The Analysis of Interval-Censored Survival Data. From a Nonparametric Perspective to a Nonparametric Bayesian Approach

M.Luz Calle i Rosingana

Memòria presentada per a aspirar al grau
de Doctor en Matemàtiques.
Dirigida per la Dra. Guadalupe Gómez Melis
Dept. d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya.

Vic, Novembre del 1996

a la meva família
i, molt especialment, a en Toni

Contents

Chapter 1

Introduction

This work concerns some problems in the area of survival analysis that arise in real clinical or epidemiological studies. In particular, we approach the problem of estimating the survival function based on interval-censored data or doubly-censored data. We will start defining these concepts and presenting a brief review of different methodologies to deal with this kind of censoring patterns.

Survival analysis is the term used to describe the analysis of data that correspond to the time from a well defined origin time until the occurrence of some particular event of interest. This event need not necessarily be death, but could, for example, be the response to a treatment, remission from a disease, or the occurrence of a symptom.

The reason why standard statistical methods are not appropriate in this setting is that the exact survival times of some subjects are sometimes not observed. The most common instance of such incomplete observation is **right censoring**. An individual is said to have a right-censored survival time when it is only known that his/her survival time exceeds some specific value. This is usually the case when the data from a study are to be analyzed at a point in time when some of the subjects have not yet experienced the event of interest. For example, in the context of a medical research where the end point is the death of a patient, the variable of interest is literally a survival time; data may be right-censored because some of the patients are alive at the end of the study period or because some of them have been lost during the follow-up.

Another form of censoring is **left censoring**, which is encountered when the actual survival time of an individual is less than what has been observed. For example, in a study of children's ability to perform some task, at the time of recruitment some children may already know how to perform that task and, therefore, the time from birth to performance of the task is, for these children, left-censored. Methods for analyzing right-censored data can be adapted to deal with left censoring (Csörgo [14], Gmez *et alt.* [34] and [35]).

In our work we are mainly interested in another type of censoring, the so-called **interval censoring**. Interval censoring arises when the time variable of interest cannot be directly observed and it is only known to have occurred during a particular interval of time. This situation is quite usual in many longitudinal studies where the event of interest, for example the occurrence of a symptom, can only be observed at the time of a medical examination. In this case, the time until occurrence of a symptom is only known to lie in the time interval between the last examination without symptoms and the first examination with symptoms. Right censoring can be viewed as a special case of interval censoring. Indeed, when data are right-censored the survival time is either known exactly or it is known to exceed the follow-up time. In the first case, the censoring interval is degenerated into a point and in the second case the censoring interval is the time interval from the end of follow-up to infinity. An example of interval censoring appears in Finkelstein and Wolfe [21]. They analyze the data from a breast cancer study where the variable of interest is time until cosmetic deterioration for a cohort of breast cancer patients who were treated with two different therapies. This variable is interval-censored because the status of the patients could only be established when the patient is examined at a medical visit. Then, the time until cosmetic deterioration is known to be some value between the time of the last examination without evidence of cosmetic deterioration and the time of the first examination where the deterioration was observed.

Right censoring has been widely studied and there are several methodologies for dealing with this kind of data, from completely parametric approaches to completely nonparametric ones. However, the techniques for analyzing interval-censored data have not been developed to the same extent.

Parametric approaches are often based on the maximum likelihood method. Under this scenario, a specific parametric model for the survival times is assumed and the estimation of the vector of parameters of the distribution based on a right-censored sample becomes straightforward. Indeed, after deriving the form of the likelihood function for the censored sample, the maximization of this function through an iterative procedure, such as Newton-Raphson method, provides the maximum likelihood estimators of the parameters. Under interval censoring the expression of the likelihood may be more complicated and, therefore, its maximization more annoying. In this case an alternative approach to obtain the maximum likelihood estimator of the vector of parameters is the expectation-maximization (E-M) algorithm [17]. The advantage of this method is that it only requires computations involving the likelihood function of an uncensored sample that is usually simpler than the likelihood function based on censored data. However, the parametric methods require the specification of the functional form that the survival function would have had in the absence of censoring.

The product-limit estimator developed by Kaplan and Meier [40] in 1958 is the pioneer work for estimating a survival curve nonparametrically. This first work, and most of the papers that followed it, considered the right-censored case. Twenty years later Turnbull, [56] and [57], proposes an extension of the product-limit estimator to deal with interval-censored data, among other censoring patterns. The main idea of Turnbull's new methodology is to establish the self-consistent equations and to solve them iteratively. Twenty more years have had to pass since the first proposal of these techniques for the practical implementation of them, presently fueled by the general availability of powerful computers.

Most of finite sample and large sample properties of the survival estimators have been established using a counting process framework. A counting process is a stochastic process adapted to a filtration and whose paths are, with probability one, right continuous, piecewise constant and have only jump discontinuities, with jumps of size $+1$. The term counting process suggests their more frequent application, that is, it will almost always denote the number of events of a certain type occurring in a given interval. Counting process methodology follows conditional arguments from where the corresponding compensators and subsequent martingales are derived. Martingale theory, mainly the large sample central limit theorem, provides the tools to derive the asymptotic properties of our survival estimates (Fleming and Harrington [24]). It is then relevant to be able to define such a filtration, that is an increasing family of sub-$\sigma-$algebras. When we are under a right censoring scheme the most natural filtration is the history of the stochastic process, and in this case the filtration at time t contains the information generated by the process on the interval $[0, t]$. However, if the random variable is interval-censored, at a given point in time we might not known whether the event of interest has occurred or not, and therefore a filtration cannot be defined. Therefore, this powerful methodology cannot be applied to the interval censoring situation.

In particular, large-sample properties such as weak convergence or strong consistency have been established for the Kaplan-Meier estimator while the asymptotic behaviour of Turnbull's estimator has only been established in special situations. In particular, we don't have consistent estimates for the variances of the survival estimates for continuous data because neither standard maximum likelihood methodology nor counting processes theory are directly applicable for interval-censored data. However, if we assume a discrete time scale, *i.e.*, if data are measured at a fixed number of points, standard maximum likelihood theory yields a consistent estimate of the variance of the estimator for the corresponding discrete distribution.

The advantages of these two approaches, parametric and nonparametric, are in general difficult to determine. On one hand, the nonparametric approach may represent an important loss of efficiency versus the use of a parametric method, if there is a scientific or empirical knowledge of the problem that justifies a model, specially if the variable is heavily censored. On the other hand, the parametric assumptions are in general difficult to assess based on a censored sample and, therefore, the use of completely parametric methodologies involves the risk of obtaining an inconsistent estimator if the parametric model does not fit suitably the data. An alternative to those opposed points of view is provided by the nonparametric Bayesian methodology. Susarla and Van Ryzin [51] derived a nonparametric Bayes estimator of the survival function for right-censored data. Their estimator is based on the class of Dirichlet processes *a priori* introduced by Ferguson [19]. They proved that the Bayes estimator includes the nonparametric Kaplan-Meier estimator as a special case and that both estimators are asymptotically equivalent. Furthermore, they proved that the nonparametric Bayes estimator has better small sample properties than the Kaplan-Meier estimator. Unfortunately, the extension of this theory to more complex censoring schemes is in general not straightforward because the corresponding nonparametric Bayes estimators are not obtainable in an explicit form. In particular, there is no generalization of the Susarla and Van Ryzin nonparametric Bayes estimator under interval censoring. For that reason, part of this work will be devoted to the derivation of a new methodology that provides a solution to this problem. This methodology is implemented by an iterative simulation procedure, the Gibbs sampler. The Gibbs sampler or, in general, the so-called Markov Chain Monte Carlo methods, provide algorithms to obtain random samples from a target distribution by simulating iteratively from conditional density functions. These methods have made a significant impact in practical statistics, since they provide numerical solution to otherwise intractable problems, specially in Bayesian analysis.

A special kind of interval censoring is found when the interval-censored variable is the origin time, and the final time is right-censored. This kind of data is called **doubly-censored** since both the initiating and the final times that define the survival or duration time of interest are censored. The case where the final event is as well interval-censored follows straightforwardly. This is typically a bivariate problem because the estimation of the duration time also involves the estimation of the initiating time and, therefore, standard univariate survival analysis techniques cannot be applied.

The early examples of this kind of censoring are found in the context of the AIDS epidemic studies. One of the most important aspects to understand the nature of the epidemic is the knowledge of the latency period distribution of AIDS. The estimation of this distribution is, however, particularly difficult, in part due to the length of the latency period but specially because the time of infection is usually unknown. Several studies to estimate this distribution are based on data provided by cohorts of haemophiliacs infected with HIV. The peculiarity of these cohorts is that, since blood samples were randomly stored in the hospitals, it is known for each individual the interval of time where the infection occurred, that is, the interval between the last negative and first positive antibody test. Therefore, the latency time is doubly-censored since its origin time is interval-censored and the final time, the time of onset of AIDS, may be right-censored.

In some studies double censoring is forced into a univariate problem by estimating the initiating time for each subject by the mid-point of the censoring interval. However, this approach is invalid unless the density of the initiating time is uniform within the censoring intervals. Other studies, Chiarotti *et alt.* [9] and [10], obtain a point estimate of the initiating time for each subject based on different parametric forms of the initiating time density. This might be a reasonable approach if the lengths of the censoring intervals are reasonably short, but, if this is not the case, and if the model is inadequate, the parametric assumption may introduce a significative bias.

A completely nonparametric methodology for analyzing doubly-censored data were first derived by De Gruttola and Lagakos [16]. They proposed an iterative algorithm to maximize the joint likelihood function for the origin time and the final time. Some practical problems related to the bivariate nature of the data were observed. To overcome these problems Gmez and Lagakos [36] proposed an alternative methodology based on maximizing two univariate likelihood functions. The method consists of two steps, in the first step they maximize the marginal likelihood function of the origin time and in the second step they maximize the conditional likelihood of the duration time given the estimated distribution of the origin time. Both methodologies are restricted to the case of a discrete time scale for the origin and the final times and, for this reason, in both cases nonidentifiability problems could arise, specially with small data sets.

The outline of my Ph.D. Thesis is the following:

Part I concerns the nonparametric approach for estimating a survival function based on doubly-censored data. In this context we propose a new algorithm for obtaining the maximum likelihood estimator of the survival function under double censoring that extends Gmez and Lagakos methodology to continuous time distributions.

In chapter 2 we first introduce the nonparametric methodologies of De Gruttola and Lagakos and the alternative two-step algorithm proposed by Gmez and Lagakos. In section 2.5 we derive the extension of Gmez and Lagakos (GL) algorithm that does not require a prior discretization of the data. This is done by adapting the self-consistent methodology for interval-censored data introduced by Turnbull [57] to the case of double censoring. The first step of the GL algorithm is easily extended to the continuous case. Indeed, this step corresponds to Turnbull's algorithm for the marginal likelihood of the interval-censored origin time. However, Turnbull's algorithm is not directly applicable for the estimation of the doubly-censored latency time distribution and, therefore, a specific procedure for maximizing the conditional likelihood is derived. In section 2.5.4 we prove that this algorithm includes the Kaplan-Meier estimator when the origin time is exactly observed for each individual. The methodology is illustrated with a cohort study of haemophiliacs that were at risk of infection with HIV in France in the early 80's. In chapter 3 we present the results of a simulation study that compares the local and global behaviour for small and moderate sample size of the algorithms studied in chapter 2.

Part II concerns the nonparametric Bayesian approaches for estimating a survival function. A new method for obtaining iteratively a nonparametric Bayes estimator of the survival function under interval censoring is proposed.

In chapters 4 and 5 we review the existing nonparametric Bayesian theory and the Markov Chain Monte Carlo methods, respectively. In particular, we describe in chapter 4 the works of Ferguson [19] and Susarla and Van Ryzin [51] for the nonparametric estimation of the survival function from complete and right-censored data, respectively, from a bayesian point of view. In chapter 5 we review the Metropolis-Hastings algorithm and the Gibbs sampler, and explain some practical techniques of inference and convergence diagnostic. In chapter 6 we propose a methodology, based on the iterative simulation method of Gibbs sampling, for obtaining the nonparametric Bayes estimator of the survival function for the case of interval censoring based on a Dirichlet process prior. The methodology is illustrated with the analysis of the data corresponding to a breast cancer study. The results of a simulation study to compare Turnbull's nonparametric method and the nonparametric Bayesian method are presented in chapter 7. On the basis of this simulation study, it appears that the use of the Bayes methodology is preferable, and specially, when the prior distribution is close to the theoretical distribution. This advantage is more important as the lenght of the censoring intervals increases. We conclude with a discussion of the results obtained and considerations on further areas of research.

All the computations have been carried out on a personal computer with a PENTIUM-S CPU at 90 MHz. The algorithms have been programmed using the C-program language. An appendix is included at the end with the programs to compute the estimators proposed in this work.

Part I

Chapter 2

Nonparametric Estimation of the Survival Function from Doubly–Censored Data

2.1 Introduction

In many longitudinal studies the interest relies on the so–called *duration time*, that is, the elapsed time between an originating event and a final event. Most statistical methods in survival analysis assume that the time to the originating event is known and allow the final time to be censored. We present here a situation where both the origin time and the final time are not directly observable. More precisely, we consider a sampling scheme where the origin time is interval-censored and the final time is right-censored. We refer to such data as *doubly–censored* data. This sampling scheme should not be confused with a different one, also referred to as doubly–censored data, where the final event is observed within a window for some subjects and left- or right-censored for others. (Turnbull [56], Chang and Yang [8]).

Doubly–censored data is found in the analysis of survival data which arise when a disease process is observed at several points in time, in general different for each patient. This scheme typically occurs in clinical trials or longitudinal studies in which there is periodic follow–up and the interest is based on both the time when a patient enters a first stage of a disease and on the elapsed time since this first stage to a second or final stage. The protocols of many clinical trials require that each patient visits the clinical center at specified successive times. At each visit, the status of the patient is examined and the occurrence of either one of two events, for instance, stage 1 and stage 2 of a given disease, is recorded. The actual visits, although scheduled in advance, are random because the patients often miss some of the appointments. As a consequence, the observation for each patient consist of the two random intervals where the first and second event have occurred. Thus, the elapsed time between the first and the second event is doubly–censored.

This sampling scheme can also be encountered in some studies of disease progression, where the only information about the initial event is obtained retrospectively, after periodical screening, providing for every individual a time–interval where the disease originated. In the context of the AIDS epidemic, several studies to estimate the latency distribution of AIDS have been based on data provided by cohorts of haemophiliacs infected with HIV. The retrospective inspection of their HIV infection status was possible because blood samples had been randomly stored in the hospitals. It was possible, then, to determine for each individual the interval where the infection had occurred, that is, the interval between the patient's last negative and first positive antibody test. Moreover, the time to AIDS was right–censored because many of the patients had not developed AIDS at the end of the study. Consequently, the latency time is doubly–censored. Note here that since the infection times may be censored into overlapping and nondisjoint intervals, methods for grouped data cannot be applied. This situation may be described, as in Frydman [25], by a three-state model :

$$\boxed{1\ \ \text{HIV}^{-}} \rightarrow \boxed{2\ \ \text{HIV}^{+}} \rightarrow \boxed{3\ \ \text{AIDS}}$$

where state 1 denotes non-infected, state 2 stands for infected and state 3 corresponds to clinical AIDS. The aim is the joint estimation of both, the distribution of time in state 1 and the distribution of time in state 2.

Nonparametric approaches to this problem have been considered by De Gruttola and Lagakos [16] and by Gómez and Lagakos [36]. De Gruttola and Lagakos propose a method (DGL in the sequel) for analyzing doubly–censored survival data in the context of the study of the progression from HIV infection to AIDS. They jointly estimate the infection time and the latency period between infection and onset of AIDS, by treating the data as a special type of bivariate survival data. Gómez and Lagakos approach this problem by developing a two-step estimation procedure (GL in the sequel). In the first step, they estimate the infection time distribution based on the marginal likelihood using the intervals where the infection is observed. Once a set of estimators for the infection probabilities is derived, they treat the interval–censored infection times as weighted exact infection times and estimate the latency distribution based on the corresponding conditional likelihood.

Other approaches to the problem have been taken by Baccheti [2] who estimates the latency time of AIDS by using an EM algorithm to maximize a penalized likelihood, by Frydman [25] who considers a three-state Markov process and develops a nonparametric maximum likelihood procedure for the estimation of the transition probabilities and the distribution functions of the times in every state. Brookmeyer and Goedert [4] and Kim, De Gruttola and Lagakos [41] propose semi–parametric procedures which allow the incorporation of covariates. Darby *et al* [15] adapt Brookmeyer and Goedert's model to fit data on the development of AIDS in haemophiliacs in the UK. Chiarotti *et alt.* [9],[10] estimates the median incubation time between HIV infection and AIDS, in a cohort of haemophiliacs in Italy, using different parametric models for the infection time and for the latency time.

We will focus our attention on the nonparametric approaches derived by De Gruttola and Lagakos [16] and by Gmez and Lagakos [36]. Gmez and Lagakos present a new algorithm as an alternative univariate methodology to overcome some of the practical problems observed with DGL algorithm. The difficulties with DGL method range from problems of convergence and speed of convergence to nonidentifiability problems. Gmez and Lagakos state that the two-step univariate methodology, GL algorithm, is more stable and converges faster than DGL algorithm. However, if the scale on which the origin and the final time are measured is too fine, problems of unstability and nonidentifiability might still remain. When this is the case, the standard approach discretizes the data into larger blocks, although this strategy may produce the lost of part of the initial information, specially with small data sets. The goal of this chapter is to extend Gmez and Lagakos methodology to overcome these difficulties. We propose a modification of the GL method in section 2.5 that makes the dimension of the problem as small as possible and avoids possible situations of nonidentifiability.

2.2 Data and Statistical Model

Following the notation in Gómez and Lagakos [36], let X and Z denote the chronological times of the originating and final events. Define the duration time to be $T = Z - X$. We wish to estimate the distribution functions, $W(x)$ and $F(t)$, of X and T under the assumption that X and T are independent random variables.

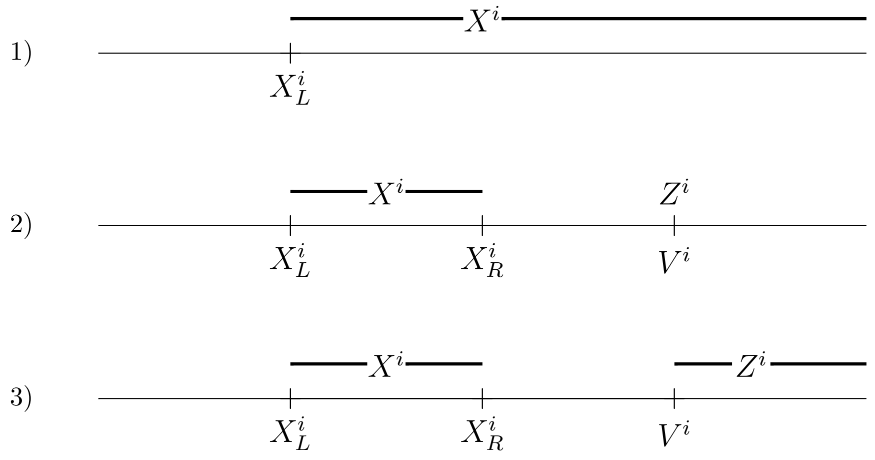
We assume that the time, X, of the originating event is interval–censored and the time, Z, of the final event is right–censored. That is, we observe the origin time X in an interval $[X_L, X_R]$ and V, the minimum between the final time Z and the time corresponding to the end of the study or the corresponding follow-up. Thus, for each subject i of a random sample of size n of a given population the observable data are of the form $(X_L^i, X_R^i, d^i, V^i, c^i)$ where d^i and c^i are the censoring indicators of the origin and final times, respectively. That is, $d^i = \mathbf{1}\{X_R^i < \infty\}$ and $c^i = 1$ if $Z^i = V^i$ and $c^i = 0$ if $Z^i > V^i$. We divide the observed data into three groups according to their censoring patterns:

1. The first group corresponds to those individuals with a right-censored origin time. In this case, $d^i = 0$ and $X_R^i = +\infty$ and this indicates that the first event had not yet occurred at the end of the study or at the time of the last follow-up. Thus, we only know that $X^i \geq X_L^i$ and have no information about the final time Z^i.

2. The second group corresponds to those individuals with an interval-censored origin time and an observed final event, that is, $d^i = 1$ and $c^i = 1$. For those individuals we know that $X_L^i \leq X^i \leq X_R^i$ and $Z^i = V^i$.

3. The last group corresponds to those individuals with an interval-censored origin time and a right-censored final time, that is $d^i = 1$ and $c^i = 0$. For those individuals we know that $X_L^i \leq X^i \leq X_R^i$ and that at the end of the study, V^i, the final event had not occurred, that is, $Z^i > V^i$.

These censoring schemes are outlined in the following diagram:



Thus, under the following assumptions

1. The origin time X and the latency time T are independent random variables,
2. the censoring scheme is noninformative in the sense that the censoring times X_L, X_R and V do not alter the following probabilities:

$$Pr(x_l \leq X \leq x_r | X_L = x_l, X_R = x_r) = Pr(x_l \leq X \leq x_r) = W(x_r) - W(x_l^-) \ ,$$

$$Pr(Z > z | V = z, c = 0, X_L, X_R) = Pr(Z > z) = 1 - F(z) \ ,$$

the **overall likelihood** based on the joint bivariate distribution of (X, T) is proportional to:

$$L_o(W, F) = \prod_{i=1}^{n} \left\{ \left[1 - W({X_L^i}^-) \right]^{1-d^i} \cdot \left[\int_{X_L^i}^{X_R^i} dW(x) \cdot dF(V^i - x) \ dx \right]^{d^i c^i} \cdot \right.$$

$$\left. \cdot \left[\int_{X_L^i}^{X_R^i} dW(x) \cdot \Big(1 - F(V^i - x) \Big) \ dx \right]^{d^i (1-c^i)} \right\}$$

where $dW(x) = W(x) - W(x^-)$ and $dF(t) = F(t) - F(t^-)$.

2.3 DGL Estimator

In De Gruttola and Lagakos [16] a discrete time scale for the origin time, say $0 < x_1 < x_2 < \ldots < x_r$, and a possible different scale for the latency time, say $0 < t_1 < t_2 < \ldots < t_s$, are assumed. This set of times will essentially induce a parametrization of the underlying distributions. Define $w_j = \text{Prob}(X = x_j)$, $f_k = \text{Prob}(T = t_k)$, $\mathbf{w} = (w_1, \ldots, w_r)$ and $\mathbf{f} = (f_1, \ldots, f_s)$.

Under the above assumptions the overall likelihood based on the joint bivariate distribution of (X, T) is proportional to:

$$L_o = L_o(\mathbf{w}, \mathbf{f}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{r} \sum_{k=1}^{s} \alpha_{jk}^i w_j f_k \right) \tag{2.1}$$

where α_{jk}^i equal 1 if $X_L^i \leq x_j \leq X_R^i$ and $V^i = x_j + t_k$ when $c^i = 1$ or
if $X_L^i \leq x_j \leq X_R^i$ $t_k > V^i - x_j$ when $c^i = 0$.

DGL method maximizes the overall likelihood L_o by a generalization of the Turnbull's [57] self–consistency algorithm to bivariate data. They define $\alpha^{\mathbf{i}} = \{\alpha_{jk}^i : 1 \leq j \leq r, 1 \leq$

$k \leq s\}$ and set I^i_{jk} equal to 1 if the true value of (X, T) for the ith individual is (x_j, t_k) and 0 otherwise. Then, the conditional expectation of I^i_{jk}, given $\alpha^{\mathbf{i}}$, is:

$$\mu^i_{jk} = \frac{\alpha^i_{jk} w_j f_k}{\sum_{l,m} \alpha^i_{lm} w_l f_m}, \tag{2.2}$$

and the corresponding marginal probabilities are

$$w^*_j = \sum_{i,k} \mu^i_{jk}/n \quad \text{and} \quad f^*_k = \sum_{i,j} \mu^i_{jk}/n \ . \tag{2.3}$$

A maximum likelihood solution, say $(\hat{\mathbf{w}}, \hat{\mathbf{f}})$, can be obtained following the iterative algorithm:

A. Choose starting values for $\mathbf{w}$ and $\mathbf{f}$.

B. Compute μ^i_{jk} from equation (2.2).

C. Compute refined estimates of $\mathbf{w}$ and $\mathbf{f}$ from (2.3).

D. Repeat steps (B) and (C) until convergence.

The maximum likelihood estimators of the distribution functions, W and F, are defined as:

$$\hat{W}(x) = \sum_{x_j \leq x} \hat{w}_j, \qquad \hat{F}(t) = \sum_{t_k \leq t} \hat{f}_k \ .$$

Furthermore, if we define the largest admissible mass points,

$$x^* = \max_{1 \leq i \leq n} \{X^i_R : X^i_R < \infty\} \quad \text{and} \quad t^* = \max_{1 \leq i \leq n} \{V^i - X^i_L \ \text{ when } \ d^i = 1 \text{ and } c^i = 1\} \ ,$$

then, $\hat{W}(x)$ puts all of its mass at values of x no greater than x^* provided that

$$x^* \geq \max_{1 \leq i \leq n} \{X^i_L\};$$

otherwise, $\hat{W}(x^*) < 1$ and $\hat{W}(x)$ is not uniquely defined for $x > x^*$.
Similarly, $\hat{F}(t)$ puts all of its mass at $t \leq t^*$ provided that

$$t^* \geq \max_{1 \leq i \leq n} \{V^i - X^i_L \text{ when } c^i = 0 \text{ and } d^i = 1\};$$

otherwise, $\hat{F}(t^*) < 1$ and $\hat{F}(t)$ is not uniquely defined for $t > t^*$.

2.4 GL Estimator

As in 2.3, a discrete time scale for the origin time and for the latency time are assumed.

The approach of Gmez and Lagakos [36] follows a two step procedure. In the first step, the infection time distribution based on the marginal likelihood is estimated. Once a set of estimators for the infection probabilities is derived, the latency distribution based on the corresponding conditional likelihood is estimated.

The marginal likelihood for $\mathbf{w}$, corresponding to the data $(X_L^i, X_R^i), \quad i = 1, \ldots, n$, is proportional to:

$$L_{\text{marg}}(\mathbf{w}) = \prod_{i=1}^{n} \left\{ \Big[\sum_{x_j = X_L^i}^{X_R^i} w_j \Big]^{d^i} \cdot \Big[1 - W(X_L^i -) \Big]^{1-d^i} \right\}, \tag{2.4}$$

and the conditional likelihood for $\mathbf{f}$, given $\mathbf{w}$, is proportional to:

$$L_c(\mathbf{f}) = \prod_{i=1}^{n} \left\{ \Big[\sum_{x_j = X_L^i}^{X_R^i} w_j \cdot dF(V^i - x_j) \Big]^{d^i c^i} \cdot \Big[\sum_{x_j = X_L^i}^{X_R^i} w_j \cdot \Big(1 - F(V^i - x_j) \Big) \Big]^{d^i(1-c^i)} \right\}. \tag{2.5}$$

FIRST STEP: Define the indicator variables $\alpha_j^i = \mathbf{1}\{x_j \in [X_L^i, X_R^i]\}$. A self-consistent equation for the infection time x_j, is given by

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_j^i w_j}{\sum_{l=1}^{r} \alpha_l^i w_l} \quad \text{for} \quad j = 1, \ldots, r \tag{2.6}$$

and a maximum likelihood solution, say $\hat{\mathbf{w}}$, can be obtained adapting Turnbull's univariate iterative algorithm:

A. Choose starting values for $\mathbf{w}$: $\mathbf{w}^{(0)} = (w_1^{(0)}, \ldots, w_r^{(0)})$.

B. Obtain improved estimates for $\mathbf{w}^{(1)}$ from equation (2.6).

C. Stop if the required accuracy has been achieved. Otherwise, return to step B with $\mathbf{w}^{(1)}$ replacing $\mathbf{w}^{(0)}$.

SECOND STEP: A self-consistent equation for the latency time f_k is given by

$$f_k = \frac{1}{n - n_0} (N_1(k) + N_2(k)) \quad \text{for} \quad k = 1, 2, \ldots, s, \qquad \text{where} \tag{2.7}$$

$$
\begin{aligned}
N_1(k) &= \sum_{i=1}^{n} d^i \cdot c^i \cdot \frac{\alpha^i_{V^i-t_k}\hat{w}_{V^i-t_k} f_k}{\sum_{j=1}^{r} \alpha^i_j \hat{w}_j dF(V^i - x_j)} \ , \\
N_2(k) &= \sum_{i=1}^{n} d^i \cdot (1-c^i) \cdot \frac{\sum_{j=1}^{r} \alpha^i_j \mathbf{1}\{V^i < x_j + t_k\} \hat{w}_j f_k}{\sum_{j=1}^{r} \alpha^i_j \hat{w}_j \{1 - F(V^i - x_j)\}}
\end{aligned}
$$

and $n_0 = \sum_{i=1}^{n}(1-d^i)$ is the number of individuals with a right-censored origin time. $N_1(k)$ and $N_2(k)$ represent the expected number of individuals that have developed the final event at time t_k among those uncensored and censored individuals, respectively. A maximum likelihood solution, say $\hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_s)$, to the self-consistency equation (2.7) can be obtained via an iterative method analogous to the one developed in the first step.

The maximum likelihood estimators of the distribution functions, W and F, are defined as in section 2.3.

2.5 Modified GL Estimator

As mentioned in the introduction, the strategy of most nonparametric methods to prevent from problems of unstability is to discretize the time scale. However, this yields to the lost of valuable information provided by the data. Even more, two different discretizations of the time scale may produce significative differences in the conclusions of a study. For this reason, we propose a modification of GL algorithm (ModGL in the sequel) that provides more stable nonparametric estimates without the need to make a priori discretization of the data. The new method makes the dimension of the problem as small as possible and avoids possible situations of unidentifiability.

ModGL estimator is obtained as the solution of a two step procedure similar to GL method. The difference between both methods is that in ModGL it is not necessary to make a priori discretization of the data. The first step of the algorithm corresponds to the nonparametric estimation of a distribution function when data are interval-censored based on the self-consistency method proposed by Turnbull [57]. In this step the marginal likelihood for W based on the censoring intervals $[X_L, X_R]$ is maximized by an iterative algorithm. The estimator obtained is denoted by $\hat{W}$. In the second step we maximize the conditional likelihood for F, given $\hat{W}$, based on solving the self-consistency equations. Turnbull's results are not directly applicable to this conditional likelihood because now we have doubly-censored and not interval-censored data.

Definition

The modified GL estimator $\hat{W}$ for the distribution function W of the origin time X is given by

$$\hat{W}(x) = \begin{cases} 0 \quad \text{if} \quad x < q_1 \\ \hat{s}_1 + \cdots + \hat{s}_k \quad \text{if} \quad p_k < x < q_{k+1} \\ 1 \quad \text{if} \quad x > p_m \end{cases}$$

and is undefined for $x \in [q_j, p_j]$, for $1 \leq j \leq m$; where $\hat{\mathbf{s}}$ satisfies the self-consistent equations (2.10) and the intervals $[q_j, p_j],\ j = 1, \ldots, m$ are defined below.
The modified GL estimator $\hat{F}$ for the distribution function F of the duration time T is given by

$$\hat{F}(t) = \begin{cases} 0 \quad \text{if} \quad t < q'_1 \\ \hat{f}_1 + \cdots + \hat{f}_k \quad \text{if} \quad p'_k < t < q'_{k+1} \\ 1 \quad \text{if} \quad t > p'_r \end{cases}$$

and is undefined for $t \in [q'_j, p'_j]$, for $1 \leq j \leq r$; where $\hat{\mathbf{f}}$ satisfies the self-consistent equations (2.18) and the intervals $[q'_j, p'_j],\ j = 1, \ldots, r$ are defined below.

2.5.1 FIRST STEP: Estimation of W based on the marginal likelihood

The marginal likelihood for W, given the observed data (X_L^i, X_R^i) is proportional to

$$L_{\text{marg}}(W) = \prod_{i=1}^{n} \left[W(X_R^i) - W(X_L^{i\,-}) \right]. \tag{2.8}$$

We will prove that the maximum likelihood estimator of W only puts mass in a set of intervals $C = \cup_{i=1}^{m} [q_i, p_i]$:

Construction of the set C

The set C is constructed from the data $\{(X_L^i, X_R^i),\ i = 1, \cdots, n\}$ as the union of disjoint closed intervals $[q_j, p_j], j = 1, \cdots, m$ satisfying the following conditions:

1. The left end point q_j lies in the set $\{X_L^i,\ i = 1, \cdots, n\}$,
2. the right end point p_j lies in the set $\{X_R^i,\ i = 1, \cdots, n\}$,
3. there is no members of $\{X_L^i\}$ or $\{X_R^i\}$ in the intervals $[q_j, p_j], j = 1, \cdots, m$ except at their end points,

4. the intervals $[q_j, p_j], j = 1, \cdots, m$ are disjoint and ordered: $q_1 \leq p_1 < q_2 \leq \cdots < q_m \leq p_m$. (Note that some of the intervals $[q_j, p_j]$ may be degenerated to a point).

The algorithm for obtaining this set of intervals is detailed at the end of this chapter.

We define $s_j = W(p_j) - W(q_j^-)$, the probability assigned to the intervals $[q_j, p_j]$ and define the vector $\mathbf{s} = (s_1, \cdots, s_m)$, where $\sum_{j=1}^m s_j = 1$.

$\hat{\mathbf{W}}$ is the maximum likelihood estimator of W

Applying Turnbull's results [57] to our special case it can be proved that:

Lemma 2.5.1 *Any distribution function W that maximizes the marginal likelihood $L_{\text{marg}}(W)$ has to be flat outside the set C.*

Lemma 2.5.2 *For fixed values of $W(p_j)$ and $W(q_j^-), 1 \leq j \leq m$, the likelihood $L_{\text{marg}}(W)$ is independent of the behaviour of W within each interval $[q_j, p_j]$.*

From these lemmas one concludes that two distributions functions that are flat outside C and with the same vector of masses $\mathbf{s}$ have the same likelihood. Therefore,

Theorem 2.5.3 *The maximization of $L_{\text{marg}}(W)$ reduces to the maximization of the function:*

$$L_X(s_1, \ldots, s_m) = \prod_{i=1}^{n} (\sum_{j=1}^{m} \alpha_j^i s_j) \tag{2.9}$$

where $s_j = W(p_j) - W(q_j^-)$ and the indicator α_j^i is defined as $\alpha_j^i = \mathbf{1}\{[q_j, p_j] \subseteq [X_L^i, X_R^i]\}$.

Self-consistent estimation

The maximization of $L_X(\mathbf{s})$ is based on the equivalence between maximum likelihood estimation and self-consistent estimation:

Theorem 2.5.4 (Turnbull) *If $\hat{\mathbf{s}}$ defines a maximum likelihood estimator for W, then $\hat{\mathbf{s}}$ satisfies the self-consistent equations given by*

$$ns_j = \sum_{i=1}^{n} \frac{\alpha_j^i s_j}{\sum_{l=1}^{m} \alpha_l^i s_l} \quad \text{for} \quad j = 1, \ldots, m. \tag{2.10}$$

And, conversely, any solution of the self-consistent equations maximizes $L_X(\mathbf{s})$.

The left-hand side of expression (2.10) represents the expected number of events that have occurred in the interval $[q_j, p_j]$ while the right-hand side corresponds to the expected number of events occurred in the same interval conditioned to the observed data.

The solution $\hat{\mathbf{s}}$ of the self-consistent equations (2.10) is obtained by the iterative procedure detailed in section 2.7.

2.5.2 SECOND STEP: Estimation of F based on the conditional likelihood

In the second step we maximize the conditional likelihood for F assuming that the cumulative distribution function for the origin time X is $\hat{W}$.

Conditional likelihood for F given $\hat{\mathbf{W}}$

Up to a proportionality constant, the **conditional likelihood** of F given $\hat{\mathbf{s}}$ is:

$$L_c(F) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{m} \alpha_j^i \hat{s}_j \left(F(V^i - q_j) - F((V^i - p_j)^-) \right) \right\}^{d^i c^i} \cdot$$

$$\left\{ \sum_{j=1}^{m} \alpha_j^i \hat{s}_j \left(1 - F(V^i - \frac{p_j + q_j}{2})^- \right) \right\}^{d^i (1-c^i)} \qquad (2.11)$$

This likelihood contains two types of factors corresponding to those individuals with an observed final time and those with a right-censored final time:

- **Contribution of an exact observation** ($c^i = 1$) **given that** $X^i \in [X_L^i, X_R^i]$

$$P(T^i = V^i - X^i \mid X^i \in [X_L^i, X_R^i]) =$$

$$\begin{aligned}
&= P(T^i = V^i - X^i,\ X^i \in [X_L^i, X_R^i]) / P(X^i \in [X_L^i, X_R^i]) = \\
&= \sum_{j=1}^{m} \alpha_j^i\ P(T^i = V^i - X^i,\ X^i \in [q_j, p_j]) / \sum_{l=1}^{m} \alpha_l^i\ P(X^i \in [q_l, p_l]) \\
&= \sum_{j=1}^{m} \alpha_j^i\ P(T^i = V^i - X^i \mid X^i \in [q_j, p_j]) \cdot P(X^i \in [q_j, p_j]) / \sum_{l=1}^{m} \alpha_l^i\ P(X^i \in [q_l, p_l]) = \\
&= \sum_{j=1}^{m} \alpha_j^i\ P(T^i \in [V^i - p_j, V^i - q_j] \mid X^i \in [q_j, p_j]) \cdot P(X^i \in [q_j, p_j]) / \sum_{l=1}^{m} \alpha_l^i\ P(X^i \in [q_l, p_l])
\end{aligned}$$

Using that X^i and T^i are independent random variables and that $P(X^i \in [q_j, p_j]) = \hat{s}_j$, this expression becomes:

$$\begin{aligned} & \sum_{j=1}^{m} \alpha_j^i \; P(T^i \in [V^i - p_j, V^i - q_j]) \cdot \hat{s}_j / \sum_{l=1}^{m} \alpha_l^i \; \hat{s}_l = \\ = \; & \sum_{j=1}^{m} \frac{\alpha_j^i \hat{s}_j}{\sum_{l=1}^{m} \alpha_l^i \hat{s}_l} \; [F(V^i - q_j) - F((V^i - p_j)^-)] \end{aligned}$$

• **Contribution of a right-censored observation** $(c^i = 0)$ **given that** $X^i \in [X_L^i, X_R^i]$

$$\begin{aligned} & P(T^i > V^i - X^i \mid X^i \in [X_L^i, X_R^i]) = \\ = \; & P(T^i > V^i - X^i, \; X^i \in [X_L^i, X_R^i]) / P(X^i \in [X_L^i, X_R^i]) = \\ = \; & \sum_{j=1}^{m} \alpha_j^i P(T^i > V^i - X^i, \; X^i \in [q_j, p_j]) / \sum_{l=1}^{m} \alpha_l^i P(X^i \in [q_l, p_l]) = \\ = \; & \sum_{j=1}^{m} \alpha_j^i P(T^i > V^i - X^i, \; X^i \in [q_j, p_j]) / \sum_{l=1}^{m} \alpha_l^i \hat{s}_l \end{aligned} \tag{2.12}$$

We decompose the probability in (2.12) as the sum of two parts:

$$\begin{aligned} & P(T^i > V^i - X^i, \; X^i \in [q_j, p_j]) = \\ = \; & P(T^i > V^i - q_j, \; X^i \in [q_j, p_j]) + \\ + \; & P(V^i - X^i < T^i < V^i - q_j, \; X^i \in [q_j, p_j]) = \\ = \; & \hat{s}_j (1 - F(V^i - q_j)) + \int_{q_j}^{p_j} \int_{V^i - x}^{V^i - q_j} dF(t) \; d\hat{W}(x) \; dt \; dx \end{aligned} \tag{2.13}$$

and approximate the integral in (2.13) by

$$\hat{s}_j \cdot (F(V^i - q_j) - F(V^i - \frac{p_j + q_j}{2})^-) \tag{2.14}$$

(see justification below). Hence, the contribution to the conditional likelihood of a right-censored observation is given by:

$$\sum_{j=1}^{m} \frac{\alpha_j^i \hat{s}_j}{\sum_{l=1}^{m} \alpha_l^i \hat{s}_l} \; \left(1 - F\left(V^i - \frac{p_j + q_j}{2}\right)^-\right)$$

Justification of the approximation (2.14):

The integral in (2.13) can only be computed if the joint distribution of X^i and T^i is known explicitly, at least, in the rectangle $\{(x,t):\ q_j \leq x \leq p_j,\ V^i - p_j \leq t \leq V^i - q_j\}$. Since X^i and T^i are assumed to be independent, it is only necessary to know their marginal distribution. We assume for simplicity that both X^i and T^i are uniformly distributed in $[q_j, p_j]$ and $[V^i - p_j, V^i - q_j]$, respectively. With this assumption, it is easy to see that the integral in (2.13) is equal to the integral obtained substituting x by the middle point of the interval $[q_j, p_j]$ (see the following figure):

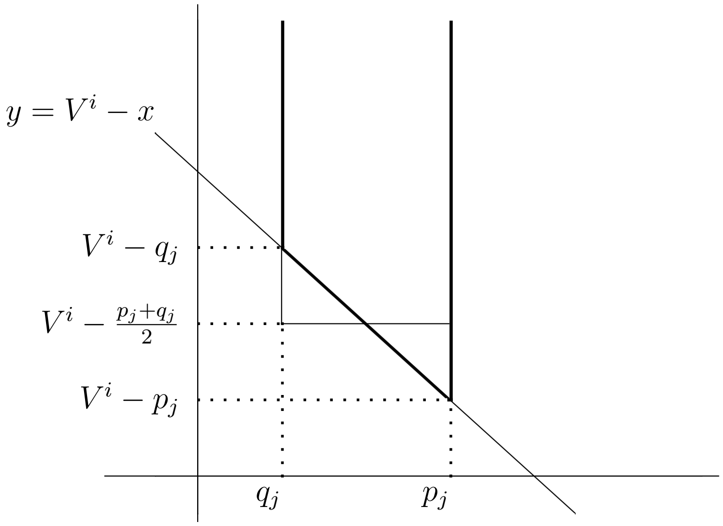


Figure 2.1: Justification of the approximation

Thus, expression (2.13) becomes

$$\begin{aligned} & \hat{s}_j(1 - F(V^i - q_j)) + \int_{q_j}^{p_j} \int_{V^i - x}^{V^i - q_j} dF(t)\ d\hat{W}(x)\ dt\ dx = \\ = \ & \hat{s}_j(1 - F(V^i - q_j)) + \int_{q_j}^{p_j} \int_{V^i - \frac{p_j+q_j}{2}}^{V^i - q_j} dF(t)\ d\hat{W}(x)\ dt\ dx = \\ = \ & \hat{s}_j(1 - F(V^i - q_j)) + \hat{s}_j(F(V^i - q_j) - F(V^i - \frac{p_j+q_j}{2})^-) = \\ = \ & \hat{s}_j \cdot (1 - F(V^i - \frac{p_j+q_j}{2})^-) \end{aligned}$$

This parametric assumption could seem a bit restrictive but it is important to note that the assumption is only made for the right-censored observations. Furthermore, the intervals $[q_j, p_j]$ and $[V^i - p_j, V^i - q_j]$ tend to be small and in many situations are degenerated into a point. In the rest of the admissible region, $\{(x,t) : x \in [q_j, p_j], t \in [V^i - q_j, +\infty)\}$, there is no parametric assumption.

Construction of the set C'

We now define the set of intervals $C' = \cup_{i=1}^{r}[q'_i, p'_i]$ where the maximum likelihood estimator of F gives positive mass.

It is useful first to unify the notation of the two factors of the conditional likelihood $F_c(t)$. With this purpose, we define for each individual the regions of its admissible latency times; that is, we define L_{ij} and R_{ij}, for $1 \leq i \leq n$ and $1 \leq j \leq m$, in the following way:

1. If $\alpha^i_j = 1$ and the ith observation is exact, $c^i = 1$, define $L_{ij} = V^i - p_j$ and $R_{ij} = V^i - q_j$.
2. If $\alpha^i_j = 1$ and the ith observation is right-censored, $c^i = 0$, define $L_{ij} = V^i - \frac{p_j + q_j}{2}$ and $R_{ij} = +\infty$.
3. If $\alpha^i_j = 0$, R_{ij} and L_{ij} are arbitrarily defined equal to 0.

Thus, the contribution of the ith observation to the conditional likelihood of F can be expressed as

$$\sum_{j=1}^{m} \frac{\alpha^i_j \hat{s}_j}{\sum_{l=1}^{m} \alpha^i_l \hat{s}_l} \left[F(R_{ij}) - F(L^-_{ij})\right] ,$$

that is , the duration time T^i lies in $[L_{ij}, R_{ij}]$ with probability $\alpha^i_j \hat{s}_j / (\sum_{l=1}^{m} \alpha^i_l \hat{s}_l)$ of having the origin X^i in the interval $[q_j, p_j]$. Therefore, the **conditional likelihood** of F is equivalently given by

$$L_c(F) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha^i_j \ \hat{s}_j \left[F(R_{ij}) - F(L^-_{ij})\right]\right]^{d^i} . \qquad (2.15)$$

We define C' as the union of disjoint intervals $[q'_1, p'_1]$, $[q'_2, p'_2], \ldots, [q'_r, p'_r]$ defined from $\{R_{ij}\}$ and $\{L_{ij}\}$ following similar steps to those used in the construction of C in the first step:

1. The left end point q'_k lies in the set $\{L_{ij},\ i = 1, \cdots, n \ \text{ and } j = 1, \cdots, m\}$,
2. the right end point p'_k lies in the set $\{R_{ij},\ i = 1, \cdots, n \ \text{ and } j = 1, \cdots, m\}$,

3. there is no members of $\{L_{ij}\}$ or $\{R_{ij}\}$ in the intervals $[q'_k, p'_k],\ k = 1, \cdots, r$ except at their end points,

4. the intervals $[q'_k, p'_k],\ k = 1, \cdots, r$ are disjoint and ordered: $q'_1 \leq p'_1 < q'_2 \leq \cdots < q'_r \leq p'_r$.

$\hat{\mathbf{F}}$ is the maximum likelihood estimator for F

Let $f_j = F(p'_j) - F((q'_j)^-)$ be the probability of the interval $[q'_j, p'_j]$ and define α^i_{jk}, the indicator of an origin time in $[q_j, p_j]$ and a duration time in $[q'_k, p'_k]$. That is,

$$\begin{aligned} \alpha^i_{jk} &= \mathbf{1}\{c^i = 1\} \cdot \mathbf{1}\left\{[q_j, p_j] \subseteq [X^i_L, X^i_R] \text{ and } [q'_k, p'_k] \subseteq [V^i - p_j, V^i - q_j]\right\} + \\ &+ \mathbf{1}\{c^i = 0\} \cdot \mathbf{1}\left\{[q_j, p_j] \subseteq [X^i_L, X^i_R] \text{ and } [q'_k, p'_k] \subseteq [V^i - \frac{p_j + q_j}{2}, +\infty)\right\}. \end{aligned} \tag{2.16}$$

Theorem 2.5.5 *The maximization of the conditional likelihood $L_c(F)$ reduces to the maximization of the function*

$$L_T(f_1, \ldots, f_r) = \prod_{i=1}^{n} \left[\sum_{k=1}^{r} \sum_{j=1}^{m} \alpha^i_{jk} \hat{s}_j f_k \right]^{d^i}. \tag{2.17}$$

This result follows from the following two lemmas.

Lemma 2.5.6 *Any distribution function F that maximizes the conditional likelihood $L_c(F)$ has to be flat outside the set C'.*

Proof.

Let F be a distribution function that increases outside $C' = \cup_{i=1}^{r}[q'_i, p'_i]$. In particular, suppose that F increases in the interval $[p'_l, q'_{l+1}]$. For construction of C', any R_{ij} in $[p'_l, q'_{l+1}]$ is smaller than any L_{ij} in this interval. Thus, there exist a real number $r_l \in [p'_l, q'_{l+1}]$ that separates the numbers R_{ij} and L_{ij} contained in $[p'_l, q'_{l+1}]$. We consider now the distribution function F^* defined equal to F outside the interval $[p'_l, q'_{l+1}]$ and constant inside it, that is:

$$F^*(t) = F(r_l),\ \forall t \in [p'_l, q'_{l+1}]$$
$$F^*(t) = F(t),\ \forall t \notin [p'_l, q'_{l+1}]$$

Then, for any $R_{ij} \in [p'_l, q'_{l+1}]$, $R_{ij} < r_l$ and since F increases in this interval, $F^*(R_{ij}) = F(r_l) > F(R_{ij})$. And, for any $L_{i'j'} \in [p'_l, q'_{l+1}]$, $L_{i'j'} > r_l$ and then $F^*(L_{i'j'}) = F(r_l) < F(L_{i'j'})$. Therefore,

$$L_c(F^*) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_j^i \ \hat{s}_j \left[F^*(R_{ij}) - F^*(L_{ij}^-) \right] \right] > \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_j^i \ \hat{s}_j \left[F(R_{ij}) - F(L_{ij}^-) \right] \right] = L_c(F)$$

and the function F cannot be a maximum of the conditional likelihood $L_c(F)$. □

Lemma 2.5.7 *For fixed values of $F(p'_j)$ and $F((q'_j)^-), 1 \leq j \leq r$, the likelihood $L_c(F)$ is independent of the behaviour of F within each interval $[q'_j, p'_j]$.*

Proof. The proof of this lemma is straightforward from the expression of $L_c(F)$.

□

Self-consistent estimation

As in the first step, the maximum solution of (2.17) is obtained through the self-consistent equations. The self-consistent equations for $\mathbf{f}$ are given by equating the expected number of observations with a duration time lying in $[q'_k, p'_k]$ and the expected number of observations with a duration time in the same interval, conditioned to the observed data; that is

$$(n - n_0) f_k = \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{m} \alpha_{jk}^i \hat{s}_j f_k}{\sum_{l=1}^{r} \sum_{j=1}^{m} \alpha_{jl}^i \hat{s}_j f_l} \right]^{d^i} \quad \text{for} \quad k = 1, \ldots, r \tag{2.18}$$

with $n_0 = \sum_{i=1}^{n} (1 - d^i)$ the number of observations with a right-censored origin time.

Now we prove the equivalence between maximum likelihood estimation and self-consistent estimation:

Theorem 2.5.8 *If $\hat{\mathbf{f}}$ is a maximum of the likelihood function $L_T(\mathbf{f})$ then $\hat{\mathbf{f}}$ satisfies the self-consistent equations (2.18). And, conversely, any solution $\hat{\mathbf{f}}$ of the self-consistent equations (2.18) maximizes $L_T(\mathbf{f})$.*

Proof.

The conditional likelihood $L_T(\mathbf{f})$ can also be expressed as

$$L_T(\mathbf{f}) = \prod_{i \in I_0} \left[\sum_{k=1}^{r} \alpha_k^i(\hat{\mathbf{s}}) f_k \right],$$

where $I_0 = \{i : \ d^i = 1\}$ and $\alpha_k^i(\hat{\mathbf{s}}) = \sum_{j=1}^m \alpha_{jk}^i \hat{s}_j$. Then, the log-likelihood is given by

$$l(\mathbf{f}) = \sum_{i \in I_0} \left[\log \left(\sum_{k=1}^r \alpha_k^i(\hat{\mathbf{s}}) f_k \right) \right].$$

Consider the function

$$H(\mathbf{u}) = l(\mathbf{f}) - \lambda(f_1 + \cdots + f_r - 1)$$

where λ is a Lagrange multiplier and $\mathbf{u} = (f_1, \cdots, f_r, \lambda)$.

We prove that the self-consistent equations (2.18) can be expressed equivalenty as

$$f_k = \left(\frac{d_k(H)}{n - n_0} + 1 \right) f_k; \qquad k = 1, \cdots, r, \tag{2.19}$$

where $d_k(H) = \dfrac{\partial}{\partial f_k}(H)$ and therefore a vector $\hat{\mathbf{f}} = (\hat{f}_1, \cdots, \hat{f}_r)$, with $\hat{f}_1 + \cdots + \hat{f}_r = 1$, is a solution of the self-consistent equations (2.18) if and only if $d_k(H) = 0, \ \forall k = 1, \cdots, r$, which is a necessary and sufficient condition to be a stationary point of $l(\mathbf{f})$.

Indeed, a necessary and sufficient condition for a vector $\mathbf{f} = (f_1, \cdots, f_r)$, with $f_1 + \cdots + f_r = 1$, to be an stationary point of $l(\mathbf{f})$ is that $\mathbf{f}$ is a solution of the following system:

$$\frac{\partial}{\partial \mathbf{u}}(H) = \frac{\partial}{\partial \mathbf{u}} \{ l(\mathbf{f}) - \lambda(f_1 + \cdots + f_r - 1) \} = 0 \ . \tag{2.20}$$

Expression (2.20) is a system of $r + 1$ equations of the form:

$$\begin{aligned} d_l(H) &= \frac{\partial}{\partial f_l} \left\{ \sum_{i \in I_0} \log(\sum_{k=1}^r \alpha_k^i(\hat{\mathbf{s}}) f_k) - \lambda(f_1 + \cdots + f_r - 1) \right\} = 0; \qquad l = 1, \cdots, r \\ d_\lambda(H) &= \frac{\partial}{\partial \lambda} \left\{ \sum_{i \in I_0} \log(\sum_{k=1}^r \alpha_k^i(\hat{\mathbf{s}}) f_k) - \lambda(f_1 + \cdots + f_r - 1) \right\} = 0. \end{aligned}$$

Computing the partial derivatives, one obtains the following equivalent system:

$$d_l(H) = \sum_{i \in I_0} \frac{\alpha_l^i(\hat{\mathbf{s}})}{\sum_{k=1}^r \alpha_k^i(\hat{\mathbf{s}}) f_k} - \lambda = 0; \qquad l = 1, \cdots, r \tag{2.21}$$

$$f_1 + \cdots + f_r = 1 \tag{2.22}$$

Multiplying each of the m equations in (2.21) by $f_l, \ l = 1, \cdots, r$ and adding them up we have

$$\sum_{l=1}^r f_l \cdot d_l(H) = \sum_{l=1}^r \sum_{i \in I_0} \frac{\alpha_l^i(\hat{\mathbf{s}}) f_l}{\sum_{k=1}^r \alpha_k^i(\hat{\mathbf{s}}) f_k} - \lambda(\sum_{l=1}^r f_l) = 0. \tag{2.23}$$

and after exchanging the order of sumation one obtains

$$\sum_{i\in I_0} \frac{\sum_{l=1}^{r} \alpha_l^i(\hat{\mathbf{s}}) f_l}{\sum_{k=1}^{r} \alpha_k^i(\hat{\mathbf{s}}) f_k} - \lambda(\sum_{l=1}^{r} f_l) = 0$$

and, since both the first term in the left expression and the sum of all f's are equal to 1, we obtain that $\lambda = \sum_{i\in I_0} 1 = n - n_0$.

Now, to prove that the solution of the self-consistent equations is a stationary point of the likelihood function we express the self-consistent equations in terms of the partial derivative of H with respect to f_k, $d_k(H)$. The right hand side of (2.18) is

$$\sum_{i\in I_0} \left\{ \frac{\alpha_k^i(\hat{\mathbf{s}})}{\sum_{l=1}^{r} \alpha_l^i(\hat{\mathbf{s}}) f_l} f_k \right\} = \left(d_k(H) + n - n_0\right) f_k; \qquad k = 1, \cdots, r. \tag{2.24}$$

The solution of the self-consistent equations will be obtained by the iterative algorithm

$$f_k^{(1)} = \left(\frac{d_k(H)}{n - n_0} + 1 \right) f_k^{(0)}; \qquad k = 1, \cdots, r.$$

It can be proved, as in Turnbull [57], that the above algorithm increases the log-likelihood $l(\mathbf{f}) = logL_T\mathbf{f}$ in each step and thus, the algorithm converges to a maximum or a saddle-point:

$$\begin{aligned}
l(\mathbf{f}^{(1)}) - l(\mathbf{f}^{(0)}) &= \sum_{k=1}^{r} (f_k^{(1)} - f_k^{(0)}) \frac{\partial l}{\partial f_k} + O(\|\mathbf{f}^{(1)} - \mathbf{f}^{(0)}\|^2) \simeq \\
&\simeq \frac{1}{n - n_0} \sum_{k=1}^{r} d_k(H) f_k^{(0)} \frac{\partial l}{\partial f_k} = \\
&= \frac{1}{n - n_0} \sum_{k=1}^{r} d_k(H) f_k^{(0)} \left(d_k(H) + n - n_0\right) = \\
&= \frac{1}{n - n_0} \sum_{k=1}^{r} d_k^2(H) f_k^{(0)} + \sum_{k=1}^{r} d_k(H) f_k^{(0)} = \\
&= \frac{1}{n - n_0} \sum_{k=1}^{r} d_k^2(H) f_k^{(0)} \geq 0
\end{aligned}$$

where the terms of second and higher order have been neglected.

□

2.5.3 Estimation of the variance of $\hat{W}$ and $\hat{F}$

When data arise from a continuous time distribution the theoretical study of the asymptotic behaviour of the nonparametric estimators becomes very difficult. As a matter of fact, distributional theory for the Turnbull's estimator has only been established in the case where data consist of left-censored, right-censored and exactly known observations (Turnbull [56], Samuelsen [48], Chang [8], Groeneboom [37]) while in the general case of interval censoring there are no results for the asymptotic distribution of the estimator. To overcome this difficulty Turnbull [57] proposes to use instead, the asymptotic results for the discrete case. That is, he considers the discretization given by the set of intervals $\{[q_j, p_j];\ j = 1, \ldots, m\}$. We will follow the same approach, that is, we estimate the covariance matrix of the vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{f}}$ using the results obtained by Gmez and Lagakos [36] in the discrete case; that is, we assume that X only puts mass on the intervals $[q_1, p_1], \ldots, [q_m, p_m]$ and that T only puts mass on the intervals $[q'_1, p'_1], \ldots, [q'_r, p'_r]$.

Covariance matrix of the vector $\hat{\mathbf{s}}$

The asymptotic covariance matrix of the vector $\hat{\mathbf{s}}$ is approximated by the inverse of the observed information matrix $\mathcal{B}(\hat{\mathbf{s}})$, where the jk term is given by

$$\mathcal{B}(j,k) = -\frac{\partial}{\partial s_k}\left(\frac{\partial}{\partial s_j}\log L_X(\mathbf{s})\right)\Big|_{\mathbf{s}=\hat{\mathbf{s}}} = \sum_{i=1}^{n}\frac{(\alpha^i_j - \alpha^i_m)(\alpha^i_k - \alpha^i_m)}{\left(\sum_{j=1}^{m}\alpha^i_j \hat{s}_j\right)^2}$$

for $j, k = 1, \ldots, m-1$, and $L_X(\mathbf{s})$ is given in (2.9).

Covariance matrix of the vector $\hat{\mathbf{f}}$

In order to take into account the variability due to the estimation of $\mathbf{s}$ we consider the conditional likelihood $L_T(\mathbf{f})$ as a function of both $\mathbf{s}$ and $\mathbf{f}$:

$$L_T(\mathbf{s},\mathbf{f}) = \prod_{i=1}^{n}\left[\sum_{k=1}^{r}\sum_{j=1}^{m}\alpha^i_{jk}s_j f_k\right]^{d^i}$$

where α^i_{jk} is defined in (2.16).

Let $Z(\mathbf{s}) = \frac{\partial}{\partial \mathbf{s}} logL_X$ be the score vector for L_X and

$$U(\mathbf{s},\mathbf{f})) = [U_1(\mathbf{s},\mathbf{f}), U_2(\mathbf{s},\mathbf{f})]' = [\frac{\partial}{\partial \mathbf{s}} logL_T, \frac{\partial}{\partial \mathbf{f}} logL_T]'$$

the score vector for L_T. Let $B(\mathbf{s})$ and $I(\mathbf{s},\mathbf{f})$ be the corresponding information matrices and $I_{11} = I_{ss}(\mathbf{s},\mathbf{f})$, $I_{12} = I_{sf}(\mathbf{s},\mathbf{f})$, $I_{21} = I_{fs}(\mathbf{s},\mathbf{f})$ and $I_{22} = I_{ff}(\mathbf{s},\mathbf{f})$ the submatrices of the information matrix I. Define $\mathcal{B}(\hat{\mathbf{s}})$ and $\mathcal{I}(\hat{\mathbf{s}},\hat{\mathbf{f}})$ as the corresponding observed information matrices.

Gmez and Lagakos [36] proved in the discrete case that $\sqrt{n}(\hat{\mathbf{f}}-\mathbf{f})$ is asymptotically normal with mean zero and covariance matrix

$$V = I_{22}^{-1} + I_{22}^{-1}\left(I_{21} - \frac{2}{n}E(U_2 Z')\right)B^{-1}I_{12}I_{22}^{-1},$$

and they estimate V by substituting the information matrices by their corresponding observed matrices, $\mathcal{B}$, $\mathcal{I}_{12}$, $\mathcal{I}_{21}$ and $\mathcal{I}_{22}$, and replacing $E(U_2 Z')$ by its empirical counterpart, $\mathcal{E}$, that is the matrix with a jk term equal to $E_{jk} = (1/n)\sum_{i=1}^{n} E_{jk}^{i}$ where E_{jk}^{i} is the contribution of the i-th individual to the matrix $U_2 Z'$.
We propose to estimate the variance of $\hat{\mathbf{f}}$ by

$$\hat{V} = \mathcal{I}_{22}^{-1} + \mathcal{I}_{22}^{-1}\left(\mathcal{I}_{21} - \frac{2}{n}\mathcal{E}\right)\mathcal{B}^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1},$$

where the jk term of $\mathcal{E}$ is given by

$$\mathcal{E}(j,k) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{\sum_{j=1}^{m}(\alpha_{jk}^{i} - \alpha_{jr}^{i})\hat{s}_j}{\sum_{k=1}^{r}\sum_{j=1}^{m}\alpha_{jk}^{i}\hat{s}_j\hat{f}_k}\right]\left[\frac{\alpha_j^i - \alpha_m^i}{\sum_{j=1}^{m}\alpha_j^i\hat{s}_j}\right]\right\}$$
$$\text{for } j = 1,\ldots,m-1 \text{ and } k = 1,\ldots,r-1$$

and the jk term of the observed information matrices are

$$\mathcal{B}(j,k) = \sum_{i=1}^{n}\frac{(\alpha_j^i - \alpha_m^i)(\alpha_k^i - \alpha_m^i)}{\left(\sum_{j=1}^{m}\alpha_j^i\hat{s}_j\right)^2},\ \text{for } j,k = 1,\ldots,m-1$$

$$\mathcal{I}_{12}(j,k) = \mathcal{I}_{21}(k,j) = -\sum_{i=1}^{n} d^i\,\frac{P_{12}^i(j,k)\cdot P^i - P_1^i(j)\cdot P_2^i(k)}{(P^i)^2},$$
$$\text{for } j = 1,\ldots,m-1 \text{ and } k = 1,\ldots,r-1 \ \text{ and}$$
$$\mathcal{I}_{22}(j,k) = \sum_{i=1}^{n} d^i\,\frac{P_2^i(j)\cdot P_2^i(k)}{(P^i)^2}\ \text{ for } j,k = 1,\ldots,r-1$$

where

$$P^i = \sum_{k=1}^{r}\sum_{j=1}^{m}\alpha_{jk}^i\hat{s}_j\hat{f}_k\ ,$$

$$P_1^i(j) = \frac{\partial}{\partial s_j}(P^i) = \sum_{k=1}^{r}(\alpha_{jk}^i - \alpha_{mk}^i)f_k\ ,\ \text{for } j = 1,\ldots,m-1\ ,$$

$$P_2^i(k) = \frac{\partial}{\partial f_k}(P^i) = \sum_{j=1}^{m}(\alpha_{jk}^i - \alpha_{jr}^i)s_j\ \text{ for } k = 1,\ldots,r-1$$

and, for $j = 1,\ldots,m-1$ and $k = 1,\ldots,r-1$,

$$P_{12}^i(j,k) = \frac{\partial}{\partial s_j}\left(\frac{\partial}{\partial f_k}(P^i)\right) = (\alpha_{jk}^i - \alpha_{jr}^i) - (\alpha_{mk}^i - \alpha_{mr}^i)\ .$$

2.5.4 Relation with the Kaplan-Meier estimator

It is interesting to note that the maximum likelihood estimator $\hat{F}$ includes the Kaplan-Meier estimator [40] as a special case:

Proposition 2.5.9 *ModGL estimator for the latency distribution reduces to the usual product–limit estimator when the origin time is exactly known.*

Proof.

If the origin time is known exactly for each individual, then $X_L^i = X_R^i$, $\forall i$ and therefore the intervals $[q_j, p_j], j = 1, \cdots m$ and $[q'_k, p'_k], k = 1, \cdots, r$ are, all of them, degenerated to a point. Say $[q_j, p_j] = x_j,\ j = 1, \cdots, m$ and $[q'_k, p'_k] = t_k,\ k = 1, \cdots, r$.

Denoting by $N_1(k)$ the number of individuals with a latency time equal to t_k and $R(k)$ the number of individuals at risk of failure at time t_k, we will prove that the product-limit estimator

$$\hat{f}_k = \hat{\lambda}_k \prod_{l=1}^{k-1} \left(1 - \hat{\lambda}_l\right),$$

where $\hat{\lambda}_l = N_1(l)/R(l)$ if $R(i) > 0$ and $\hat{\lambda}_l = 0$ if $R(i) = 0$ is the hazard at time t_l, is a solution of the self-consistent equations (2.18) for $\mathbf{f}$

$$(n - n_0) f_k = \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{m} \alpha_{jk}^i \hat{s}_j f_k}{\sum_{l=1}^{r} \sum_{j=1}^{m} \alpha_{jl}^i \hat{s}_j f_l} \right]^{d^i} \quad \text{for} \quad k = 1, \ldots, r.$$

These equations can be expressed as the sum of two factors corresponding to those individuals with known final time and to those individuals with a right-censored final time, that is,

$$(n - n_0) f_k = N_1(k) + N_2(k)\ ,\ k = 1, \cdots, r \tag{2.25}$$

where the process $N_1(k)$ is defined as above and the process $N_2(k)$ represents the expected number of individuals with latency times equal to t_k who have not experienced a final event by the end of the study.

If the origin time is observed for each individual i, then $n_0 = 0$ and α_{jk}^i indicates an origin time equal to x_j and a latency time equal to t_k. That is, if $c^i = 1$, $\alpha_{jk}^i = \mathbf{1}\{X^i = x_j\}\mathbf{1}\{V^i - x_j = t_k\}$ and, if $c^i = 0$, $\alpha_{jk}^i = \mathbf{1}\{X^i = x_j\}\mathbf{1}\{V^i - x_j < t_k\}$. Then, the process $N_1(k)$ reduces to

$$N_1(k) = \sum_{i=1}^{n} c^i \sum_{j=1}^{m} \mathbf{1}\{X^i = x_j\}\mathbf{1}\{V^i - x_j = t_k\}\ ,$$

and $N_2(k)$ becomes

$$N_2(k) = \sum_{i=1}^{n}(1-c^i)\sum_{j=1}^{m}\mathbf{1}\{X^i = x_j\}\sum_{l=1}^{k-1}\mathbf{1}\{V^i - x_j = t_l\}\frac{f_k}{1-F(t_l)}.$$

If we define $M(l) = \sum_{i=1}^{n}(1-c^i)\sum_{j=1}^{m}\mathbf{1}\{X^i = x_j\}\mathbf{1}\{V^i - x_j = t_l\}$ the number of individuals censored at time t_l, then $N_2(k)$ can be expressed as

$$N_2(k) = \sum_{l<k} M(l)\frac{f_k}{1-F(t_l)}$$

and the self-consistent equations become

$$nf_k = N_1(k) + \sum_{l<k} M(l)\frac{f_k}{1-F(t_l)}, \tag{2.26}$$

or equivalently

$$N_1(k) = \left(n - \sum_{l<k}\frac{M(l)}{1-F(t_l)}\right) f_k \ . \tag{2.27}$$

We have to prove that $\hat{f}_k$ satisfies equation (2.27). Using that $1-\hat{F}(t_k) = \prod_{l=1}^{k-1}(1-\hat{\lambda}_l)$, the above equation becomes

$$N_1(k) = \frac{n\prod_{l=1}^{k-1}(1-\hat{\lambda}_l) - M(1)\prod_{l=2}^{k-1}(1-\hat{\lambda}_l) - \ldots - M(k-1)}{\prod_{l=1}^{k-1}(1-\hat{\lambda}_l)} \cdot \hat{f}_k. \tag{2.28}$$

Now, from the fact that $R(k) = R(k-1) - N_1(k-1) - M(k-1)$, it is easy to prove by induction that:

$$R(k) = n\prod_{l=1}^{k-1}(1-\hat{\lambda}_l) - M(1)\prod_{l=2}^{k-1}(1-\hat{\lambda}_l) - \ldots - M(k-1)$$

and hence, equation (2.28) reduces to

$$N_1(k) = \frac{R(k)}{\prod_{l=1}^{k-1}(1-\hat{\lambda}_l)} \cdot \hat{f}_k.$$

or, equivalently

$$\hat{f}_k = \hat{\lambda}_k \prod_{l=1}^{k-1}\left(1-\hat{\lambda}_l\right)$$

as we wanted to prove.

□

2.6 Application to a Cohort Study of Haemophiliacs

To illustrate the methodologies considered in this chapter, we analyze the data given in De Gruttola and Lagakos [16] of a cohort of haemophiliacs that were at risk of infection with the human inmunodeficiency virus, HIV. The cohort corresponds to 262 patients that were treated at the Hpital Kremlin Bictre and the Hpital Coeur des Yvelines in France since 1978 and were at risk of infection from the contaminated blood factor they received for their disease. Serum samples were routinely stored and subsequently they could be tested for presence of HIV antibodies. The data was divided in two subsets: 105 patients in the heavily-treated group, that is in the group of patients who received at least 1,000 $\mu g/kg$ of blood factor for at least one year between 1982 and 1985, and 157 patients in the lightly-treated group, corresponding to those patients who received less than 1,000 $\mu g/kg$ in each year. By August 1988, 197 patients had become infected (97 in the heavily-treated group and 100 in the lightly-treated group) and 43 of these had developed clinical symptoms of AIDS (29 in the heavily-treated group and 14 in the lightly-treated group). The comparison of the two treatment groups could allow an indirect evaluation of the effects of different viral doses on the risk of infection and on the risk of AIDS once infected.

The observations, based on a discretization of the time axis into 6-month intervals, are of the form $(X_L^i, X_R^i, d^i, V^i, c^i)$. X_L^i and X_R^i are the chronologic times of the patient's last negative and first positive antibody test, respectively, d^i stands for the infection indicator, V^i denotes the chronologic time of first clinical symptom of AIDS when $c^i = 1$ and, for those individuals who had not developed AIDS at the end of the study ($c^i = 0$), V^i is the time of the last blood sample tested.

In this example, it is difficult to appreciate the advantages of the modified GL estimator because the data in this study were reported after a discretization of the time scale.

The estimators for the infection times of seroconversion obtained by the three methods, DGL, GL and ModGL, are displayed in figures 2.2 and 2.3, corresponding to the heavily-treated group and the lightly-treated group, respectively. The three estimators for $W(x)$ are very similar. Comparing these two figures we see that there is a difference between the distribution of infection times in the two treatment groups. The heavily-treated group presents shorter times of infection than the lightly-treated group. For instance, while in the heavily-treated group half of the patients were infected before 1985, in the lightly-treated group the median is obtained one year later.

Figures 2.4 and 2.5 give the estimated cumulative distribution function of the latency

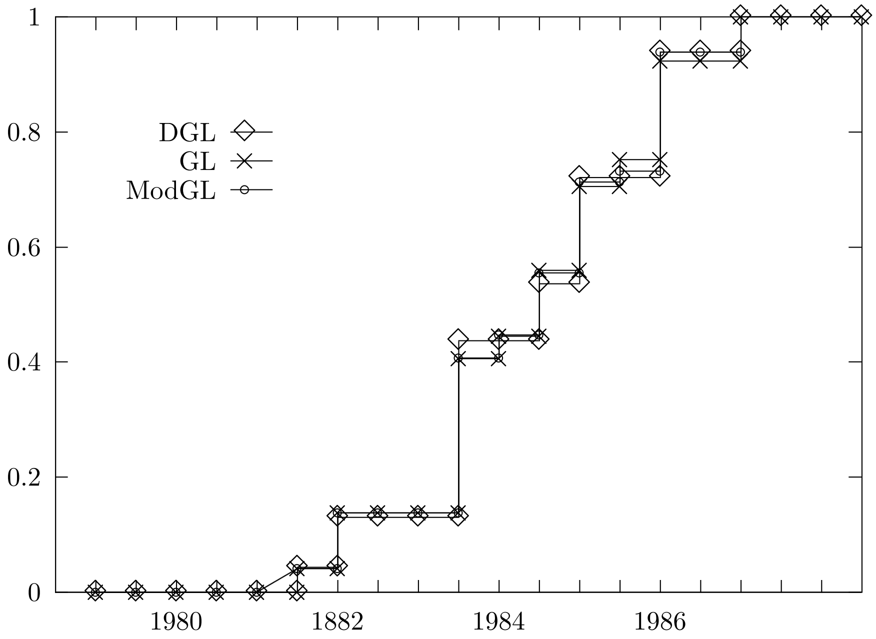


Figure 2.2: Estimated cumulative distribution function of time of HIV seroconversion for heavily-treated group.

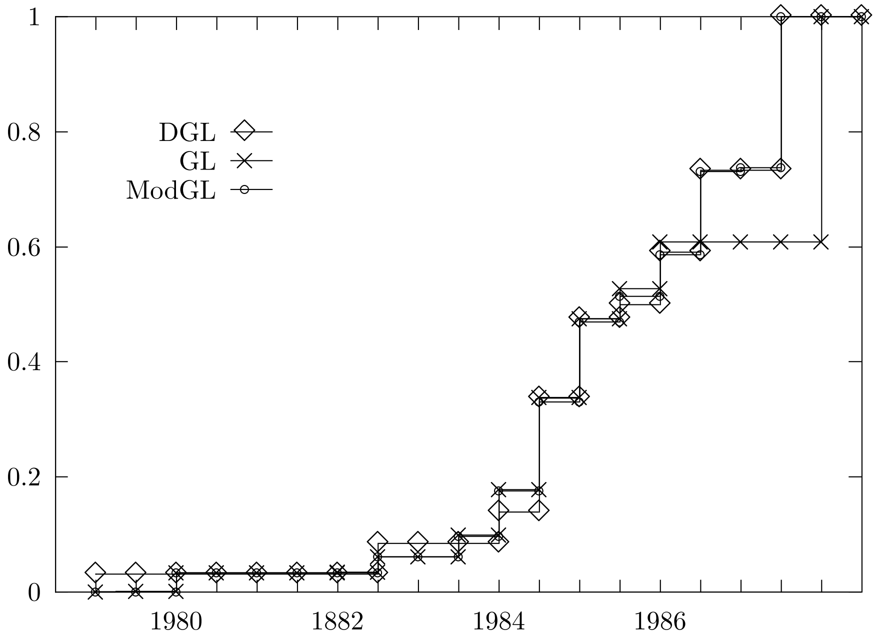


Figure 2.3: Estimated cumulative distribution function of time of HIV seroconversion for lightly-treated group.

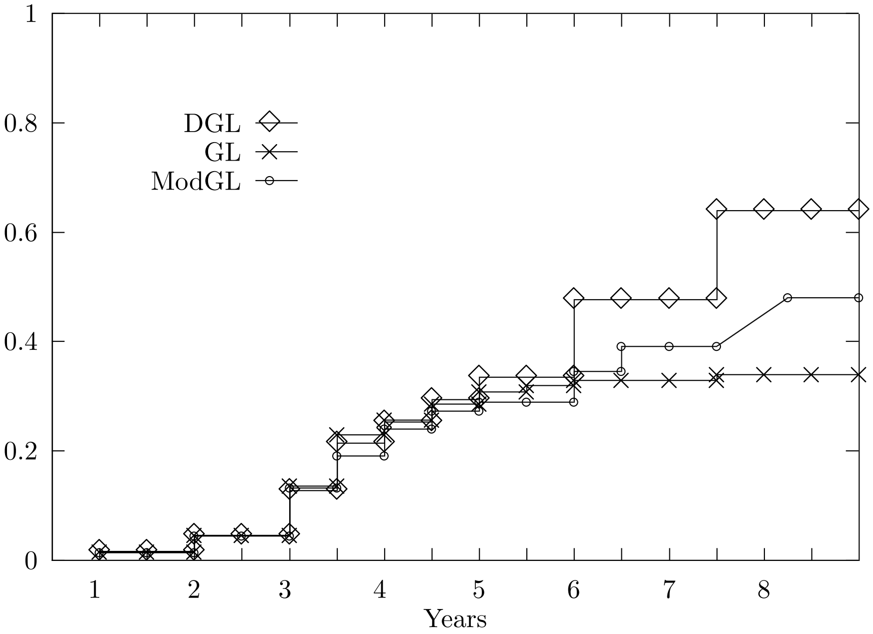


Figure 2.4: Estimated cumulative distribution function of latency time between HIV seroconversion and onset of symptoms for heavily-treated group.

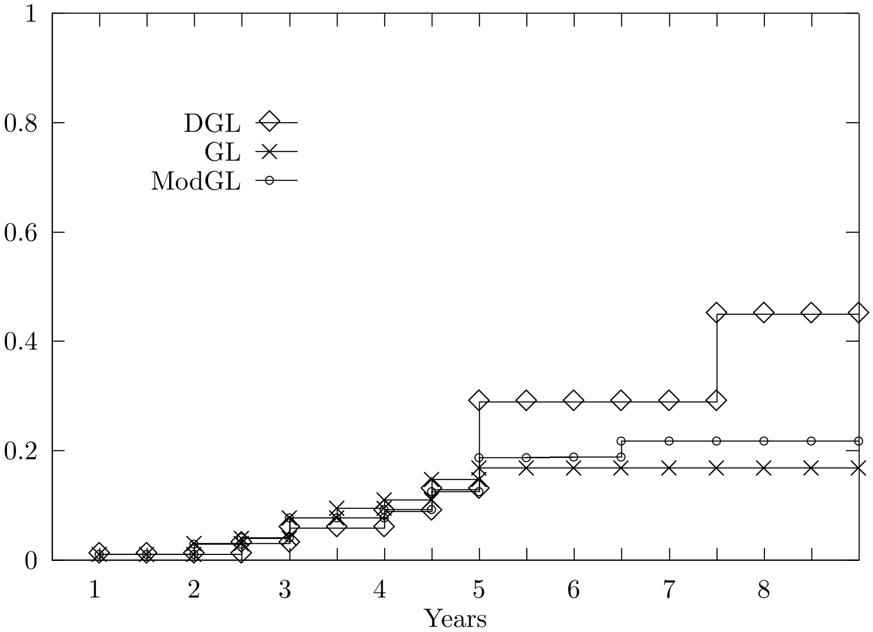


Figure 2.5: Estimated cumulative distribution function of latency time between HIV seroconversion and onset of symptoms for lightly-treated group.

times for the two groups. The estimators are very similar for the first 5 years and differ thereafter. We find here again differences between the two treatment groups. The heavily-treated group seems to have shorter latency times than the other group of patients. However, the interpretation of these results must be done carefully because of the small number of patients who developed AIDS. The data, as reported in De Gruttola and Lagakos [16], and the numerical results obtained with ModGL estimator are presented at the end of this section.

Data

Observations for 262 hemophilia patients by amount of blood factor received. Numbers in parentheses denote multiplicities. Censored times of disease denoted by +.

X_L	X_R	V	X_L	X_R	V	X_L	X_R	V
				Heavily treated				
15	∞	(2)	16	∞	(3)	17	∞	(3)
10	11	21+	1	16	21+	12	13	21+
13	15	21+	14	16	21+	12	14	21+
14	15	21+	13	16	21+	14	15	21+
13	15	21+	9	12	21+	14	15	21+
1	11	21+	12	14	21+	11	12	21+
15	16	21+	15	16	21+	1	13	21+
10	11	21+	5	7	21+	5	7	21+
15	15	21+	14	15	21+	12	13	21+
12	13	21+	1	14	21+	14	15	21+
10	11	21+	10	11	21+	8	10	21+
15	16	21+	9	10	21+	10	12	21+
1	14	21+	1	15	21+	1	13	21+
14	15	21+	3	15	21+	12	13	21+
14	15	21+	9	10	21+	14	15	21+
15	16	21+	1	15	21+	1	14	21+
11	13	21+	10	11	20+	1	7	21+
9	12	21+	1	11	21+	12	13	21+
13	14	21+	10	15	21+	13	15	21+
1	12	21+	7	10	21+	1	15	21+
9	12	21+	7	15	21+	14	16	21+
11	13	21+	11	13	21+	11	13	21+
1	6	21+	8	15	21+	10	11	21+
12	13	21+	7	9	21+	12	13	16
9	13	18	13	14	18	9	12	18
3	14	17	10	11	15	14	15	16
7	9	21	12	13	20	13	14	16
1	7	13	3	7	17	10	11	16
13	15	18	10	12	19	5	7	12
9	11	18	1	10	11	9	13	15
5	8	13	10	11	16	13	15	18
1	7	16	10	12	16	10	12	17
8	10	15	9	12	21	10	12	17
10	14	16						

X_L	X_R	V	X_L	X_R	V	X_L	X_R	V
Lightly treated								
1	∞		15	∞	(19)	16	∞	(31)
17	∞	(10)	18	∞				
10	15	21+	12	14	21+	1	15	21+
1	15	21+	1	15	21+	10	12	21+
1	16	21+	15	16	21+	3	10	21+
8	15	21+	8	13	21+	1	12	21+
13	14	21+	5	11	21+	14	16	21+
1	11	21+	9	14	21+	8	16	21+
11	12	21+	1	17	21+	1	18	21+
1	15	21+	11	16	21+	8	12	21+
9	13	21+	1	15	21+	13	14	21+
9	14	21+	1	5	21+	1	16	21+
12	15	21+	9	12	21+	13	15	21+
4	11	21+	1	16	21+	1	15	21+
14	15	21+	1	12	21+	14	15	21+
1	14	21+	6	13	21+	13	14	21+
15	16	21+	7	12	21+	12	14	21+
12	14	21+	1	13	21+	12	13	21+
13	15	21+	15	16	21+	1	15	21+
13	15	21+	8	16	21+	10	12	21+
14	15	21+	11	15	21+	13	15	21+
3	16	21+	6	8	21+	15	16	21+
11	14	21+	13	14	21+	12	14	21+
7	10	21+	1	12	21+	1	15	21+
12	13	21+	1	15	21+	10	16	21+
11	14	21+	1	14	21+	12	13	21+
9	14	21+	12	14	21+	11	12	20+
1	11	21+	1	16	21+	12	13	21+
14	15	21+	1	15	21+	15	16	21+
11	12	13	13	13	21	13	14	20
10	12	20	6	12	16	1	12	15
1	3	8	11	14	21	1	5	6
10	11	20	7	13	17	12	13	17
6	13	21	11	14	16			

Results for the heavily-treated group

Estimated survival function of time of HIV seroconversion of heavily-treated group using ModGL estimator and 95% confidence intervals

k	p_k	q_{k+1}	$1-\hat{W}(x)$ (s.d.)	Lower	Upper
1	0	5.00	1	–	–
2	6.00	7.00	0.9591 (0.0035)	0.952	0.965
3	7.00	8.00	0.8624 (0.0037)	0.855	0.869
4	8.00	9.00	0.8624 (0.0037)	0.855	0.869
5	9.00	10.00	0.8623 (0.0037)	0.855	0.869
6	10.00	11.00	0.6052 (0.0085)	0.588	0.621
7	11.00	12.00	0.5497 (0.0070)	0.535	0.563
8	12.00	13.00	0.4424 (0.0073)	0.428	0.456
9	13.00	14.00	0.2930 (0.0058)	0.281	0.304
10	14.00	15.00	0.2595 (0.0066)	0.246	0.272
11	15.00	16.00	0.0618 (0.0024)	0.057	0.066
12	16.00	17.00	0.0615 (0.0024)	0.056	0.066

Estimated survival function of latency time between HIV seroconversion and onset of AIDS for heavily-treated group and 95% confidence intervals

k	p'_k	q'_{k+1}	$1-\hat{F}(t)$ (s.d.)	Lower	Upper
1	0	1.00	1	–	–
2	1.00	2.00	0.9858 (0.0013)	0.983	0.988
3	2.00	3.00	0.9858 (0.0013)	0.983	0.988
4	3.00	4.00	0.9553 (0.0024)	0.950	0.960
5	4.00	5.00	0.9553 (0.0024)	0.950	0.960
6	5.00	6.00	0.8637 (0.0042)	0.855	0.871
7	6.00	7.00	0.8090 (0.0046)	0.800	0.818
8	7.00	8.00	0.7605 (0.0061)	0.748	0.772
9	8.00	9.00	0.7277 (0.0050)	0.718	0.737
10	9.00	10.00	0.7115 (0.0064)	0.699	0.724
11	10.00	11.00	0.7097 (0.0366)	0.638	0.781
12	11.00	12.00	0.6546 (0.0221)	0.611	0.698
13	12.00	13.00	0.6017 (0.0531)	0.497	0.706
14	13.00	14.00	0.5995 (0.0531)	0.495	0.703
15	14.00	15.50	0.5099 (0.0491)	0.414	0.606

Results for the lightly-treated group

Estimated survival function of time of HIV seroconversion of lightly-treated group using ModGL estimator and 95% confidence intervals

k	p_k	q_{k+1}	$1-\hat{W}(x)$ (s.d.)	Lower	Upper
1	0	3.00	1	–	–
2	3.00	5.00	0.9666 (0.0015)	0.963	0.969
3	5.00	8.00	0.9666 (0.0015)	0.963	0.969
4	8.00	10.00	0.9389 (0.0025)	0.934	0.944
5	10.00	11.00	0.9031 (0.0040)	0.895	0.911
6	11.00	12.00	0.8246 (0.0045)	0.816	0.833
7	12.00	13.00	0.6700 (0.0044)	0.661	0.678
8	13.00	14.00	0.5307 (0.0049)	0.521	0.540
9	14.00	15.00	0.4862 (0.0045)	0.477	0.495
10	15.00	16.00	0.4135 (0.0043)	0.405	0.422
11	16.00	17.00	0.2688 (0.0051)	0.259	0.279
12	17.00	18.00	0.2625 (0.0197)	0.224	0.301

Estimated survival function of latency time between HIV seroconversion and onset of AIDS for lightly-treated group and 95% confidence intervals

k	p'_k	q'_{k+1}	$1-\hat{F}(t)$ (s.d.)	Lower	Upper
1	0	1.00	1	–	–
2	1.00	2.00	0.9895 (0.0008)	0.988	0.991
3	2.00	3.00	0.9895 (0.0008)	0.988	0.991
4	3.00	4.00	0.9704 (0.0017)	0.967	0.974
5	4.00	5.00	0.9592 (0.0043)	0.950	0.967
6	5.00	6.00	0.9221 (0.0053)	0.912	0.932
7	6.00	7.00	0.9221 (0.0053)	0.912	0.932
8	7.00	8.00	0.9072 (0.0057)	0.896	0.918
9	8.00	9.00	0.8747 (0.0057)	0.863	0.886
10	9.00	10.00	0.8122 (0.1474)	0.523	1
11	10.00	11.00	0.8115 (0.1841)	0.450	1
12	11.00	12.00	0.8115 (0.1841)	0.450	1
13	12.00	13.00	0.7821 (0.1666)	0.455	1
14	13.00	18.00	0.7820 (0.0493)	0.685	0.878

2.7 ModGL Algorithm

The purpose of **ModGL** algorithm is to compute the proposed Modified Gmez and Lagakos's estimators for doubly-censored data. That is, the algorithm provides both, a nonparametric estimator of the survival function for the interval-censored origin time and an estimator of the survival function for the doubly-censored latency time.

The algorithm implements the two step procedure given in section 2.5. In the first step the estimator of the survival function of the origin time is obtained by solving iteratively the self-consistent equations (2.10). Using the estimators obtained in step 1, the second step computes the estimator of the survival function of the latency time by solving iteratively the self-consistent equations (2.18).

ALGORITHM

FIRST STEP:

A. Construct the set of identifiable intervals $[q_1, p_1], \ldots, [q_m, p_m]$, defined in section 2.5.1, using the subroutine INTERVQP.

B. Choose starting values for $\mathbf{s}$: $\mathbf{s}^{(0)} = (s_1^{(0)}, \ldots, s_m^{(0)})$.

C. Obtain improved estimates for $\mathbf{s}^{(1)}$ from the self-consistent equations:

$$ns_j^{(1)} = \sum_{i=1}^{n} \frac{\alpha_j^i s_j^{(0)}}{\sum_{l=1}^{m} \alpha_l^i s_l^{(0)}} \quad \text{for} \quad j = 1, \ldots, m.$$

D. Stop if the required accuracy has been achieved.
Otherwise, return to step C with $\mathbf{s}^{(1)}$ replacing $\mathbf{s}^{(0)}$.

SECOND STEP:

A. Construct the set of identifiable intervals $[q_1', p_1'], \ldots, [q_r', p_r']$, defined in section 2.5.2, using the subroutine INTERVQP.

B. Choose starting values for $\mathbf{f}$: $\mathbf{f}^{(0)} = (f_1^{(0)}, \ldots, f_r^{(0)})$.

C. Obtain improved estimates for $\mathbf{f}^{(1)}$ from the self-consistent equations:

$$(n-n_0)f_k^{(1)} = \sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{m}\alpha_{jk}^{i}\hat{s}_j f_k^{(0)}}{\sum_{l=1}^{r}\sum_{j=1}^{m}\alpha_{jl}^{i}\hat{s}_j f_l^{(0)}}\right]^{d^i} \quad \text{for} \quad k=1,\ldots,r.$$

D. Stop if the required accuracy has been achieved.
Otherwise, return to step C with $\mathbf{f}^{(1)}$ replacing $\mathbf{f}^{(0)}$.

Subroutine INTERVQP:

The purpose of this subroutine is to construct the set $C = \cup_{j=1}^{m}[q_j, p_j]$ of identifiable intervals. It takes as input the maximum number of intervals in C, and the observed vectors X_L and X_R. The output are the vectors q_j and p_j that define the identifiable intervals $[q_j, p_j],\ j=1,\ldots,m$.

ALGORITHM

1. Construction of the partition of $\mathrm{I\!R}^+$, $0 \le a_1 < \cdots < a_{k+1} \le +\infty$, generated by the observed data $(X_L^i, X_R^i),\ i=1,\ldots,n$.

2. Each $a_l,\ l=1\ldots,k$ is classified into one of the following three categories by an indicator, **ind**:

   ind$(a_l) = 1$ if a_l is a right end-point of a censoring interval and, at the same time, it is the left end-point of another censoring interval.

   ind$(a_l) = 2$ if a_l is a left end-point of a censoring interval.

   ind$(a_l) = 3$ if a_l is a right end-point of a censoring interval.

   That is,
   if $a_l \in \{X_L^i, i=1\ldots,n\} \cap \{X_R^i, i=1\ldots,n\}$ then **ind**$(a_l)=1$
   if $a_l \in \{X_L^i, i=1\ldots,n\} \cap \{X_R^i, i=1\ldots,n\}^{\mathrm{c}}$ then **ind**$(a_l)=2$
   if $a_l \in \{X_L^i, i=1\ldots,n\}^{\mathrm{c}} \cap \{X_R^i, i=1\ldots,n\}$ then **ind**$(a_l)=3$

3. Then, the intervals $[q_j, p_j]$ are constructed as follows:
   For $l=1,\cdots,k$
   if **ind**$(a_l)=1$ then $q_j = p_j = a_l$
   if **ind**$(a_l)=2$ and **ind**$(a_{l+1})=3$ then $q_j = a_l$ and $p_j = a_{l+1}$

Auxiliary Subroutines

We have used the subroutines **crear_vector**, **crear_intvector** and **crear_matriu** to allocate memory to a vector, to a vector with integer components and to a matrix, respectively.

Subroutines **lliberar_vector**, **lliberar_intvector** and **lliberar_matriu** leave up the memory allocated to a vector, to a vector with integer components and to a matrix, respectively.

Chapter 3

Comparison of Nonparametric Methodologies for Double Censoring. A simulation Study

3.1 Introduction

The two-step estimation procedure of Gmez and Lagakos (GL) [36] was proposed as an alternative to those problems observed with the method derived by DeGruttola and Lagakos (DGL) [16]. To improve the behaviour of GL method and to avoid some problems of unstability we have proposed in section 2.5 an extension of it that allows continuous time distributions, the ModGL method. From a theoretical point of view the two-step estimation procedures, both GL and ModGL algorithms, lead to estimators not as efficient as those obtained from DGL because while DGL maximizes the joint likelihood, GL and ModGL maximize the marginal and the conditional likelihoods separately. Our goal now is to show that despite of these considerations, for small and moderate sample sizes, both approaches, DGL and ModGL, behave similarly , while ModGL is computationally more efficient and therefore should be preferred. A simulation study is carried out to compare ModGL estimator to DGL estimator for the latency distribution when data are doubly–censored. First results of this simulation study are given in [6].

3.2 Design and Implementation of the Simulation Study

Simulations are performed for a large variety of scenarios. Several options are considered for the sample size, for the distribution of X and T and for the percentage of right censoring.

Random variables. Two parametric models are assumed for the origin time X, namely, a uniform distribution in the interval $[1, 16]$ and a Weibull distribution with scale and shape parameters equal to 10 and to 3, respectively. The latency time T is assumed to follow a Weibull distribution with scale parameter equal to 10 and shape parameter b equal to 0.5, 1 and 4, corresponding to a decreasing, constant and increasing hazard function, respectively.

Generation of observations. After X^i and T^i have been randomly generated from one of the models considered above, the observable data $(X_L^i, X_R^i, d^i, V^i, c^i),\ \ i = 1, \ldots, n$ are constructed in two independent steps for samples of size equal to 30, 50 and 100.

In the first step, the random intervals $[X_L^i, X_R^i]$ are constructed containing X^i via a mechanism that mimics those longitudinal studies where there is periodical follow–up. The intervals arise from regularly scheduled visits but patients might miss some of the appointments. In particular, we consider a situation in which 30% of the patients attend all the visits, having each of them a censoring interval of length 1 unit (1 month, 6 months, 1 year, ...), 30% of the patients miss 1 visit in the interval of interest and have therefore a censoring interval of length 2 units, 20% miss 2 visits and have censoring intervals of length 3 units, 10% of the patients miss 3 visits and their intervals are of length 4 units and the remainder 10% have an interval of length 5 units.

In the second step (V^i, c^i) are generated in the following way: First, we compute the final time Z^i as the sum of the origin time X^i and the latency time T^i: $Z^i = X^i + T^i$. Then, we construct the observed values V^i as the minimum between Z^i and a constant C and define the censoring indicator $c^i = \mathbf{1}\{Z^i \leq C\}$. For every run, the constant C is computed so that the given percentage of censoring, p, is achieved. This percentage p is taken to be equal to $10\%, 30\%$, 50% and 70%.

Design of the experiment. 500 replications of the process are carried out for each possible scenario. Based on the observable data the maximum likelihood estimators $\hat{W}$ and $\hat{F}$ are computed following DGL and ModGL procedures. Convergence of the algorithms is declared for a tolerance equal to 0.0001.

Evaluation. The overall performance of the estimator $\hat{F}(t)$ is studied by means of two statistics:

$$D1 = \left(\int_0^{t^*} (\hat{F}(t) - F(t))^2 \; dt \right)^{(1/2)}, \qquad D2 = \sup_{0 \leq t \leq t^*} |\hat{F}(t) - F(t)|$$

that measure the distance between the estimator $\hat{F}(t)$ and the theoretical distribution $F(t)$ for values of t between 0 and the maximum admissible time t^*. The mean and the standard deviation of D_1 and D_2 are computed for each run. As a measure of the local performance of the estimator $\hat{F}$, we compute its deciles for each run. Then, the bias and mean squared error (MSE) for each method are compared.

3.3 Results and discussion

The comparison of both methods has not been done for a percentage of censoring equal to 70% because DGL algorithm has given many problems of convergence at this level of censoring.

We first consider the results concerning the global performance of both estimators, that is, comparisons of the L_2 distance and the supremum norm distance for the two estimators. These results are reported in tables 3.1 to 3.4. In each table we provide the mean and the standard error (in parentheses) of the corresponding measures based on 500 replications. Those situations where ModGL method has a larger mean distance have been printed in boldface. Tables 3.1 and 3.2 corresponds to a uniform origin time and tables 3.3 and 3.4 to a Weibull origin time.

By examining these tables we observe that:

- When the origin time is uniform, if the latency time has increasing hazard ($b = 4$) the mean L_2 distance of ModGL estimator is in every case slightly smaller than the distance of DGL estimator. In this case, the advantage of ModGL method is clear, especially for small sample sizes ($n = 30, n = 50$). If a constant or decreasing hazard are considered the advantage of ModGL is still observed. Only in one situation ModGL behaves worse that DGL with respect to the mean L_2 distance (table 3.1). Similar results are obtained when the supremum norm distance is considered, while in this case there are 5 instances where DGL performs slightly better that ModGL (table 3.2).

- An analogous behaviour is found when the origin time is Weibull. We don't have reasons to prefer DGL algorithm when the hazard is increasing. Although in 12 out of 18 possibilities DGL has smaller mean distance when a constant or decreasing hazard are considered, the difference between them remains not significative.

The local behaviour of the estimators is shown in figures 3.1 to 3.6 where the ratio between the mean squared errors of the deciles of DGL and the deciles of ModGL have been graphically displayed. If this ratio is greater than 1, ModGL estimator should be preferred to DGL estimator. The figures have been restricted to a sample size equal 30. Sample sizes equal 50 and 100 yield analogous results. As the illustrations show, ModGL method performs better for every decile if the hazard is increasing, achieving ratios near 2 (Figures 3.1 and 3.4). For an exponential latency time the ratio is approximately equal to 1.2 and quite homogeneous over the entire distribution and for the three levels of censoring (Figures 3.2 and 3.5). Finally, when $b = 0.5$ there is no evidence of the superiority of one method (Figures 3.3 and 3.6). From a bias point of view both methods behave similarly. Both estimators slightly underestimate the deciles of the Weibull distribution when a non–decreasing hazard is considered and they slightly overestimate them when the hazard is decreasing.

We have observed also with this simulation study that ModGL algorithm is computationally more efficient. In particular, ModGL algorithm is three times faster than DGL and it converges even for a 70% of right censoring.

We conclude that both estimators perform very similarly with small and moderate sample sizes and therefore the two-step algorithm of ModGL can be considered a good computational alternative to the DGL method.

Table 3.1

Mean L2 distance between the estimator and the theoretical distribution when it is assumed a Uniform origin time and a Weibull latency time with shape parameter *b*

DGL: DeGruttola & Lagakos estimator *ModGL: Modified Gomez & Lagakos estimator*

			n=30	
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.3653(0.11)	0.4426(0.14)	0.6010(0.22)
ModGL	10	0.3157(0.11)	0.4195(0.14)	0.5695(0.24)
DGL	30	0.4212(0.13)	0.4855(0.16)	0.5618(0.18)
ModGL	30	0.3450(0.12)	0.4399(0.14)	0.5117(0.17)
DGL	50	0.4772(0.15)	0.5536(0.16)	0.5286(0.17)
ModGL	50	0.3759(0.13)	0.5048(0.12)	**0.5438**(0.12)

			n=50	
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.3191(0.09)	0.3552(0.11)	0.4956(0.15)
ModGL	10	0.2836(0.09)	0.3324(0.11)	0.4523(0.16)
DGL	30	0.3758(0.10)	0.4160(0.13)	0.4874(0.15)
ModGL	30	0.3034(0.10)	0.3745(0.11)	0.4409(0.13)
DGL	50	0.4270(0.12)	0.4963(0.14)	0.4924(0.14)
ModGL	50	0.3300(0.11)	0.4360(0.11)	0.4849(0.09)

			n=100	
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2707(0.07)	0.2654(0.08)	0.4034(0.11)
ModGL	10	0.2500(0.07)	0.2474(0.08)	0.3506(0.12)
DGL	30	0.3252(0.08)	0.3408(0.10)	0.4029(0.11)
ModGL	30	0.2603(0.08)	0.3014(0.08)	0.3621(0.07)
DGL	50	0.3694(0.09)	0.4429(0.13)	0.4581(0.14)
ModGL	50	0.2793(0.08)	0.3634(0.08)	0.4309(0.06)

Table 3.2

Mean suprem norm between the estimator and the theoretical distribution when it is assumed a Uniform origin time and a Weibull latency time with shape parameter b

DGL: DeGruttola & Lagakos estimator *ModGL: Modified Gomez & Lagakos estimator*

		n=30		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2776(0.07)	0.2029(0.05)	0.2549(0.02)
ModGL	10	0.2310(0.07)	0.1851(0.05)	0.2104(0.05)
DGL	30	0.3175(0.09)	0.2304(0.06)	0.2666(0.03)
ModGL	30	0.2547(0.08)	0.2198(0.05)	**0.2690**(0.05)
DGL	50	0.3553(0.10)	0.2879(0.07)	0.2941(0.04)
ModGL	50	0.2732(0.08)	0.3070(0.06)	**0.3638**(0.04)

		n=50		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2483(0.06)	0.1675(0.04)	0.2482(0.00)
ModGL	10	0.2132(0.06)	0.1516(0.04)	0.1912(0.04)
DGL	30	0.2847(0.07)	0.1952(0.05)	0.2565(0.02)
ModGL	30	0.2284(0.06)	0.1890(0.04)	0.2498(0.04)
DGL	50	0.3186(0.08)	0.2555(0.06)	0.2818(0.03)
ModGL	50	0.2460(0.07)	**0.2709**(0.04)	0.3416(0.04)

		n=100		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2111(0.04)	0.1320(0.03)	0.2466(0.00)
ModGL	10	0.1915(0.04)	0.1209(0.03)	0.1731(0.03)
DGL	30	0.2494(0.05)	0.1594(0.04)	0.2486(0.00)
ModGL	30	0.1988(0.05)	0.1579(0.03)	0.2342(0.03)
DGL	50	0.2801(0.06)	0.2300(0.05)	0.2753(0.03)
ModGL	50	0.2129(0.06)	**0.2428**(0.03)	**0.3233**(0.02)

Table 3.3

Mean L2 distance between the estimator and the theoretical distribution when it is assumed a Weibull origin time and a Weibull latency time with shape parameter b.

DGL: DeGruttola & Lagakos estimator ModGL: Modified Gomez & Lagakos estimator

		n=30		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.3473(0.10)	0.4336(0.14)	0.5767(0.23)
ModGL	10	0.3102(0.10)	0.4139(0.14)	0.5667(0.24)
DGL	30	0.3679(0.11)	0.4443(0.15)	0.4848(0.18)
ModGL	30	0.3353(0.11)	0.4442(0.13)	**0.5075**(0.15)
DGL	50	0.4061(0.13)	0.4752(0.17)	0.4595(0.19)
ModGL	50	0.3696(0.13)	**0.5187**(0.12)	**0.5532**(0.11)

		n=50		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.3028(0.08)	0.3454(0.11)	0.4529(0.16)
ModGL	10	0.2779(0.09)	0.3309(0.11)	0.4463(0.16)
DGL	30	0.3200(0.09)	0.3768(0.12)	0.4076(0.16)
ModGL	30	0.2960(0.09)	**0.3800**(0.10)	**0.4468**(0.12)
DGL	50	0.3586(0.11)	0.4083(0.15)	0.3892(0.16)
ModGL	50	0.3301(0.11)	**0.4596**(0.10)	**0.5014**(0.08)

		n=100		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2551(0.07)	0.2573(0.08)	0.3469(0.12)
ModGL	10	0.2454(0.07)	0.2495(0.08)	**0.3475**(0.12)
DGL	30	0.2643(0.07)	0.2995(0.10)	0.3122(0.12)
ModGL	30	0.2519(0.07)	**0.3156**(0.08)	**0.3729**(0.07)
DGL	50	0.2855(0.08)	0.3249(0.13)	0.3093(0.14)
ModGL	50	0.2707(0.08)	**0.3904**(0.08)	**0.4531**(0.06)

Table 3.4

Mean suprem norm between the estimator and the theoretical distribution when it is assumed a Weibull origin time and a Weibull latency time with shape parameter b.

DGL: DeGruttola & Lagakos estimator *ModGL: Modified Gomez & Lagakos estimator*

		n=30		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2634(0.07)	0.2010(0.05)	0.2501(0.07)
ModGL	10	0.2282(0.06)	0.1849(0.05)	0.2166(0.05)
DGL	30	0.2782(0.07)	0.2229(0.06)	0.2685(0.07)
ModGL	30	0.2483(0.07)	**0.2364**(0.06)	**0.2794**(0.05)
DGL	50	0.3075(0.09)	0.2728(0.08)	0.3089(0.08)
ModGL	50	0.2751(0.09)	**0.3436**(0.06)	**0.3843**(0.04)

		n=50		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2339(0.06)	0.1648(0.04)	0.2206(0.06)
ModGL	10	0.2091(0.05)	0.1522(0.04)	0.1986(0.04)
DGL	30	0.2486(0.06)	0.1912(0.04)	0.2432(0.06)
ModGL	30	0.2228(0.06)	**0.2054**(0.04)	**0.2643**(0.05)
DGL	50	0.2751(0.08)	0.2343(0.07)	0.2777(0.07)
ModGL	50	0.2488(0.07)	**0.3084**(0.05)	**0.3663**(0.04)

		n=100		
Method	% cens	b=4	b=1	b=0.5
DGL	10	0.2000(0.04)	0.1285(0.03)	0.1960(0.05)
ModGL	10	0.1889(0.04)	0.1206(0.03)	0.1806(0.04)
DGL	30	0.2071(0.05)	0.1538(0.04)	0.2171(0.05)
ModGL	30	0.1935(0.05)	**0.1784**(0.03)	**0.2488**(0.03)
DGL	50	0.2225(0.06)	0.1912(0.06)	0.2473(0.06)
ModGL	50	0.2077(0.05)	**0.2778**(0.03)	**0.3479**(0.03)

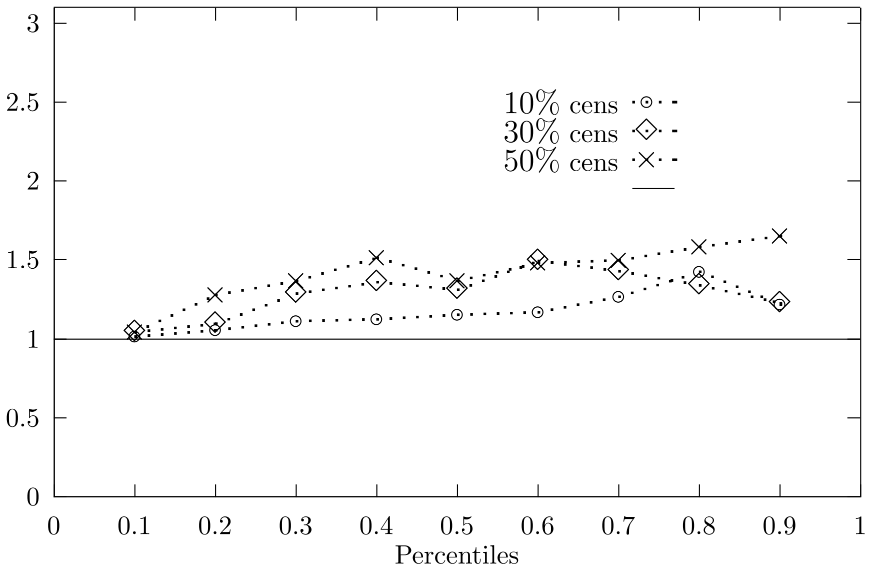


Figure 3.1: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Uniform and Latency time Weibull with $b = 4$)

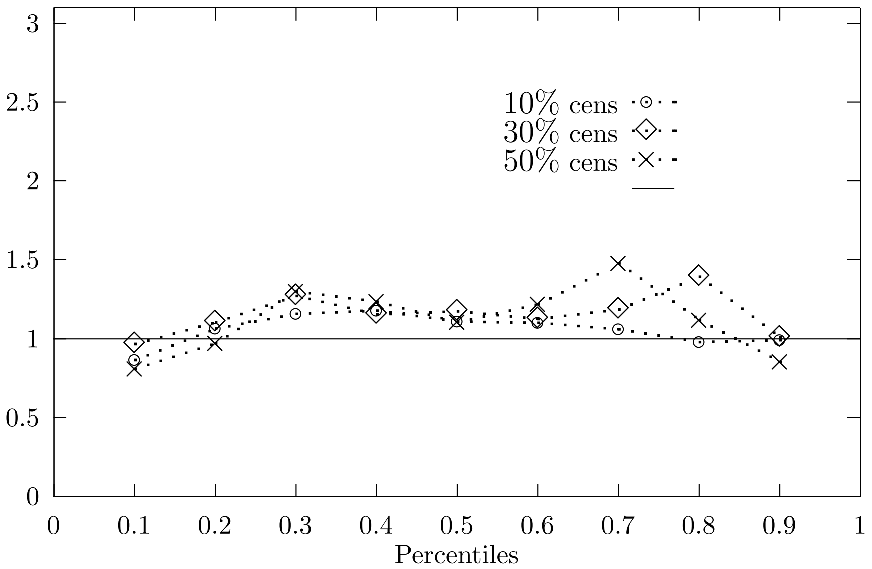


Figure 3.2: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Uniform and Latency time Weibull with $b = 1$)

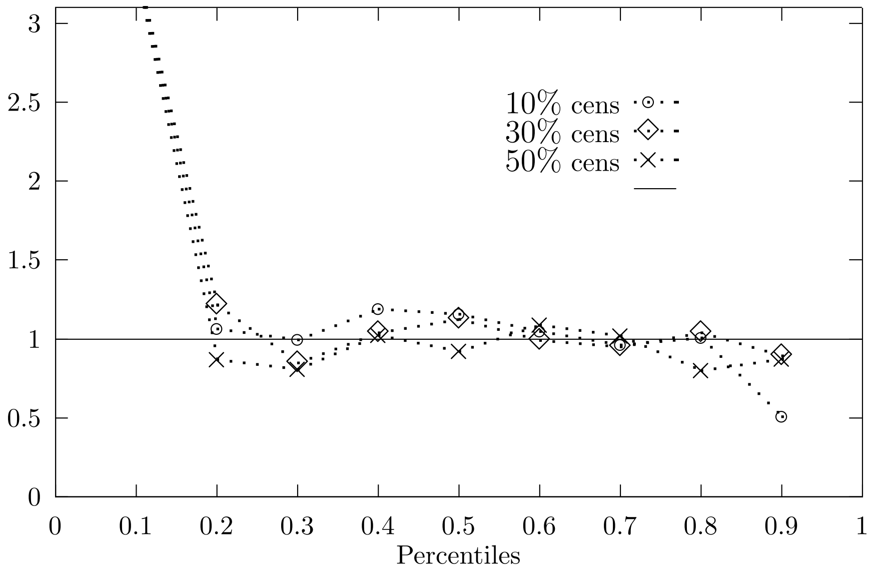


Figure 3.3: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Uniform and Latency time Weibull with $b = 0.5$)

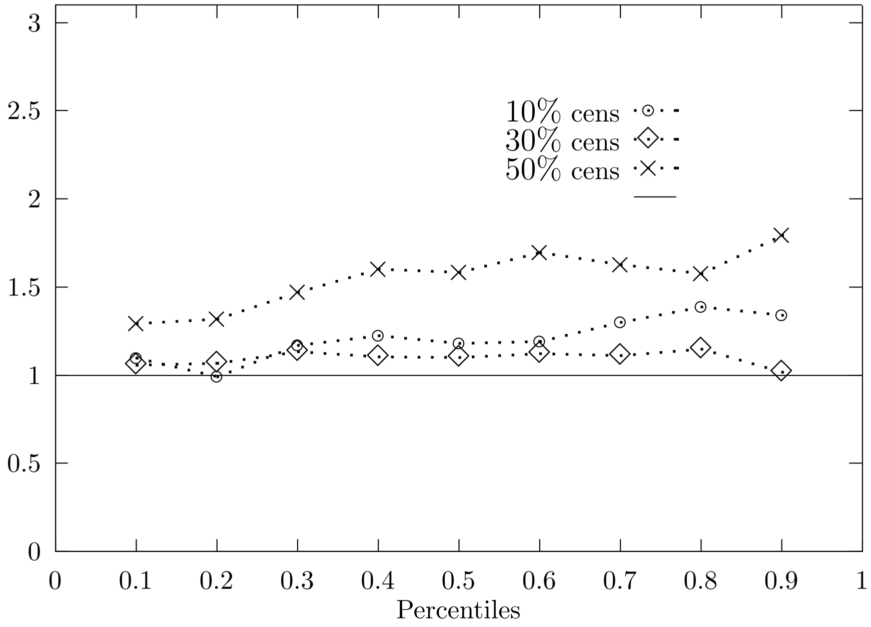


Figure 3.4: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Weibull(10,3) and Latency time Weibull with $b = 4$)

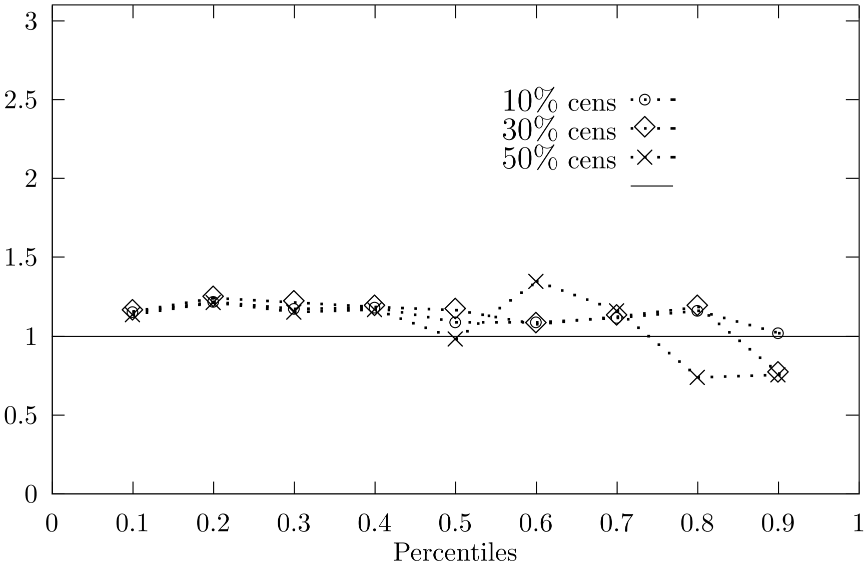


Figure 3.5: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Weibull(10,3) and Latency time Weibull with $b = 1$)

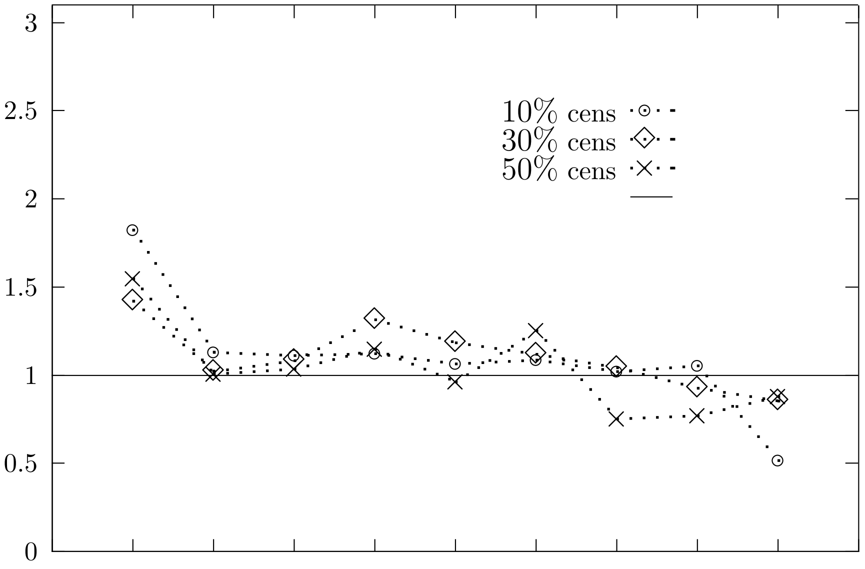


Figure 3.6: *Relative MSE of percentiles of two estimators of the latency distribution (DGL/ModGL)* (Origin time Weibull(10,3) and Latency time Weibull with $b = 0.5$)

Part II

Chapter 4

Nonparametric Bayesian Estimation of a Survival Function

4.1 Introduction

We consider the problem of estimating an unknown survival function based on censored data. When there are no reasons that indicate any specific parametric model to hold, the nonparametric approach may be appropriate. In the presence of right censoring, the most widely used nonparametric estimator of a survival curve is the Kaplan and Meier estimator [40]. Kaplan and Meier consider several nonparametric estimators of the survival function and show that one of them, the product limit estimator, is in fact a maximum likelihood estimator. There are other nonparametric methodologies for more complex censoring schemes, such as, Turnbull's algorithm [57] when there is interval-censored data or DGL [16] and GL [36] in the presence of double censoring.

Sometimes, however, there are clear reasons that indicates that a certain parametric family is adequate. In those cases, a nonparametric approach may represent a loss of efficiency versus a parametric estimator. The problem with the parametric methodology is that the models only hold approximately and with complex censoring schemes the parametric assumptions are difficult to assess. Then, with the parametric methodology one runs the risk of obtaining an inconsistent estimator if the model is not correctly specified.

In most situations it is difficult to decide in favour of one of those opposed points of view, parametric or nonparametric methods. We consider as an alternative the nonparametric Bayesian approach that allows the incorporation of prior information about the modelization of the problem but, at the same time, reduces the unfortunate consequences of an incorrect parametric assumption. Ferguson's paper [19] pioneers the approach followed in this chapter when the sample data is complete. He introduces a class of prior distributions, the so-called Dirichlet processes, that are an important tool for the treatment of nonparametric statistical problems from a Bayesian point of view. Using this class of prior distributions Ferguson finds the Bayesian estimator for the survival function under the squared error loss. This estimator is a mixture of the prior guess and of the empirical survival function.

Susarla and Van Ryzin [51] use the class of Dirichlet processes to obtain a nonparametric Bayesian estimator of the survival function when data are right-censored. The resulting estimator includes, as a limiting case, the Kaplan-Meier estimator and it is shown to be preferable in many situations, specially with heavy censoring. Ferguson and Phadia [20] extend the results of Susarla and Van Ryzin to a more general class of prior distributions, namely, the processes that are neutral to the right. This class of processes includes among others the Dirichlet process and the gamma process.

The incorporation of covariates in the estimation of the survival function is approached in the papers of Kalbfleisch [39] and Burridge [5]. They propose a Bayesian method to analyze the Cox proportional hazards model [13] by treating the cumulative hazard function as a gamma process.

In this chapter we describe the works of Ferguson [19] and Susarla and Van Ryzin [51] [52] for complete and right-censored data, respectively, and prove some asymptotical results of the nonparametric Bayes estimator when the data are completely observed.

4.2 The Dirichlet Distribution

The Dirichlet is a well known distribution because it is the conjugate prior for the parameters of the multinomial distribution.

Definition 4.2.1 *The random vector $(X_1, \cdots, X_k)$ is said to have a* **Dirichlet distribution** *with parameters $(\alpha_1, \cdots, \alpha_k)$, and is denoted by $\mathcal{D}(\alpha_1, \cdots, \alpha_k)$, with α_j positive numbers for all j, if the joint distribution of the first $(k-1)$ components has density*

$$f(x_1, \cdots, x_{k-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \left(\prod_{j=1}^{k-1} x_j^{\alpha_j - 1} \right) \left(1 - \prod_{j=1}^{k-1} x_j \right)^{\alpha_k - 1}$$

over the $(k-1)$-dimensional simplex S defined by

$$S = \{(x_1, \cdots, x_{k-1}) : x_j \geq 0, \; \sum_{j=1}^{k-1} x_j \leq 1\}$$

and $x_k = 1 - \sum_{j=1}^{k-1} x_j$.

Properties:

Assume that $(X_1, \cdots, X_k) \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$ and let $\alpha = \sum_{i=1}^{k} \alpha_i$.

1. When $k = 2$, the Dirichlet distribution reduces to the Beta distribution $\mathcal{B}e(\alpha_1, \alpha_2)$.

2. Let $Y_1, \cdots, Y_k$ be independent gamma random variables with parameters $\alpha_i > 0$ and $\beta = 1$, for $i = 1, \ldots, k$, respectively and define

$$X_i = Y_i / \sum_{j=1}^{k} Y_j \quad \text{for } i = 1, \cdots, k-1$$

$$\text{and} \quad X_k = 1 - \sum_{i=1}^{k-1} X_i,$$

then $(X_1, \cdots, X_k) \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$.
This important property provides an efficient method for sampling from the Dirichlet distribution.

3. If $k_1, \cdots, k_l$ are integers such that $0 < k_1 < k_2 < \cdots < k_l = k$, then

$$(\sum_{i=1}^{k_1} X_i, \sum_{i=k_1+1}^{k_2} X_i, \cdots, \sum_{i=k_{l-1}+1}^{k_l} X_i) \sim \mathcal{D}(\sum_{i=1}^{k_1} \alpha_i, \sum_{i=k_1+1}^{k_2} \alpha_i, \cdots, \sum_{i=k_{l-1}+1}^{k_l} \alpha_i).$$

This result follows form the additivity property of the gamma distribution.

4. The marginal distributions of each $(X_1, \ldots, X_k)$ are Beta, that is,
$X_j \sim \mathcal{B}e(\alpha_j, \alpha - \alpha_j)$ for every $j = 1, \ldots, k$.

4.3 Dirichlet Processes

Nonparametric models are characterized by the specification of a probability distribution on an infinite-dimensional space. The random probability measures in this infinite-dimensional space can be thought as a stochastic processes with index set A, a σ-field of subsets of the sample space X. In this context the distribution of a random probability measure P is determined by the specification of the joint distribution of $(P(A_1), \dots, P(A_k))$ for all k and for all partition $(A_1, \dots, A_k)$ of X, provided that some consistent properties are satisfied.

Definition 4.3.1 *Let X be the sample space and $\mathcal{A}$ a σ-field of subsets.*
A **random probability measure** *on the measurable space $(\Omega, \mathcal{A})$ is a stochastic process $\{P(A),\ A \in \mathcal{A}\}$ such that:*

1. *$P(A)$ is a random variable with values in $[0,1]$, $A \in \mathcal{A}$,*
2. *$P(\Omega)$ is degenerate at 1,*

   and

3. *P is finitely additive in distribution, i.e.,*
   if $(B'_1, \cdots, B'_j)$ and $(B_1, \cdots, B_k)$ are measurable partitions, and if $(B'_1, \cdots, B'_j)$ is a refinement of $(B_1, \cdots, B_k)$ with $B_1 = \cup_1^{r_1} B'_i, B_2 = \cup_{r_1+1}^{r_2} B'_i, \cdots, B_k = \cup_{r_{k-1}+1}^{j} B'_i$, then the distribution of

$$(\sum_1^{r_1} P(B'_i), \sum_{r_1+1}^{r_2} P(B'_i), \cdots, \sum_{r_{k-1}+1}^{j} P(B'_i))$$

   is identical to the distribution of $(P(B_1), \cdots, P(B_k))$:

The notion of a Dirichlet process is introduced by Ferguson [19] who constructs a random probability measure, P, by defining the joint distribution of the random variables $(P(A_1), \cdots, P(A_k))$ for every k and for any sequence of measurable sets $(A_1, \cdots, A_k)$.

Definition 4.3.2 *Let α be a finite non-null measure on $(\Omega, \mathcal{A})$. A stochastic process indexed by elements of $\mathcal{A}$, $\{P(A),\ A \in \mathcal{A}\}$ is said to be a* **Dirichlet process** *on $(\Omega, \mathcal{A})$ with parameter α, denoted by $\mathcal{D}(\alpha)$, if for every $k = 1, 2, \cdots$ and for any measurable partition $(A_1, \cdots, A_k)$ of X, the random vector $(P(A_1), \cdots, P(A_k))$ has a Dirichlet distribution with parameter $(\alpha(A_1), \cdots, \alpha(A_k))$.*

Ferguson proved that a Dirichlet process is a random probability measure, that is, P verifies the conditions in definition 4.3.1. The distribution of P is a probability on $([0,1]^{\mathcal{A}}, \sigma(\mathcal{B}^{\mathcal{A}})$, where $[0,1]^{\mathcal{A}}$ represents the space of all functions from $\mathcal{A}$ into $[0,1]$ and $\sigma(\mathcal{B}^{\mathcal{A}})$ represents the σ-field generated by the field of cylinder sets in $[0,1]^{\mathcal{A}}$. The main result in Ferguson [19], restricted to the measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathbb{R}$ denotes the real line and $\mathcal{B}$ the σ-field of Borel sets, is that, if the prior process is a Dirichlet process then the posterior process given a random sample is also a Dirichlet process, with an updated parameter measure.

Theorem 4.3.3 (Ferguson) *If P is a Dirichlet process on $(\Omega, \mathcal{A})$ with parameter α, and if $(X_1, \cdots, X_n)$ is a sample of size n from P, then the posterior distribution of P given $(X_1, \cdots, X_n)$ is also a Dirichlet process on $(\Omega, \mathcal{A})$ with parameter $\alpha + \sum_1^n \delta_{X_i}$, where δ_x denotes the measure giving mass one to the point x.*

4.4 Elements of decision theory

In decision theory a **game** (Ω, A, L) has the following elements: Ω is the set of the possible states of nature θ, A is the set of actions available to the statistician and L is a loss function which defines the loss $L(\theta, a) \in \mathbb{R}$ which a statistician suffers if he takes action a when the true state of nature is θ.

A **statistical decision problem** is a game (Ω, A, L) whose result x lies in a sample space $\mathcal{X}$ and is randomly distributed with a density $p(x|\theta)$ which depends on the state $\theta \in \Omega$. On the basis of the result x, the statistician chooses an action $d(x) \in A$, resulting in a random loss $L(\theta, d(x))$. The **risk function** is the expectation of the loss over all possible outcomes of the experiment:

$$R(\theta, d) = E\{L(\theta, d(x))\} = \int L(\theta, d(x)) p(x|\theta)\ dx.$$

A **decision rule** is any function d for which $R(\theta, d)$ exists and is finite for all $\theta \in \Omega$. If there are prior beliefs about θ which can be expressed in terms of a prior density $p(\theta)$, the **Bayes risk** of the decision rule d is defined as the expectation of $R(\theta, d)$ over all possible values of θ, that is,

$$r(d) = E\{R(\theta, d)\} = \int R(\theta, d) p(\theta)\ d\theta.$$

A **Bayes decision rule** d is defined as the decision rule which minimizes the Bayes risk $r(d)$. It is easy to prove that this happens when the decision rule d is chosen so that the posterior expected loss of $d(x)$, $E\{L(\theta, d(x))|x\}$ is minimum for all x:

$$\begin{aligned} r(d) &= \int R(\theta, d)p(\theta)\ d\theta = \int\int L(\theta, d(x))p(x|\theta)p(\theta)\ dx\ d\theta = \\ &= \int\int L(\theta, d(x))p(x, \theta)\ dx\ d\theta = \int \left(\int L(\theta, d(x))p(\theta|x)\ d\theta\right)\ p(x)\ dx = \\ &= \int \left(E\{L(\theta, d(x))|x\}\right) p(x)\ dx \end{aligned}$$

then, $r(d)$ is minimized if $E\{L(\theta, d(x))|x\}$ is minimum for all x.

A **Bayes estimator** of the state θ is the Bayes decision rule $d(x)$ for a given result x.

4.5 Bayesian Inference From Complete Data

In this section we derive the Bayes estimator $\hat{S}(t)$ of the random survival function $S(t) = 1 - F(t) = P(t, +\infty)$ that minimizes the squared error loss:

$$L(\hat{S}, S) = \int_0^{+\infty} (\hat{S}(t) - S(t))^2\ dw(t),$$

where w is a nonnegative and nondecreasing function on $(0, +\infty)$

Proposition 4.5.1 *The Bayes estimator $\hat{S}$ of the survival function S under a prior process P is the posterior mean of S with respect to the posterior distribution of P given a sample $x = (x_1, \cdots, x_n)$; that is, $\hat{S}(t) = E\{S(t)|x\}$.*

Proof. We find the decision rule $\hat{S}$ that minimizes the posterior expected loss over all possible samples denoted by x:

$$E\left\{L(\hat{S}, S)|x\right\} = \int_0^{+\infty} E\left\{\left(\hat{S}(t) - S(t)\right)^2 |x\right\}\ dw(t)$$

where E denotes expectation with respect to the posterior distribution of P. This expression is minimized when the mean-squared error is minimum:

$$E\left\{\left(\hat{S}(t) - S(t)\right)^2 |x\right\} = E\left\{\left(\hat{S}(t) - E\{S(t)|x\} + E\{S(t)|x\} - S(t)\right)^2 |x\right\} =$$

$$= E\left\{\left(\hat{S}(t) - E\{S(t)|x\}\right)^2 |x\right\} + \tag{4.1}$$

$$+2E\left\{\left((\hat{S}(t) - E\{S(t)|x\})(E\{S(t)|x\} - S(t))\right) |x\right\} + \tag{4.2}$$

$$+E\left\{(E\{S(t)|x\} - S(t))^2 |x\right\}\ . \tag{4.3}$$

Expression (4.2) becomes equivalent to $2(\hat{S}(t) - E\left\{S(t)|x\right\})E\left\{(E\left\{S(t)|x\right\} - S(t))\,|x\right\}$ which is obviously equal to 0, and expression (4.3) is the posterior variance of $S(t)$, $Var\{S(t)|x\}$. Therefore, the minimum is achieved if expression (4.1) is zero, hence when $\hat{S}(t) = E\left\{S(t)|x\right\}$. □

Now we use the above result to get the Bayes estimator of S when the prior process P is Dirichlet.

Bayes estimator of the survival function $S(t)$ under a Dirichlet process prior

Let $X_1, \cdots, X_n$ be a random sample of size n from a Dirichlet process P of parameter measure α. Then, the posterior distribution of P given the observations is again a Dirichlet process, $\mathcal{D}(\alpha + \sum_1^n \delta_{X_i})$. Therefore, the posterior distribution of $S(t) = P(t, +\infty)$ is a beta distribution, $\mathcal{B}e((\alpha + \sum_1^n \delta_{X_i})(t, +\infty), (\alpha + \sum_1^n \delta_{X_i})(-\infty, t])$, for each t. Thus, from proposition 4.5.1, the **Bayes estimator** $\hat{S}_\alpha(t)$ of the survival function is the first moment of this beta distribution:

$$\hat{S}_\alpha(t) = \quad E\left\{S(t)|X_1, \cdots, X_n\right\} = \frac{\alpha(t, +\infty) + \sum_1^n \delta_{X_i}(t, +\infty)}{\alpha(\mathrm{I\!R}) + n} \tag{4.4}$$

Interpretation of the parameter measure of the Dirichlet process

To get an interesting interpretation of the parameter measure α we derive now the special case in which there is no data available. If P is a Dirichlet process of parameter measure α, the posterior distribution of P is also $\mathcal{D}(\alpha)$. The random variable $S(t) = P(t, +\infty)$ is therefore distributed as a beta distribution $\mathcal{B}e(\alpha(t, +\infty), \alpha(-\infty, t])$, for each t. Thus, from proposition 4.5.1, the **Bayes estimator** $\hat{S}_0(t)$ of the survival function in the *no-sample problem* is the first moment of this distribution. That is

$$\hat{S}_0(t) = E\left\{S(t)\right\} = \alpha(t, +\infty)/\alpha(\mathrm{I\!R}) \ .$$

In this case the Bayes estimator $\hat{S}_0(t)$ can be interpreted as the prior guess of the unknown survival function $S(t)$, since it has been obtained in the absence of any observation. Therefore, if we denote by $\beta = \alpha(\mathrm{I\!R})$ the measure of the real line, the parameter measure α can be expressed as $\alpha(t, +\infty) = \beta \cdot \hat{S}_0(t)$. That is, the measure α is the prior survival function $\hat{S}_0(t)$ weighted by a measure of faith in this prior guess, β.

With this notation, the Bayes estimator (4.4) can be expressed as a linear combination of the prior guess at S and of the empirical survival function, with respective weights

$\beta/(\beta+n)$ and $n/(\beta+n)$:

$$\hat{S}_\alpha(t) = \left(\frac{\beta}{\beta+n}\right)\hat{S}_0(t) + \left(\frac{n}{\beta+n}\right) S_n(t|X_1,\cdots,X_n) \tag{4.5}$$

where

$$S_n(t|X_1,\cdots,X_n) = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(t,+\infty)$$

is the empirical survival function. The ratio β/n represents the relative weight of the prior guess to the empirical distribution. If β is large compared to n, little weight is given to the observations. If β is small compared to n, little weight is given to the prior guess at S. As β tends to zero (the ”noninformative” Dirichlet prior), the Bayes estimator converges to the empirical survival function.

4.5.1 Asymptotic behaviour of the Bayes estimator

We now study the asymptotic behaviour of $\hat{S}_\alpha(t)$ as an estimator of $S(t)$ in a nondecision theoretic setup. We first examine the mean-squared error consistency and weak convergence of the nonparametric Bayes estimator.

Theorem 4.5.2 *The Bayes estimator $\hat{S}_\alpha(t)$ is mean-squared consistent, that is* $E\left\{\left(\hat{S}_\alpha(t) - S(t)\right)^2\right\}$ *tends to zero as $n \to \infty$.*

Proof. Mean-squared consistency can be proved by expressing the Bayes estimator as a linear combination of the prior guess $\hat{S}_0(t)$ and the empirical survival function $S_n(t)$ as in equation (4.5). Indeed,

$$E\left\{\left(\hat{S}_\alpha(t) - S(t)\right)^2\right\} = E\left\{\left(\left(\frac{\beta}{\beta+n}\right)\hat{S}_0(t) + \left(\frac{n}{\beta+n}\right)S_n(t) - S(t)\right)^2\right\} =$$

$$E\left\{\left(\left(\frac{\beta}{\beta+n}\right)(\hat{S}_0(t) - S(t)) + \left(\frac{n}{\beta+n}\right)(S_n(t) - S(t))\right)^2\right\} =$$

$$\left(\frac{\beta}{\beta+n}\right)^2 E\left\{\left(\hat{S}_0(t) - S(t)\right)^2\right\} +$$

$$+\left(\frac{n}{\beta+n}\right)^2 E\left\{(S_n(t) - S(t))^2\right\} + \tag{4.6}$$

$$+2\frac{\beta n}{(\beta+n)^2}(\hat{S}_0(t) - S(t))E\left\{(S_n(t) - S(t))\right\} = \tag{4.7}$$

$$= \left(\frac{\beta}{\beta+n}\right)^2(\hat{S}_0(t) - S(t))^2 + \frac{n^2}{(\beta+n)^2}\frac{1}{n}S(t)(1-S(t))$$

where (4.6) is the variance of the empirical survival curve $S_n(t)$ and expression (4.7) is zero because $E\left\{S_n(t)\right\} = S(t)$.

□

Theorem 4.5.3 $\sqrt{n}(\hat{S}_\alpha(t) - S(t))$ *follows, asymptotically, a mean zero normal distribution with variance* $S(t)(1 - S(t))$*.*

Proof. The proof is straightforward from the expression of $\hat{S}_\alpha(t)$ as a linear combination of the prior survival $\hat{S}_0(t)$ and the empirical survival function $S_n(t)$ as in (4.5):

$$\sqrt{n}(\hat{S}_\alpha(t) - S(t)) = \frac{\sqrt{n}\beta}{\beta + n}(\hat{S}_0(t) - S(t)) + \frac{n}{\beta + n}\sqrt{n}(S_n(t) - S(t))$$

and from the asymptotical properties of the empirical distribution $S_n(t)$. □

4.6 Bayesian Inference From Right-Censored Data

When data is right-censored, Susarla and Van Ryzin [51] propose a nonparametric Bayesian estimator of the survival function $S(t)$ based on a Dirichlet process prior.

Let $X_1, \cdots, X_n$ be a random sample with distribution function F and let $Y_1, \cdots, Y_n$, be independent and identically distributed random variables from a distribution function G. Assume that X_i is right-censored by Y_i. The observed data are then of the form: $(Z_i, d_i),\ i = 1, \cdots, n$, where

$$Z_i = min\left\{X_i, Y_i\right\} \text{ and } d_i = \mathbf{1}\left\{X_i \leq Y_i\right\}\ .$$

For simplicity, and without loss of generality, data are rearranged so that the first k pairs $(Z_i, \delta_i),\ i = 1, \cdots, k$, are the exact observations, while the rest $n-k$ pairs correspond to the censored observations.

The estimator of the survival function $S(t) = 1 - F(t)$ is obtained by Susarla and Van Ryzin in two steps. Assuming a Dirichlet process *a priori* on $\mathbb{R}^+$ of parameter measure α, they first show that the conditional distribution of F given the exact observations $(Z_1, 1), \cdots (Z_k, 1)$ is a Dirichlet process of parameter $\alpha^* = \alpha + \sum_1^k \delta_{Z_i}$. Secondly, they consider $X_{k+1}, \cdots, X_n$ a random sample from the process $\mathcal{D}(\alpha^*)$ and find the conditional moments of $S(t)$ given the censored data $(Z_i, 0),\ i = k+1, \cdots, n$.

The results can be stated in the following theorem:

Theorem 4.6.1 (Susarla and Van Ryzin) *The conditional moments of $S(t)$ given a right-censored sample are of the form:*

$$E[(S(t))^p|(\delta, \mathbf{Z})] = \prod_{s=0}^{p-1} \left\{ \frac{\alpha(t,\infty)+s+N^+(t)}{\alpha(\mathbb{R}^+)+s+n} \prod_{i\in I} \left\{ \frac{\alpha[Z_i,\infty)+s+N^+(Z_i)+\lambda_i}{\alpha[Z_i,\infty)+s+N^+(Z_i)} \right\}^{1-d_i} \right\}$$

where $I = \{i : Z_i \leq t$ and i is the first subscript among tied censored Z's$\}$ and $\lambda_i =$ number of observations at Z_i, $i = 1, \cdots, n$, and $N^+(t) =$ number of observations (censored or not) $> t$.

Bayes estimator of the survival function under right censoring

Taking $p = 1$, one obtains the **Bayes estimator** for the survival function $S(t)$ based on right-censored data:

$$\hat{S}_\alpha(t) = \frac{\alpha(t,\infty)+N^+(t)}{\beta+n} \prod_{i\in I} \left\{ \frac{\alpha[Z_i,\infty)+N^+(Z_i)+\lambda_i}{\alpha[Z_i,\infty)+N^+(Z_i)} \right\}^{1-d_i} \tag{4.8}$$

In the censored data case the Bayes estimator $\hat{S}_\alpha(t)$ cannot be expressed as a linear combination of the prior distribution and the maximum likelihood estimator, but the interpretation of β/n as the relative weight of those distributions is still useful. It is clear from (4.8) that if no weight is given to the data, that is if $n \to 0$ and $\beta > 0$, $\hat{S}_\alpha(t)$ reduces to the prior survival $\hat{S}_0(t) = \alpha(t,+\infty)/\beta$. On the other hand, if no weight is given to the prior distribution the Bayes estimator reduces to the Kaplan-Meier estimator:

Proposition 4.6.2 *If $\beta = \alpha(\mathbb{R}^+) \to 0$ and $n > 0$, the Bayes estimator reduces to the Kaplan-Meier estimator.*

Proof. As $\alpha(\mathbb{R}) \to 0$, the Bayes estimator $\hat{S}_\alpha(t)$ tends to

$$L = \frac{N^+(t)}{n} \prod_{i\in I} \left\{ \frac{N^+(Z_i)+\lambda_i}{N^+(Z_i)} \right\}^{1-d_i} .$$

The first term $N^+(t)/n$ can be expressed as a product where the consecutive terms simplify:

$$\frac{N^+(t)}{n} = \prod_{i\in I} \frac{N^+(Z_i)}{N^+(Z_i)+\lambda_i} .$$

Then

$$\begin{aligned} L &= \frac{N^+(t)}{n} \prod_{i\in I} \left\{ \frac{N^+(Z_i)+\lambda_i}{N^+(Z_i)} \right\}^{1-d_i} = \\ &= \prod_{i\in I} \left\{ \frac{N^+(Z_i)}{N^+(Z_i)+\lambda_i} \right\} \cdot \left\{ \frac{N^+(Z_i)+\lambda_i}{N^+(Z_i)} \right\}^{1-d_i} \end{aligned}$$

The terms corresponding to censored observations ($d_i = 0$) simplify and therefore the product have to be computed only for the uncensored observations:

$$\begin{aligned} L &= \prod_{i \in I} \left\{ \frac{N^+(Z_i)}{N^+(Z_i) + \lambda_i} \right\}^{d_i} = \\ &= \prod_{i \in I} \left\{ 1 - \frac{\lambda_i}{N^+(Z_i) + \lambda_i} \right\}^{d_i} \end{aligned}$$

and this corresponds to the Kaplan-Meier estimator in the presence of ties. □

It is interesting to note here that the explicit derivation of the Bayes estimator (4.8) has been possible only because in the presence of right censoring we know for each t the exact number of observations that have failed until that time, and therefore, it is possible to define the counting process $N^+(t)$. However, this interesting property does not hold for other censoring schemes. In an interval censoring scheme we may not know how many observations have failed in a given interval. For that reason, in general it is difficult to obtain an explicit form for the Bayes estimator of a survival function.

Next theorems state the asymptotic behaviour of the Susarla and Van Ryzin [52] Bayes estimator.

Theorem 4.6.3

1. *For a fixed t such that $S(t) > 0$, the Bayes estimator $\hat{S}_\alpha(t)$ is mean-squared consistent with*
$$E[\hat{S}_\alpha(t) - S(t)]^2 = O(n^{-1})$$
2. *The asymptotic distribution of $\sqrt{n}(\hat{S}_\alpha(t) - S(t))$ is normal with mean zero and variance*
$$(S(t))^2 \int_0^t [S^2(u)(1 - G(u))]^{-1} \, dF(u) \ .$$

4.6.1 Comparison with the Kaplan-Meier estimator

An important consequence of the above results is that the Bayes estimator $\hat{S}_\alpha(t)$ and the Kaplan-Meier estimator are asymptotically equivalent, having the same consistency properties and identical asymptotical variance. Hence, why do we need to constructthe nonparametric Bayes estimator if the Kaplan-Meier estimator is easier to compute and provides equivalent answers ? There are at least two advantages of the Bayes estimator over the Kaplan-Meier estimator. The first is that the Bayes estimator makes more use of the censored data; that is, one needs all the observations to calculate the Bayes estimator, while to calculate the Kaplan-Meier estimator one only needs the number of censored observations between two uncensored observations. In fact, it is possible to recover from the estimated survival curve, the actual observations pairs $\{(Z_i, d_i), i = 1, \cdots, n\}$. More precisely, the Bayes estimator is a function of the full sufficient statistic $\{(Z_i, d_i), i = 1, \cdots, n\}$, while the Kaplan-Meier estimator is not. A second advantage of the Bayes estimator is that it smoothes the nonparametric estimator by shrinking it toward a smooth survival curve. In fact, if $\alpha(.)$ is continuous, strictly decreasing and differentiable, the Bayes estimator pieces together strictly decreasing, differentiable curves which join in a continuous but nondifferentiable manner at each censored observation. One disadvantage in smoothing is the possible introduction of bias in the estimator by the prior parametric family that has been chosen.

To study this question thoroughly, Rai, Susarla and Van Ryzin [47] carry out a simulation study to compare the performance of these estimators for small sample sizes. In this study the prior survival $\hat{S}_0(t)$ and the weight β are chosen using the following *empirical Bayes* approach:

The prior survival is specified from a certain parametric family of survival curves, $\hat{S}_0(t; \boldsymbol{\theta})$, and the vector of parameters $\boldsymbol{\theta}$ is estimated from the data by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$.

After having chosen a prior survival $\hat{S}_0(t; \hat{\boldsymbol{\theta}})$, they argue that the weight β should increase as the sample size in order to keep constant the relative weight β/n between the prior survival function and the empirical survival function. That is, they suggest taking $\beta(n) = O(n^k)$. However, with such a choice of β, the consistency of the Bayes estimator is not ensured. To obtain a consistent estimator it is necessary to take $\beta(n) = O(n^k)$ with $0 < k < 1$. In particular, the suggested choice of Rai, Susarla and Van Ryzin [47] is to take $k = 1/2$, that is $\beta(n) = c \cdot \sqrt{n}$, to give enough weight to smoothing the nonparametric estimator toward the parametric family but without loosing the good asymptotical properties.

In this simulation study, four estimators of the survival function are compared. The first estimator is totally parametric, is the maximum likelihood estimator obtained assuming an exponential survival function. The second and third are the consistent and unconsistent Bayes estimators, $\beta = \sqrt{n}$ and $\beta = n$, respectively, with an exponential prior survival curve. The final estimator is the Kaplan-Meier product limit estimator. These estimators are compared by means of three different norms, considering four percentages of censoring, and for the case where the true survival curve is exponential and when the true survival curve is not exponential, but gamma.

The results of this simulation study show that:

- The consistent mean-squared Bayes estimator with $\beta = \sqrt{n}$ appears to have definitely better small sample properties than the Kaplan-Meier estimator with no loss in large sample properties.
- The large sample advantages of the Kaplan-Meier estimator over the biased inconsistent Bayes estimator with $\beta = n$ may not show up, particularly with heavy censoring, until fairly large samples are taken.

Chapter 5

Markov Chain Monte Carlo Methods

5.1 Introduction

Markov Chain Monte Carlo (MCMC) methods includes a variety of iterative simulation methods to generate values from a sequence of distributions that converge to a desired *target distribution*. Such computational algorithms have made a significative impact in practical statistics, specially in bayesian analysis, since they provide numerical solutions to otherwise intractable problems. With the classical Monte Carlo methods one generates independent samples directly from the target distribution in such a way that the empirical distribution of the sample approximates it. However, it is rare that independent samples from an arbitrary distribution can be obtained directly. The usual strategy is then to sample from a distribution that is close, in some sense, to the target distribution and that is easy to sample. Then, the distribution of interest is approximated by an appropriately weighted empirical distribution from this sample. Importance Sampling and Rejection-Acceptance are typically the methods used to obtain independent samples from a distribution similar to the distribution of interest. This has been the usual way of exploring distributions from a sampling approach. However, for many complex models, such as hierarchical models in Bayesian analysis or models involving missing or censored data, such direct strategies are not feasible. In those cases where one is unable to obtain independent samples it may be appropriate to use dependent samples generated using iterative simulation methods. The idea behind such iterative simulation methods (MCMC methods) is to obtain a Markov Chain of simulated values from a Markov process whose invariant distribution is the target distribution. If the simulation process is iterated a large number of times, the simulated values can be used to summarize features of the distribution of interest. Iterative simulation procedures are typically less efficient and require larger samples than direct simulation methods, however MCMC methods are applicable in a wider range of problems.

There exist different algorithms to construct Markov chains with an specific invariant distribution. We consider here the most widely used Markov Chain simulation methods, namely, the Metropolis-Hastings algorithm and the Gibbs sampler. The Gibbs sampler was first developed by Geman and Geman [31] in the context of image-processing. The models used in that paper were Markov random fields involving Gibbs distributions, from where the Gibbs sampler takes its name. The roots of the MCMC methods can be found earlier in the works of Metropolis *et alt.* [44] and Hastings [38] who derived similar algorithms from Markov processes. However, these methodologies do not become a widely used technique until 1990 when Gelfand and Smith [27] extend the theory of Geman and Geman [31] to continuous distributions and show how to use their method in a wide range of statistical problems. There is already an extensive literature concerning MCMC methods. Among them, Chib, S. and Greenberg, E. [11] and Casella, G. and George, E.I. [7] are good introductory papers of the Metropolis-Hastings algorithm and of the Gibbs sampler, respectively. For a more rigorous mathematical discussion of the MCMC methods Tierney's paper [55] is appropriate.

In section 5.2 we present some definitions and results on Markov chains that are useful to understand how the iterative simulation methods work. The Metropolis-Hastings algorithm and the Gibbs Sampler are introduced in sections 5.3 and 5.4, respectively. In the last section we make some comments about some techniques for making inference and monitoring convergence from iterative simulation methods.

5.2 Some definitions and theoretical results on Markov Chains

We present here some basic concepts and results on Markov chains that are useful to understand how the iterative simulation methods work.

Suppose that we are observing random variables $X_0, X_1, \cdots$ which are the successive states of a system. We call this system a Markov chain if the probabilities for passing into the next state are completely determined by the present state of the system. More precisely,

Definition 5.2.1 *The random variables* $X_0, X_1, \cdots$ *are a* **Markov chain** *if for all*

$x_0, x_1, \cdots, x_n$ and any measurable event A

$$P(X_{n+1} \in A | X_n = x_n, \cdots X_0 = x_0) = P(X_{n+1} \in A | X_n = x_n),$$

or equivalently, if the conditional densities satisfy

$$f(x_{n+1}|x_n, \cdots, x_0) = f_n(x_{n+1}|x_n).$$

In this case,

$$P(X_{n+1} \in A | X_n = x) = P(x, A)$$

are called the **transition probabilities** *and represent the probability of moving from x to a point in A.*

Definition 5.2.2 *Any probability distribution π is a limiting or* **equilibrium distribution** *of a chain if*

$$\lim_{n\to\infty} P(X_n \in A | X_0 = x) = \pi(A).$$

Definition 5.2.3 *Any probability distribution π satisfying*

$$\pi(A) = \int_\Omega P(x, A)\pi(dx)$$

for all measurable set A is called an **invariant distribution** *of the chain.*

Definition 5.2.4 *A Markov chain with invariant distribution π is* **irreducible** *if, for any initial state, the probability of entering any set to which π assigns positive probability is positive.*

Definition 5.2.5 *A Markov chain is* **periodic** *if there are portions of the state space it can only visit at certain regularly spaced times; otherwise, the chain is aperiodic.*

The goal of the MCMC methods is to create a Markov chain whose equilibrium distribution is the distribution of interest. The following result ensures that this objective can be achieved by constructing an irreducible and aperiodic Markov chain with the target distribution as its invariant distribution.

Theorem 5.2.6 *If a chain has a proper invariant distribution π and it is irreducible and aperiodic, then*

1) *π is the unique invariant distribution and is also the equilibrium distribution of the chain, that is, $\lim_{n\to\infty} P(X_n \in A|X_0 = x) = \pi(A)$.*

2) *(Ergodic theorem) If $f(.)$ is a real valued function such that $E_\pi\{|f(X)|\} < \infty$, then $\dfrac{1}{n}\sum\limits_{t=1}^{n} f(X_t) \to E_\pi\{f(X)\}$, almost surely, as $n \to \infty$, where $E_\pi\{f(X)\}$ is the expectation of $f(X)$ with respect to π.*

This result is important in practice since most output from the iterative simulation methods will be summarized in terms of the statistic $\dfrac{1}{n}\sum\limits_{t=1}^{n} f(X_t)$.

Proposition 5.2.7 *A sufficient condition for a given distribution π to be the invariant distribution of a chain is that the transition probabilities of the chain satisfy the* **reversibility condition***:*

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

Proof. Indeed, for any measurable set A, the transition probabilities with respect to Lebesgue measure on $\mathbb{R}^d$ can be expressed as

$$P(x,A) = \int_A P(x,y)\ dy + r(x)\mathbf{1}\{x \in A\}$$

where the first term is the probability of passing from x to a different state in A and $r(x) = 1 - \int_\Omega P(x,y)\ dy$ is the probability that the chain remains at x. Then,

$$\begin{aligned}
&\int_\Omega P(x,A)\pi(dx) = \\
&= \int_\Omega \left[\int_A P(x,y)\ dy\right] \pi(x)\ dx + \int_\Omega r(x)\mathbf{1}\{x \in A\}\pi(x)\ dx = \\
&= \int_A \left[\int_\Omega P(x,y)\pi(x)\ dx\right]\ dy + \int_A r(x)\pi(x)\ dx = \qquad (5.1)\\
&= \int_A \left[\int_\Omega P(y,x)\pi(y)\ dx\right]\ dy + \int_A r(x)\pi(x)\ dx = \\
&= \int_A (1-r(y))\pi(y)\ dy + \int_A r(x)\pi(x)\ dx = \\
&= \int_A \pi(y)\ dy = \pi(A)
\end{aligned}$$

where in (5.1) we have used the reversibility condition. □

5.3 The Metropolis-Hastings Algorithm

Let $\pi(\boldsymbol{\theta})$ be the distribution of interest with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \Omega$ the vector of parameters, where the components of $\boldsymbol{\theta}$ could be themselves vectors. The Metropolis-Hastings algorithm constructs a Markov chain $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots, \boldsymbol{\theta}^t, \cdots$ whose invariant distribution is $\pi(\boldsymbol{\theta})$. The algorithm is similar to the acceptance-rejection method in the sense that a new candidate is accepted or rejected according to a given probability α.

If the chain is currently at $\boldsymbol{\theta}^{t-1}$ at time $t-1$, the next state $\boldsymbol{\theta}^t$ is chosen by first sampling a candidate value $\boldsymbol{\theta}'$ from an arbitrary transition probability function $q(\boldsymbol{\theta}^{t-1}, \cdot)$ and this candidate is accepted with probability $\alpha(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}')$, that is, the algorithm can be summarized as

Given a starting point $\boldsymbol{\theta}^0$, for $t = 1, 2, \cdots$:

- Sample $\boldsymbol{\theta}'$ from the distribution $q(\boldsymbol{\theta}^{t-1}, \cdot)$.
- Set
$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}' & \text{with probability } \alpha(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}') \\ \boldsymbol{\theta}^{t-1} & \text{otherwise.} \end{cases}$$

The probability α is determined in such a way that the chain satisfies the reversibility condition:

Theorem 5.3.1 *The transition probability from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, given by:*

$$P(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} q(\boldsymbol{\theta}, \boldsymbol{\theta}') \cdot \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') & \textit{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}' \\ 0 & \textit{if } \boldsymbol{\theta} = \boldsymbol{\theta}' \end{cases}$$

where

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \quad min \left\{ \dfrac{\pi(\boldsymbol{\theta}') q(\boldsymbol{\theta}', \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\theta}')}, 1 \right\} & \textit{if } \pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0 \\ 1 & \textit{if } \pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\theta}') = 0 \ . \end{cases}$$

satisfies the reversibility condition and therefore the target distribution is an invariant distribution of the chain.

Proof.

Indeed, if the transition probability $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ satisfies the reversibility condition, then the chain obtained from q has $\pi(\boldsymbol{\theta})$ as its invariant distribution. Nevertheless, it is most likely that for an arbitrary q the reversibility condition is not fulfilled and, therefore, we find values $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ for which, for example,

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}') > \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}',\boldsymbol{\theta}).$$

Intuitively this means that it is more likely to go from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ than from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$. To balance this situation the moves from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ will be reduced by introducing a probability $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') < 1$ of accepting the move and, conversely, the moves from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$ will be accepted with probability $\alpha(\boldsymbol{\theta}',\boldsymbol{\theta}) = 1$.

Then, the reversibility condition of the transition probability

$$\pi(\boldsymbol{\theta})P(\boldsymbol{\theta},\boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}',\boldsymbol{\theta})$$

becomes

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}')\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}',\boldsymbol{\theta}).$$

Therefore, the probability of move is defined as

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') = \begin{cases} \min \left\{ \dfrac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}',\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}')}, 1 \right\} & \text{if } \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}') > 0 \\ 1 & \text{if } \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}') = 0\ . \end{cases}$$

□

If, in addition, the transition probability function q is selected in such a way that the Markov chain is irreducible and aperiodic, the chain converges to the target distribution π. Tierney [55] suggests different choices of the transition probability q that may be useful in practical simulation studies.

5.4 The Gibbs Sampler

The Gibbs Sampler algorithm is a special case of the Metropolis-Hastings algorithm where the components of $\boldsymbol{\theta}$ are updated one by one. Indeed, the Gibbs sampler consists in sampling iteratively from the **full conditional distributions**, where the k^{th} full conditional distribution, denoted by

$$\pi(\theta_k|\boldsymbol{\theta}_{-k}) = \pi(\theta_k|\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_d)\ ,$$

is the distribution of the k-th component of $\boldsymbol{\theta}$ conditioned on all the remaining components.

Then, given an arbitrary set of starting values $\boldsymbol{\theta}^0 = (\theta_1^0, \cdots, \theta_d^0)$ the Gibbs sampler algorithm proceeds making successively random draws from the full conditional distributions as follows:

ALGORITHM:

```
    Sample $\theta_1^{(i)}$ from $\pi(\theta_1|\theta_2^{(i-1)}, \cdots, \theta_d^{(i-1)})$
    Sample $\theta_2^{(i)}$ from $\pi(\theta_2|\theta_1^{(i)}, \theta_3^{(i-1)}, \cdots, \theta_d^{(i-1)})$
    Sample $\theta_3^{(i)}$ from $\pi(\theta_3|\theta_1^{(i)}, \theta_2^{(i)}, \theta_4^{(i-1)}, \cdots, \theta_d^{(i-1)})$
    $\cdots$
    Sample $\theta_d^{(i)}$ from $\pi(\theta_d|\theta_1^{(i)}, \cdots, \theta_{d-1}^{(i)})$
```

This loop completes the ith iteration of the Gibbs sampler generating the vector $\boldsymbol{\theta}^i = (\theta_1^i, \cdots, \theta_d^i)$. Repeating this process we get the sequence $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^t, \cdots$ which is a realization of a Markov chain with transition probability from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$ given by the product of the full conditional distributions.

$$P(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}) = \prod_{l=1}^{d} \pi(\theta_l^{(t+1)}|\theta_1^{(t+1)}, \cdots, \theta_{l-1}^{(t+1)}, \theta_{l+1}^{(t)}, \cdots, \theta_d^{(t)}).$$

Under mild regularity conditions, Geman and Geman [31] showed that the following results hold:

Theorem 5.4.1 *The joint distribution of* $(\theta_1^i, \cdots, \theta_d^i)$ *converges geometrically to* $\pi(\theta_1, \cdots, \theta_d)$, *as* $i \to \infty$.

Theorem 5.4.2 (ergodic theorem) *For any measurable function* f *of* $(\theta_1, \cdots, \theta_d)$ *whose expectation exists,*

$$\frac{1}{k}\sum_{i=1}^{k} f(\theta_1^i, \cdots, \theta_d^i) \to E(f(\theta_1, \cdots, \theta_d)), \text{ almost surely, as } k \to \infty \ .$$

As mentioned before, the Gibbs sampler is a special case of the Metropolis-Hastings algorithm. For each iteration of the Metropolis-Hastings algorithm, it is necessary to perform d steps of the Gibbs sampler, corresponding to the d components of $\boldsymbol{\theta}$. In each step the jth component of $\boldsymbol{\theta}$ is updated. The function $q(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*)$ of the Metropolis-Hastings algorithm is defined as the conditional density of $\boldsymbol{\theta}_j$ given the other components.

That is

$$q(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*) = \begin{cases} \pi(\theta_j^* | \boldsymbol{\theta}_{-j}^{t-1}) & \text{if } \boldsymbol{\theta}_{-j}^* = \boldsymbol{\theta}_{-j}^{t-1} \\ 0 & \text{otherwise.} \end{cases}$$

where $\boldsymbol{\theta}_{-j}^{t-1}$ represents the $(d-1)$ dimensional vector whose components are the components of $\boldsymbol{\theta}$, except for $\boldsymbol{\theta}_j$, at their current values:

$$\boldsymbol{\theta}_{-j}^{t-1} = (\theta_1^t, \cdots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \cdots, \theta_d^{t-1}).$$

In the Gibbs sampler, the value obtained for θ_j in each step is always accepted because the probability $\alpha(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*)$ of accepting a new candidate for θ_j is always 1. Indeed,

$$\alpha(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*) = \frac{\pi(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{t-1})}{\pi(\boldsymbol{\theta}^{t-1}) q(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*)} = \frac{\pi(\boldsymbol{\theta}^*) \pi(\theta_j^{t-1} | \boldsymbol{\theta}_{-j}^{t-1})}{\pi(\boldsymbol{\theta}^{t-1}) \pi(\theta_j^* | \boldsymbol{\theta}_{-j}^{t-1})} = \frac{\pi(\boldsymbol{\theta}_{-j}^{t-1})}{\pi(\boldsymbol{\theta}_{-j}^{t-1})} = 1.$$

5.5 Inference and Convergence diagnostics

The values obtained by iterative simulation methods will be used to make inferences about some aspects of the target distribution. However there are two main difficulties in making such inferences. The first problem is to know how long has to be the simulated sequence. The second problem is the possible correlation between draws that may cause inefficiencies in simulations.

One possibility is to base all inference on *one long run* of the Markov chain and use time-series results to monitor convergence. The first k values of the sequence have to be discarded to assure that the effect of the starting value can be ignored. The number k is usually called the *burn-in* of *warm-up*. This strategy is followed by different authors such as Geyer [32] or Raftery and Lewis [46]. The basic difficulty of this approach is that the sequence may remain for a long time in a small subset of the sample space heavily influenced by the starting distribution.

As an alternative, Gelman and Rubin [30] propose the use of m *independent sequences* and the *burn-in* k to be the first point at which the densities appear to be the same. The main reasons for using multiple chains are to allow variance estimation from the independent chains, to reduce correlations in the total sample and to aid in detecting problems in the simulation. Cowles and Carlin [12] present a thorough comparative review of convergence diagnostics. The strategy proposed by Gelman and Rubin to make inference from multiple sequences can be summarized as:

Simulating multiple sequences from an overdispersed distribution

In order to avoid that the chain remains for a long time in a small region of the sample space, they suggest to simulate independent sequences with starting values drawn from an overdispersed distribution. Such starting distribution may be obtained for instance by importance resampling from an approximate distribution.

Then, simulate independently m sequences ($m \geq 2$) of length $2n$, with starting points drawn from the starting distribution. To diminish the effect of the starting values they propose to discard the first n iterations of each sequence and use only the last n.

Monitoring convergence

A first approach for detecting lack of convergence of the chains is to plot the sample trace of the different sequences in the same graphic and see if they can be distinguished or, on the contrary, they appear to be the same.

A more quantitative method, inspired in the analysis of variance, is to form an overestimate and an underestimate of the variance of the target distribution, with the property that the estimates will be roughly equal at convergence but not before. Since it is a method based on the normal-theory, it is best to transform the scalar estimands to be approximately normal (for example, take logarithms of all-positive quantities and logits of quantities that lie between 0 and 1).

For each scalar of interest x we have nm simulated values $x_{ij},\ i = 1, \cdots, n;\ j = 1, \cdots, m$ corresponding to the n valid iterations of the m independent sequences. From these values, we calculate the variance between the m sequence means, $\bar{x}_{.j}$

$$B/n = \sum_{j=1}^{m} (\bar{x}_{.j} - \bar{x}_{..})^2/(m-1) \quad \text{where} \quad \bar{x}_{.j} = \sum_{i=1}^{n} x_{ij}/n \quad \text{and} \quad \bar{x}_{..} = \sum_{j=1}^{m} \bar{x}_{.j}/m$$

and the average of the m within-sequence variances

$$W = \sum_{j=1}^{m} s_j^2/m \quad \text{where} \quad s_j^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2/(n-1).$$

The target mean, $E(x)$, might be estimated by the sample mean $\bar{x}_{..}$, and the posterior target variance $\text{var}(x)$ might be estimated by a weighted average of W and B,

$$\widehat{\text{var}}(x) = \frac{n-1}{n} W + \frac{1}{n} B$$

which overestimates the target variance $\text{var}(x)$ if the starting distribution is overdispersed but $E\{\widehat{\text{var}}(x)\} = \text{var}(x)$ if $n \to \infty$ or if the starting distribution equals the target distribution. On the other hand, the within variance W based on finite sequences underestimates the target variance and, as $n \to \infty$, the expectation of W approaches $\mathrm{var}(x)$. Then, convergence of the iterative simulation can be studied by the ratio of the current variance estimate and the within-sequence estimate

$$\hat{R} = \frac{\hat{\mathrm{var}}(x)}{W}$$

which tends to 1 in the limit $n \to \infty$. If $\hat{R}$ is not near 1 for all scalar estimands of interest the simulations should be continued.

A very useful software for analysing the output from the Gibbs sampler is the program CODA [3] *"Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output"*. This program implements different graphical analysis, and convergence diagnostic tests. In particular, it provides the Gelman and Rubin's convergence diagnostics mentioned before, that is:

- Plots of the sample trace for each variable.
- Plots of Gelman and Rubin's factor $\hat{R}$.

Chapter 6

Nonparametric Bayesian Estimation from Interval-Censored Data

6.1 Introduction

In chapter 4 we presented the nonparametric Bayesian analysis as an appropriate alternative to the estimation of a survival function in the presence of censoring. With this approach it is possible to incorporate prior believes about the survival function without the need of assuming restrictive parametric models. However, the extension of this theory to complex censoring schemes requires complicated computations that are in general not affordable in an explicit way. Our goal is to obtain the nonparametric Bayes estimator of the survival function when there is interval censoring by means of an iterative simulation method.

As suggested in Smith and Roberts [50], the Gibbs sampler is a very useful method in problems involving incomplete data. If the missing data are reintroduced in the model as further unknowns, the implementation of the algorithm leads in general to more tractable situations. In fact, we will use a version of the Gibbs sampler, the Data Augmentation algorithm [54]. The basic idea behind this algorithm is to augment the *observed data* y by a quantity z, which will be referred to as *latent data*. It is assumed that given both y and z one can calculate or sample from the augmented data posterior $\pi(\theta|y, z)$. Then, to obtain the distribution of interest, $\pi(\theta|y)$, the algorithm proceeds by generating iteratively values z from the predictive distribution $\pi(z|y)$ and values of θ from the augmented posterior $\pi(\theta|y, z)$.

In the case of **censoring**, this strategy would correspond to introduce the censored data as additional parameters. In each iteration of the Gibbs sampler the censored data are updated analogously for the other parameters. More precisely, let us assume that Y is the random variable of interest and assume a parametric model with vector parameter $\boldsymbol{\theta}$ for Y. In the presence of censoring a sample of Y can be expressed as $y = (y_{obs}, y_{cens})$ where $y_{obs} = (y_1, \dots, y_s)$ are exactly observed, and the remaining data $y_{cens} = (y_{s+1}, \dots, y_n)$ are censored, and are only known to lie in some regions defined by the sampling data. If we define y_{cens} as further unknowns, with $\boldsymbol{\theta}$ and y_{cens} together constituting the augmented unknowns, the distribution of interest, that is, the observed posterior $\pi(\theta|y_{obs})$ is obtained by generating successively from the corresponding full conditional distributions:

$$\pi(y_{cens}|\boldsymbol{\theta}, y_{obs}, \text{Data}) = \pi(y_{cens}|\boldsymbol{\theta}, \text{Data})$$

$$\pi(\boldsymbol{\theta}|y_{obs}, y_{cens}, \text{Data}) = \pi(\boldsymbol{\theta}|y)$$

The first conditional is the joint distribution of the censored observations given $\boldsymbol{\theta}$ and the data. This is typically a truncated distribution in the regions specified by the data. With the introduction of the censored data y_{cens} in the model, the second conditional distribution is simply the joint posterior of $\boldsymbol{\theta}$ if there is no censoring. Thus, the Gibbs sampler here proceeds in two steps: a first step where each censored value is imputed by an uncensored one, obtaining as a result a complete data set, and a second step where the original parameters are updated in the usual way, conditioning on the imputed values.

6.2 Inference from Interval-Censored Data using the Gibbs Sampler

Let X be the random variable of interest with distribution function W. Let us assume that the variable X is subject to an interval censoring mechanism. In this situation the observable data for each individual i are of the form (X_L^i, X_R^i), denoting that $X^i \in [X_L^i, X_R^i]$.

Our goal is to obtain the nonparametric Bayes estimator $\hat{S}(t)$ of the random survival function $S(t) = 1 - W(t) = P(X > t)$ based on the interval-censored data, under the squared error loss

$$L(\hat{S}, S) = \int_0^{+\infty} (\hat{S}(t) - S(t))^2 \, dw(t) \ ,$$

where w is a weight function, and assuming a Dirichlet process of parameter measure α as a prior distribution for W.

In proposition 4.5.1, the Bayes estimator of the survival function under the squared error loss was shown to be its posterior expectation, that is, $\hat{S}(t) = E\left\{S(t)|x\right\}$, where now x denotes the observed data $\{(X_L^i, X_R^i),\ i = 1, \ldots, n\}$ and E denotes expectation with respect to the posterior distribution of the process P.

Our approach consists on obtaining, iteratively, a random sample from the posterior distribution of $S(t)$ and estimate its posterior expectation $E\left\{S(t)|x\right\}$ by the sample mean. In order to do that we use the Data Augmentation strategy. This process is carried out in a finite number of times, say $0 = t_0 < t_1 < \ldots < t_r = +\infty$ that are fixed in advance.

6.2.1 Nonparametric Bayes Estimator

Let the vector of probabilities $\mathbf{w} = (w_1, \ldots, w_r)$ where $w_j = P(X \in (t_{j-1}, t_j])$, $j = 1, \ldots, r$ and let $\delta_j^i = \mathbf{1}\{X^i \in (t_{j-1}, t_j]\}$ that indicates in which time interval has occurred the event of interest.

We propose $\hat{S}(t_j) = 1 - \sum_{s \leq j} \hat{w}_s$ the nonparametric Bayes estimator of the survival function at time t_j, where $\hat{w}_j,\ j = 1, \ldots, r$ are the sample means of the simulated vectors from a Dirichlet distribution, that are obtained iteratively from the following algorithm, that will be referred as NPBE algorithm:

0) Define starting values for $\mathbf{w}$: $\mathbf{w}^0 = (w_1^0, \ldots, w_r^0)$

1) For each individual $i = 1, \ldots, n$, generate the random vector $(\delta_1^i, \ldots, \delta_r^i)$ from a truncated multinomial of sample size 1 and parameters $(w_1^0, \ldots, w_r^0)$. Compute $n_j = \sum_{i=1}^n \delta_j^i$, the number of X's in each interval $(t_{j-1}, t_j]$.

2) Generate $\mathbf{w}^1 = (w_1^1, \ldots, w_r^1)$ from a Dirichlet distribution of parameters $(\alpha_1 + n_1, \ldots, \alpha_r + n_r)$, where $\alpha_j = \alpha((t_{j-1}, t_j])$.

3) Replace $\mathbf{w}^0$ by $\mathbf{w}^1$ and return to (1).

Note that $\hat{\mathbf{w}}$ estimates the mean of the posterior distribution of $\mathbf{w}$ given the data, but other quantities such as the median or other percentiles could be also empirical estimated from the simulated sample.

We now proceed to justify this choice.

Gibbs algorithm. First approach

A natural way of implementing the Gibbs algorithm for obtaining a sample from the posterior distribution of $S(t_j),\ j=1,\dots,r$, given the censored data, would be to perform successively a two steps algorithm where the first step would correspond to the imputation of uncensored values for each individual and the second step would correspond to sampling from the known posterior distribution of $S(t_j)$ given the complete data obtained in the first step. Since, if the prior distribution is a Dirichlet process, the posterior distribution of $S(t)$ given the complete data is a beta distribution, the algorithm would be:

1. Sample $X^1,\cdots,X^n$ from a Dirichlet process $W=1-S$ with the restriction of $X^i \in [X_L^i, X_R^i]\ , i=1,\dots,n$.
2. For $j=1,\cdots,r$, sample $S(t_j)$ from a beta distribution $\mathcal{B}e(\alpha^*(t_j,+\infty),\alpha^*(-\infty,t_j])$ where $\alpha^*(A)=\alpha(A)+\sum_1^n \delta_{X^i}(A)$ for any subset A of the real line.

Although the algorithm is intuitive and straightforward, it involves the generation of a sample from a Dirichlet process which is not an easy step. Doss [18] uses the results of Sethuraman [49] to provide an algorithm for approximately simulate a Dirichlet process. However, this approach introduces an unnecessary difficulty to the problem. Since our goal is to obtain the posterior distribution of $S(t_j)$, we are mainly interested in getting into step (2). Step (1) is only an auxiliary step that is necessary because with the imputation of values for each censored observation step (2) can be carried out conditionally to a complete data set. For this reason we propose an alternative algorithm based on the introduction of new auxiliary variables.

Modification of the initial algorithm

Consider the partition of the real line given by $0=t_0<t_1<\dots<t_r=+\infty$ and define the vector of probabilities for each interval $(t_{j-1},t_j]$, say $\mathbf{w}=(w_1,\dots,w_r)$, where $w_j=P(X\in(t_{j-1},t_j])$.

With the introduction of the variable of probabilities $\mathbf{w}$ in the model, the r simulations in step (2) can be reduced to only one draw from a Dirichlet distribution. Indeed, the random variable $S(t_j)$ can be expressed as $S(t_j)=1-\sum_{l\leq j} w_l$ and the r simulations from a beta distribution reduce to sampling a vector $\mathbf{w}=(w_1,\dots,w_r)$, with $\sum w_j=1$, from a Dirichlet distribution.

Aside from that, we can use an important property of the Dirichlet processes:

Result 6.2.1 *If P is a Dirichlet process, for each measurable set A, the posterior distribution of $P(A)$ given a sample $X^1, \cdots, X^n$ from P, only depends on the number of X's that fall in A and not on where they fall within or outside of A.*

Then, to obtain the posterior distribution of $S(t_j), j = 1, \cdots, r$, or equivalently, the posterior distribution of the vector $\mathbf{w} = (w_1, \ldots, w_r)$, it is only necessary to know the number n_j of X's falling in each interval $(t_{j-1}, t_j]$, for $j = 1, \ldots, r$. Thus, instead of working with the sample $X^1, \cdots, X^n$ we introduce as a new auxiliary variable the vector $\mathbf{n} = (n_1, \cdots, n_r)$.

Full conditionals of the proposed NPBE algorithm

The parameters of interest are the components of vector $\mathbf{w} = (w_1, \ldots, w_r)$ that define $S(t_j) = 1 - \sum_{l \leq j} w_l$. To obtain its posterior distribution under a censored sample, we introduce the vector $\mathbf{n} = (n_1, \ldots, n_r)$ as further unknowns, where n_j is the number of X's falling into $(t_{j-1}, t_j]$. With $\mathbf{w}$ and $\mathbf{n}$ together as the augmented unknowns, the full conditionals are:

1. The posterior distribution of $\mathbf{n}$ given $\mathbf{w}$ and the data
$$f(n_1, \ldots, n_r | \mathbf{w}, [X_L^i, X_R^i], i = 1, \ldots, n) \ .$$
2. The posterior distribution of $\mathbf{w}$ given $\mathbf{n}$ and the data
$$f(w_1, \ldots, w_r | \mathbf{n}, [X_L^i, X_R^i], i = 1, \ldots, n) \ .$$

Thus, the proposed NPBE algorithm is obtained through iterative simulation from these conditional distributions. Now we proceed to obtain the form of these distributions:

(1) The posterior distribution of n given w and the data.

For each individual i, we consider the vector $\boldsymbol{\delta}^i = (\delta_1^i, \ldots, \delta_r^i)$ where
$\delta_j^i = \mathbf{1}\{X^i \in (t_{j-1}, t_j]\}$ indicates in which time interval has occurred the event of interest. $\boldsymbol{\delta}^i$ is a vector such that every component equals zero, except one. We assume that the prior distribution of $\boldsymbol{\delta}^i = (\delta_1^i, \ldots, \delta_r^i)$ conditioned to $\mathbf{w}$ is a multinomial distribution of sample size 1

$$f(\delta_1^i, \ldots, \delta_r^i | \mathbf{w}) = \prod_{j=1}^{r} w_j^{\delta_j^i} \text{ where } \sum_{j=1}^{r} \delta_j^i = 1 \ ,$$

that is the natural distribution in the nonparametric context.

Then, for each observation $i = 1, \ldots, n$, the conditional distribution of $(\delta_1^i, \ldots, \delta_r^i)$ given the data $[X_L^i, X_R^i]$, and given the vector $\mathbf{w}$, is:

$$f(\delta_1^i, \ldots, \delta_r^i | [X_L^i, X_R^i], \mathbf{w}) = f(\delta_1^i, \ldots, \delta_r^i, [X_L^i, X_R^i] | \mathbf{w}) / f([X_L^i, X_R^i] | \mathbf{w}). \qquad (6.1)$$

The numerator in (6.1) is

$$f(\delta_1^i, \ldots, \delta_r^i, [X_L^i, X_R^i] | \mathbf{w}) = \begin{cases} f(\delta_1^i, \ldots, \delta_r^i | \mathbf{w}) & \text{if } \sum_{j=1}^r \beta_j^i \cdot \delta_j^i = 1, \\ 0, & \text{otherwise} \end{cases}$$

where $\beta_j^i = \mathbf{1}\{(t_{j-1}, t_j] \subset [X_L^i, X_R^i]\}$.

The denominator in (6.1) is

$$\begin{aligned} & f([X_L^i, X_R^i] | \mathbf{w}) = P(X_L^i \leq X^i \leq X_R^i | \mathbf{w}) = \\ = \quad & \sum_{j=1}^r \beta_j^i P(X^i \in (t_{j-1}, t_j] | \mathbf{w}) = \sum_{j=1}^r \beta_j^i w_j \end{aligned}$$

Therefore, the posterior distribution of $(\delta_1^i, \ldots, \delta_r^i)$ has the following expression

$$f(\delta_1^i, \ldots, \delta_r^i | [X_L^i, X_R^i], \mathbf{w}) = \begin{cases} \dfrac{w_1^{\delta_1^i} \cdots w_r^{\delta_r^i}}{\sum_{j=1}^r \beta_j^i w_j} & \text{if } \sum_{j=1}^r \beta_j^i \cdot \delta_j^i = 1, \\ 0, & \text{otherwise} \end{cases} \qquad (6.2)$$

which corresponds to a truncated multinomial of sample size 1 and parameters $(w_1, \ldots, w_r)$.

After sampling independently for each individual $i = 1, \cdots, n$ the vector $\boldsymbol{\delta}^i = (\delta_1^i, \ldots, \delta_r^i)$ from n truncated multinomials, the vector $\mathbf{n}$ is obtained by computing the sum of them, that is, $n_j = \sum_{i=1}^n \delta_j^i$.

(2) The posterior distribution of w given n and the data.

Under the assumption of a Dirichlet process prior $\mathcal{D}(\alpha)$, the prior distribution of $\mathbf{w}$ is a Dirichlet distribution $\mathcal{D}(\alpha_1, \ldots, \alpha_r)$:

$$p(\mathbf{w}) = p(w_1, \ldots, w_r) = C \prod_{j=1}^r w_j^{\alpha_j - 1}$$

where C is the normalized constant and α_j is the mass given to the interval $(t_{j-1}, t_j]$ by the measure α. If α is expressed as $\alpha(t) = \beta \cdot S_0(t)$, where S_0 represents the prior guess of the survival function and β represents the degree of concentration of the true survival function around S_0, the parameters of the Dirichlet distribution have the form

$$\alpha_j = \beta(S_0(t_{j-1}) - S_0(t_j)) \ .$$

Note that, given the number of X's falling into each interval, the distribution of $\mathbf{w}$ is independent of the data. That is,

$$f(\mathbf{w}|n_1, \ldots, n_r, [X_L^i, X_R^i]) = f(\mathbf{w}|n_1, \ldots, n_r) = \frac{f(n_1, \ldots, n_r, \mathbf{w})}{f(n_1, \ldots, n_r)}$$

and this is proportional to the numerator:

$$\begin{array}{rl} & f(n_1, \ldots, n_r, \mathbf{w}) = f(n_1, \ldots, n_r|\mathbf{w}) \cdot p(\mathbf{w}) = \\ = & \left(\prod_{j=1}^r w_j^{n_j}\right) \cdot \left(C \cdot \prod_{j=1}^r w_j^{\alpha_j - 1}\right) = C \cdot \prod_{j=1}^r w_j^{n_j + \alpha_j - 1} \end{array}$$

with $w_r = 1 - w_1 - \ldots - w_{r-1}$. Thus, the posterior distribution of $\mathbf{w}$ given $\mathbf{n}$ and the data is a Dirichlet distribution of parameters $\alpha_j + n_j, \ j = 1, \ldots, r$.

□

6.2.2 Implementation of the NPBE algorithm

To implement the proposed algorithm it is necessary to specify the parameter measure of the Dirichlet process prior. The parameter measure α can be expressed as $\alpha(t, +\infty) = \beta \cdot \hat{S}_0(t)$, where $\hat{S}_0(t)$ is the prior guess of the survival function $S(t)$ and β is a measure of faith in the prior guess. Thus, it is necessary to choose a parametric model for $\hat{S}_0(t)$ and a constant β. This election may be done by modelling the prior knowledge of the problem or following an empirical Bayes approach. It is also necessary to specify the grid of times $t_0 \leq t_1 \leq \ldots \leq t_r$ where the survival function is going to be estimated. Our proposal to consider the partition of the real line given by the different end-points of the censoring intervals $[X_L^i, X_R^i], i = 1, \ldots, n$.

After that, the proposed NPBE algorithm proceeds as follows: We construct M independent sequences consisting on $2k$ successive simulations from the corresponding full conditional distributions (section 6.2.1).

ALGORITHM

```
• For each sequence $m = 1, \ldots, M$:

    0) Define starting values for $\mathbf{w}$: $\mathbf{w_m}^0 = (w^0_{m1}, \ldots, w^0_{mr})$

• For each iteration $l = 1, \ldots, 2k$:

    1) For each individual $i = 1, \ldots, n$

            Generate $(\delta^i_1, \ldots, \delta^i_r)$ from a truncated multinomial of sample size 1
            and parameters $(w^0_{m1}, \ldots, w^0_{mr})$.
            Compute $n_j = \sum_{i=1}^n \delta^i_j$, the number of $X$'s in each interval $(t_{j-1}, t_j]$.

    2) Generate $\mathbf{w_m}^1 = (w^1_{m1}, \ldots, w^1_{mr})$ from a Dirichlet distribution $(\alpha_1 + n_1, \ldots, \alpha_r + n_r)$.

    3) Replace $\mathbf{w}^0$ by $\mathbf{w}^1$ and return to (1).
```

As mention in chapter 5, after all the process has been performed, we discard the first half iterations of each sequence in order to diminish the effect of the starting values. Therefore, we have $k \cdot M$ vectors $\mathbf{w}^l_m = (w^l_{m1}, \ldots, w^l_{mr})$ where $l = 1, \ldots, k$ is the iteration index and $m = 1, \ldots, M$ is the sequence index. We estimate the vector of probabilities $\mathbf{w}$ by the sample mean of these $k \cdot M$ vectors:

$$\hat{w}_j = \frac{1}{kM} \sum_{l=1}^{k} \sum_{m=1}^{M} w^l_{jm} \ , \ j = 1, \ldots, r$$

and the survival function at time t_j by

$$\hat{S}(t_j) = 1 - \sum_{s \leq j} \hat{w}_s \ .$$

Steps (1) and (2) of this algorithm are detailed below:

6.2.3 Sampling from a Product of Truncated Multinomials

For each individual i, the probability of its censoring interval $[X^i_L, X^i_R]$ is
$p_i = P(X \in [X^i_L, X^i_R]) = \sum_{j=1}^r \beta^i_j w_j$, where $\beta^i_j = \mathbf{1}\{(t_{j-1}, t_j] \subset [X^i_L, X^i_R]\}$.
We consider the interval of lenght $A_i = [0, p_i]$ and divided it into a partition of intervals of length equal to each positive mass: $\beta^i_1 w_1, \beta^i_2 w_2, \ldots, \beta^i_r w_r$.

We generate a random number from a uniform $[0, p_i]$ distribution and observe to which subinterval belongs. If, the number generated belongs to the subinterval of length equal to w_k, this would mean that the event of interest has occurred in $(t_{k-1}, t_k]$, and therefore we define $\delta_k^i = 1$ and $\delta_j^i = 0, \ \forall j \neq k$.

ALGORITHM

For each individual i:

- Generate x from a Uniform$[0, p_i]$ where $p_i = \sum_{j=1}^{r} \beta_j^i w_j$:

  Generate u from a Uniform[0,1].

  Compute $x = u \cdot (\beta_1^i w_1 + \cdots + \beta_r^i w_r)$

- Locate the position of x and define the components of $\boldsymbol{\delta}^i$:

```
$tsup = \beta_1^i w_1;$
$k = 0;$
while $(x > tsup)\{$
  $k = k + 1;$
  $tsup = tsup + \beta_k^i w_k;$
$\}$
$\delta_k^i = 1;$
```

- Compute the number of events in each interval $(t_{j-1}, t_j]$:

  $n_j = \sum_{i=1}^{n} \delta_j^i, \quad j = 1, \cdots, r$

6.2.4 Sampling from a Dirichlet Distribution

There are different methods for sampling from a Dirichlet distribution. The different approaches basically fall into four categories:

1. The multivariate extension of Jhnk's Method,

2. a transformation based on the gamma distribution,

3. a transformation based on the beta distribution and

4. the acceptance-rejection method.

A simulation study [45] comparing the performance of these methods showed that: Jhnk's method is not efficient for large values of the parameters of the Dirichlet distribution, the acceptance-rejection method is not applicable to the entire permissible range of parameter values, and the efficiency, in terms of computational time, of the methods based on the transformation of the gamma and the beta distributions are very similar. We have therefore used the approach based on the gamma distribution.

Transformation Based on Gamma Variables

This method is based on the relationship between the Dirichlet and the gamma distribution:

Result 6.2.2 *If $Z_1 \sim$ Gamma $(\alpha_1, 1)$ $Z_2 \sim$ Gamma $(\alpha_2, 1), \cdots, Z_k \sim$ Gamma $(\alpha_k, 1)$ then the random vector $(Y_1, \ldots, Y_k)$ follows a Dirichlet distribution with parameters $(\alpha_1, \cdots, \alpha_k)$ where*

$$Y_1 = Z_1/(Z_1 + \cdots + Z_k),\ Y_2 = Z_2/(Z_1 + \cdots + Z_k), \cdots, Y_{k-1} = Z_{k-1}/(Z_1 + \cdots + Z_k)\ .$$

In particular, to generate a vector $\mathbf{w} = (w_1, \ldots, w_r)$ from a Dirichlet distribution of parameters $n_j + \alpha_j$, we generate r values $y_1, \ldots, y_r$ from r Gamma distributions of parameters $n_1 + \alpha_1, \ldots, n_r + \alpha_r$, respectively if $n_j + \alpha_j > 0$, and we assign $y_j = 0$ otherwise. Then, the components of the vector $\mathbf{w}$ are defined as

$$w_j = \frac{y_j}{\sum_{i=1}^r y_i}, \text{ for } \ j = 1, \ldots, r-1$$

and $w_r = 1 - w_1 - \cdots - w_{r-1}$.

6.3 Illustration

To illustrate the methodology proposed, we analyze the data that appeared in Finkelstein and Wolfe [21] corresponding to a breast cancer retrospective study. The objective of the study was to compare the long-term cosmetic effect in early breast cancer patients who were treated with primary radiation therapy and adjuvant chemotherapy to those treated with radiotherapy alone. It was known that adjuvant chemotherapy improved the overall survival but there was clinical evidence that it affected negatively the rate of deterioration of the cosmetic state. This study was carried out to verify this fact. For this analysis, the indicator of a negative overall cosmetic appearance was breast retraction since it was one of the least subjective possible measures of cosmetic deterioration. The visits for each patients were arranged every 4 to 6 months. The observed data for an individual i in this study is of the form $(L_i, R_i]$ meaning that at time L_i the patient had shown no deterioration, but in the next visit, at time R_i, breast retraction was present. Since some patients did not keep all the appointments, the data became interval-censored and the methods for grouped data could not be applied.

The data for $n = 95$ patients, $n_1 = 49$ in the radiotherapy and chemotherapy group and $n_2 = 47$ in the radiotherapy group, are presented below.

Data

Observed data of the breast cancer study

L_i	R_i	L_i	R_i	L_i	R_i	L_i	R_i
radiotherapy and chemotherapy							
48	60	8	12	0	22	24	31
17	27	17	23	24	30	16	24
13	$+\infty$	11	13	16	20	18	25
17	26	32	$+\infty$	23	$+\infty$	44	48
14	17	0	5	5	8	12	20
11	$+\infty$	33	40	31	$+\infty$	13	39
19	32	34	$+\infty$	13	$+\infty$	16	24
35	$+\infty$	15	22	11	17	22	32
10	35	30	34	13	$+\infty$	10	17
8	21	4	9	11	$+\infty$	14	19
4	8	34	$+\infty$	30	36	18	24
16	60	35	39	21	$+\infty$	11	20
48	$+\infty$						

L_i	R_i	L_i	R_i	L_i	R_i	L_i	R_i
radiotherapy alone							
46	50	45	$+\infty$	6	10	0	7
46	$+\infty$	46	$+\infty$	7	16	17	$+\infty$
7	14	37	44	0	8	4	11
15	$+\infty$	11	15	22	$+\infty$	46	$+\infty$
46	$+\infty$	25	37	46	$+\infty$	26	40
46	$+\infty$	27	34	36	44	46	$+\infty$
36	48	37	$+\infty$	40	$+\infty$	17	25
46	$+\infty$	11	18	38	$+\infty$	5	12
37	$+\infty$	0	5	18	$+\infty$	24	$+\infty$
36	$+\infty$	5	11	19	35	17	25
24	$+\infty$	32	$+\infty$	33	$+\infty$	19	26
37	$+\infty$	34	$+\infty$	36	$+\infty$		

We start analyzing the data corresponding to those patients who were treated with adjuvant chemotherapy. For this group of patients we consider four different estimators of the survival function for the time to breast retraction. The four estimators we compare are:

1. the maximum likelihood estimator assuming an exponential survival curve,
2. the nonparametric estimator obtained with Turnbull's algorithm,
3. the nonparametric Bayes estimator with a prior survival $\hat{S}_0(t)$ exponentially distributed and with $\beta = n$ =number of patients=49, and
4. the nonparametric Bayes estimator with a prior survival $\hat{S}_0(t)$ exponentially distributed and with $\beta = \sqrt{n} = 7$.

The Bayes estimators were obtained through the implementation of the Gibbs sampling scheme described in section 6.2 taking $M = 5$ independent sequences and $i = 2000$ iterations in each sequence. These estimators were computed in a grid $0 = t_0 < t_1 < \ldots < t_r = +\infty$ corresponding to the partition of the real line induced by the different left and right end-points, L_i and R_i, for $i = 1, \ldots, n$, of the censoring intervals.

The goal of this first analysis is mainly to illustrate the behaviour of the different approaches, parametric, nonparametric and Bayesian. For that reason we have considered an exponential survival function *'a priori'* because, though it is clear that this parametric

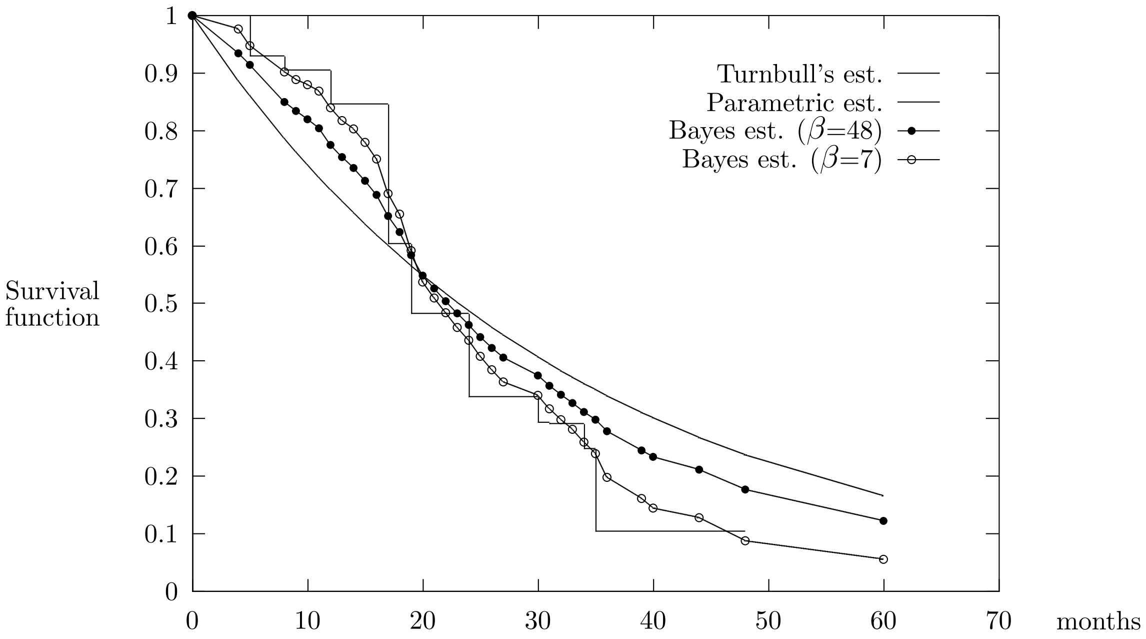


Figure 6.1: Four estimators of the survival time to cosmetic deterioration for breast cancer patients treated with chemotherapy.

model does not fit suitably the data, it allows us to emphasize the differences between the three methodologies.

The four estimators of the survival function are plotted in figure 6.1. In this figure we can see how the Bayes estimators lie between the parametric estimator and the nonparametric Turnbull's estimator. Indeed, the Bayes estimator can be interpreted as the result of 'shrinking' the nonparametric estimator towards the parametric family assumed *'a priori'*. For that reason, as β increases, the resulting Bayes estimator is closer to the parametric model and, conversely, it approaches the nonparametric estimator as we diminish the parameter β, that is, as we diminish the faith on the prior guess.

For a more realistic analysis of the data we have considered *'a priori'* a more flexible parametric model, the Weibull model, and a parameter $\beta = \sqrt{n}$ to ensure a consistent estimator. The parameters of the Weibull distribution have been obtained through maximum likelihood based on the censored data. The results for the chemotherapy group are shown in figure 6.2 while, in figure 6.3 are the resulting estimators for the radiotherapy group. The Bayes estimator identifies smaller intervals of time while the Turnbull's estimator leaves large intervals of time where the form of the survival function is completely unknown. For instance, in the radiotherapy group, in figure 6.3, the survival function

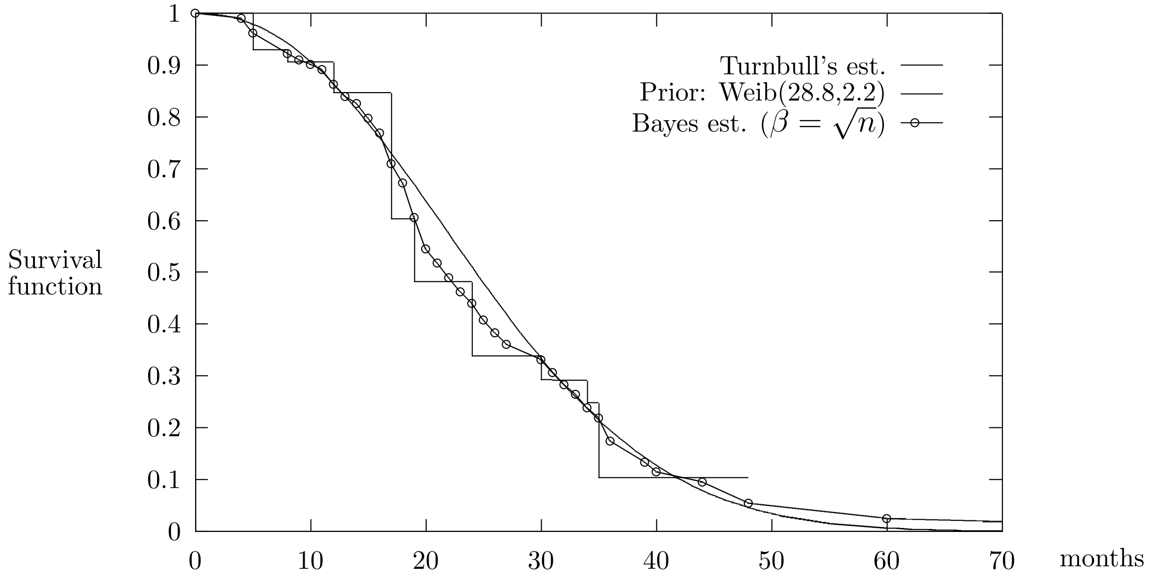


Figure 6.2: Time to cosmetic deterioration for breast cancer patients. Chemotherapy group.

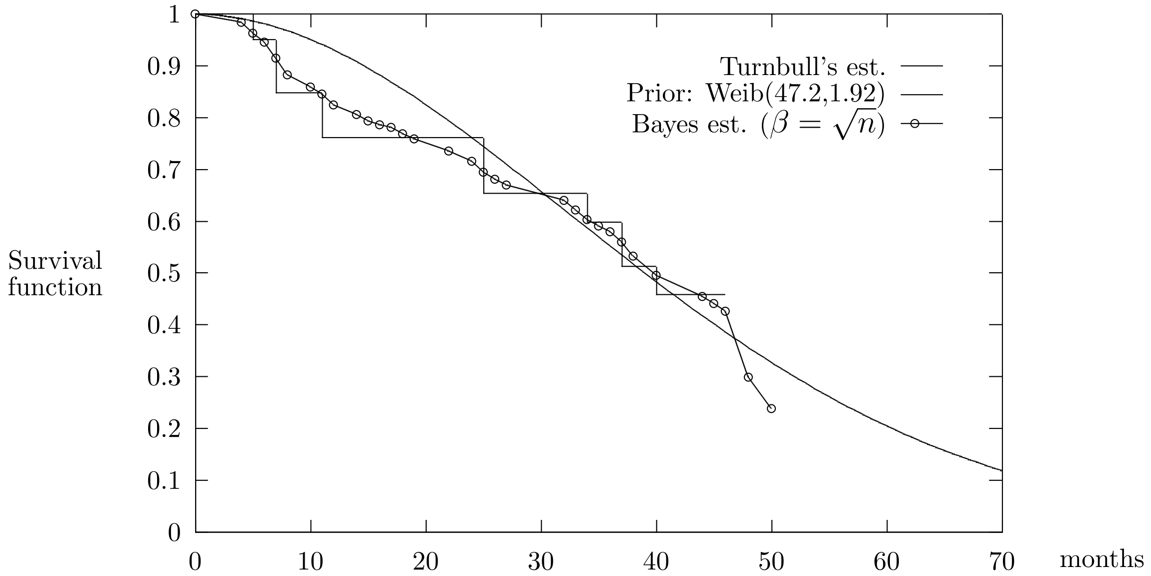


Figure 6.3: Time to cosmetic deterioration for breast cancer patients. Radiotherapy group.

from month 11 to month 25, more than one year, is completely unidentified.

Another practical advantage of the nonparametric Bayes estimators is shown in figure 6.4 where the two treatment groups are compared. The difference between these two groups becomes more evident when smoother curves are drawn. However, both, nonparametric and Bayes approaches yield to similar conclusions, that is, the chemotherapy in addition to previous radiotherapy increases the hazard of breast retraction.

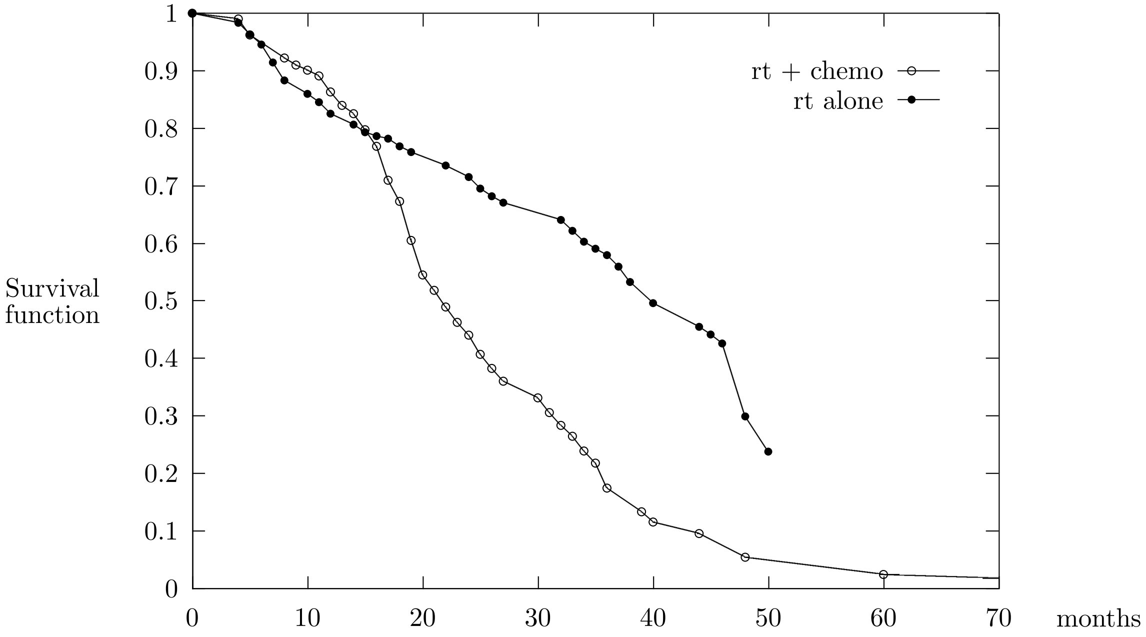


Figure 6.4: Time to cosmetic deterioration for breast cancer patients by treatment group

6.3.1 Convergence diagnostic

Convergence of the Gibbs sampler has been stablished both graphically and numerically using the program CODA [3] *"Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output"*.

We present the results of the analysis of convergence for the chemotherapy group. Similar results were obtained for the other group of patients.

We have studied the convergence of five components of the 36-dimensional vector $(w_1, \ldots, w_r)$, where $w_j = S(t_{j-1}) - S(t_j)$, in such a way that with this analysis the tails of the distribution as well as the center are covered. The components considered are those corresponding to the intervals of time $I_1 = (5, 8],\ I_2 = (15, 16], I_3 = (21, 22],\ I_4 = (33, 34]$ and $I_5 = (48, 60]$, because these intervals contain the 5th, 25th, 50th, 75th and 95th quantiles of the distribution, respectively. For these components, we have used the diagnostic methods proposed by Gelman and Rubin (see section 5):

1. Plots of the sample trace for each variable.
2. Plots of Gelman and Rubin's factor $\hat{R}$.

These diagnostic methods are efficient when the initial values are overdispersed with respect to the target distribution. For that reason we have considered the following 5 initial situations:

1. $w(I_1) = 0.6, w(I_2) = 0.0001$ and $w(I_3) = 0.1$
2. $w(I_2) = 0.6, w(I_3) = 0.0001$ and $w(I_4) = 0.1$
3. $w(I_3) = 0.6, w(I_4) = 0.0001$ and $w(I_5) = 0.1$
4. $w(I_4) = 0.6, w(I_5) = 0.0001$ and $w(I_1) = 0.1$
5. $w(I_5) = 0.6, w(I_1) = 0.0001$ and $w(I_2) = 0.1$

and the rest of the mass equally distributed between the other 33 components. For each component of interest $w(I_k),\ k = 1, \ldots, 5$, we have plotted in the left column of figure 6.5 the traces of its first 100 values of the 5 independent sequences obtained iteratively with the Gibbs sampler. In the right column of the same figure there are the corresponding traces but for the last 100 iterations. It seems clear from these pictures that convergence is achieved almost inmediately and the behaviour of the first 100 iterations is identical to the behaviour of the last 100, both, in central tendency and variability.

Convergence is confirmed numerically with the computation of the Gelman and Rubin's shrink factor, that compares an overestimate of the variance, $\hat{\mathrm{var}}(x)$, with the within-sequence variance, W:

$$\hat{R} = \frac{\hat{\mathrm{var}}(x)}{W}.$$

In figure 6.6 there are the plots of the shrink factors computed at every iteration for each component in study. These plots have been produced by splitting the chain for each variable into a number of segments, in this case, each chain has been divided into 50 segments as follows: the first segment contains the first 20 iterations, the second segment contains the first 40 iterations, . . ., the k^{th} segment contains the first $k \cdot 20$ iterations. The median and 97.5% quantile of the shrink factors in each segment are plotted against the maximum iteration number for the segment. The plots in figure 6.6 show that for every component, the shrink factor stabilizes around 1 very quickly, what again indicates that convergence has been achieved.

Figure 6.5: Plots of the sample trace for each variable

Figure 6.6: Plots of Gelman and Rubin's factor

Chapter 7

The Nonparametric Perspective versus the Bayesian Approach. A Simulation Study

7.1 Introduction

The goal of this chapter is the comparison of the estimators of the survival function when data are interval-censored under the following three methodologies:

1. The nonparametric Bayesian estimation,

2. the nonparametric Turnbull's estimation and

3. the parametric maximum likelihood estimation.

The first approach has been developed in chapter 6, the second approach has been developed in section 2.5 of chapter 2 and the third approach corresponds to the standard maximum likelihood methodology. The first qualitative advantage of the first approach over the classical nonparametric methodology is the possibility of incorporate prior knowledge about the problem in study. Furthermore, the Bayesian approach builds an smoother estimator that facilitates the analysis and further interpretation of the results. However, the derivation of finite and large sample properties for the proposed nonparametric Bayes estimator is difficult. For this reason we have carried out a simulation study to compare the overall performance of the three estimators.

7.2 Design and implementation of the Monte Carlo study

Two series of Monte Carlo experiments have been conducted. 500 samples of size 30, 50 or 100 from one of five parametric models have been simulated. The two series of simulations correspond to two different mechanism to construct the censoring intervals.

7.2.1 Random variables

The data were simulated from the following distributions:

a) **Exp(10):** An exponential distribution with mean equal to 10 and with density function and cumulative distribution function given by

$$f(x) = \frac{1}{10}e^{-x/10},\ x > 0 \quad \text{and} \quad F(x) = 1 - e^{-x/10}\ .$$

b) **Weib(10,2):** A Weibull distribution with scale and shape parameter equal to 10 and to 2, respectively, and with density function and cumulative distribution function given by

$$f(x) = \frac{2x}{10^2}e^{-(x/10)^2},\ x > 0 \quad \text{and} \quad F(x) = 1 - e^{-(x/10)^2}\ .$$

c) **Gam(5,2):** A Gamma distribution with shape and scale parameter equal to 5 and to 2, respectively, and with density function given by

$$f(x) = \frac{1}{\Gamma(5)2^5}x^4e^{-x/2},\ x > 0\ .$$

The cumulative distribution function $F(x)$ does not admit a closed form expression.

d) **Weib(10,8):** A Weibull distribution with scale and shape parameter equal to 10 and to 8, respectively, and with density function and cumulative distribution function given by

$$f(x) = \frac{8x}{10^8}e^{-(x/10)^8}\ x > 0 \quad \text{and} \quad F(x) = 1 - e^{-(x/10)^8}\ .$$

e) **Gam(20,0.5):** A Gamma distribution with shape and scale parameter equal to 20 and to 1/2, respectively, and with density function given by

$$f(x) = \frac{2^{20}}{\Gamma(20)}x^{19}e^{-2x},\ x > 0\ .$$

The cumulative distribution function $F(x)$ does not admit a closed form expression.

The choice of the Weibull and Gamma models is by no means exhaustive but it represents the most common models in the context of survival analysis studies. The parameters were chosen to obtain a mean value near 10 and a range between 0 and 20, aproximately.

In our simulation study we have always considered the exponential distribution for both the prior distribution in the Bayesian approach and the parametric model in the parametric estimation. Thus, in every situation we deal with three distributions, namely, the theoretical distribution, the prior distribution and the so-called parametric distribution.

Case (a) corresponds to the situation where the theoretical distribution, the prior distribution and the parametric distribution are all exponential. On the other hand, an opossite situation is considered in cases (d) and (e) where the choice of the parameters of the Weibull and Gamma distributions implies cumulative distributions that are very far from exponentiality. Cases (b) and (c) are taken as an intermediate step between the other situations. In these two last cases, the corresponding cumulative distribution functions are not very far from the exponential curve.

7.2.2 Sampling mechanisms

The random samples from the exponential and Weibull distributions were generated from the Inverse Probability Method. This method cannot be used to generate the Gamma distribution because there is no closed expression for its distribution function. For this reason, the random samples have been generated as follows (Fishman [23]):

$$X = -\beta \log \left(\prod_{j=1}^{\alpha} U_j \right) ,$$

where $U_j,\ j = 1, \ldots, \alpha$ are uniform [0,1] variates.
(This algorithm provides samples from a Gamma distribution with integer shape parameter).

7.2.3 Generation of the censoring intervals

To construct the censoring intervals $[X_L^i, X_R^i]$ we have proceed as in 3.2. After generating X^i, $i = 1, \ldots, n$ from one of the models considered for X, the random intervals $[X_L^i, X_R^i]$ are constructed containing X^i via a mechanism that mimics those longitudinal studies where there is periodical follow–up. The intervals arise from regularly scheduled visits and patients might miss some of the appointments.

To study the effect of censoring, we have consider two different situations. In the first case the censoring intervals have a mean lenght of approximately 2.5 units. The second case corresponds to a heavy censoring situation with censoring intervals with a mean length of approximately 7 units.

First censoring mechanism

30% of the patients attend all the visits, having each of them a censoring interval of length 1 unit (1 month, 6 months, 1 year, ...), 30% of the patients miss 1 visit in the interval of interest and have therefore a censoring interval of length 2 units, 20% miss 2 visits and have censoring intervals of length 3 units, 10% of the patients miss 3 visits and their intervals are of length 4 units and the remainder, 10%, have an interval of length 5 units.

Second censoring mechanism

35% of the patients miss 4 visits, having each of them a censoring interval of length 5 units, 35% of the patients miss 7 visits in the interval of interest and have therefore a censoring interval of length 8 units and the remainder, 30%, have an interval of length 10 units.

7.2.4 Performance of the estimators

To compare the overall performance of the estimators we have computed the L_2 distance between the estimator $\hat{F}(x)$ and the correct distribution of X, $F(x)$, for each run. We have approximate the L_2 distance by:

$$\left(\int_0^{p_{0.95}} (\hat{F}(x) - F(x))^2 \, dx \right)^{(1/2)}$$

where the integral has been computed numerically. The mean and standard deviation of these distances for the 500 runs are reported in tables 7.1 to 7.5.

Computation of the Gamma distribution function

The theoretical distribution function of the Gamma variable X with and integer shape parameter α and scale parameter β has been approximated by the distribution of a Poisson process with mean $1/\beta$. Then, the distribution function of X is given by

$$F(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{(x/\beta)^k e^{-x/\beta}}{k} \ .$$

Computation of the estimators of the distribution function

The estimator $\hat{F}(x)$ was computed from one of the following methods:

1. The nonparametric Bayes estimator has been computed using the Gibbs sampler algorithm implemented in the C-language program called **GIBBSIC.C**. As explained in chapter 6, the estimator is computed for a finite number of times $0 = t_0 < t_1 < \ldots < t_r = +\infty$. For values of $t \in (t_{j-1}, t_j)$ the estimator has been computed by exponential interpolation.
   To obtain a simple estimate of the parameter of the prior exponential distribution we have first transformed the data into right-censored by taking the middle point of the finite censoring intervals as the exact of observation. Then the parameter of the exponential is obtained by the maximum likelihood estimator of these right-censored data:
$$\hat{\theta} = \frac{r}{\sum_{i=1}^{n} d^i (L_i + R_i)/2 + \sum_{i=1}^{n} (1 - d^i) L_i},$$
$$\text{where} \quad d^i = \mathbf{1}\{R_i < \infty\} \quad \text{and } r = \sum_{i=1}^{n} d^i \ .$$

2. The nonparametric Turnbull's estimator has been computed using the C-program language **ICTURNB.C**. In those intervals were the estimator is not defined we have performed a linear interpolation.

3. The maximum likelihood estimator of the parameter of the exponential distribution has been computed with the E-M algorithm [17] as follows:

E-M algorithm for computing the maximum likelihood estimator of an exponential distribution based on interval censored data.

We assume that X_i is exponentially distributed with mean equal to $1/\theta$, that is, its density function is $f(x) = \theta e^{-\theta x}$ and its survival function is $S(x) = e^{-\theta x}$. We assume also that X_i is interval-censored and, therefore, the observed data are of the form (L_i, R_i), meaning that $L_i \leq X_i \leq R_i$. The log likelihood function based on an uncensored exponential sample $X = (X_1, \ldots, X_n)$ is

$$l_0(\theta, X) = \log \prod_{i=1}^{n} \theta e^{-\theta X_i} = n \log \theta - \theta \sum_{i=1}^{n} X_i \ .$$

The two steps of the E-M algorithm for obtaining the maximum likelihood estimator of θ are:

Expectation step. Given the current estimate θ_k of θ, calculate the conditional expectation of the complete data log-likelihood given the observed data:

$$Q(\theta, \theta_k) = E[l_0(\theta; X)|X_{obs}, \theta_k]$$

In the case of the exponential distribution and interval-censored data this conditional expectation becomes

$$\begin{aligned} Q(\theta, \theta_k) &= E[l_0(\theta; X)|L_i \leq X_i \leq R_i, \theta_k] = \\ &= n \log \theta - \theta \sum_{i=1}^{n} E[X_i|L_i \leq X_i \leq R_i, \theta_k] = \\ &= n \log \theta - \theta \sum_{i=1}^{n} \frac{(L_i + 1/\theta_k)e^{-\theta_k L_i} - (R_i + 1/\theta_k)e^{-\theta_k R_i}}{e^{-\theta_k L_i} - e^{-\theta_k R_i}} \end{aligned}$$

where the last expression is obtained by computing the expectation of a truncated exponential in the interval $[L_i, R_i]$.

Maximization step. Determine a new estimate θ_{k+1} as the value of θ that maximizes $Q(\theta, \theta_k)$.

The maximization step is done by deriving $Q(\theta, \theta_k)$ with respect to θ and equating to zero. This gives the following iterative procedure for obtaining the MLE of θ:

$$\theta_{k+1} = n / \sum_{i=1}^{n} \frac{(L_i + 1/\theta_k)e^{-\theta_k L_i} - (R_i + 1/\theta_k)e^{-\theta_k R_i}}{e^{-\theta_k L_i} - e^{-\theta_k R_i}}$$

7.3 Results and discussion

The results of the simulation study are displayed in tables 7.1 to 7.5. Each entry in these tables corresponds to the mean of the L_2 distance between the estimator and the survival function of the 500 iterations. The standard error is given in parenthesis.

The analysis of the results may be divided into two cases:

a) The case in which the theoretical distribution is exponential and, therefore, the parametric distribution and the prior guess of the Bayes estimator coincide with the true distribution.

In this case, reported in table 7.1, we can see that the minimum distance L_2 is achieved by the parametric estimator. This result was to be expected because the parametric assumption was true. It is not surprising either that the Bayes estimator performs better than the nonparametric Turnbull's estimator, because the Bayes estimator 'shrinks' the nonparametric estimator towards the true survival curve. This relative advantage of the Bayes estimator over Turnbull's one increases considerably with the second censoring mechanism, that is, when the censoring intervals have a larger lenght. In this case, the lost of efficiency of the Turnbull's estimator is significative, while the nonparametric Bayes estimator reduces this lost of efficiency.

b) The second set of simulations corresponds to the case in which the true distribution is not exponential and, therefore, the distribution of the parametric estimator and the prior guess of the Bayes estimator differ from the theoretical distribution. (This situation is reported in tables 7.2 to 7.5).

   In this case, the Bayes estimator performs always better than both the maximum likelihood estimator assuming exponentiality and the nonparametric Turnbull's estimator. Indeed, the parametric estimator performs now clearly worse than the others because it only uses the observed data to estimatethe mean of the distribution.

   With the first censoring mechanism, that is, when the censoring intervals are short, the advantage of the Bayes estimator over the Turnbull's estimator is more important in those cases where the prior distribution is not far from the true survival curve, as it happens when the true distribution is Weibull(10,2) (Table 7.2) or Gamma(5,2) (Table 7.3). In the other cases, Table 7.4 and Table 7.5, where there is an important discrepancy between the prior distribution and the true distributions, both estimators perform similarly, with a small advantage of the Bayes estimator that is very likely to arise from the fact that the Bayes estimator is smoother than Turnbull's estimator.

   When we consider larger censoring intervals, as in the second censoring mechanism, the advantage of the Bayes estimator increases significatively, specially with small sample sizes (30 or 50). For a larger sample size, $n = 100$, the performance of the Bayes estimator is still better than that of the Turnbull's estimator when the prior survival curve is relatively close to the true survival curve (the Weibull(10,2) distribution in Table 7.2 and the Gamma(5,2) distribution in Table 7.3).

With an illustrative purpose we present graphics 7.1, 7.2 and 7.3 corresponding to the estimated distribution function following the maximum likelihood, Turnbull's and nonparametric Bayes method for the cases a, b and d (section 7.2.1). These three examples have been based on the results of an arbitrary simulated run. It is inmediate to apreciate that the nonparametric Bayes estimator always provides smoother and more treatable curves, even when the prior is far from the theoretical distribution.

Based on these simulations, we believe that the gain of using the proposed nonparametric Bayes estimator instead of the nonparametric Turnbull's estimator when data are interval-censored is very important. This is specially true when the lenght of the censoring intervals is large, as it will be the case in most of the real examples involving interval-censored data.

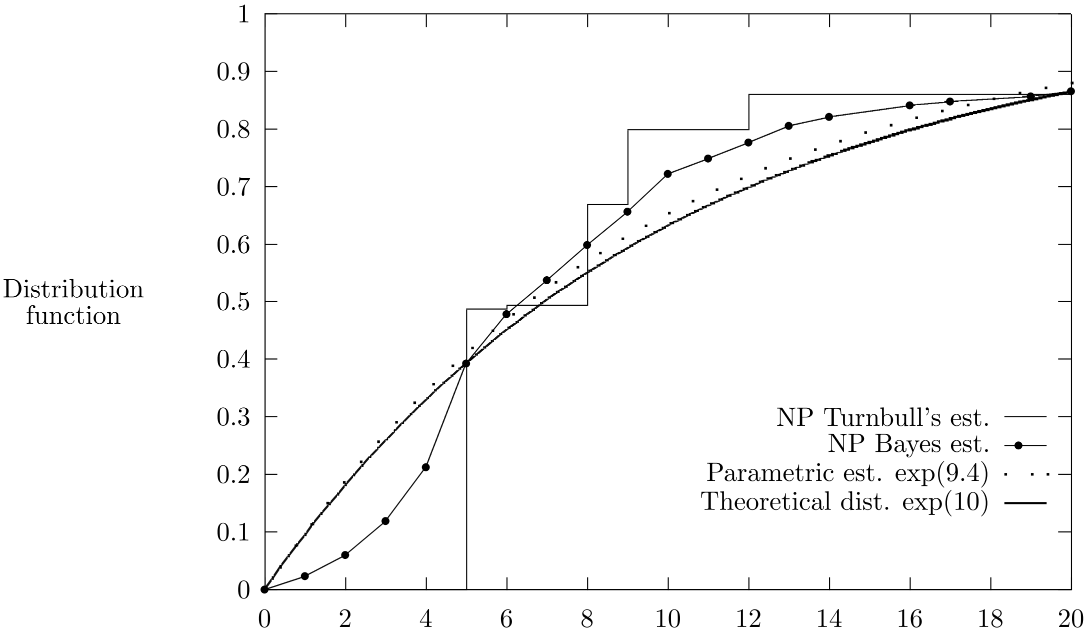


Figure 7.1: Three estimators of an exponential distribution function of mean 10

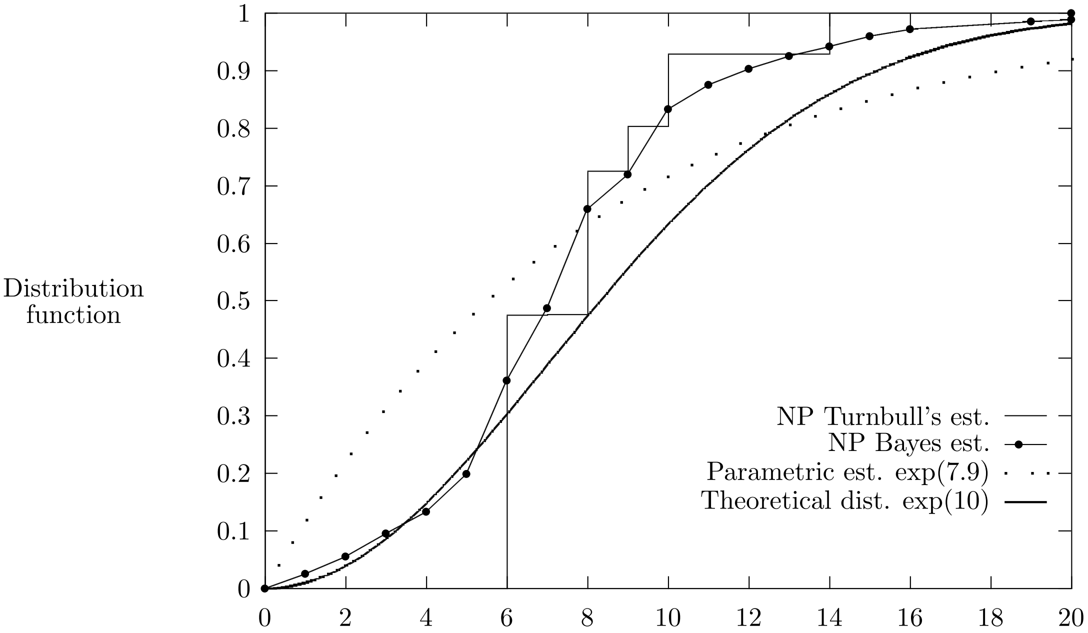


Figure 7.2: Three estimators of a Weibull distribution function of parameters 10 and 2

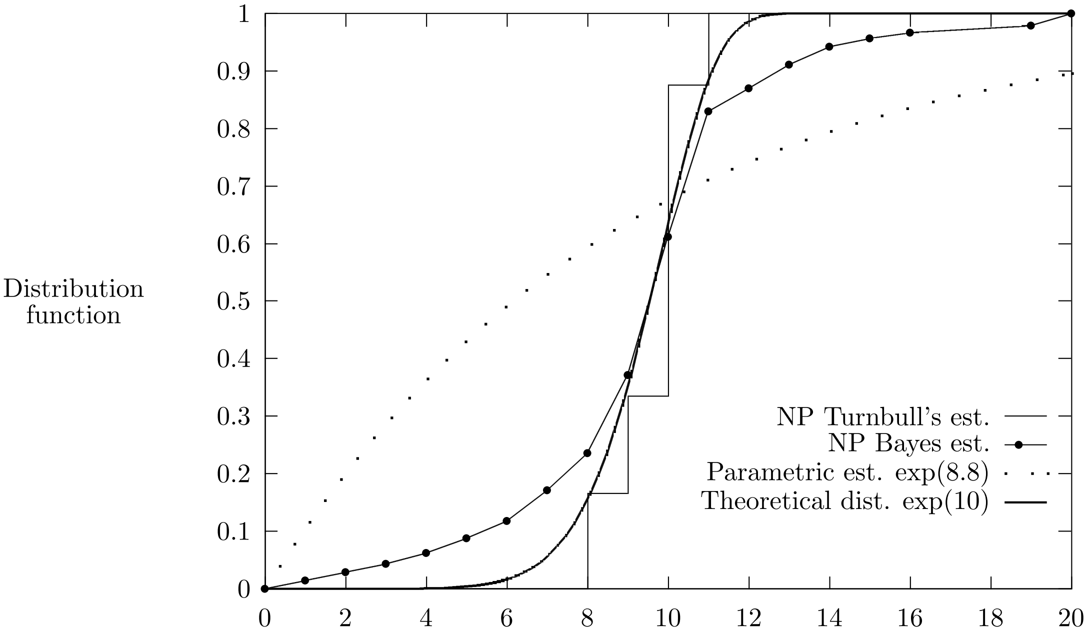


Figure 7.3: Three estimators of a Weibull distribution function of parameters 10 and 3

Table 7.1

Mean L_2 distance between the following three estimators and an exponential survival function of mean 10

1) Nonparametric Bayes estimator with exponential prior.
2) Nonparametric Turnbull's estimator.
3) Parametric maximum likelihood estimator assuming exponentiality.

(μ is the mean lenght of the censoring intervals)

Exp(10)		sample size		
		$n = 30$	$n = 50$	$n = 100$
1rst. censoring	Bayes est.	0.3268(0.13)	0.2827(0.11)	0.2250(0.09)
mechanism	Turnbull's est.	0.4432(0.11)	0.3690(0.10)	0.2900(0.08)
$\mu \sim 2.5$	Parametric est.	0.2586(0.18)	0.2010(0.14)	0.1517(0.11)
2nd. censoring	Bayes est.	0.3409(0.20)	0.3346(0.05)	0.3141(0.02)
mechanism	Turnbull's est.	0.6817(0.10)	0.5981(0.04)	0.5232(0.04)
$\mu \sim 7$	Parametric est.	0.1920(0.14)	0.1527(0.11)	0.1189(0.08)

Table 7.2

Mean L_2 distance between the following three estimators and a Weibull survival function of parameters 10 and 2

1) Nonparametric Bayes estimator with exponential prior.
2) Nonparametric Turnbull's estimator.
3) Parametric maximum likelihood estimator assuming exponentiality.

(μ is the mean lenght of the censoring intervals)

Weib(10,2)		sample size		
		$n = 30$	$n = 50$	$n = 100$
1rst. censoring mechanism $\mu \sim 2.5$	Bayes est.	0.2256(0.11)	0.1990(0.06)	0.1649(0.09)
	Turnbull's est.	0.3764(0.10)	0.3180(0.11)	0.2695(0.09)
	Parametric est.	0.6074(0.06)	0.6147(0.05)	0.6045(0.04)
2nd. censoring mechanism $\mu \sim 7$	Bayes est.	0.2925(0.12)	0.2713(0.12)	0.2390(0.10)
	Turnbull's est.	0.5117(0.11)	0.4510(0.06)	0.3585(0.08)
	Parametric est.	0.5928(0.07)	0.5974(0.05)	0.5816(0.04)

Table 7.3

Mean L_2 distance between the following three estimators and a Gamma survival function of parameters 5 and 2

1) Nonparametric Bayes estimator with exponential prior.
2) Nonparametric Turnbull's estimator.
3) Parametric maximum likelihood estimator assuming exponentiality.

(μ is the mean lenght of the censoring intervals)

Gam(5,2)		sample size		
		$n = 30$	$n = 50$	$n = 100$
1rst. censoring	Bayes est.	0.3230(0.13)	0.2853(0.11)	0.2521(0.08)
mechanism	Turnbull's est.	0.3768(0.11)	0.3189(0.10)	0.2692(0.07)
$\mu \sim 2.5$	Parametric est.	0.7618(0.06)	0.7564(0.05)	0.7506(0.04)
2nd. censoring	Bayes est.	0.3496(0.10)	0.2371(0.05)	0.2222(0.05)
mechanism	Turnbull's est.	0.6646(0.13)	0.5370(0.10)	0.3411(0.05)
$\mu \sim 7$	Parametric est.	0.7662(0.07)	0.7579(0.05)	0.7467(0.04)

Table 7.4

Mean L_2 distance between the following three estimators and a Weibull survival function of parameters 10 and 8

1) Nonparametric Bayes estimator with exponential prior.
2) Nonparametric Turnbull's estimator.
3) Parametric maximum likelihood estimator assuming exponentiality.

(μ is the mean lenght of the censoring intervals)

Weib(10,8)		sample size		
		$n = 30$	$n = 50$	$n = 100$
1rst. censoring mechanism $\mu \sim 2.5$	Bayes est.	0.4092(0.03)	0.3333(0.06)	0.3038(0.04)
	Turnbull's est.	0.4806(0.05)	0.3402(0.07)	0.3200(0.05)
	Parametric est.	1.1450(0.01)	1.1280(0.01)	1.1270(0.00)
2nd. censoring mechanism $\mu \sim 7$	Bayes est.	0.5479(0.11)	0.5198(0.05)	0.3046(0.08)
	Turnbull's est.	0.5970(0.13)	0.5506(0.06)	0.3222(0.07)
	Parametric est.	1.1376(0.04)	1.1498(0.03)	1.1482(0.02)

Table 7.5

Mean L_2 distance between the following three estimators and a Gamma survival function of parameters 20 and 0.5

1) Nonparametric Bayes estimator with exponential prior.
2) Nonparametric Turnbull's estimator.
3) Parametric maximum likelihood estimator assuming exponentiality.

(μ is the mean lenght of the censoring intervals)

Gam(20,5)		sample size		
		$n = 30$	$n = 50$	$n = 100$
1rst. censoring	Bayes est.	0.3530(0.10)	0.3180(0.08)	0.2839(0.01)
mechanism	Turnbull's est.	0.3626(0.10)	0.3198(0.08)	0.2935(0.01)
$\mu \sim 2.5$	Parametric est.	1.0675(0.03)	1.0671(0.02)	1.0658(0.02)
2nd. censoring	Bayes est.	0.5693(0.14)	0.4127(0.05)	0.2848(0.07)
mechanism	Turnbull's est.	0.6386(0.17)	0.5069(0.06)	0.3104(0.07)
$\mu \sim 7$	Parametric est.	1.0841(0.05)	1.0824(0.04)	1.0793(0.03)

Chapter 8

Discussion and Future Areas of Research

In this work we have approached two alternative nonparametric methodologies for dealing with interval censoring. The first is the classical nonparametric maximum likelihood analysis and the second is the nonparametric Bayesian methodology. The first methodology has been extended to deal with the case of double censoring.

The main feature of the nonparametric maximum likelihood methodology, that make it so appealing, is its robustness, due to the weak set of assumptions required for its validity. This approach is appropriate when either very little is known about the underlying distribution, or alternatively, when the problem is extremely complex to be modelled. Apart from that, this analysis is interesting because it provides the maximum likelihood estimator of the distribution, and therefore maximum likelihood theory can be used to derive its large sample properties. Therefore, for large samples and when the censoring intervals are not very wide, this approach provides an excellent approximation of the theoretical distribution.

However, with small samples and under heavy censoring, the classical nonparametric methods are known to be very inefficient since all the estimation about the form of the underlying distribution is built from the poor information provided by the data. In this case, the alternative nonparametric Bayesian methodology represents an important gain in efficiency since it allows the inclusion in the analysis of prior knowledge of the problem. This prior knowledge can arise from medical, biological or any other experimental study. This intuitive improvement of the Bayes approach has been established empirically by a simulation study.

The first inmediate future research is the development of the nonparametric Bayesian methodology for a doubly-censored pattern. We believe that these new results will be quite straightforward from the current work.

So far we have only approached univariate problems without considering other variables. However, these methodologies would be of more aplicability if they were extended to deal with multivariate problems. Indeed, the objective of most of the survival studies is to obtain predictors for the survival time based on a set of covariables. So the next step in our research will be towards the comparison of two survival curves and the derivation, if possible, of a quantitative way of assessing their difference. Furthermore, the possibility of fitting a Cox's proportional hazard model when data are interval-censored is a very attractive goal. The extension of Kalbfleisch's work [39] to the interval-censored situation together with the extension of Finkelstein's nonparametric proposal [22] and Frydman's semiparametric estimation [26] is another area of future research.

Bibliography

[1] Andersen, P.K. and Gill, R.D. (1982) *Cox's regression model for counting processes: A large-sample study*, Annals of Statistics, **10**, 1100-1120.

[2] Bacchetti, P. (1990) *Estimating the Incubation period of AIDS by Comparing Population Infection and Diagnosis Patterns*, Journal of the American Statistical Association, **85**, 1002-1008.

[3] Best, N.G., Cowles, M.K. and Vines, S.K. (1995) *CODA Manual version 0.30*, MRC Biostatistics unit, Cambridge, UK.

[4] Brookmeyer, R. and Goedert, J.J. (1989) *Censoring in an Epidemic with an Application to Hemophilia-Associated AIDS*, **45**, 325-335.

[5] Burridge, J. (1981) *Empirical Bayes Analysis of Survival Time Data*, Journal of the Royal Statistical Society, Series B **43**, 65-75.

[6] Calle, M.L. and Gmez, G. (1994) *A Comparison Study for Interval-Right-Censored Survival Data*, Technical Report No. DR 94/04, Dept. d'Estadstica i Investigaci Operativa. Universitat Politcnica de Catalunya.

[7] Casella, G. and George, E. I. (1992) *Explaining the Gibbs Sampler*, The American Statistician, **46**, 167-174.

[8] Chang, M. N. and Yang, G. L. (1987) *Strong consistency of a nonparametric estimator of the survival function with doubly censored data*, Annals of Statistics, **16**, 1536-1547.

[9] Chiarotti, F. and *alt.* (1992) *Effects of Different Parametric Estimates of Seroconversion time on Analysis of Progression to AIDS among Italian HIV-Positive Haemophiliacs* , Statistics in Medicine, **11**, 591-601.

[10] Chiarotti, F. and *alt.* (1994) *Median Time From Seroconversion to AIDS in Italian HIV-Positive Haemophiliacs: Different Parametric Estimates*, Statistics in Medicine, **13**, 163-175.

[11] Chib, S. and Greenberg, E. (1995) *Understanding the Metropolis-Hastings Algorithm*, The American Statistician, **49**, 327-335.

[12] Cowles, M.K. and Carlin, B.P. (1995) *Markov chain Monte Carlo Convergence Diagnostics: A Comparative Review*, Journal of the American Statistical Association, (to appear).

[13] Cox, D.R. (1972) *Regression models and life tables*, Journal of the Royal Statistical Society, Series B **34**, 187-220.

[14] Csörgo, S. and Horvth, L. (1980) *Random censorship from the left*, Studia Sc. Math. Hung., **15**, 397-491.

[15] Darby, S.C., Doll, R. and Thakrar, B. (1990) *Time From Infection With HIV to Onset of AIDS in Patients with Haemophilia in the UK*, Statistics in Medicine, **9**, 681-689.

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

[17] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) *Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)*, Journal of the Royal Statistical Society, Series B **39**, 1-38.

[18] Doss, H. (1994) *Bayesian Nonparametric Estimation For Incomplete Data via Successive Substitution Sampling*, The Annals of Statistics, **22**, 1763-1786.

[19] Ferguson, T.S. (1973) *A Bayesian Analysis of Some Nonparametric Problems*, The Annals of Statistics, **1**, 209-230.

[20] Ferguson, T.S. and Phadia, E. G. (1979) *Bayesian Nonparametric Estimation Based on Censored Data*, The Annals of Statistics, **7**, 163-186.

[21] Finkelstein, D.M. and Wolfe, R.A. (1985) *A Semiparametric model for Regression Analysis of Interval-Censored Failure Time Data*, Biometrics, **41**, 933-945.

[22] Finkelstein, D.M. (1986) *A Proportional Hazards Model for Interval-Censored Failure Time Data*, Biometrics, **42**, 845-854.

[23] Fishman, G.S. (199?) *Concepts and Methods in Discrete Event Digital Simulation* Wiley-Interscience Publication.

[24] Fleming, T.R. and Harrington, D.P. (1991) *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley-New York.

[25] Frydman, H. (1992) *A Nonparametric Estimation Procedure for a Periodically Observed Three-state Markov Process, with Application to Aids*, Journal of the Royal Statistical Society, Series B **54**, 853-866.

[26] Frydman, H. (1995) *Semiparametric Estimation in a Three-state Duration-Dependent Markov Model From Interval-Censored Observations with Application to AIDS Data*, Biometrics, **51**, 502-511.

[27] Gelfand, A.E. and Smith, F.M. (1990) *Sampling-Based Approaches to Calculating Marginal Densities*, Journal of the American Statistical Association, **85**, 398-409.

[28] Gelfand, A.E., Smith, F.M. and Tai-Ming L. (1992) *Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling*, Journal of the American Statistical Association, **87**, 523-532.

[29] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., (1995), *Bayesian Data Analysis*. Chapman and Hall.

[30] Gelman, A. and Rubin, D.B. (1992) *Inference from Iterative Simulation Using Multiple Sequences*, Statistical Science, **7**, 457-472.

[31] Geman, S. and Geman, D. (1984) *Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **6**, 721-741.

[32] Geyer, C.J. (1992) *Practical Markov Chain Monte Carlo*, Statistical Science, **7**, 473-511.

[33] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

[34] Gmez, G., Juli, O. and Utzet, F. (1992) *Survival Analysis for Left Censored Data*, Survival Analysis: State of the Art (Editors: J.P. Klein and P. Goel), 269-288. Kluwer Academic Publishers. Dordrecht. Holanda.

[35] Gmez, G., Juli, O. and Utzet, F. (1994) *Asymptotic Properties of the Left Kaplan-Meier Estimator*, Communications in Statistics: Theory and Methods, **23**, 123-135.

[36] Gómez, G. and Lagakos, S. W. (1994) *Estimation of the Infection Time and Latency Distribution of AIDS with Doubly Censored Data*, Biometrics, **50**, 204-212.

[37] Groeneboom, P. and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar, Band 19, Birkhauser, New York.

[38] Hastings, W. K. (1970), *Monte Carlo Sampling Methods Using markov Chains and Their Applications*, Biometrika, **57**, 97-109.

[39] Kalbfleisch, J.D. (1978) *Non-parametric Bayesian Analysis of Survival Time Data*, Journal of the Royal Statistical Society, Series B **40**, 214-221.

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

[41] Kim, M., De Gruttola, V. and Lagakos, S.W. (1993), *Analyzing Doubly Censored Data with Covariates, with Application to AIDS*, Biometrics, **49**, 13-22.

[42] Kuo, L. (1991) *Sampling Based Approach to Computing Nonparametric Bayesian Estimators with Doubly Censored Data*, Computing Science and Statistics. Proceedings of the 23rd. Symposium on the Interface, 612-615.

[43] Kuo, L. and Smith, F.M. (1992) *Bayesian Computations in Survival Models via the Gibbs Sampler*, Survival Analysis: State of the Art, 11-24.

[44] Metropolis, N. *et alt.* (1953), *Equations of State Calculations by Fast Computing Machines*, Journal of Chemical Physics, **21**, 1087-1091.

[45] Narayanan, A. (1990) *Computer Generation of Dirichlet Random Vectors*, J. Statist. Comput. Simul., **36**, 19-30.

[46] Raftery, A. E. and Lewis, S. (1992), *How many iterations in the Gibbs Sampler ?*, Bayesian Statistics 4, (Eds. J. Bernardo, J. Berger, A.P. David and A.F.M. Smith), Oxford University Press, Oxford, 763-773.

[47] Rai, K. , Susarla, V. and Van Ryzin, J. (1980) *Shrinkage Estimation in Nonparametric Bayesian Survival Analysis: A Simulation Study*, Commun. Statist.-Simula. Computa., **3**, 271-298.

[48] Samuelsen, S. O. (1989), *Asymptotic theory for non-parametric estimators form doubly censored data*, Scandinavian Journal of Statistics, **16**, 1-21.

[49] Sethuraman, J. (1994), *A constructive definition of Dirichlet priors*, Statist. Sinica, **4**, 639-650.

[50] Smith, A.F.M. and Roberts, G.O. (1993) *Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods*, Journal of the Royal Statistical Society, Series B **55**, 3-23.

[51] Susarla, V. and Van Ryzin, J. (1976) *Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations*, Journal of the American Statistical Association, **71**, 897-902.

[52] Susarla, V. and Van Ryzin, J. (1978) *Large Sample Theory for a Bayesian Nonparametric Survival Curve Estimator Based on Censored Samples*, The Annals of Statistics, **6**, 755-768.

[53] Tanner, M. A. (1993), *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer-Verlag, New York.

[54] Tanner, M. A. and Wong, W.H. (1987), *The calculation of posterior distributions by data augmentation*, Journal of the American Statistical Association, **82**, 528-540.

[55] Tierney, L. (1994) *Markov Chains for Exploring Posterior Distributions*, The Annals of Statistics, **22**, 1701-1762.

[56] Turnbull, B.W. (1974) *Nonparametric estimation of a survivorship function with doubly censored data*, Journal of the American Statistical Association, **69**, 169-173.

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