . \tag{3.2}$$

Define the empirical survival distribution of the inspections $Y_1, \cdots, Y_n$

$$\tilde{S}_Y^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n}[Y_i > t],$$

the left censoring process

$$N_L(t) = \sum_{i=1}^{n}[Y_i \le t\ ,\ \delta_i = 1],$$

and the right censoring process

$$N_R(t) = \sum_{i=1}^{n}[Y_i \le t\ ,\ \delta_i = 0].$$

Let $W_1^{(n)}(t) = \frac{1}{n}(N_L(\infty) - N_L(t))$ and $W_2^{(n)}(t) = \frac{1}{n}(N_R(\infty) - N_R(t))$. N_L and N_R are counting processes. Note that $N_L(\infty)$ and $N_R(\infty)$ are finite, and $N_L(\infty)+N_L(\infty)=n$. We denote the integrals $\int_{(t,\infty)}$ by $\int_{t+}^{\infty}$, $\int_{[t,\infty)}$ by $\int_{t}^{\infty}$, and $\int_{[0,t]}$ by $\int_0^t$. Define $\frac{0}{0} = 0$. Let λ be the Lebesgue measure on $[0, \infty)$. A function on $[0, \infty)$ is said to be a sub-survival function if it is nonnegative non-increasing, and right continuous. It is called a survival function if additionally its values at 0 and ∞ are respectively, 1 and 0.

We construct the estimators of S and S_Y, through the following two equations (3.3) and (3.4) (see Remarks below).

$$W_1^{(n)}(t) + \int_{t+}^{\infty}\frac{1-S^{(n)}(t)}{1-S^{(n)}(s)}dW_1^{(n)}(s) = \int_{t+}^{\infty}S_Y^{(n)}(s-)\,dS^{(n)}(s)\ , \tag{3.3}$$

$$W_2^{(n)}(t) = -\int_{t+}^{\infty}S^{(n)}(s)dS_Y^{(n)}(s), \tag{3.4}$$

with $S^{(n)}(0) = S_Y^{(n)}(0) = 1$, $t \in [0, \infty)$.

In the next section, we will show that a solution $S^{(n)}$ exists and that it is a self-consistent estimator of S.

Remark:

(1) An obvious approach to obtaining estimators of S and S_Y is to replace all functions in (3.1) and (3.2) by their estimators, and search for a solution to the equations. For this approach, we would solve for $S^{(n)}$ and $S_Y^{(n)}$ from the equations:

$$W_1^{(n)}(t) = -\int_{t+}^{\infty}(1-S^{(n)}(s))dS_Y^{(n)}(s) \ ,$$

$$W_2^{(n)}(t) = -\int_{t+}^{\infty}S^{(n)}(s)dS_Y^{(n)}(s) .$$

This approach can be applied to right censorship and double censorship. Chang and Yang (1987) utilize this method in the latter case. However, in our case of interval censoring, if we solve for $S^{(n)}$ and $S_Y^{(n)}$ from these equations, then $S_Y^{(n)}$ will be the empirical of S_Y and $S^{(n)}$ will not be a survival function. The modification made in (3.3) leads to the appropriate solution of $S^{(n)}$ as a survival function.

(2) The solution $S_Y^{(n)}$ from (3.3) and (3.4) is not the empirical $\tilde{S}_Y^{(n)}$. Later in Section 3.3 we give its relationship to the empirical $\tilde{S}_Y^{(n)}$.

3.2 Self-Consistency

Given $\{ (Y_i, \delta_i): i=1,2,\cdots,n \} = \mathscr{A}_n$, computation of the conditional expectation $E(\frac{1}{n}\sum_{i=1}^{n}[T_i > t] \mid \mathscr{A}_n)$ gives $\iint \{\frac{F(t\wedge u)}{F(u)}[x \le u] + \frac{F(t)-F(t\wedge u)}{1-F(u)}[x > u]\} dP_n(x,u)$, where $P_n(x,u) = \frac{1}{n}\sum_{i=1}^{n}[T_i \le x, Y_i \le u]$, F = 1-S. A self-consistent estimator $S^{(n)}$ of S is a solution of the equation (self-consistency equation)

$$S^{(n)}(t) = E(\frac{1}{n}\sum_{i=1}^{n}[T_i > t] \mid \mathscr{A}_n),$$

where the right-hand side is evaluated at $S^{(n)}$. In this section we show the equivalence between the self-consistent estimator and the solution to the equations (3.3) and (3.4).

Theorem 3.1

Given the observations $\{ (Y_i, \delta_i)$: $i=1,2,\cdots,n \}$, *if* $S^{(n)}$ *is a self-consistent estimator of* S, *then there exists a survival function* $S_Y^{(n)}$, *such that* $S^{(n)}$ *and* $S_Y^{(n)}$ *are the solution of the equations* (3.3) *and* (3.4). *Conversely if* $S^{(n)}$ *and* $S_Y^{(n)}$ *are the solution of the equations* (3.3) *and* (3.4), *then* $S^{(n)}$ *is a self-consistent estimator of* S.

Proof: Suppose $S^{(n)}$ is a self-consistent estimator of S, then

$$1-S^{(n)}(t) = E_{S^{(n)}}\{ F_n(t) \mid Y_1,\cdots,Y_n;\delta_1,\cdots,\delta_n \}, \qquad (3.5)$$

where $F_n(t) = \frac{1}{n}\sum_{i=1}^{n}[T_i \le t]$ is the empirical of $F = 1\text{-}S(t)$, and $E_{S^{(n)}}$ is the expectation under the assumption that T_i have the survival distribution $S^{(n)}$ for any i. So

$$1\text{-}S^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n} E_{S^{(n)}}\{[T_i \le t] \mid Y_i,\delta_i\},$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{\frac{1-S^{(n)}(t\wedge Y_i)}{1-S^{(n)}(Y_i)}\delta_i + \frac{S^{(n)}(t\wedge Y_i)-S^{(n)}(t)}{S^{(n)}(Y_i)}(1-\delta_i)\},$$

$$= \frac{1}{n}[N_L(t) + \int_{t+}^{\infty}\frac{1-S^{(n)}(t)}{1-S^{(n)}(s)}dN_L(s) + \int_0^t \frac{S^{(n)}(s)-S^{(n)}(t)}{S^{(n)}(s)}dN_R(s)]. \qquad (3.6)$$

Recall the definition of $W_1^{(n)}$ and $W_2^{(n)}$, and that $N_L(\infty) + N_R(\infty) = n$. We get

$$S^{(n)}(t) = W_1^{(n)}(t) + W_2^{(n)}(t) + \int_{t+}^{\infty}\frac{1-S^{(n)}(t)}{1-S^{(n)}(s)}dW_1^{(n)}(s) - \int_0^t \frac{S^{(n)}(t)}{S^{(n)}(s)}dW_2^{(n)}(t). \qquad (3.7)$$

Define the estimator of S_Y by

$$S_Y^{(n)}(t) = 1 + \int_0^t \frac{1}{S^{(n)}(s)}dW_2^{(n)}(s), \quad t \in [0,\infty). \qquad (3.8)$$

The definition of $S_Y^{(n)}$ is valid, since $S^{(n)}(t) \ge W_2^{(n)}(t)$ by (3.7), and $\frac{0}{0} = 0$. Also, $S_Y^{(n)}$ is a right continuous non-increasing function on $[0, \infty)$ with $S_Y^{(n)}(0) = 1$.

Differentiating (3.8) yields

$$dS_Y^{(n)}(s) = \frac{1}{S^{(n)}(s)} dW_2^{(n)}(s), \text{ or } dW_2^{(n)}(s) = S^{(n)}(s) dS_Y^{(n)}(s). \tag{3.9}$$

By integration, this leads to (3.4).

Let $Y_{(n+1)} > Y_{(n)}$ be an arbitrary point to put on the remaining mass of $S^{(n)}$ and $S_Y^{(n)}$. Then $S^{(n)}(\infty)=S_Y^{(n)}(\infty) = 0$. This does not affect the self-consistency of $S^{(n)}$. So (3.7) becomes

$$S^{(n)}(t) = W_1^{(n)}(t) - \int_{t+}^{\infty} S^{(n)}(s) dS_Y^{(n)}(s) + \int_{t+}^{\infty} \frac{1-S^{(n)}(t)}{1-S^{(n)}(s)} dW_1^{(n)}(s) - S^{(n)}(t)\int_0^t dS_Y^{(n)}(s),$$

with $S_Y^{(n)}(0) = 1$. Using the fact that

$$d(UV) = VdU + U_{-}dV \tag{3.10}$$

for discontinuous functions U and V, this becomes

$$W_1^{(n)}(t) + \int_{t+}^{\infty} \frac{1-S^{(n)}(t)}{1-S^{(n)}(s)} dW_1^{(n)}(s) = -\int_{t+}^{\infty} S_Y^{(n)}(s-) dS^{(n)}(s), \tag{3.11}$$

which is (3.3). Hence $S^{(n)}$ satisfies equations (3.3) and (3.4).

Conversely, the entire argument above can be reversed. Suppose $S^{(n)}$ and $S_Y^{(n)}$ are the solutions of (3.3) and (3.4) with $S^{(n)}(0) = S_Y^{(n)}(0) = 1$ and $S^{(n)}(\infty) = S_Y^{(n)}(\infty) = 0$. Then we will get (3.8) from (3.4), and substituting it in (3.3) will lead to (3.7), which is the self-consistency equation (3.5).

The estimators $S^{(n)}$ and $S_Y^{(n)}$ defined through (3.3) and (3.4) is specified only at the observed points $\{Y_i\colon 1\le i\le n\}$. We extended $S^{(n)}$ and $S_Y^{(n)}$ to R^+ by making them right continuous step functions, which jump only at Y_i's. Unlike the right censoring and double censoring cases, even if we just concern ourselves with its values at the observed points, the self-consistent estimator from interval censored data is not unique. Both Turnbull (1976) and Groeneboom (1992) obtain a MLE by maximizing $\prod_{i=1}^{n} \{F(Y_i)\}^{\delta_i}\{1-F(Y_i)\}^{1-\delta_i}$. They show the estimator to be self-consistent and unique.

The following example shows that there exists a self-consistent estimator which is not the MLE.

Example 3.1

The following data, (1, 1), (2, 0), (3, 1), (4, 1), and (5, 0), n=5, with corresponding intervals (0, 1], (2, ∞), (0, 3], (0, 4] and (5, ∞) were analyzed using the APL program to solve the self-consistency equations. Two different estimators shown below were obtained (assume $Y_{(6)}= 6$ is the point to put on the remaining mass). It can be shown that both $S_1^{(n)}$ (Figure 3.1) and $S_2^{(n)}$ (Figure 3.2) are solutions to (3.3) and (3.4) and that $S_1^{(n)}$ is the MLE. Hence there exists self-consistent estimator, namely $S_2^{(n)}$, which is not MLE.

$$\begin{aligned} S_1^{(n)}(t) &= 1, \quad t \in [0, 1); \\ &= 1/2, t \in [1, 3); \\ &= 1/3, t \in [3, 6); \\ &= 0, \quad t \in [6, \infty); \end{aligned} \qquad \begin{aligned} S_2^{(n)}(t) &= 1, \quad t \in [0, 1); \\ &= 2/5, t \in [1, 6); \\ &= 0, \quad t \in [6, \infty). \end{aligned}$$

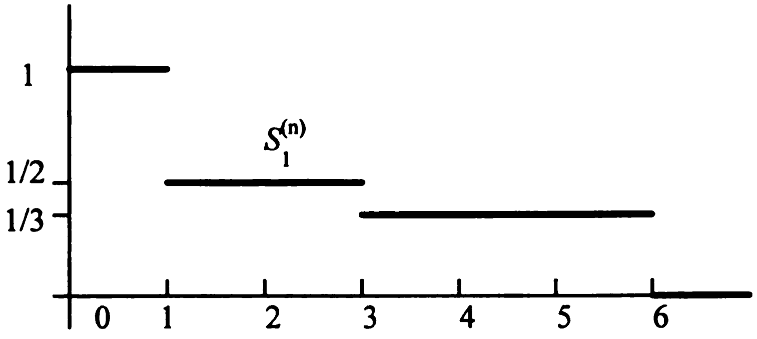

Figure 3.1

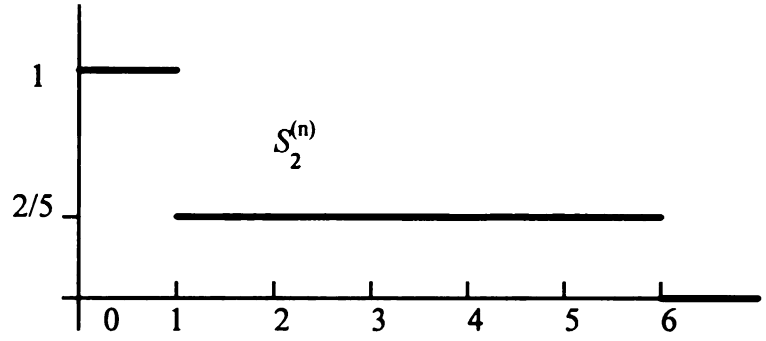

Figure 3.2

3.3 Strong Consistency

We already defined $S, S_Y, W_1, W_2, S^{(n)}, S_Y^{(n)}, W_1^{(n)}, W_2^{(n)}$, and $\tilde{S}_Y^{(n)}$. They are survival or sub-survival functions on [0, ∞), and their relationships are given in (3.1) - (3.4).

A self-consistent estimator $S^{(n)}$ is a non-increasing right continuous step function. Let $J_n = \{ t_k\!: \; k = 1, 2, \cdots, m\}$ be the collection of jump points of $S^{(n)}$. Let $0 = t_0 < t_1 <, \cdots, < t_m < \infty$. Then $\{ t_k\!: \; k = 1, 2, \cdots, m\} \subset \{ Y_i : \; i = 1, 2, \cdots, n\}$. Define

$$V_1^{(n)}(t) = W_1^{(n)}(t) + \int_{t+}^{\infty} \frac{1 - S^{(n)}(t)}{1 - S^{(n)}(s)} dW_1^{(n)}(s). \tag{3.12}$$

Then (3.3) becomes

$$V_1^{(n)}(t) = -\int_{t+}^{\infty} S_Y^{(n)}(s-) dS^{(n)}(s). \tag{3.13}$$

Since $W_1^{(n)}(t) + W_2^{(n)}(t) = \tilde{S}_Y^{(n)}(t)$, and using (3.9), we get

$$\begin{aligned} dW_1^{(n)}(t) &= d\tilde{S}_Y^{(n)}(t) - dW_2^{(n)}(t), \\ &= d\tilde{S}_Y^{(n)}(t) - S^{(n)}(t) dS_Y^{(n)}(t), \\ &= d(\tilde{S}_Y^{(n)}(t) - S_Y^{(n)}(t)) + (1 - S^{(n)}(t)) dS_Y^{(n)}(t). \end{aligned} \tag{3.14}$$

Differentiating both sides of (3.12) and using (3.10) yields

$$dV_1^{(n)}(t) = dW_1^{(n)}(t) - dS^{(n)}(t) \int_t^{\infty} \frac{dW_1^{(n)}(s)}{1 - S^{(n)}(s)} - \frac{1 - S^{(n)}(t)}{1 - S^{(n)}(t)} dW_1^{(n)}(t),$$

$$= -dS^{(n)}(t)\int_t^{\infty}\frac{dW_1^{(n)}(s)}{1-S^{(n)}(s)},$$

$$= -dS^{(n)}(t)\int_t^{\infty}\frac{d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))}{(1-S^{(n)}(s))}+dS^{(n)}(t)\int_t^{\infty}\frac{1-S^{(n)}(s)}{1-S^{(n)}(s)}dS_Y^{(n)}(s),$$

$$= -dS^{(n)}(t)\int_t^{\infty}\frac{d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))}{(1-S^{(n)}(s))}+dS^{(n)}(t)S_Y^{(n)}(t-). \tag{3.15}$$

Now using $dV_1^{(n)}(t) = S_Y^{(n)}(t-)dS^{(n)}(t)$ by (3.13), we get

$$dS^{(n)}(t)\int_t^{\infty}\frac{d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))}{(1-S^{(n)}(s))}=0\ ,\ t\geq 0. \tag{3.16}$$

Since { t_k: $k = 1, 2, \cdots, m$} are the points of jump of $S^{(n)}$, so $dS^{(n)}(t_k) \neq 0$ and (3.16) implies

$$\int_{t_k}^{\infty}\frac{d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))}{(1-S^{(n)}(s))}=0,\ k=1,2,\cdots,m. \tag{3.17}$$

Hence

$$\int\limits_{[t_{k-1},t_k)}\frac{d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))}{(1-S^{(n)}(s))}=0.$$

Since $S^{(n)}$ is constant on $[t_{k-1}, t_k)$,

$$\frac{1}{1-S^{(n)}(t_{k-1})}\int\limits_{[t_{k-1},t_k)}d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))=0\ , \text{ which implies}$$

$$\int_{t_k}^{\infty}d(\tilde{S}_Y^{(n)}(s)-S_Y^{(n)}(s))=0. \tag{3.18}$$

Therefore

$$S_Y^{(n)}(t_k-)=\tilde{S}_Y^{(n)}(t_k-),\quad k=1,2,\cdots,\ m\,. \tag{3.19}$$

That is the left limits of $S_Y^{(n)}$ and $\tilde{S}_Y^{(n)}$ agree at every point of jump of $S^{(n)}$.

To obtain the strong consistency of a sequence of self-consistent estimators $S^{(n)}$, we need to assign some condition on J_n which is the points of jump of $S^{(n)}$. Define

$A_{t,\varepsilon} = \{ \omega: \exists\, N$, such that for any $n > N, J_n(\omega)\cap(t - \varepsilon , t + \varepsilon)\neq \varnothing\}$.

Condition C: $P(A_{t,\varepsilon}) = 1$ for any $t > 0$ and every $\varepsilon > 0$.

The above condition also means that almost surely for any $t > 0$ and $\varepsilon > 0$, there exists $N = N(t, \varepsilon)$ such that for any $n > N$, $J_n(\omega)\cap(t - \varepsilon , t + \varepsilon)\neq \varnothing$. Namely, almost surely, in any neighborhood of t, we can always find points of jump of $S^{(n)}$ for n large enough.

Theorem 3.2

Under Condition C, $S^{(n)}$ uniformly converges to S almost surely. That is

$P(\lim\limits_{n\to\infty} \sup\limits_{t\in[0,\infty)} \left|S^{(n)}(t) - S(t)\right| = 0) = 1.$

Since $\{S^{(n)}(t)\}$and $\{S_Y^{(n)}(t)\}$ are series of survival functions, by Helly's theorem, there exists a sub-sequence $\{S^{(n')}(t),\ S_Y^{(n')}(t)\}$ and sub-survival functions $S^0(t)$ and $S_Y^0(t)$, such that $S^{(n')} \to S^0$ at continuous points of S^0, and $S_Y^{(n')} \to S_Y^0$ at continuous points of S_Y^0. With probability 1, $W_1^{(n)}(t)\to W_1(t)$, $W_2^{(n)}(t)\to W_2(t)$, and $\tilde{S}_Y^{(n)}(t)\to S_Y(t)$ uniformly for $t \in[0, \infty)$, since W_1 , W_2 and S_Y are continuous. Hence without loss of generality, we may assume uniform convergence on the whole space Ω. Further since we will show every sub-sequence has the same limit, we will also assume $\{n'\} = \{n\}$. To prove the theorem, we need a few lemmas.

Lemma 3.1

$S_Y^0(t)$ is continuous on $[0, \infty)$.

Proof: From (3.8), $S_Y^{(n)}(t) = 1 + \int_0^t \frac{1}{S^{(n)}(s)} dW_2^{(n)}(s)$, $t \in [0, \infty)$.

For continuity points of $S_Y^0(t)$ $0\leq s_1 < s_2$, we have

$$-(S_Y^{(n)}(s_2)-S_Y^{(n)}(s_1)) = -\int_{(s_1,s_2]} \frac{1}{S^{(n)}(s)} dW_2^{(n)}(s),$$

$$\le \frac{-(W_2^{(n)}(s_2)-W_2^{(n)}(s_1))}{S^{(n)}(s_2)},$$

$$\le \frac{-(W_2^{(n)}(s_2)-W_2^{(n)}(s_1))}{W_2^{(n)}(s_2)}, \tag{3.20}$$

where the last inequality follows from $S^{(n)}(s_2) \ge W_2^{(n)}(s_2)$ by (3.7). Letting $n \to \infty$

$$-(S_Y^0(s_2)-S_Y^0(s_1)) \le \frac{-(W_2(s_2)-W_2(s_1))}{W_2(s_2)}. \tag{3.21}$$

Hence S_Y^0 is continuous on [0, ∞) by the continuity of W_2. So, $S_Y^{(n)}(t) \to S_Y^0(t)$ uniformly for $t \in [0, \infty)$.

Lemma 3.2

$$W_2(t) = -\int_t^\infty S^0(s)\,d\,S_Y^0(s). \tag{3.22}$$

Proof: Let t be a continuity point of S^0, by (3.4),

$$W_2^{(n)}(t) = -\int_{t+}^\infty S^{(n)}(s)\,d\,S_Y^{(n)}(s),$$

$$= -\int_{t+}^\infty d[S^{(n)}(s)S_Y^{(n)}(s)] + \int_{t+}^\infty S_Y^{(n)}(s-)dS^{(n)}(s),$$

$$= S^{(n)}(t)S_Y^{(n)}(t) + \int_{t+}^\infty S_Y^{(n)}(s-)dS^{(n)}(s).$$

Lemma 3.1 implies $S_Y^{(n)} \to S_Y^0$ uniformly on [0, ∞). Hence

$$W_2^{(n)}(t) \xrightarrow{n\to\infty} S^0(t)S_Y^0(t) + \int_{t+}^\infty S_Y^0(s)dS^0(s),$$

$$= -\int_t^\infty S^0(s)\,d\,S_Y^0(s).$$

Since the continuity points of S^0 are dense in [0, ∞) and W_2 is continuous, therefore $W_2(t) = -\int_t^\infty S^0(s)dS_Y^0(s)$, $t \in [0, \infty)$.

Lemma 3.3

$S_Y^{(n)}$ converges to S_Y, that is $S_Y^0 \equiv S_Y$.

Proof: For $t \in [0, \infty)$, let $s_n \in J_n$ such that $|s_n - t| = \min\{|t_k - t|; t_k \in J_n\}$. Then under Condition C, $s_n \to t$. Remember $S_Y^{(n)}(s_n -) = \tilde{S}_Y^{(n)}(s_n -)$, so

$$S_Y^{(n)}(t) - S_Y(t) = [S_Y^{(n)}(t) - S_Y^{(n)}(s_n -)] + [\tilde{S}_Y^{(n)}(s_n -) - S_Y(s_n -)] + [S_Y(s_n -) - S_Y(t)],$$
$$\overset{\text{def.}}{=} I_{1n} + I_{2n} + I_{3n}$$

$I_{1n} \to 0$, by $S_Y^{(n)} \to S_Y^0$ uniformly and the continuity of S_Y^0 .

$I_{2n} \to 0$, by $\tilde{S}_Y^{(n)} \to S_Y$ uniformly on $[0, \infty)$

$I_{3n} \to 0$, by the continuity of S_Y .

So, $S_Y^{(n)}(t) \to S_Y(t)$ for $t \in [0, \infty)$.

Hence $S_Y^0 \equiv S_Y$ on $[0, \infty)$, since both S_Y^0 and S_Y are continuous.

Proof of theorem 3.2:

By (3.2), Lemma 3.2 and Lemma 3.3, we have $W_2(t) = -\int_t^\infty S(s)\,dS_Y(s)$ and $W_2(t) = -\int_t^\infty S^0(s)\,dS_Y(s)$, $t \in [0, \infty)$. That is $\int_t^\infty (S^0(s) - S(s))\,dS_Y(s) = 0$, $t \in [0, \infty)$. So S^0 $= S$ a.s. $[\lambda]$. But S^0 is right continuous and S is continuous, so $S^0 \equiv S$ on $[0, \infty)$. That is $S^{(n)} \to S$, the continuity of S makes the convergence uniform. Therefore

$$P(\lim_{n\to\infty} \sup_{t\in[0,\infty)} |S^{(n)}(t) - S(t)| = 0\ ) = 1.$$

Groeneboom (1992) studied the maximum likelihood estimator (MLE) of S, which he called the nonparametric MLE (NPMLE). He shows that the NPMLE is strongly consistent. We prove here that if $S^{(n)}$ is the NPMLE, then $S^{(n)}$ must satisfies Condition C and therefore strongly consistent. This is stated below.

Theorem 3.3

If $S^{(n)}$ is the MLE of S, then $S^{(n)}$ satisfies Condition C.

Proof: See Appendix.

3.4 Weak Convergence

In this section, we discuss the weak convergence of self-consistent estimators. Now assume *T* and *Y* are defined on [0, 1] (or defined on [0, *M*] with *M* > 0). Define

$$u_1^{(n)}(t) = \sqrt{n}\,(S^{(n)}(t) - S(t)),$$
$$u_2^{(n)}(t) = \sqrt{n}\,(S_Y^{(n)}(t) - S_Y(t)),$$
$$q_1^{(n)}(t) = \sqrt{n}\,(W_1^{(n)}(t) - W_1(t)),$$
$$q_2^{(n)}(t) = \sqrt{n}\,(W_2^{(n)}(t) - W_2(t)),$$
$$R^{(n)}(t) = \sqrt{n}\,(\tilde{S}_Y^{(n)}(t) - S_Y(t)).$$

It is easy to see that the processes $q_1^{(n)}(t)$, $q_2^{(n)}(t)$ and $R^{(n)}(t)$ converge weakly to a Gaussian process, since they are sum of i.i.d. random variables modified by $n^{-1/2}$. We will now discuss the weak convergence of $u_1^{(n)}(t)$ and $u_2^{(n)}(t)$. Taking the derivative of (3.2) and (3.4), we have $dW_2(t) = S(t)dS_Y(t)$ and $dW_2^{(n)}(t) = S^{(n)}(t)dS_Y^{(n)}(t)$. Subtracting one from the other, then

$$dq_2^{(n)}(t) = S^{(n)}(t)du_2^{(n)}(t) + u_1^{(n)}(t)dS_Y(t), \text{ or}$$

$$du_2^{(n)}(t) = \frac{dq_2^{(n)}(t)}{S^{(n)}(t)} - \frac{u_1^{(n)}(t)}{S^{(n)}(t)}dS_Y(t). \tag{3.25}$$

Assume that the largest observation is right censored to make 1 as the last jump point of $S^{(n)}$. Integrate both sides of (3.25) to get

$$u_2^{(n)}(t-) = \int_0^{t-} \frac{dq_2^{(n)}(s)}{S^{(n)}(s)} - \int_0^{t-} \frac{u_1^{(n)}(s)}{S^{(n)}(s)} dS_Y(s).$$

By (3.19), $S_2^{(n)}(t_k-) = \tilde{S}_Y^{(n)}(t_k-)$, so $u_2^{(n)}(t_k-) = R^{(n)}(t_k-) = \int_0^{t_k-} dR^{(n)}(s)$. Hence

$$\int_0^{t_k-} \frac{u_1^{(n)}(s)}{S^{(n)}(s)} dS_Y(s) = \int_0^{t_k-} \frac{dq_2^{(n)}(s)}{S^{(n)}(s)} - \int_0^{t_k-} dR^{(n)}(s). \tag{3.26}$$

Applying the same technique we used to obtain (3.18) we may remove $S^{(n)}(s)$ in the denominator,

$$\int_0^{t_k-} u_1^{(n)}(s) dS_Y(s) = \int_0^{t_k-} [dq_2^{(n)}(s) - S^{(n)}(s)dR^{(n)}(s)],\ k = 1, 2, \cdots, m. \tag{3.27}$$

The following lemma will be used to find the asymptotic variance.

Lemma 3.4

For any bounded measurable functions h_1 *and* h_2 *on* $[0,\ 1]$, *the variance of the following integral is*

$$\text{Var } \{\int_0^t h_1(s)dq_1^{(n)}(s) + \int_0^t h_2(s)dq_2^{(n)}(s)\}$$

$$= -\int_0^t h_1^2(s)dW_1(s) - \int_0^t h_2^2(s)dW_2(s) - \{\int_0^t h_1(s)dW_1(s) + \int_0^t h_2(s)dW_2(s)\}^2,\ 0 \le t \le 1.$$

Proof: Since$\{Y_i,\ \delta_i\}$ are i.i.d., and by the definition of $q_1^{(n)}, q_2^{(n)}, W_1^{(n)}$ and $W_2^{(n)}(t)$, we have

$$\text{Var } \{\int_0^t h_1(s)dq_1^{(n)}(s) + \int_0^t h_2(s)dq_2^{(n)}(s)\}$$

$$= \text{Var } \{\sqrt{n}\int_0^t h_1(s)dW_1^{(n)}(s) + \sqrt{n}\int_0^t h_2(s)W_2^{(n)}(s)\},$$

$$= \text{Var } \{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{h_1(Y_i)[\delta_i = 1, Y_i \le t] + h_2(Y_i)[\delta_i = 0, Y_i \le t]\}\},$$

$$= \text{Var } \{h_1(Y_i)[\delta_i = 1, Y_i \le t] + h_2(Y_i)[\delta_i = 0, Y_i \le t]\},$$

$$= \text{E } \{h_1(Y_i)[\delta_i = 1, Y_i \le t] + h_2(Y_i)[\delta_i = 0, Y_i \le t]\}^2$$

$$- \{\text{E } \{h_1(Y_i)[\delta_i = 1, Y_i \le t] + h_2(Y_i)[\delta_i = 0, Y_i \le t]\}\}^2,$$

$$\overset{\text{def.}}{=} I_1 - I_2.$$

The cross-product term of I_1 is 0 when expanding, so

$$I_1 = E\{h_1^2(Y_i)[\delta_i = 1, Y_i \le t]\} + E\{h_1^2(Y_i)[\delta_i = 1, Y_i \le t]\},$$

$$= -\int_0^t h_1^2(s)dW_1(s) - \int_0^t h_2^2(s)dW_2(s).$$

By the same principle, $I_2 = \{\int_0^t h_1(s)dW_1(s) + \int_0^t h_2(s)dW_2(s)\}^2$.

Theorem 3.4

Let $S^{(n)}$ be the self-consistent estimators of S satisfying Condition C, then

(i) $\sqrt{n}\int_0^1 (S^{(n)}(s) - S(s))dS_Y(s)$ *converges weakly to a normal distribution with mean* 0 *and variance* $-\int_0^1 S(s)(1-S(s))dS_Y(s)$.

Further, if Y has positive continuous density g, then

(ii) $\sqrt{n}\int_0^1 (S^{(n)}(s) - S(s))ds$ *converges weakly to a normal distribution with mean* 0 *and variance* $\int_0^1 \frac{S(s)(1-S(s))}{g(s)}ds$.

Proof: (i) Taking t_k=1 in (3.27) yields

$\sqrt{n}\int_0^1 (S^{(n)}(s) - S(s))dS_Y(s) = \int_0^1 [dq_2^{(n)}(s) - S^{(n)}(s)dR^{(n)}(s)]$, which converges weakly to a normal distribution since $S^{(n)}$ is strongly consistent.

The variance is given by the asymptotic variance of $\int_0^1 [dq_2^{(n)}(s) - S(s)dR^{(n)}(s)]$ = $-\int_0^1 S(s)dq_1^{(n)}(s) + \int_0^1 (1-S(s))dq_2^{(n)}(s)$, which is obtained by using Lemma 3.4 as

$$\text{Var}\ \{-\int_0^1 S(s)dq_1^{(n)}(s) + \int_0^1 (1-S(s))dq_2^{(n)}(s)\}$$

$$= -\int_0^1 S^2(s)dW_1(s) - \int_0^1 (1-S(s))^2 dW_2(s) + \{-\int_0^1 S(s)dW_1(s) + \int_0^1 (1-S(s))dW_2(s)\}^2,$$

$$= -\int_0^1 S(s)(1-S(s))dS_Y(s).$$

(ii) Define the function g_n on [0, 1] by $g_n(s) = g(t_{k-1})$ for $s \in [t_{k-1}, t_k)$, $k = 1, 2, \cdots, m$. Then almost surely $\frac{1}{g_n} - \frac{1}{g} \to 0$ uniformly on [0, 1] since g is continuous. So

$$\begin{aligned}
\int_0^1 u_1^{(n)}(s)ds &= -\int_0^1 u_1^{(n)}(s)\frac{1}{g(s)}dS_Y(s), \\
&= -\int_0^1 u_1^{(n)}(s)\frac{1}{g_n(s)}dS_Y(s) + o_p(1), \\
&= -\sum_{k=1}^{m}\frac{1}{g_n(s)}\int_{[t_{k-1},t_k)} u_1^{(n)}(s)dS_Y(s) + o_p(1), \\
&= -\sum_{k=1}^{m}\frac{1}{g_n(s)}\int_{[t_{k-1},t_k)} [dq_2^{(n)}(s) - S^{(n)}(s)dR^{(n)}(s)] + o_p(1), \\
&= -\int_0^1 \frac{1}{g_n(s)}[dq_2^{(n)}(s) - S^{(n)}(s)dR^{(n)}(s)] + o_p(1), \\
&= -\int_0^1 \frac{1}{g(s)}[dq_2^{(n)}(s) - S(s)dR^{(n)}(s)] + o_p(1).
\end{aligned}$$

The desired result follows since $-\int_0^1 \frac{1}{g(s)}[dq_2^{(n)}(s) - S(s)dR^{(n)}(s)]$ converges weakly to a normal distribution. The asymptotic variance is obtained by the same method as in (i).

Groeneboom (1992) proved the weak convergence of MLE under certain conditions. The following theorems are proved by using Groeneboom's Lemma 5.4 (restated here as Lemma 3.5). Assume T and Y have bounded densities f and g respectively, satisfying $f(t) \geq \alpha > 0$, $g(t) \geq \alpha > 0$, for $t \in [0, 1]$ and g has a bounded derivative on [0, 1].

Lemma 3.5 (Groeneboom)

Let $\alpha_n = M\, n^{-1/3}$, *where* $M > 0$ *is a constant. If* $S^{(n)}$ *is the MLE, then for any* $t_0 \in$ [0, 1],

(i) *The probability that* $S^{(n)}$ *does not have a jump in an interval of the form* $[t_0-\alpha_n, t_0+\alpha_n]$ *can be made arbitrarily small.*

(ii) $$\sup_{t\in[t_0-\alpha_n,t_0+\alpha_n]} \left|S^{(n)}(t) - S(t_0)\right| = O_p(n^{-1/3}).$$

The following theorem presents the relationship among $S_Y^{(n)}$, $\tilde{S}_Y^{(n)}$, and S_Y.

Theorem 3.5

If $S^{(n)}$ *and* $S_Y^{(n)}$ *are solution of* (3.3) *and* (3.4), *with* $S^{(n)}$ *being the MLE, then*

(i) $\sqrt{n}(S_Y^{(n)}(t) - \tilde{S}_Y^{(n)}(t)) \xrightarrow{P} 0$, $t \in [0, 1]$;

(ii) $u_2^{(n)}(t) = \sqrt{n}(S_Y^{(n)}(t) - S_Y(t))$ *converges weakly to a normal distribution with mean* 0 *and variance* $S_Y(t)(1 - S_Y(t))$.

Proof: For any $t \in (0, 1)$, let $t_n > t$ such that $S^{(n)}(s) = S^{(n)}(t_n)$ for $s \in [t, t_n]$ and $S_Y^{(n)}(t_n) = \tilde{S}_Y^{(n)}(t_n)$. Then $t_n - t = O_p(n^{-1/3})$ by (i) of Lemma 3.5. Define

$P_n(t,y) = \frac{1}{n}\sum_{i=1}^{n}[T_i \le t, Y_i \le y]$, which is the empirical of the pairs (T_i, Y_i), and

$P(t, y) = P(T \le t, Y \le y) = (1-S(t))(1-S_Y(y))$.

Then $S_Y^{(n)}(t) = -\int_{t+}^{1} \frac{1}{S^{(n)}(s)} dW_2^{(n)}(s)$, $W_2^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n}[T_i > Y_i, Y_i > t] = \int_0^1\int_{t+}^1 [s > y] dP_n(s,y)$,

and $\tilde{S}_Y^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n}[Y_i > t] = \int_0^1\int_{t+}^1 dP_n(s,y)$. So

$S_Y^{(n)}(t_n) - S_Y^{(n)}(t) = \int_{(t,t_n]} \frac{1}{S^{(n)}(s)} dW_2^{(n)}(s) = -\frac{1}{S^{(n)}(t_n)}\int_0^1\int_{(t,t_n]}[s > y] dP_n(s,y)$, and

$\tilde{S}_Y^{(n)}(t_n) - \tilde{S}_Y^{(n)}(t) = -\int_0^1\int_{(t,t_n]} dP_n(s,y)$.

Therefore

$$S_Y^{(n)}(t)-\tilde{S}_Y^{(n)}(t)=-\int_0^1\int_{(t,t_n]}\{1-\frac{[s>y]}{S^{(n)}(t_n)}\}dP_n(s,y),$$

$$=\frac{1}{S^{(n)}(t_n)}\int_0^1\int_{(t,t_n]}\{[s>y]-S^{(n)}(t_n)\}dP_n(s,y),$$

$$=\frac{1}{S^{(n)}(t_n)}\int_0^1\int_{(t,t_n]}\{[s>y]-S^{(n)}(t_n)\}dP(s,y)$$

$$+\frac{1}{S^{(n)}(t_n)}\int_0^1\int_{(t,t_n]}\{[s>y]-S^{(n)}(t_n)\}d(P_n-P)(s,y),$$

$$\overset{\text{def.}}{=} I_{1n}+I_{2n}\,.$$

$$I_{1n}=-\frac{1}{S^{(n)}(t)}\int_{(t,t_n]}\{S(s)-S^{(n)}(t_n)\}dS_Y(s),$$

$$=\frac{1}{S^{(n)}(t)}\{S^{(n)}(t_n)\int_{(t,t_n]}dS_Y(s)-\int_{(t,t_n]}S(s)dS_Y(s)\},$$

$$=\frac{1}{S^{(n)}(t)}\{(S^{(n)}(t_n)-S(\theta))(S_Y(t)-S_Y(t_n))\}\quad(\text{where }\theta\in(t,t_n]),$$

$$=O_p(n^{-\frac{2}{3}}),$$

since $S^{(n)}(t_n)-S(\theta)=O_p(n^{-\frac{1}{3}})$ by (ii) of Lemma 3.5, and $S_Y(t)-S_Y(t_n)=O_p(n^{-\frac{1}{3}})$. Also $I_{2n}=o_p(n^{-1/2})$ by P_n- $P=O_p(n^{-1/2})$ and t_n-$t=O_p(n^{-1/3})$. So

$$\sqrt{n}(S_Y^{(n)}(t)-\tilde{S}_Y^{(n)}(t))=o_p(1).$$

(ii) follows from (i) since we know $\sqrt{n}(\tilde{S}_Y^{(n)}(t)-S_Y(t))$ converges weakly to a Gaussian process.

Under these conditions, we can use Lemma 3.5 to extend the result of Theorem 3.4 to integrals over [0, t], for any $t\in(0,1)$.

Theorem 3.6

Let $S^{(n)}$ be the MLE of S, then

(i) $\sqrt{n}\int_0^t (S^{(n)}(s)-S(s))dS_Y(s)$ *converges weakly to a normal distribution with mean* 0 *and variance* $-\int_0^t S(s)(1-S(s))dS_Y(s)$.

(ii) $\sqrt{n}\int_0^t (S^{(n)}(s)-S(s))ds$ *converges weakly to a normal distribution with mean* 0 *and variance* $\int_0^t \frac{S(s)(1-S(s))}{g(s)}ds$.

Proof: We will show for any $t \in (0,1)$

$$\int_0^t u_1^{(n)}(s)dS_Y(s)-\int_0^t [dq_2^{(n)}(s)-S(s)dR^{(n)}(s)] \xrightarrow{P} 0, \qquad (3.28)$$

since $\int_0^t [dq_2^{(n)}(s)-S(s)dR^{(n)}(s)]$ converges weakly to a Gaussian process.

By (i) of Lemma 3.5, for $t \in (0,1)$, there exists a jump point $t_k > t$, such that $t_k - t = O_p(n^{-1/3})$. Then by (3.27) and using the same principle as in the proof of Theorem 3.5,

$$\int_0^t u_1^{(n)}(s)dS_Y(s)-\int_0^{t_k-} u_1^{(n)}(s)dS_Y(s) = -\sqrt{n}\int_{(t,t_k)} (S^{(n)}(s)-S(s))dS_Y(s) = o_p(1), \text{ and}$$

$$\int_0^{t_k-} [dq_2^{(n)}(s)-S^{(n)}(s)dR^{(n)}(s)]-\int_0^t [dq_2^{(n)}(s)-S^{(n)}(s)dR^{(n)}(s)] = o_p(1).$$

(3.28) follows from the strong consistency of $S^{(n)}$. Hence $\sqrt{n}\int_0^t (S^{(n)}(s)-S(s))dS_Y(s)$ converges weakly to a normal distribution. Having proved weak convergence, the variance and (ii) can be obtained in the same way as in Theorem 3.4.

Chapter 4

BAYESIAN ESTIMATION

4.1 Introduction

The problem of nonparametrically estimating a survival function from censored data has been addressed by a number of authors. For right censored data, Susarla & Van Ryzin (1976) gave a nonparametric Bayes solution to the estimator under squared error loss using Dirichlet process prior. The resulting Bayes estimator was shown to reduce to the PL estimator in the cases where $\alpha(R^+) \to 0$, where $\alpha(\,\cdot\,)$ is a finite measure on R^+ which serves as a parameter of the Dirichlet process prior.

In this chapter, with right censored data, we find another expression for the Bayes estimator, which is related to the self-consistency equation. This approach sheds more light on the connection between the Bayes estimator and the Kaplan-Meier's PL estimator (also the MLE). For interval censored data, we also present a Bayes estimator, but we find that in this case the Bayes estimator may not reduce to the MLE as $\alpha(R^+) \to 0$.

Let $(R^+,\ \mathscr{B},\ \mathscr{P}\,)$ be the probability space, where $R^+ = (0, \infty)$, and $\mathscr{B}$ is the Borel σ-field on $(0, \infty)$. P is a random measure. We say $P \in \mathscr{D}(\alpha)$ to mean P is a Dirichlet process on $(\,R^+,\ \mathscr{B})$ with parameter α. See Ferguson(1973) for the definition of Dirichlet process. $\mathscr{E}(P(A)) = \dfrac{\alpha(A)}{\alpha(R^+)}$, for $A \in \mathscr{B}$. $T_1, \cdots, T_n$ is called a sample of size n from P, if

$$\mathscr{P}\{T_1 \in C_1, \cdots, T_n \in C_n \mid P(A_1), \cdots, P(A_m); P(C_1), \cdots, P(C_n)\} = \prod_{i=1}^{n} P(C_i),\ a.s.$$

We assume $T_1, \cdots, T_n$ are the lifetimes with survival $S(t)$, and further that they are a sample from P, $S(t) = P(\ t\ , \infty\ )$. $P \in \mathscr{D}(\alpha)$. The Bayes estimator is always under the squared error loss $L(\hat{S},S)=\int_0^\infty(\hat{S}(u)-S(u))^2dw(u)$, where w is a weight function and $\hat{S}(u)$ is a estimator of $S(u)$.

Ferguson (1973) proved:

(i) *The conditional distribution of P given* $T_1,\cdots,T_n$ *is a Dirichlet process with parameter*

$$\beta=\alpha+\sum_{i=1}^n\delta_{T_i}\ .$$

(ii) *Let* $S(t)=P(\,t\,,\infty)$, *then, given* $T_1,\cdots,T_n$, *the Bayes estimator of* $S(t)$ *is*

$$\begin{aligned}\hat{S}_B(t) &= \mathcal{E}(S(t)\,|\,T_1,\cdots,T_n),\\ &=\frac{\alpha(t,\infty)}{\alpha(R)+n}+\frac{1}{\alpha(R)+n}\sum_{i=1}^n[T_i>t],\\ &=\frac{\alpha(t,\infty)}{\alpha(R)+n}+\frac{n}{\alpha(R)+n}S_n(t).\end{aligned}\tag{4.1}$$

where $S_n(t)=\frac{1}{n}\sum_1^n[T_i>t]$ *is the empirical of* $S(t)$, *and* $[T_i>t]$ *is the indicator of* $(T_i>t)$.

4.2 Right Censoring

We now consider right censored data, where $T_1,\cdot\cdot\cdot,T_n$ are the lifetimes and $Y_1,\cdot\cdot\cdot,Y_n$ are the censoring times. We observe $\{(Z_i,\delta_i)\,;\,i=1,\cdots,n\}$, with $Z_i=T_i\wedge Y_i$, $\delta_i=[T_i\le Y_i]$, $i=1,\cdot\cdot\cdot,n$. Without loss of generality, assume$\delta_1=\cdots=\delta_k=1$, $\delta_{k+1}=\cdots=\delta_n=0$ and that $Z_1,\cdots,Z_n$ are all distinct. Let

$N_U(t)=\sum_{i=1}^n[Z_i\ge t,\delta_i=1]$ be the number of uncensored observations in $[t,\infty)$, and

$N_C(t)=\sum_{i=1}^{n}[Z_i \geq t, \delta_i = 0]$ be the number of censored observations in $[t, \infty)$.

Let $N(t)=N_U(t)+N_C(t)$, $N^+(t)=N(t+)$, then $N(t)$ is right continuous counting process. Actually $N(t)$ is the number at risk at time t.

To find the Bayes estimator under such right censored data, Susarla & Van Ryzin use the following lemma and theorem.

Lemma 4.1

Given $(Z_l, 1), \cdots (Z_k, 1)$, *the conditional distribution of P is a Dirichlet process with parameter* $\beta_1 = \alpha + \sum_{i=1}^{k} \delta_{Z_i}$.

Theorem 4.1 (Susarla & Van Ryzin)

Consider $T_{k+1}, \cdots, T_n$ *as a sample of size n-k from* $\mathscr{D}(\beta_1)$, *which is obtained by Lemma 4.1. Then under condition* $(Z_{k+1}, 0), \cdots, (Z_n, 0)$, *the Bayes estimator of survival function S(t) is*

$$\hat{S}(t)=\frac{\alpha(t,\infty)+N^+(t)}{\alpha(R)+n}\prod_{i=1}^{n}\left\{\frac{\alpha(Z_i,\infty)+N(Z_i)}{\alpha(Z_i,\infty)+N^+(Z_i)}\right\}^{[Z_i \leq t, \delta_i = 0]} . \tag{4.2}$$

Remark:

From (4.2) , by taking $\alpha(R^+) \to 0$, Susarla and Van Ryzin showed that the limit of $\hat{S}(t)$ is the Kaplan and Meier's PL estimator

$$\hat{S}_{PL}(t)=\prod_{i=1}^{n}\left\{1-\frac{d(-N(Z_i))}{N(Z_i)}\right\}^{[Z_i \leq t, \delta_i = 1]} .$$

The following theorem enable us to see more about the relation between the Bayes estimator and the Kaplan & Meier's PL estimator. The proof of lemma 4.2 is in Appendix.

Lemma 4.2

Let $\hat{S}(t) = \mathscr{E}\{S(t) \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)\}$. *Then*

$$\mathscr{P}\{T_j > t \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)\} = [Z_j > t]\,,\ \text{if}\ \delta_j = 1;$$

$$= \frac{\hat{S}(t \vee Z_j)}{\hat{S}(Z_j)},\ \text{if}\ \delta_j = 0;\ j = 1, \cdots, n.$$

Theorem 4.2

Given $\{(Z_i,\, \delta_i)\,;\, i = 1,\cdots, n\}$, $\hat{S}(t)$, *the Bayes estimator of* $S(t)$ *is the solution of the following equation*

$$\hat{S}(t) = \frac{1}{\alpha(R^+)+n}\{\alpha(t,\infty) + N^+(t) + \int_0^t \frac{\hat{S}(t)}{\hat{S}(s)} d(-N_C(t))\}. \tag{4.3}$$

Proof: By (4.1) and Lemma 4.2,

$$\hat{S}(t) = \mathscr{E}\{S(t) \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)\},$$

$$= \mathscr{E}\{\, \mathscr{E}(S(t) \mid T_1,\cdots,T_n) \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)\},$$

$$= \mathscr{E}\{\frac{\alpha(t,\infty)}{\alpha(R^+)+n} + \frac{1}{\alpha(R^+)+n}\sum_{j=1}^{n}[T_j > t] \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)\},$$

$$= \frac{\alpha(t,\infty)}{\alpha(R^+)+n} + \frac{1}{\alpha(R^+)+n}\sum_{j=1}^{n} \mathscr{P}(T_j > t \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n)),$$

$$= \frac{1}{\alpha(R^+)+n}\{\alpha(t,\infty) + \sum_{j=1}^{k}[T_j > t] + \sum_{j=k+1}^{n} \mathscr{P}(T_j > t \mid (Z_1,\delta_1),\cdots,(Z_n,\delta_n))\},$$

$$= \frac{1}{\alpha(R^+)+n}\{\alpha(t,\infty)+N_U(t+)+\sum_{j=k+1}^{n}\frac{\hat{S}(t\vee Z_j)}{\hat{S}(Z_j)}\},$$

$$= \frac{1}{\alpha(R^+)+n}\{\alpha(t,\infty)+N_U(t+)+N_C(t+)+\int_0^t\frac{\hat{S}(t)}{\hat{S}(s)}d(-N_C(s))\},$$

$$= \frac{1}{\alpha(R^+)+n}\{\alpha(t,\infty)+N^+(t)+\int_0^t\frac{\hat{S}(t)}{\hat{S}(s)}d(-N_C(s))\}.$$

Remark:

Since$\hat{S}_{PL}(t)$ is the solution of equation

$$\hat{S}_{PL}(t)=\frac{1}{n}\{N^+(t)+\int_0^t\frac{\hat{S}_{PL}(t)}{\hat{S}_{PL}(s)}d(-N_C(s))\}, \tag{4.4}$$

from (4.3), it is obvious that the limit of $\hat{S}(t)$ as $\alpha(R^+) \to 0$ is the Kaplan-Meier's PL estimator, In other words, the Kaplan-Meier's PL estimator can be obtained from the self-consistent equation $\hat{S}(t)=E_{\hat{S}}\{S_n(t)\,|\,(Z_1,\delta_1),\cdots,(Z_n,\delta_n)\}$, and the Bayes estimator can be obtained from the self-consistent equation $\hat{S}(t)=E_{\hat{S}}\{\hat{S}_B(t)\,|\,(Z_1,\delta_1),\cdots,(Z_n,\delta_n)\}$, where $S_n(t)$ is the empirical of $S(t)$, and $\hat{S}_B(t)=\dfrac{\alpha(t,\infty)}{\alpha(R)+n}+\dfrac{n}{\alpha(R)+n}S_n(t)$.

4.3 Interval Censoring

For interval censoring case 1, we observe $\{(Y_i, \delta_i): i = 1,\cdots, n\}$. Let $A_i=\begin{cases}(0,Y_i] & \text{if } \delta_i=1,\\ (Y_i,\infty) & \text{if } \delta_i=0.\end{cases}$ be the observed intervals. We want to find the Bayes estimator of $S(t)$, $\hat{S}(t)= \mathcal{E}\{S(t)\,|(Y_1,\delta_1),\cdots,(Y_n,\delta_n)\}$.

Let $Y_{(1)} < \cdots < Y_{(n)}$ be the ranked Y_i's, and let $\delta_{(i)}$ and $A_{(i)}$ be the δ and A corresponding to $Y_{(i)}$. Then, under squared error loss, using the method of Lemma 1 in Susarla & Van Ryzin (1976), the Bayes estimator is

$$\begin{aligned}\hat{S}(t) &= \mathcal{E}(S(t)|(Y_1,\delta_1),\cdots,(Y_n,\delta_n)),\\ &= \mathcal{E}(S(t)|T_1 \in A_1,\cdots,T_n \in A_n),\\ &= \frac{\mathcal{E}(P(t,\infty)\prod_{i=1}^{n} P(A_{(i)}))}{\mathcal{E}(\prod_{i=1}^{n} P(A_{(i)}))} \overset{\text{def.}}{=} \frac{I_1}{I_2}. \end{aligned} \tag{4.5}$$

Now let us calculate I_1 and I_2.

Let $B_1 = (0, Y_{(1)}], B_2 = (Y_{(1)}, Y_{(2)}],\cdots, B_n = (Y_{(n-1)}, Y_{(n)}], B_{n+1} = (Y_{(n)}, \infty)$, and let $\alpha_1 = \alpha(B_1)$, $\alpha_2 = \alpha(B_2)$, $\cdots$, $\alpha_n = \alpha(B_n)$, $\alpha_{n+1} = \alpha(B_{n+1})$. Then $A_i = \bigcup_{j=1}^{i} B_j$ if $\delta_i = 1$, and $A_i = R^+ - \bigcup_{j=1}^{i} B_j$ if $\delta_i = 0$. $B_1 , \cdots, B_n , B_{n+1}$ is a partition of R^+.

$P \in \mathcal{D}(\alpha)$, the density of $(P(B_1),\cdots,P(B_n))$ is

$$\frac{1}{D(\alpha_1,\cdots,\alpha_{n+1})} \cdot \prod_{i=1}^{n} x_i^{\alpha_i - 1} \cdot (1-\sum_{i=1}^{n} x_i)^{\alpha_{n+1}-1},$$

where $D(\alpha_1,\cdots,\alpha_{n+1}) = \dfrac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{n+1})}{\Gamma(\alpha_1+\cdots+\alpha_{n+1})}$, $\alpha_1 + \cdots + \alpha_{n+1} = \alpha(R^+)$.

Suppose $t = Y_{(k)}$, then

$I_2 \cdot D(\alpha_1,\cdots,\alpha_{n+1})$

$$= \int_{x_1+\cdots+x_n \le 1} \prod_{i=1}^{n} f_i(x_1,\cdots,x_i,\delta_{(i)}) \cdot \prod_{i=1}^{n} x_i^{\alpha_i - 1} \cdot (1-\sum_{i=1}^{n} x_i)^{\alpha_{n+1}-1} dx_n \cdots dx_1, \tag{4.6}$$

and

$I_1 \cdot D(\alpha_1,\cdots,\alpha_{n+1})$

$$= \int_{x_1+\cdots+x_n\le 1} \prod_{i=1}^{n} f_i(x_1,\cdots,x_i,\delta_{(i)})(1-\sum_{j=1}^{k} x_j)\cdot \prod_{i=1}^{n} x_i^{\alpha_i-1}\cdot(1-\sum_{i=1}^{n} x_i)^{\alpha_{n+1}-1} dx_n\cdots dx_1, \qquad (4.7)$$

$$\text{where } f_i(x_1,\cdots,x_i,\delta_{(i)}) = \begin{cases} \sum_{j=1}^{i} x_j & \text{if } \delta_{(i)} = 1; \\ 1-\sum_{j=1}^{i} x_j & \text{if } \delta_{(i)} = 0. \end{cases}$$

Hence we have the following theorem.

Theorem 4.3

Given $(Y_1,\delta_1),\cdots,(Y_n,\delta_n)$, *the Bayes estimator of* $S(t)$ *is*

$$\hat{S}(t) = \mathcal{E}(S(t)\,|\,(Y_1,\delta_1),\cdots,(Y_n,\delta_n)) = \frac{I_1}{I_2},$$

where I_1 *and* I_2 *are decided by* (4.7) *and* (4.6).

The calculation of I_1 and I_2 is done by using

$$\int_0^c x^{\gamma-1}(c-x)^{\eta-1}dx = c^{\gamma+\eta-1}B(\gamma,\eta) \text{ for } 0\le c\le 1 \text{ and } \gamma,\ \eta > 0,$$

where $B(\gamma,\eta) = \dfrac{\Gamma(\gamma)\Gamma(\eta)}{\Gamma(\gamma+\eta)}$, and $\Gamma(a+1) = a\Gamma(a)$ for $a > 0$.

Even in the simplest cases, the expression for I_1 and I_2 can be quite complicated. For example, if all δ's are 0, then

$$I_2 = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{n+1})}\cdot\frac{\Gamma(\alpha_{n+1}+1)\Gamma(\alpha_n)}{\Gamma(\alpha_{n+1}+\alpha_n+1)}\cdots\frac{\Gamma(\alpha_{n+1}+\cdots\alpha_2+n)\Gamma(\alpha_1)}{\Gamma(\alpha_{n+1}+\cdots+\alpha_1+n)},$$

$$= \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n)}\alpha(Y_{(n)},\infty)(\alpha(Y_{(n-1)},\infty)+1)\cdots(\alpha(Y_{(1)},\infty)+n-1),$$

$$= \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n)}\alpha(Y_{(n)},\infty)(\alpha(Y_{(n-1)},\infty)+N^+(Y_{(n-1)}))\cdots(\alpha(Y_{(1)},\infty)+N^+(Y_{(1)})).$$

If there exists at least one δ equal to 1, then I_2 is a summation of the forms above. Such as, suppose $\delta_{(n-1)} = 1$, then

$$I_2 = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n-1)}\alpha(Y_{(n)},\infty)(\alpha(Y_{(n-2)},\infty)+N^+(Y_{(n-2)})-1)\cdots(\alpha(Y_{(1)},\infty)+N^+(Y_{(1)})-1)$$

$$-\frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n)}\alpha(Y_{(n)},\infty)(\alpha(Y_{(n-1)},\infty)+N^+(Y_{(n-1)}))\cdots(\alpha(Y_{(1)},\infty)+N^+(Y_{(1)})).$$

In each case, I_1 will have a similarly complicated expression.

Theorem 4.3 tells us a way to calculate the Bayes estimator $\hat{S}(t)$, but it is too tedious to find an explicit expression. The following example show that given a Dirichlet process prior α, with nonnull probability, there exist Bayes estimator $\hat{S}(t)$, which does not converge to the MLE as $\alpha(R^+) \to 0$.

Example 4.1

Let y_1, y_2, y_3 and y_4 be fixed real numbers with $0 < y_1 < y_2 < y_3 < y_4$. Consider a Dirichlet prior such that

$\alpha[0, y_1) = \alpha[y_1, y_2) = \alpha[y_2, y_3) = \alpha[y_3, y_4) = \alpha[y_4, \infty) = x$. $\alpha(R^+) = 5x$, where $0 \le x \le \infty$.

Let Y be a discrete random variable taking values on y_1, y_2, y_3 and y_4. Let T be the lifetime with Dirichlet prior α. Suppose the intervals observed are (y_1, ∞), (y_2, ∞), $(0, y_3]$, and (y_4, ∞), $n = 4$. The probability of getting such observations is $P(\, Y_1 = y_1,\ T_1 > Y_1;\ Y_2 = y_2,\ T_2 > Y_2;\ Y_3 = y_3,\ T_3 \le y_3;\ Y_4 = y_4,\ T_4 > Y_4\,)$, which is positive. Then the Bayes estimator $\hat{S}$, when $t = y_3$, is

$$\hat{S}(t) = \mathcal{E}\{\, S(t) \mid X_1 \in (y_1, \infty), X_2 \in (y_2, \infty), X_3 \in (0, y_3], X_4 \in (y_4, \infty)\},$$

$$= \frac{\mathcal{E}(P(t,\infty)P(0,y_3]\prod_{i\neq 3} P(y_i,\infty)}{\mathcal{E}(P(0,y_3]\prod_{i\neq 3} P(y_i,\infty))} \overset{\text{def.}}{=} \frac{I_1}{I_2}.$$

Rewrite I_1 and I_2,

$I_1 = \mathcal{E}(P(y_3,\infty)\prod_{i\neq 3} P(y_i,\infty)) - \mathcal{E}(P(y_3,\infty)\prod_{i} P(y_i,\infty))$, and

$I_2 = \mathcal{E}(\prod_{i\neq 3} P(y_i,\infty)) - \mathcal{E}(\prod_{i} P(y_i,\infty))$.

Then, each terms in I_1 and I_2 can be calculated by integration as in Susarla & Van Ryzin (1976). Denote $\alpha = \alpha(R^+)$.

$$I_1 = \frac{\Gamma(\alpha)}{\Gamma(\alpha+4)}x(2x+1)(3x+2)(4x+3) - \frac{\Gamma(\alpha)}{\Gamma(\alpha+5)}x(2x+1)(2x+2)(3x+3)(4x+4),$$

$$I_2 = \frac{\Gamma(\alpha)}{\Gamma(\alpha+3)}x(3x+1)(4x+2) - \frac{\Gamma(\alpha)}{\Gamma(\alpha+4)}x(2x+1)(3x+2)(4x+3).$$

So, $\hat{S}(t) = \frac{I_1}{I_2} = \frac{(x+1)(3x+2)(4x+3)(5x+4)-(x+1)(2x+2)(3x+3)(4x+4)}{(3x+1)(4x+2)(5x+3)(5x+4)-(2x+1)(3x+2)(4x+3)(5x+4)}$, and

$\lim_{x\to 0}\hat{S}(t) = 13/22.$

On the other hand, the MLE is $S^{(n)}(y_1) = S^{(n)}(y_2) = 1$, $S^{(n)}(y_3) = S^{(n)}(y_4) = 1/2$. As $x \to 0$, so that $\alpha(R^+) \to 0$, $\hat{S}(t)$ does not converge to 1/2. Hence, $\hat{S}$ does not converge to the MLE as $\alpha(R^+) \to 0$.

Chapter 5

SIMULATIONS

In this chapter, we investigate the performance of our estimation procedures through simulation studies from known distributions.

a. Parametric assumptions

We just consider the 2-inspection interval censorship model described in Chapter 2. The two inspection times are W_1, W_2 ($W_1 < W_2$) are generated through independent exponential rv's V_1, V_2 with mean 1; That is, W_1=V_1, W_2=V_1+V_2. The survival time T of interest is also assumed exponential (mean 1/θ) and independent of (V_1, V_2).

The stopping number $N(t) = \min\{ k>0: T \leq W_k \leq t \}$($N(t)$= +∞, if there is no such k) of (2.1), has values 1, 2, and +∞. Its distribution is obtained from (2.2) - (2.4):

$$P[N(t)=1] = \frac{\theta}{(\theta+1)} - e^{-t}(1-\frac{1}{(\theta+1)}e^{-\theta t}),$$

$$P[N(t)=2] = \frac{1}{(\theta+1)}\{te^{-(\theta+1)t} + \frac{\theta}{(\theta+1)}(1-e^{-(\theta+1)t})\} - \frac{1}{\theta}(e^{-t} - e^{-(\theta+1)t}),$$

$$P[N(t)=\infty] = 1 - P[N(t) \leq 2].$$

Allowing $t \to \infty$, leads to consideration of an unrestricted follow up period. This is the case (Case2) considered by Groeneboom and Wellner (1992), With t infinite in (2.1) the random variable N (≡ $N(\infty)$), has three values 1, 2 and +∞ corresponding to (1) $T \in (0, W_1]$, or (2) $T \in (W_1, W_2]$, or (3) $T > W_2$. The distribution of N is then

$$P[N=1] = \frac{\theta}{\theta+1}, P[N=2] = \frac{\theta}{(\theta+1)^2}, P[N=\infty] = \frac{1}{(\theta+1)^2}.$$

Thus $\frac{\theta}{\theta+1}$ is the proportion of left censoring, $\frac{\theta}{(\theta+1)^2}$ the proportion of interval censoring and $\frac{1}{(\theta+1)^2}$ the proportion of right censoring. The objective of our simulations is to compare the nonparametric estimator of the survival distribution $S(t) = e^{-\theta t}$, $t > 0$ of T, based on the Turnbull scheme, with that obtained through maximum likelihood estimation of the parameter θ. The data are generated from n independent triples (V_{1i}, V_{2i}, T_i), $1 \leq i \leq n$.

b. Maximum likelihood estimation of θ

The information on T may be written in the form $T \in (L, R]$, where L=0, $R=V_1$ when N=1; and $L=V_1$, $R=V_1+V_2$ when N=2; and $L=V_1+V_2$, $R= +\infty$ when $N=\infty$; The SAS LIFEREG procedure is used to estimate θ form the data $\{ T_i \in (L_{1i}, R_i)$: $1 \leq i \leq n\}$.

c. Turnbull method for estimation of $S(t)$

From the data $\{ T_i \in (L_i, R_i]$: $1 \leq i \leq n\}$, Turnbull (1976) provided an algorithm for the (nonparametric) estimation of $S(t)$. He constructed a set of disjoint intervals $\{(q_j, p_j]$: $1 \leq j \leq m\}$, where q_j's and p_j's lie in the sets $\{L_i$: $1 \leq i \leq n\}$ and $\{R_i$: $1 \leq i \leq n\}$. The likelihood is proportional to

$$L(S) = \prod_{i=1}^{n} (S(L_i) - S(R_i)). \tag{5.1}$$

Define the vector of probability $\boldsymbol{s} = (s_1, \cdots, s_m)$ by $s_j = S(q_j) - S(p_j)$. The problem of maximizing (5.1) reduces to one of maximizing

$$L(S) = \prod_{i=1}^{n} (\sum_{j=1}^{m} \alpha_{ij} s_j), \tag{5.2}$$

where $\alpha_{ij} = 1$ if $(q_j, p_j] \subseteq (L_i, R_i]$, 0 otherwise.

For $1 \le i \le n$, $1 \le j \le m$, let $I_{ij} = 1$ if $T_i \in (q_j,\ p_j]$ and 0 otherwise. Because of the censoring the value of I_{ij} may not be known, however its expectation is given by

$$E\,[\,I_{ij}\,] = \alpha_{ij}\, s_j \,/ \sum_{k=1}^{m} \alpha_{ik} s_k \ = \mu_{ij}\,(s). \tag{5.3}$$

If we treated (5.3) as observed rather than expected frequencies, the proportion of observations in intervals is $(\sum_{i=1}^{n} \mu_{ij}(\boldsymbol{s}))\,/\,n\ = \pi_j$. We say that the vector of probability $\boldsymbol{s}$ is self-consistent if

$$s_j = \pi_i(s_1, \cdots, s_m) \ \ (1 \le j \le m). \tag{5.4}$$

A self-consistent estimator of $\boldsymbol{s}$ is defined to be any solution of the simultaneous equations (5.4). The form of (5.4) immediately suggests a iterative procedure for finding the solution.

(i) Obtain initial estimates s_j^0 $(1 \le j \le m)$. This can be any set of positive numbers summing to unit.

(ii) Evaluated $\mu_{ij}\,(\,\boldsymbol{s}^0)$ for $1 \le j \le m$.

(iii) Obtain improved estimates s_j^1 by setting

$s_j^1 = \pi_j(\,\boldsymbol{s}^0)$ for $1 \le j \le m$.

(iv) Return to step (ii) with $\boldsymbol{s}^1$ replacing $\boldsymbol{s}^0$, etc.

(v) Stop when the required accuracy has been achieved.

d. Results

1. Our first simulation study involves 10 replications of size n = 100 from each of the exponential distributions with θ= .5, 1, 2, 3. The distribution of N is denoted by $p = (\frac{\theta}{\theta+1}, \frac{\theta}{(\theta+1)^2}, \frac{1}{(\theta+1)^2})$ and $\hat{p}$ denotes the observed proportion of left, interval, and right censorship, averaged over the 10 replications. Table 5.1 summarizes these results

for the four distribution that we generated. Even though the number of replication is small, the degree of closeness of p to $\hat{p}$ is very satisfactory.

2. For each distribution, maximum likelihood estimate $\hat{\theta}$ is obtained from the SAS LIFEREG procedure for each replication. In Table 5.2 we input the average of these estimates over the 10 replications.

3. For the nonparametric maximum likelihood estimation of $S(t)$ we use the Turnbull self-consistency algorithm. APL program were written to efficiently carry out the task of solving the self-consistency equations.

For each replication, the Turnbull estimator $\hat{S}(t)$ is close to a step function, with gaps, in the sense that it is undefined on some intervals. We extended the definition of $\hat{S}$ to these intervals by extending consecutively the steps from the left of the curve to the next jump in $\hat{S}$. These step functions were then averaged over the 10 replications to obtain an estimator of $S(t)$.

The figures illustrate the Turnbull estimator based on a single replication, and that obtained by averaging across the 10 replications. The true survival curve is $S(t) = e^{-2t}$; t>0.

	θ=0.5		θ=1		θ=2		θ=3	
Censoring	p	$\hat{p}$	p	$\hat{p}$	p	$\hat{p}$	p	$\hat{p}$
Left (N=1)	.333	.317	.500	.510	.667	.655	.750	.744
Interval (N=2)	.222	.220	.250	.230	.222	.224	.188	.191
Right (N=∞)	.444	.463	.250	.260	.111	.121	.062	.065

Table 5.1

θ	$\hat{\theta}$
.5	.4969
1	.9978
2	1.9678
3	2.9299

Table 5.2

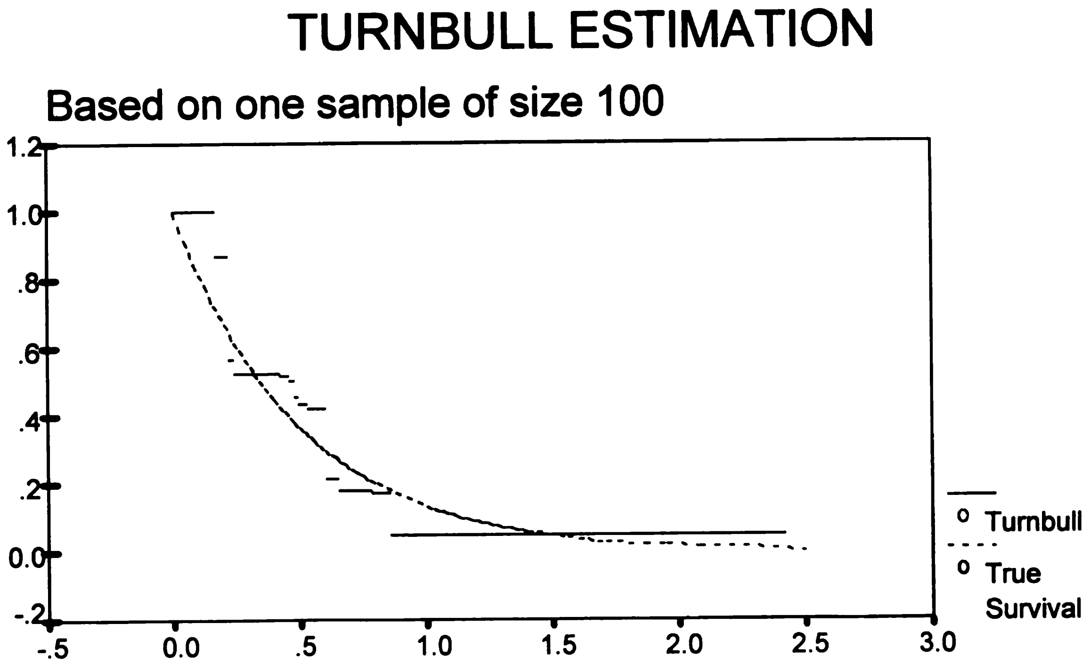

Figure 5.3

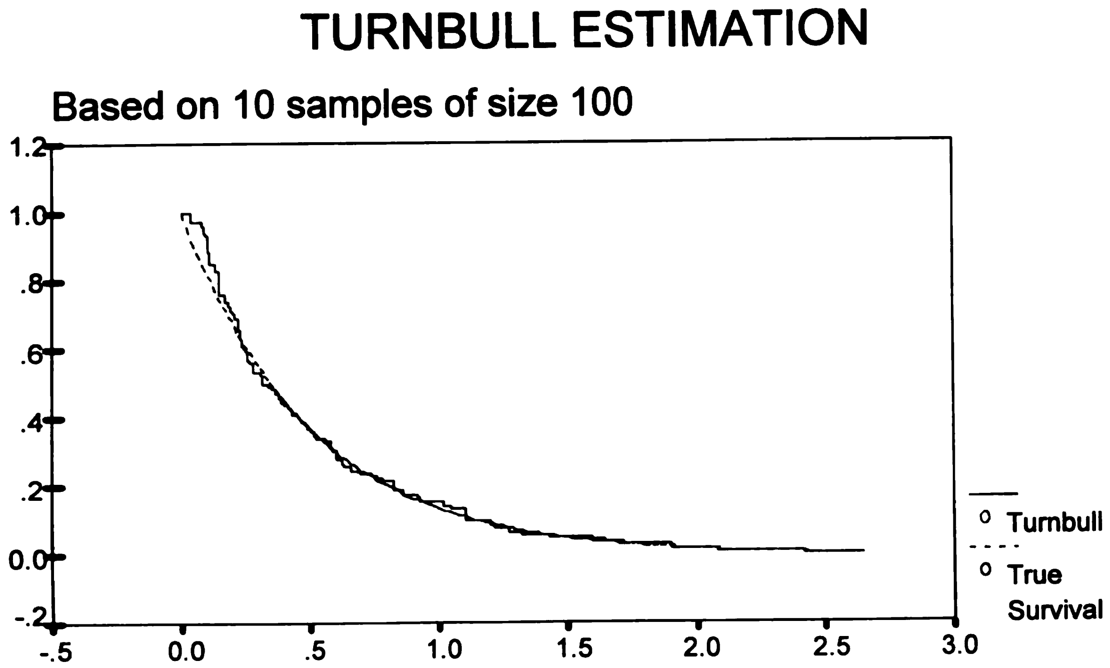
TURNBULL ESTIMATION
Based on 10 samples of size 100
1.2
1.0
.8
.6
.4
.2
0.0
-.2
-.5
0.0
.5
1.0
1.5
2.0
2.5
3.0
o Turnbull
o True
Survival

Figure 5.4

APPENDIX

Proof of theorem 3.3:

Let $F^{(n)} = 1 - S^{(n)}$, then $F^{(n)}$ is the MLE of F. Let $Y_{(1)} < \cdots < Y_{(n)}$ be the ranked Y_i's, and let $\delta_{(i)}$ be the δ corresponding to $Y_{(i)}$. Groeneboom (1992) showed that the value of $F^{(n)}$ at $Y_{(i)}$ is the left derivative of H^* at i, where H^* is the convex minorant of the points $(i, \sum_{j\le i}\delta_{(j)})$ on $[0, n]$. Call a point $\tau \in \{ Y_i : i = 1, 2, \cdots, n \}$ a vertex of the convex minorant if it is a point such that $F^{(n)}(\tau) < F^{(n)}(Y_i)$ for any $Y_i > \tau$, that is H^* changes its slope at k if $\tau = Y_{(k)}$.

We know $F^{(n)} \to F^0$ on the continuous points of F^0. If we could show F^0 is strictly increasing, then Condition C will be satisfied. Suppose F^0 is not strictly increasing, there exist s_1 and s_2, $s_1 < s_2$, such that $F^0(s_1) = F^0(s_2)$. For our convenience, assume F^0 is continuous at s_1 and s_2, $F^0(s) < F^0(s_1)$ for any $s < s_1$, and $F^0(s) > F^0(s_2)$ for any $s > s_2$. If F^0 is not continuous at s_1 and/or s_2, we will consider $s_1+\varepsilon$ and/or $s_2-\varepsilon$ for small ε, and the proof is similar. Let $s_{1n} = \max \{ \tau \le s_1 : \tau \text{ is a vertex of } F^{(n)} \}$ and $s_{2n} = \min \{ \tau \ge s_2 : \tau \text{ is a vertex of } F^{(n)}\}$, then $s_{1n} \to s_1$ and $s_{2n} \to s_2$, since $F^{(n)} \to F^0$. Therefore, since H^* is convex minorant, we have

$$F^{(n)}(s_1) \le \frac{\text{the number of } \delta_i\text{'s being 1 in } (s_{1n}, s_{2n}]}{\text{the number of } \delta_i\text{'s in } (s_{1n}, s_{2n}]} \le F^{(n)}(s_2). \tag{3.23}$$

As $n \to \infty$, $F^{(n)}(s_1) \to F^0(s_1)$, $F^{(n)}(s_2) \to F^0(s_2) = F^0(s_1)$. The numerator of (3.23) becomes

$$\begin{aligned}\frac{1}{\text{n}}(\text{the number of } \delta_i\text{'s being 1 in } (s_{1n}, s_{2n}]) &= \frac{1}{\text{n}}\sum_{i=1}^{n}[T_i \le Y_i, Y_i \in (s_{1n}, s_{2n}]], \\ &= -\int_{s_{1n}}^{s_{2n}} dW_1^{(n)}(y), \\ &\xrightarrow{n\to\infty} \int_{s_1}^{s_2} F(y)dF_Y(y),\end{aligned}$$

since $W_1^{(n)} \to W_1$ and $W_1(t) = \int_t^{\infty} F(s) dF_Y(s)$ by (3.1). The denominator of (3.23) becomes

$$\frac{1}{\text{n}}(\text{the number of } \delta_i\text{'s in } (s_{1n}, s_{2n}]) = \frac{1}{\text{n}} \sum_{i=1}^{n} [Y_i \in (s_{1n}, s_{2n}]],$$

$$\xrightarrow{n \to \infty} \int_{s_1}^{s_2} dF_Y(y).$$

So, as $n \to \infty$, we have

$$\frac{\int_{s_1}^{s_2} F(y) dF_Y(y)}{\int_{s_1}^{s_2} dF_Y(y)} = F^0(s_1). \tag{3.24}$$

By Mean Value Theorem of integral, there exists $\theta \in (s_1, s_2)$ such that $F(\theta) = F^0(s_1)$. But H^* is convex minorant, so

$$\frac{\text{the number of } \delta_i\text{'s being 1 in } (s_{1n}, \theta)}{\text{the number of } \delta_i\text{'s in } (s_{1n}, \theta)} \geq F^{(n)}(s_1).$$

Following the same argument as obtaining (3,24) as $n \to \infty$,

$$\frac{\int_{s_1}^{\theta} F(y) dF_Y(y)}{\int_{s_1}^{\theta} dF_Y(y)} \geq F^0(s_1) = F(\theta).$$

By the Mean Value Theorem of integral again, there exists $\theta_1 \in (s_1, \theta)$ such that $F(\theta_1) \geq F^0(s_1) = F(\theta)$, which is contrary to the fact that F is strictly increasing. Therefore F^0 has to be strictly increasing.

Proof of Lemma 4.2:

The lemma is trivially true when $\delta_j = 1$. We now assume all δ's are 0, otherwise we could apply lemma 4.1 so that the new prior would become a Dirichlet process with parameter β_1. We want to prove

$$\mathscr{P}(T_j > t|(Z_1,0),\cdots,(Z_n,0)) = \frac{\hat{S}(t \vee Z_j)}{\hat{S}(Z_j)},$$

or equivalently

$$\frac{\mathscr{E}(P(t \vee Z_j,\infty)\prod_{i \neq j} P(Z_i,\infty))}{\mathscr{E}(P(Z_j,\infty)\prod_{i \neq j} P(Z_i,\infty))} = \frac{\mathscr{E}(P(t \vee Z_j,\infty)\prod_{i} P(Z_i,\infty))}{\mathscr{E}(P(Z_j,\infty)\prod_{i} P(Z_i,\infty))}. \tag{4.0}$$

Denote the LHS of (4.0) by $\frac{A}{B}$, denote the RHS of (4.0) by $\frac{C}{D}$.

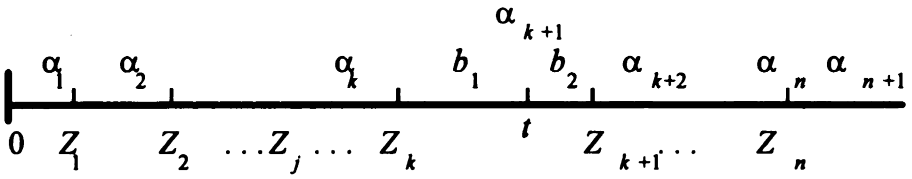

Figure 4.1

Without loss of generality, we assume $Z_1 < \cdots < Z_n$. (4.0) is obviously true if $t \leq Z_j$, so let $t \in (Z_k, Z_{k+1}]$, and let $B_1 = (0, Z_1]$, $B_2 = (Z_1, Z_2], \cdots, B_n = (Z_{n-1}, Z_n]$, $B_{n+1} = (Z_n, \infty)$, and $\alpha_1 = \alpha(B_1)$, $\alpha_2 = \alpha(B_2), \cdots, \alpha_n = \alpha(B_n)$, $\alpha_{n+1} = \alpha(B_{n+1})$.

Let $b_1 = \alpha(Z_k, t]$ and $b_2 = \alpha(t, Z_{k+1}]$, then $b_1 + b_2 = \alpha_{k+1}$ (see Figure 4.1). Use

$$\int_0^c x^{\gamma-1}(c-x)^{\eta-1}dx = c^{\gamma+\eta-1}B(\gamma,\eta) \quad \text{for } 0 \leq c \leq 1 \text{ and } \gamma, \eta > 0,$$

where $B(\gamma,\eta) = \frac{\Gamma(\gamma)\Gamma(\eta)}{\Gamma(\gamma+\eta)}$, and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$. Then, we have

$$A = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n)} \alpha_{n+1}(\alpha_{n+1} + \alpha_n + 1) \cdots (\alpha_{n+1} + \cdots + \alpha_{k+2} + n+1-(k+2))$$

$$(\alpha_{n+1} + \cdots + \alpha_{k+2} + b_2 + n+1-(k+2)+1)(\alpha_{n+1} + \cdots + \alpha_{k+1} + n+1-(k+1)+1) \cdots$$

$$(\alpha_{n+1} + \cdots + \alpha_{j+2} + n+1-(j+2)+1)(\alpha_{n+1} + \cdots + \alpha_j + n+1-j) \cdots (\alpha_{n+1} + \cdots + \alpha_2 + n-1),$$

$$C = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n+1)}\ \alpha_{n+1}(\ \alpha_{n+1}\ + \alpha_n+1) \cdots (\ \alpha_{n+1}+ \cdots + \alpha_{k+2}+ n+1\text{-}(k+2))$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_{k+2}+ b_2 +n+1\text{-}(k+2)+1\ )(\ \alpha_{n+1}+ \cdots + \alpha_{k+1}+ b_2 +n+1\text{-}(k+1)+1\ ) \cdots$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_{j+2}+n+1\text{-}(j+2)+1\ )(\ \alpha_{n+1}+ \cdots + \alpha_{j+1}+ n+1\text{-}(j+1) +1)$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_j+ n+1\text{-}j +1) \cdots (\ \alpha_{n+1}+ \cdots +\ \alpha_2+n),$$

$$B = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n)}\ \alpha_{n+1}(\ \alpha_{n+1}\ + \alpha_n+1) \cdots\ (\ \alpha_{n+1}+ \cdots + \alpha_{k+1}+ n+1\text{-}(k+1)\ ) \cdots$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_{j+2}+n+1\text{-}(j+2)\ )\ (\ \alpha_{n+1}+ \cdots + \alpha_{j+1}+ n+1\text{-}(j+1)\ )$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_j+ n+1\text{-}j\ ) \cdots (\ \alpha_{n+1}+ \cdots +\ \alpha_2+n\text{-}1),$$

and

$$D = \frac{\Gamma(\alpha(R^+))}{\Gamma(\alpha(R^+)+n+1)}\ \alpha_{n+1}(\ \alpha_{n+1}\ + \alpha_n+1) \cdots\ (\ \alpha_{n+1}+ \cdots + \alpha_{k+1}+ n+1\text{-}(k+1)\ ) \cdots$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_{j+2}+n+1\text{-}(j+2)\ )\ (\ \alpha_{n+1}+ \cdots + \alpha_{j+1}+ n+1\text{-}(j+1) +1)$$

$$(\ \alpha_{n+1}+ \cdots + \alpha_{j+1}+ n+1\text{-}(j+1)\ )\ (\ \alpha_{n+1}+ \cdots + \alpha_j+ n+1\text{-}j+1\ ) \cdots (\ \alpha_{n+1}+ \cdots +\ \alpha_2+n).$$

So,

$$\frac{A}{C}\cdot\frac{\Gamma(\alpha(R^+)+n+1)}{\Gamma(\alpha(R^+)+n)}$$

$$= \frac{(\alpha_{n+1}+\cdots+\alpha_j+n+1-j)\cdots(\alpha_{n+1}+\cdots+\alpha_2+n-1)}{(\alpha_{n+1}+\cdots+\alpha_{j+1}+n+1-(j+1)+1)(\alpha_{n+1}+\cdots+\alpha_j+n+1-j+1)\cdots(\alpha_{n+1}+\cdots+\alpha_2+n)},$$

$$= \frac{\prod_{i=1}^{j-1}(\alpha(Z_i,\infty)+N^+(Z_i))}{\prod_{i=1}^{j}(\alpha(Z_i,\infty)+N(Z_i))};$$

and

$$\frac{B}{D}\cdot\frac{\Gamma(\alpha(R^+)+n+1)}{\Gamma(\alpha(R^+)+n)}$$

$$=\frac{(\alpha_{n+1}+\cdots+\alpha_j+n+1-j)\cdots(\alpha_{n+1}+\cdots+\alpha_2+n-1)}{(\alpha_{n+1}+\cdots+\alpha_{j+1}+n+1-(j+1)+1)(\alpha_{n+1}+\cdots+\alpha_j+n+1-j+1)\cdots(\alpha_{n+1}+\cdots+\alpha_2+n)},$$

$$=\frac{\prod_{i=1}^{j-1}(\alpha(Z_i,\infty)+N^+(Z_i))}{\prod_{i=1}^{j}(\alpha(Z_i,\infty)+N(Z_i))}.$$

Hence, we get $\frac{A}{C}=\frac{B}{D}$, that is $\frac{A}{B}=\frac{C}{D}$.

REFERENCES

CHANG M.N. & YANG G.L. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* 1987, **15**, 1536-1547.

CHANG M.N. Weak Convergence of a Self-consistent Estimator of the Survival Function with Doubly Censored Data. *Ann. Statist.*, 1987, **5**, 1536-1547.

FERGUSON T.S. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1973, **1**, 209-230.

GARDINER J.C. & SUSARLA V. Weak convergence of a Bayesian nonparametric estimator of the survival function under progressive censoring. *Statistics and Decisions*, 1983, **C**, 257-268.

GROENEBOOM P. & WELLNER J.A. Information bounds and nonparametric maximum likelihood estimation. Birkhäuser Verlag, Basel. 1992.

KAPLAN E.L. & MEIER P. Nonparametic Estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 1958, **53**, 457-481.

LAWLESS J.E. Statistical models and Methods for Lifetime Data. Wiley, New York, 1982.

LEIDERMAN P.H., Babu D., et. al. African infant precocity and some social influences during the first year. *Nature*, 1973, **242**, 247-249.

PECKHAM C.S. Children born to women with HIV-I infection: Natural history and risk of transmission. The European Collaborative Study. *Lancet*, 1991, **337**, 253-260.

PETO R. Experimental survival curves for interval-censored data. *Appl. Statist.*, 1973, **22**, 86-91.

SAMUELSON S.O. Asymptotic theory for nonparametric estimators from doubly censored data. *J. Statist.*, 1989, **16**, 1-21.

SUSARLA V. & VAN RYZIN J. Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations. *J. Amer. Statist. Assoc.*, 1976, **61,** 897-902.

TURNBULL B.W. The empirical distribution function from arbitrarily grouped, censored and truncated data. *J. Royal Statist. Soc. Ser. B.,* 1976, **38**, 290-295.