Statistical Papers 45, 139-173 (2004)

Statistical Papers

Survey article

Frequentist and Bayesian approaches for interval-censored data

Guadalupe Gómez[1], M. Luz Calle[2], and Ramon Oller[3]

[1] Departament d'Estadística, Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain
[2] Departament d'Informàtica i Matemàtica, Universitat de Vic, Sagrada Familia 7, 08500 Vic, Spain
[3] Departament de Matemàtica i Informàtica, Universitat de Vic, Sagrada Familia 7, 08500 Vic, Spain

Received: December 6, 2001; revised version: October 9, 2002

Interval censoring appears when the event of interest is only known to have occurred within a random time interval. Estimation and hypothesis testing procedures for interval-censored data are surveyed. We distinguish between frequentist and Bayesian approaches. Computational aspects for every proposed method are described and solutions with S–Plus, whenever are feasible, are mentioned. Three real data sets are analyzed.
Key Words: AIDS; Bayesian inference; Hypothesis testing; Interval censoring; Nonparametric methods; Permutational tests.

1 Introduction

Survival analysis is used in various fields for analyzing data involving the duration between two events. It is also known as *event history analysis, lifetime data analysis, reliability analysis* or *time to event analysis.* A key characteristic that distinguishes survival analysis from other areas in statistics is that survival data are usually censored. Censoring occurs when information about the survival time of some individuals is incomplete. Different circumstances can produce different types of censoring, such as, right-censored data, left-censored data and interval-censored data. This paper is devoted to this last censoring scheme.

Interval censoring mechanisms arise when the event of interest cannot be directly observed and it is only known to have occurred during a random interval of time. In this situation, the only information about the survival time T is that it lies between two observed times L and R. We find in the articles of Peto (1973) and Turnbull (1976) the first approach to the estimation of the distribution function when data are interval-censored. These authors consider closed intervals, $[L, R]$, so that exact observations are taking into account. We find in the literature other censoring mechanisms closely related to the concept of interval censoring as introduced by Peto and Turnbull. For example, if the

event is only known to be larger or smaller than an observed monitoring time, the data conforms to the current status model or interval-censored data, case 1. In experiments with two monitoring times, U and V with $U < V$, where it is only possible to determine whether the event of interest occurs before the first monitoring time ($T \leq U$), between the two monitoring times ($U < T \leq V$), or after the last monitoring time ($T > V$), the observable data is known as interval–censored data, case 2. A natural extension of case 1 and case 2 models is the case k model, where k is a fixed number of monitoring times. Schick and Yu (2000) discuss an extended case k model where the number of monitoring times is random. In all these censoring schemes the intervals are semi–closed and non–censored observations are not considered. Yu *et al.* (2000) generalize the case 2 model so that exact observations are allowed. More than 150 papers have been published, since those first two pioneering papers, focusing on different cases of interval–censored data, deriving theoretical properties for the estimators or dealing with regression problems where the response is interval-censored.

Examples of time–to–event data, and in particular of interval–censored data, arise in diverse fields, such as biology, demography, economics, engineering, epidemiology, medicine and public health. Although an *adhoc* analysis is required to analyze interval–censored data, the lack of statistical software packages for this type of censoring has driven many researchers to use methods which do not take into account the random nature of these intervals. Many researchers use imputation techniques, especially right–point or mid–point imputation, which may generate biased results. Furthermore, we should remark that theoretical and computational results using the techniques we present here could be different if we treat intervals as closed or semi–closed. The continuous nature of the variables would induce us to think that such a precision is not important. However, as it is exposed in Ng (2002), different interpretations of the intervals lead to different likelihood functions, which in turn could imply different nonparametric maximum likelihood estimates.

Section 2 gives a list of illustrations where interval-censored data are encountered. Furthermore, we detail along the paper three interval-censored situations which have been analyzed by the authors and will illustrate some of the methods. The data sets, as well as some of the *adhoc* programs, can be downloaded from `www-eio.upc.es/seccio_fme/research/GRASS` or can be obtained directly from the authors. In the next three sections, we present an overview of different statistical methods to analyze interval–censored data. The estimation of the survival function, or other related functions, could be accomplished either via a frequentist approach, in Sections 3 and 4, or through the Bayesian paradigm, in Section 5. Both approaches have important advantages and drawbacks and the decision of the most suitable approach is in general difficult to determine. For each of these two approaches, nonparametric models, where no distributional assumptions are made, as well as parametric models are developed. The particular case of doubly–censored data is discussed in Subsection 3.2. For the sake of inferential completeness we have developed in Section 4 the nonparametric problem of the comparison of two or more interval–censored samples.

The aim of this paper is to put together, using a common perspective and notation, the existing literature on interval censoring. While most of the results have been already published, as is cited throughout the text, we provide additional technical justifications for some of the theoretical results (Lemma 1, Lemma 2 and Theorem 1). The justification for the construction of the likelihood given in Proposition 1 is new, as it is the way that the permutational tests are presented in Section 4. Finally, we are not aware of whether the Bayesian parametric approach given in Subsection 5.1, although straigthforward, has been presented elsewhere.

We would like to mention here that in the writing of this paper, many other documents have crossed our work. In particular, we are aware that the research concerning interval–censored data case 1, 2 or k, as well as semiparametric regression models, is only briefly commented. We have focused on the more general interval censoring scheme, considering closed intervals, and in the development of the paper from the applied point of view. For that reason, we have included in all the sections a computational subsection describing either our own *adhoc* programs or the implementation with S–Plus.

2 Examples of Interval–Censored Survival Data

In this section we start reviewing different real situations where interval-censored data have been encountered. Peto (1973) reports data from annual surveys on 196 girls for which sexual maturity development, at the time of each survey, were recorded. Development was complete in some girls before the first survey, some girls were lost to follow–up before the last survey and before development was complete, and some girls had not completed development at the last survey. An estimator for the proportion who were not yet mature as a function of age was required. This is the first paper, to the best of our knowledge, where interval–censored data have been analyzed.

Interval-censored data are quite usual in longitudinal studies where subjects in the study are not monitored continuously and instead the event of interest is detectable only at specific times of observation, for example, at the time of a medical examination. We find this type of censoring in a great variety of scenarios. Finkelstein (1986) studies regression analysis methods for interval-censored data to analyze data from a breast cancer study where patients were followed for cosmetic response to therapy. Although patients were scheduled to be seen, at clinic visits, every 4 to 6 months, the fact was that after completion of primary irradiation treatment, or for those who were geographically remote, the intervals between visits were wider. For this study the data on the time of failure were recorded as an interval such as $(L, R]$, meaning that at L months the patient had shown no change, but by R months, the cosmetic state of her breast had deteriorated. The objective of the analysis is to compare the patients who receive adjuvant chemotherapy to those who did not and to determine whether chemotherapy affects the rate of deterioration of the cosmetic state. Another instance is exemplified by Smith *et al.* (1997) who, while investigating

occupational exposure to tuberculosis, encounters interval-censored data because the exact data of tuberculosis were unavailable and they had to rely on the time interval defined by the tuberculin skin test conversion.

In the context of the AIDS epidemic we find many instances where interval-censored data have been reported. Kooperberg and Clarkson (1997) analyze evidence of precancer from an ongoing study of the natural history of anal dysplasia in gay men who are enrolled in the AIDS Prevention Project in Seattle. The data are as well interval-censored because the precise time between two interviews when the precancerous condition was developed is unknown. Yu *et al.* (2000) analyze the distribution of the time to clinical relapse to ovarian cancer based on a clinical trial where a tumor marker is available. Those patients with high (or low) values are closely monitored. The paper by Goggins and Finkelstein (2000) focuses on the analysis of multivariate interval–censored data corresponding to a study of an opportunistic infection in HIV–infected individuals. The presence for the infection agent was tested both in the blood and in the urine at scheduled clinic visits. The failure times are censored into the interval between the last negative test and the first positive test. Since often patients missed several visits, the censoring intervals are overlapping and of varying lengths and methods for grouped data are not appropriate.

Animal tumourigenicity experiments result in another special type of interval-censored data. The goal of such studies is to analyze the effect of a suspected carcinogen on the time to tumour onset when the onset times cannot be observed. Rather, animals die or are sacrificed at predetermined time intervals, and are examined for the presence or absence of a tumour. If the tumours are irreversible, the observed death times (natural and sacrifices) provide left- and right-censored observations on the time until tumour onset (Gómez and Julià, 1990, Gómez and Van Ryzin, 1992). This type of data is an special instance of what we have defined as current status data where a unique monitoring time –in this case the natural death or the sacrifice– is considered for each individual.

Interval–censored data is also encountered in demographical studies where the use of a retrospective survey and population register data permits numerous applications of event–history analysis. Courgeau and Najim (1996) exploit interval–censored techniques to estimate the distribution of migrations or job changes over time based on Demographic Panel Surveys, and on surveys on social, geographical and wealth mobility in the 19^{th} and 20^{th} centuries in France. The data they analyze are interval–censored because, concerning residential or occupational mobility, they only know that a move has occurred between two censuses or family events.

Last but not least, interval censoring might occur together with left truncation. Different authors have approached this problem. Among others, Pan and Chappell (2002) approach this problem while comparing the probabilities of losing functional independence for male and female seniors.

3 Frequentist Approach

3.1 Nonparametric Methods

One of the first papers approaching the interval–censored situation is due to Peto (1973) who reports data from annual surveys on sexual maturity development of girls. Peto proposes a method based on maximizing the log–likelihood by a suitable constrained Newton–Raphson programmed search. Few years later, Turnbull (1976) approaches the more general problem of the analysis of arbitrarily grouped, censored and truncated data and derives an algorithm to obtain the nonparametric estimator of the distribution function. This algorithm can be applied, in particular, to deal with interval-censored situations. Few more years elapsed before these methods were applied in different setups, but these two pioneers papers are today the seed of most of the practical results. Among other papers we mention a couple. Gentleman and Geyer (1994) provide standard convex optimization techniques to maximize the likelihood function and to check the unicity of the solution; Böhning *et al.* (1996) view the problem from the perspective of a mixing problem of indicator functions and propose to use their statistical package C.A.MAN to compute the nonparametric estimator. The nonparametric estimator for the distribution function that these authors propose is a discrete distribution function that maximizes the likelihood over the set of discrete distributions that are piecewise constant between a finite set of points that depend on the observations. Since these estimators are step functions, their behaviour is quite unsmooth and sometimes they lack of interpretability, mainly when comparing survival curves. In the remainder of this section, we describe and illustrate the nonparametric methodology. We start, in Subsection 3.1.1, giving a theoretical justification for the construction of the likelihood function under noninformative censoring. We develop Turnbull's self–consistency method in Subsection 3.1.2, providing additional details of the proofs of his results. The asymptotic behaviour of the proposed estimators is discussed in Subsection 3.1.3. The last two subsections contain a discussion of computational aspects and an illustration.

3.1.1 Definition. Notation. Estimability. Likelihood

Let T be the random variable of interest. In our setting T is a positive random variable representing the time until the occurrence of a certain event $\mathcal{E}$ with unknown right-continuous distribution function $W(t) = \text{Prob}\{T \leq t\}$, survival function $S(t) = 1 - W(t)$ and density function $w(t)$, if it exists. In a study of n items or individuals, their potential times to $\mathcal{E}$, namely, $T_1, \ldots, T_n$, are unknown and instead we observe intervals that contain the unobserved values of $T_1, \ldots, T_n$. Let $\mathcal{D} = \{[L_i, R_i], 1 \leq i \leq n\}$ be the interval–censored survival data where L_i is the last observed time for the i^{th} individual before the event $\mathcal{E}$ has occurred and R_i indicates the first time the event $\mathcal{E}$ has been observed. We are in fact formally observing random censoring vectors (L_i, R_i), $i = 1, \ldots, n$, comming from a joint density function, $f_{[L,R]}(l, r; \gamma)$, such that $L \leq R$ with probability 1. Denote by $f_{[T,L,R]}(t, l, r; W, \gamma)$ the joint density of the unobserved

vector (T, L, R) and note that is such that $L \leq T \leq R$ with probability 1.

We suppose that censoring occurs noninformatively in the sense that for any t, l, r such that $l \leq t \leq r$, the conditional density of T given L and R, $f_{[T|L,R]}(t|l, r; W, \gamma)$, satisfies

$$f_{[T|L,R]}(t|l, r; W, \gamma) = \frac{dW(t)}{W(r) - W(l-)}, \tag{1}$$

where we define $W(t-) = \lim_{\Delta \to 0+} W(t - \Delta)$. That is, censoring times L and R do not anticipate events.

Proposition 1
Assume that we have a unique individual for which we have observed the failure time T falling inside the random interval $[l, r]$. If censoring occurs noninformatively, the contribution to the likelihood of this individual is proportional to $\int_l^r dW(t)$.

Proof. We first prove that there exists a function K such that the conditional density of (L, R) given T is such that for any t, l, r with $l \leq t \leq r$, then

$$f_{[L,R|T]}(l, r|t; \gamma) = K(l, r; \gamma).$$

Indeed, for any t, l, r such that $l \leq t \leq r$, following the usual rules for conditional densities and the noninformative condition (1), we have

$$\begin{aligned} f_{[L,R|T]}(l, r|t; W, \gamma) &= \frac{f_{[T,L,R]}(t, l, r; W, \gamma)}{dW(t)} = \frac{f_{[T|L,R]}(t|l, r; W, \gamma) f_{[L,R]}(l, r; \gamma)}{dW(t)} \\ &= \frac{dW(t) f_{[L,R]}(l, r; \gamma)}{(W(r) - W(l-)) dW(t)} = \frac{f_{[L,R]}(l, r; \gamma)}{W(r) - W(l-)} = K(l, r; \gamma) \end{aligned}$$

It is then obvious that the contribution to the likelihood of an individual whose failure time is observed to fall within the interval $[l, r]$ is given by

$$f_{[L,R]}(l, r; \gamma) = \int_l^r f_{[L,R|T]}(l, r|t; W, \gamma) dW(t) = K(l, r; \gamma) \int_l^r dW(t).$$

□

Hence, if censoring occurs noninformatively and if the law governing L and R does not involve any of the parameters of interest, we can base our inferences on the likelihood function $L(W|\mathcal{D})$ given by

$$\begin{aligned} L(W|\mathcal{D}) &= \prod_{i=1}^n \int_{L_i}^{R_i} dW(u_i) = \prod_{i=1}^n [W(R_i) - W(L_i^-)] \\ &= \prod_{i=1}^n [S(L_i^-) - S(R_i)] = \prod_{i=1}^n \text{Prob}\{L_i \leq T_i \leq R_i\}. \end{aligned} \tag{2}$$

3.1.2 Self-consistency equations. Maximum likelihood estimation

The goal is to find a monotonically increasing function $W(t)$ which maximizes the overall likelihood function (2). The resulting estimator might not be unique because the likelihood for an interval-censored observation depends only on the difference between the survival values at the end-points of that interval and not at all on the detailed behaviour within the interval.

In what follows we describe Turnbull's self–consistency method. We start constructing the set of intervals where the mass is concentrated. From the sets $\mathcal{L} = \{L_i, 1 \leq i \leq n\}$ and $\mathcal{R} = \{R_i, 1 \leq i \leq n\}$ we can derive all the distinct closed intervals whose left and right end-points lie in the sets $\mathcal{L}$ and $\mathcal{R}$ respectively and which contain no members of $\mathcal{L}$ or $\mathcal{R}$ other than at their left and right endpoints respectively. Let these intervals, known as Turnbull's intervals, be written in order as $\mathcal{I} = \{[q_1, p_1], [q_2, p_2], \ldots, [q_m, p_m]\}$. We illustrate this construction with the following example.

Example: Suppose that the following $n = 6$ intervals have been observed $\mathcal{D} = \{[L_i, R_i], 1 \leq i \leq 6\} = \{[0,1], [4,6], [2,6], [0,3], [2,4], [5,7]\}$. Then, Turnbull's intervals are given by $\mathcal{I} = \{[q_1, p_1] = [0,1], [q_2, p_2] = [2,3], [q_3, p_3] = [4,4], [q_4, p_4] = [5,6]\}$.

Lemma 1 (Turnbull) *Any distribution function which increases outside Turnbull's intervals $\mathcal{I}$ cannot be a maximum likelihood estimator of W. Thus, it suffices to consider only distribution curves which are horizontal everywhere except in the intervals $\mathcal{I}$ and which increase in some or all of these intervals.*

Proof. Let W be a distribution function which increases outside Turnbull's intervals. Assume, without loss of generality, that W is horizontal everywhere except in the interval (p_l, q_{l+1}) and in the intervals $\mathcal{I}$. By the construction of Turnbull's intervals, the only possible members of $\mathcal{L}$ or $\mathcal{R}$ between p_l and q_{l+1} are necessarily such that all the right end-points are smaller than all the left end-points. Let r_l be a point in (p_l, q_{l+1}) that is greater than all the right and less than all the left end-points in (p_l, q_{l+1}). We can then construct a distribution function W^* which is equal to W everywhere except in (p_l, q_{l+1}) where is defined as $W^*(t) = W(r_l)$ for every $t \in (p_l, q_{l+1})$. The factors $W(R) - W(L)$ in the likelihood can be one of the following three types:

1. If $L < R < p_l$ or $q_{l+1} < L < R$ then $W(R) - W(L^-) = W^*(R) - W^*(L^-)$
2. If $p_l < R < r_l < q_{l+1}$ then $W(R) - W(L^-) \leq W(r_l) - W(L^-) = W^*(R) - W^*(L^-)$
3. If $p_l < r_l < L < q_{l+1}$ then $W(R) - W(L^-) \leq W(R) - W(r_l) = W^*(R) - W^*(L^-)$

We illustrate the second situation in Figure 1. By construction, if $R \in (p_l, q_{l+1})$ then $W(R) < W^*(R)$.

Thus, we conclude that $L(W^*|\mathcal{D}) \geq L(W|\mathcal{D})$ and that W cannot be a maximum likelihood estimator of W.

□

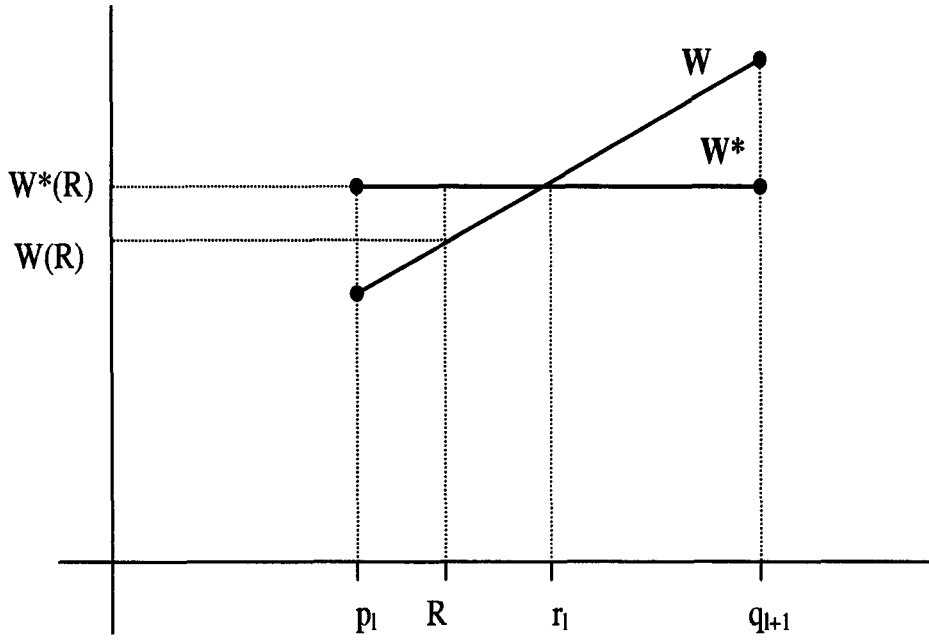

Figura 1: Graphical illustration of the second situation in the proof of Lemma 1

Lemma 2 (Turnbull) *The total likelihood is a function only of the amount that the distribution curve increases in the intervals $\mathcal{I}$ and is independent of how the increase actually occurs, so the estimated distribution curve is unspecified in each $[q_j, p_j]$ and is well defined and flat between these intervals. Note that while estimating the distribution function W, we are as well estimating the survival function $S = 1 - W$.*

Denoting by $w_j = W(p_j) - W(q_j^-) = \text{Prob}\{q_j \leq T \leq p_j\}$ the weight of the j^{th} interval, $j = 1, \ldots, m-1$, $w_m = 1 - \sum_{j=1}^{m-1} w_j$, Lemmas 1 and 2 define equivalence classes that enable us to write down $L(W|\mathcal{D})$ as

$$L_T(w_1, \ldots, w_{m-1}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_j^i [W(p_j) - W(q_j^-)] \right) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_j^i w_j \right) \quad (3)$$

where the indicator $\alpha_j^i = \mathbf{1}\{[q_j, p_j] \subseteq [L_i, R_i]$ expresses whether or not the interval $[q_j, p_j]$ is contained in $[L_i, R_i]$. The vectors $(w_1, \ldots, w_m)$ define equivalence classes on the space of distribution functions W which are flat outside $\cup_{j=1}^{m} [q_j, p_j]$. Therefore, the maximum will be at best unique only up to equivalence classes and the problem of maximizing $L(W|\mathcal{D})$ has been reduced to the finite-dimensional problem of maximizing a function of $w_1, \ldots, w_{m-1}$ subject to the constraints $w_j \geq 0$ and $1 - \sum_{j=1}^{m-1} w_j \geq 0$.

The total likelihood, as a function of $w_1, \ldots, w_{m-1}$, is strictly convex (except on the boundaries of the constrained region on which the likelihood func-

tion is zero), so the values of $w_1, \ldots, w_{m-1}$ that maximize it are unique. Let $(\hat{w}_1, \ldots, \hat{w}_m)$ be the maximizing solution of (3). **Turnbull's nonparametric estimator** $\hat{W}$ for W is given by

$$\hat{W}(t) = \begin{cases} 0 & \text{if} \quad t < q_1 \\ \hat{w}_1 + \cdots + \hat{w}_k & \text{if} \quad p_k \leq t < q_{k+1}, \quad 1 \leq k \leq m-1 \\ 1 & \text{if} \quad t \geq p_m \end{cases} \tag{4}$$

and is not specified for $t \in [q_j, p_j]$, for $1 \leq j \leq m$. Therefore $\hat{W}$ is an increasing step function, with $m+1$ horizontal lines with gaps in between and the way in which $\hat{W}$ increases inside these gaps is arbitrary. Note that only the total probability assigned by W to the intervals $[q_j, p_j]$ can be identified.

The variances and covariances of the non zero $\hat{w}_k$ are given by the inverse of the second derivatives matrix of the loglikelihood (7) with respect to $w_1, \ldots, w_{m-1}$. However, there is no yet theoretical justification for this procedure, the problem being a violation of the usual assumption of a fixed number of unknown parameters that remains unchanged with increasing the sample size.

We now introduce the concept of self–consistency and give its equivalence with the property of maximum likelihood. $\widehat{\boldsymbol{w}}$ **is a self–consistent estimate of** $\boldsymbol{w}$**=** $(w_1, \ldots, w_m)$ if

$$\hat{w}_j \quad = \quad \mathbf{E}_{\widehat{\boldsymbol{w}}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{q_j \leq T_i \leq p_j\} \mid \mathcal{D} \right].$$

In other words, solving the conditional expectation equation, a **self-consistent estimator** of $(w_1, \ldots, w_m)$ is defined to be any solution of the following simultaneous equations:

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} w_j \quad 1 \leq j \leq m. \tag{5}$$

Define $\mu_j^i(w_1, \ldots, w_m)$ and $\tau_j(w_1, \ldots, w_m)$ as

$$\begin{aligned} \mu_j^i(w_1, \ldots, w_m) &= \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} w_j \\ \tau_j(w_1, \ldots, w_m) &= \frac{1}{n} \sum_{i=1}^{n} \mu_j^i(w_1, \ldots, w_m). \end{aligned} \tag{6}$$

Remark: Note that the terms $\sum_{l=1}^{m} \alpha_l^i w_l$ correspond to the sum of probabilities associated with the i^{th} individual.

Lemma 3 *We introduce the logarithm of the likelihood (3) as*

$$l(\boldsymbol{w}) = \log(L_T(w_1, \ldots, w_{m-1})) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \alpha_j^i w_j \right). \tag{7}$$

The directional derivative $d_j(\boldsymbol{w})$ *of* $l(\boldsymbol{w})$ *defined as*

$$d_j(\boldsymbol{w}) = \frac{\partial l(\boldsymbol{w})}{\partial w_j} - \sum_{k=1}^{m} w_k \frac{\partial l(\boldsymbol{w})}{\partial w_k}$$

satisfies

$$\tau_j(\boldsymbol{w}) = w_j \left(1 + \frac{d_j(\boldsymbol{w})}{n}\right), \; j = 1, \cdots, m,$$

where τ_j *has been defined in (6).*

Proof. Notice that the directional derivative $d_j(\boldsymbol{w})$ corresponds to

$$\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} l\left(\frac{w_1}{1+\epsilon}, \ldots, \frac{w_j + \epsilon}{1+\epsilon}, \ldots, \frac{w_m}{1+\epsilon}\right),$$

which considers the effect of increasing the j^{th} component by a small positive amount ϵ and divides all the components by $1+\epsilon$ in order to keep the sum equal to 1. That is,

$$\begin{aligned} d_j(\boldsymbol{w}) &= \frac{\partial l(\boldsymbol{w})}{\partial w_j} - \sum_{k=1}^{m} w_k \frac{\partial l(\boldsymbol{w})}{\partial w_k} = \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} - \sum_{k=1}^{m} w_k \sum_{i=1}^{n} \frac{\alpha_k^i}{\sum_{l=1}^{m} \alpha_l^i w_l} \\ &= \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} - \sum_{i=1}^{n} \frac{\sum_{k=1}^{m} \alpha_k^i w_k}{\sum_{l=1}^{m} \alpha_l^i w_l} = \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} - n. \end{aligned}$$

It follows that,

$$\begin{aligned} 1 + \frac{d_j(\boldsymbol{w})}{n} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l}, \\ \left(1 + \frac{d_j(\boldsymbol{w})}{n}\right) w_j &= \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_j^i}{\sum_{l=1}^{m} \alpha_l^i w_l} w_j = \frac{1}{n} \sum_{i=1}^{n} \mu_j^i(\boldsymbol{w}) = \tau_j(\boldsymbol{w}). \end{aligned}$$

□

Theorem 1 (Turnbull)

1. *If* $\hat{\boldsymbol{w}}$ *is a maximum likelihood estimator for* W*, then* $\hat{\boldsymbol{w}}$ *satisfies the self-consistent equations (5).*
2. *Conversely, the solution* $\hat{\boldsymbol{w}}$ *of the self-consistent equations (5) is the non-parametric maximum likelihood estimator of* $\boldsymbol{w}$ *provided that* $d_j(\boldsymbol{w}) \leq 0$ *whenever* $w_j = 0$.

Proof.

1. The maximization of $l(\boldsymbol{w})$ can be considered as a concave programming problem with linear constraints. Thus, the Kuhn-Tucker conditions (Gentleman and Geyer, 1994) are necessary and sufficient for optimality, that is, $\boldsymbol{w}$ is a maximum likelihood estimate if and only if, for every j, either $d_j(\boldsymbol{w}) = 0$ or $d_j(\boldsymbol{w}) \leq 0$ when $w_j = 0$. Henceforth, it is obvious that the MLE are self-consistent.

2. If $\boldsymbol{w}$ is a self-consistent solution, it satisfies

$$\left(1 + \frac{d_j(\boldsymbol{w})}{n}\right) w_j = w_j, \tag{8}$$

hence if $w_j > 0$, it follows that $d_j(\boldsymbol{w}) = 0$ and if $w_j = 0$ since we are assuming that $d_j(\boldsymbol{w}) \leq 0$ the Kuhn-Tucker conditions are fulfilled and $\boldsymbol{w}$ is a maximum likelihood estimator.

□

Example continued: The likelihood corresponding to the previous 6 intervals is given by

$$\begin{aligned} L_T(w_1, w_2, w_3, w_4) &= \prod_{i=1}^{6} \left(\sum_{j=1}^{4} \alpha_j^i [W(p_j) - W(q_j -)] \right) \\ &= (w_1)(w_3 + w_4)(w_2 + w_3 + w_4)(w_1 + w_2)(w_2 + w_3)(w_4), \end{aligned}$$

and the maximizing solution is found at the point $(\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{3}{8})$. Thus Turnbull's nonparametric estimator $\hat{W}$ for W is given by

$$\hat{W}(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{1}{4} & \text{if } 1 \leq t < 2 \\ \frac{1}{2} = \frac{1}{4} + \frac{1}{4} & \text{if } 3 \leq t < 4 \\ \frac{5}{8} = \frac{1}{4} + \frac{1}{4} + \frac{1}{8} & \text{if } 4 \leq t < 5 \\ 1 & \text{if } t \geq 6 \end{cases}$$

3.1.3 Asymptotic behaviour

Turnbull derived self-consistent equations for a very general censoring scheme and in particular for a very general definition of interval censoring as it is described in the introduction. Since Turnbull's self-consistent equation is not of the form of an integral equation, the study of its large sample properties have not been very fruitful. Yu *et al* (2000) prove that Turnbull's estimator is strongly consistent under the assumption that the support of the vector (L, R) is finite, and that censoring occurs noninformatively in the sense described in (1). The assumption concerning the support of (L, R) is reasonable since it means that

the support of the inspection times is finite, which in practice is true because most follow–up studies are recorded on a discrete time scale and the total study period is finite. The asymptotic distributional behaviour of Turnbull's estimator has not been yet established.

Several authors prove consistency of the generalized maximum likelihood estimator for interval–censored data, case 2, when there are only a finite number of inspection times x_j, $j = 1, \ldots, m$ in any finite interval (Gentleman and Geyer, 1994), or under the assumption that the vector (L, R) is discrete but that W is arbitrary (Yu *et al*, 1998).

The asymptotical properties of the nonparametric maximum likelihood estimator (NPMLE) when data are interval-censored, case 1 or 2, are largely discussed in Groeneboom and Wellner (1992). They propose the convex minorant algorithm for computing the nonparametric maximum likelihood of the distribution function and prove that if T is a continuous random variable and the interval window is independent of T, then the NPML estimator is consistent. Concerning asymptotic normality, Yu *et al* (1998) obtain for interval-censored data, case 2, the joint asymptotic normality of the generalized maximum likelihood estimate at the usual rate $\sqrt{n}$ for the points in $\mathcal{A} = \{a \in \mathbb{R} : \mathrm{P}(\mathbf{L} = \mathbf{a}) + \mathrm{P}(\mathbf{R} = \mathbf{a}) > \mathbf{0}\}$.

3.1.4 Computational aspects

So far, most of the analysis that involve interval–censored data, have been done with software specifically developed by the corresponding authors. The program, `ICTURNBULL.C`, used for the analysis illustrated in this paper, has been written in C-language and requires a rectangular data file consisting on 2 columns and $n+1$ rows, where n is the sample size. The first row includes the sample size and the number that play the role of infinity (we usually use 9999). The following n rows include the left and the right endpoint of the censoring interval for each individual.

S–Plus version 6 for Linux or 2000 for Windows provides a new set of commands to perform survival analysis with interval–censored data. The algorithm used by this software considers semi–closed intervals $(L, R]$ where $L < T \leq R$ and incorporates exact, right-censored, and left-censored data. A vector `ident` containing the identification of the n individuals under study is first defined. The object `censor.codes` assigns a numerical value to each individual to distinguish whether the observation is exact (`censor.codes=1`), right-censored (`censor.codes=0`), left-censored data (`censor.codes=2`) or interval-censored (`censor.codes=3`). Vectors `lower` and `upper` contain the lower and the upper limit, respectively of the intervals. An object from the type `data frame` is then constructed as follows: `int.data <-- data.frame(ident, lower, upper, censor.codes)`. This is the object that the new procedure `kaplanMeier` needs in order to estimate the survival function using Turnbull's method, that is,
`surv.est<--kaplanMeier(censor(lower,upper,censor.codes)~1, data=int.data)`

Remark: It is important to note that the original analysis by Turnbull, and the one used along this paper, where the intervals are closed ($[L, R]$ meaning

$L \leq T \leq R$), cannot be done straightforwardly using the above S–plus procedure. One, not very elegant, way of taking advantage of S–plus procedures is to redefine the lower vector subtracting a small quantity, say 0.001, and reinterpreting then. Plots of the estimated survival function can be obtained by either `plot(surv.est)` or `plot.kaplanMeier(surv.est)`.

3.1.5 Illustration 1

Intravenous drug addiction and the human immunodeficiency virus (HIV) infection are two recent and closely related epidemics. In an attempt to estimate the elapsed time to HIV–infection since they enter the intravenous drug users risk group, the presence of interval–censored data shows again.

The cohort is based on the 306 (240 male and 66 female) intravenous drug users entering the detoxification unit of the Germans Trias i Pujol Hospital in Badalona (Spain), between February 1987 and November 1997 which have started intravenous drug use between 1986 and 1991. The following variables were available for most of the patients: date of birth, date of first IV–drugs use, date of last negative HIV antibody test, date of the first positive HIV antibody test. Three exclusive and exhaustive subcohorts were defined. The *seroconvertor subcohort* consists on the 29 patients (9.5%) for whom information on a negative HIV test and a positive HIV test was available and these two dates define the interval where the HIV–infection has occurred. Thus, the infection time for these patients is interval–censored. The *HIV–positive subcohort*, or *seroprevalent subcohort*, consists on the 121 patients (39.5%) that arrived HIV–positive to the detoxification unit. The infection time is in this case left–censored in the interval of time between the date starting at risk for HIV–infection and the earliest positive HIV test. The *HIV-negative subcohort* consists on the 156 patients (51%) that arrived HIV–negative to the detoxification unit and remained HIV–negative at the date of their last antibody test. The infection time in this subcohort is right–censored, the lower limit of the censoring interval being the time of the last negative HIV test and the upper limit is infinity.

Figure 2, computed via the S–Plus software, shows the estimated survival function for the failure time variable defined as the number of months elapsed time to HIV–infection since the patients enter the intravenous drug risk group, both for men and for women. We observe that men tend to spend more time infection free than women. The statistical significance of this difference, along with differences of age and year of first IV-drug use are studied in Subsection 3.3. A larger data set including patients who started intravenous drug use before 1986 or after 1991 was analyzed using nonparametric Bayesian techniques by the authors in Gómez *et al.* (2000).

3.2 Doubly–Censored Data

Most statistical methods in survival analysis assume that the time to the originating event is known and allow the final time to be censored. Here we consider a situation where the origin time is interval-censored and the final time is right-

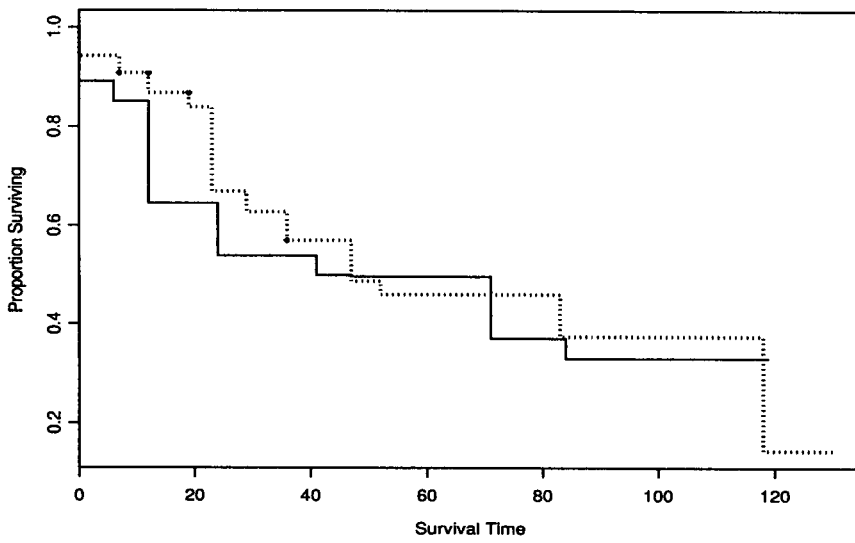

Figura 2: Probabilities of being HIV–infection free for women (line) and men (dotted line)

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 (Chang and Yang, 1987).

Under the assumption that there is a discrete time scale both for the origin time and for the latency time, De Gruttola and Lagakos (1989) propose a method 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. An alternative approach is proposed by Gómez and Lagakos (1994) who develop a two-step estimation procedure. 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. Gómez and Calle (1999) propose a modification of Gómez and Lagakos algorithm which does not require the discretization of the data.

3.2.1 Gómez and Calle estimator

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, respectively, under the assumption that X and T are independent random variables. We assume that X is interval–censored in $[L, R]$ and that Z is right–censored. Let V be 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 $(L_i, R_i, d_i, V_i, c_i)$ where d_i and c_i are the censoring indicators of the origin and final times, respectively. That is, $d_i = \mathbf{1}\{R_i < \infty\}$ and $c_i = 1$ if $Z_i = V_i$ and $c_i = 0$ if $Z_i > V_i$.

The procedure is based on the following two steps. In the first step straightforward Turnbull's method is applied. This produces the following set of intervals $\{[q_1, p_1], \ldots, [q_m, p_m]\}$ where the W distribution assigns its mass. The corresponding estimator for the distribution is denoted by $\hat{W}$. Denote by $\hat{w}_j = \text{Prob}(q_j \leq T \leq p_j)$, $1 \leq j \leq m$. For the second step, where again discreteness of T is removed, a new set of intervals has to be defined where the distribution F is identifiable. Denoting by $L_{ij} = V_i - p_j$ and $R_{ij} = V_i - q_j$ when $c_i = 1$ and by $L_{ij} = V_i - \frac{p_j + q_j}{2}$ and $R_{ij} = \infty$ when $c_i = 0$, the conditional likelihood can be written as

$$L_c(F|\hat{W}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_j^i \, \hat{w}_j \left[F(R_{ij}) - F(L_{ij}^-) \right] \right]^{d_i}.$$

The reader is addressed to Gómez and Calle (1999) for further details. Note that here, as in the univariate case, a set of intervals, $[q_1', p_1']$, $[q_2', p_2']$, $\ldots$, $[q_r', p_r']$, where F places its mass can be defined. These intervals are obtained from the different $\{R_{ij}\}$ and $\{L_{ij}\}$ in the same way as Turnbull's intervals.

The maximum likelihood estimator for $(f_1, \ldots, f_k)$, where $f_j = F(p_j') - F((q_j')^-)$ is the probability of the interval $[q_j', p_j']$, is obtained as the solution of the self-consistent equations

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

with $n_0 = \sum_{i=1}^{n} (1 - d_i)$ the number of observations with a right-censored origin time and α_{jk}^i, the indicator of an origin time in $[q_j, p_j]$ and a duration time in $[q_k', p_k']$.

3.2.2 Computational aspects

The methodology has been implemented in a C-language program, `MODGL.C`, which is available from the authors upon request. The program requires a data file consisting on a first row which contains the sample size n and the number that plays the role of infinity (we usually use 9999) and n rows each one containing the values of (L_i, R_i, V_i, c_i) for each individual.

3.2.3 Illustration 2

In the study of the chronological time of the HIV infection, De Gruttola and Lagakos (1989) analyze a French cohort of hemophilia patients who were infected with HIV in the early 1980's. The cohort corresponds to 262 patients that were treated at the Hôpital Kremlin Bicêtre and the Hôpital 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. Two group of patients were distinguished: 105 patients in the heavily-treated group, that is those 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.

Since blood samples from these individuals were periodically collected and stored, they could be retrospectively tested to determine a time interval during which the infection occurred. The time of infection for these patients is then interval–censored, the infection is only known to have occurred in the interval of time specified by the last negative and the first positive assessment. Because the latency period between infection with HIV and the development of AIDS can be very long, many of the hemophiliacs infected at that time still had not developed AIDS by the end of the study. Hence, both the initiating and terminating events that determine the latency period can be censored in the same individual.

The observations, based on a discretization of the time axis into 6-month intervals, are of the form $(L_i, R_i, d_i, V_i, c_i)$. L_i and 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.

We apply Gómez and Calle procedure to each one of the two groups of this data set and we obtain estimators for the distribution to the time to HIV–infection and for the latency distribution. Figure 3 gives the estimated cumulative distribution function of the latency times for the two groups. The estimators are very similar for the first 3 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 were analyzed by the authors in Gómez and Lagakos (1994) and Gómez and Calle (1999). In this paper, the data are analyzed as well in Subsec-

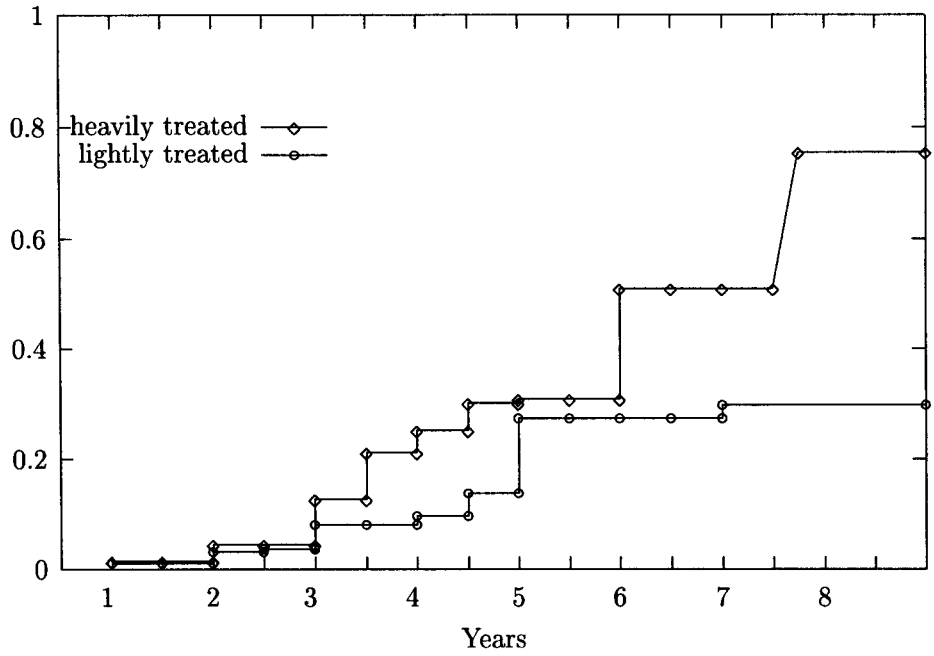

Figura 3: Estimated cumulative distribution function of latency time between HIV seroconversion and onset of symptoms for heavily-treated group and lightly-treated group.

tion 5.1 to illustrate a Bayesian regression model for interval–censored data.

3.3 Parametric regression models

An effective and standard approach to analyze interval-censored survival data when a parametric model is appropriate is to use maximum likelihood estimation. Let T be a positive random variable representing the time until the occurrence of a certain event $\mathcal{E}$ with unknown right-continuous distribution function $W(t;\theta) = \text{Prob}\{T \leq t;\theta\}$, density function $w(t;\theta)$ and unknown finite–dimensional parameter θ. The potential times $T_1, \ldots, T_n$ of n individuals are unknown and we assume here that, as usual, we have interval–censored survival data $\mathcal{D} = \{[L_i, R_i], \quad 1 \leq i \leq n\}$ such that $L_i \leq T_i \leq R_i$. Under the noninformative assumption (1), the parametric likelihood for W, given $\mathcal{D}$, is proportional to

$$L(\theta|\mathcal{D}) = \prod_{i=1}^{n} [W(R_i;\theta) - W(L_i-;\theta)] = \prod_{i=1}^{n} \int_{L_i}^{R_i} w(u_i;\theta)du_i. \tag{9}$$

Lindsey and Ryan (1998) develop a piecewise exponential model for the interval–censored case. In order to do that they break the time scale into J

intervals and assume a constant hazard within each. This model has the advantage that as J increases it becomes more nonparametric in nature. This method can be extended so that covariate effects are accommodated using proportional hazards. Standard likelihood theory can be used if the number of intervals is not too large. Although no standard statistical packages consider this model, the EM algorithm is easily implemented as the authors describe in the appendix.

Lindsey (1998) investigates the effect of ignoring interval censoring for parametric modeling. To this end he fits different parametric families and accommodates regression equations both for the location and dispersion parameters. His conclusions –somehow arguable– are that for parametric models interval censoring can often be ignored and the midpoint of the interval used instead in the likelihood function.

The decision between a parametric or a nonparametric approach is not easy. On one hand, if there is scientific or empirical knowledge of the problem that justifies a model, the nonparametric approach may represent an important loss of efficiency versus the use of a parametric method, 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. Therefore, the use of completely parametric methodologies involves the risk of deriving inconsistent estimators for the parameters of interest and if the parametric model does not fit suitably the data, this might lead to inaccurate conclusions. However, among other features, the parametric approach has the advantage that it provides the means to predict different parameter based quantities for a longterm (*.i.e.,* the percentage of HIV–infected individuals who will be AIDS-free). It also permits the description of the hazard function at different times and are useful for point and variance estimation of relative percentiles. However, all the inferences will depend upon the assumption of the model, and there are not, yet, goodness-of-fit tests to check how suitable is the parametric model when data are interval–censored. A large number of papers have acknowledged the interval–censored nature of the data and have used parametric regression models to analyze the data. It is is worth mentioning, in the context of the AIDS epidemic, the papers by Brookmeyer and Goedert (1989) and Muñoz and Xu (1996).

There has been also recent work on estimation from semiparametric regression models with interval-censored data. Semiparametric models such as the proportional hazards model or the proportional odds model treat the baseline hazard function, or the survival baseline function, as a nuisance parameter. Younes and Lachin (1997) present a flexible family of link-based regression models with time-independent covariates. Their model yields the proportional hazards model and the proportional odds model as special cases. Kooperberg and Clarkson (1997) introduce a methodology for hazard regression in which linear or cubic splines and their tensor products are used to estimate the conditional log–hazard function based on a great variety of censoring scenarios that include interval–censored data and time dependent covariates. Goetghebeur and Ryan (2000) propose a semiparametric approach that while retaining some of the appealing features of Kooperberg and Clarkson's smoothing method, it reduces to a standard Cox proportional hazards model in the absence of interval censoring.

3.3.1 Computational aspects

Several parametric families can be framed into a log-linear model, $\log T = \mu + \beta Z + \sigma W$, where W stands for the error distribution, for which standard maximum likelihood theory can be used. S–plus new release develops the `censorReg` routine which provides a way of fitting the above log-linear model for interval–censored data, accepting, among others, the Weibull, extreme value, normal, log–normal, logistic and log–logistic as the error distributions..

Following the example for the data frame `int.data` in Subsection 3.1, the S–plus procedure `censorReg` can be used as follows:

```
int.data.censor<-censor(lower,upper,censor.codes)
cens.mod<-censorReg(int.data.censor~1,dist="weibull",data=int.data)
```

This command fits a Weibull model (without covariates). An extensive output will be given using `summary(cens.mod)`. Plots to judge the goodness-of fit can be obtained via `plot(cens.mod)` or `probplot(cens.mod)`. In particular, the command `probplot6(cens.mod)` produces 6 probability plots for the maximum likelihood fitting of 6 different distributions. Once an error distribution has been chosen, the procedure `censorReg` can be used, as well, to incorporate several covariates.

3.3.2 Illustration 1

We reanalyze again the time to HIV–infection choosing a regression parametric model including the covariates age, gender, and year of first intravenous drug use. The data were reasonably well fitted by the log–logistic distribution and age was found to be significant at a 95% level ($p = 0.0345$). The parameter β in the log–linear model, when only age is taken as covariate, is estimated to be equal to 0.0686, with a 95% confidence interval given by $[0.0365, 0.133]$. As a consequence, older people tend to be HIV–infection free for longer number of months. For instance, the median number of months until HIV–infection for a 35 years old intravenous drug user is $\exp((35 - 20) \cdot 0.0686) \approx 2.8$ times the median number of months of a 20 years old individual.

4 Hypothesis testing

One important question that arises in many survival studies is to establish if there are differences in the survival times among different groups of individuals. While many k–sample tests have been developed when data are uncensored or right–censored, research for interval–censored data is still ongoing. Most approaches to this problem try to generalize these known tests to the interval–censored framework. In Mantel (1967) we find an interval–censored data version of the Wilcoxon test, in Peto and Peto (1972) we find a different extension of the Wilcoxon test and an extension of the Log–rank test and in Fay and Shih (1998) we find an interval–censored data form of the t–test. The main characteristic of these articles is the use of permutational distributions. The difficulty of finding

the distribution of the test statistic is avoided with this permutational approach. Other approaches assume that the collection of possible interval endpoints is discrete. This assumption ensures a finite number of parameters in the log–likelihood which allows to find test statistics with known asymptotic distribution, see for example Finkelstein (1986) and Petroni and Wolfe (1994).

4.1 Permutational tests

We introduce now the permutational approach to the k–sample problem. Let T be the time to the event of interest. Assume that we have k groups of data, $G_1, \ldots, G_k$ with respective sample sizes $n_1, \ldots, n_k$. Define $W_1, \ldots, W_k$ the distribution functions of T under each one of these groups. The k–sample problem establishes a test between $H_0 : W_1 = \cdots = W_k$ and $H_a : W_i \neq W_j$ for some i,j. Denote by $\mathbf{z}_i$ a vector of covariates representing to which group the i^{th} observation belongs. In the two sample problem, the usual choice of the covariate is $\mathbf{z}_i = \alpha_i^{(2)}$ where $\alpha_i^{(2)}$ is an indicator function that is equal to 0 if the individual belongs to group G_1 and 1 if it belongs to group G_2. When we have k groups many choices of $\mathbf{z}_i$ are possible, for instance,

$$\mathbf{z}_i = \left(\frac{\alpha_i^{(1)}}{\sqrt{n_1}}, \frac{\alpha_i^{(2)}}{\sqrt{n_2}}, \ldots, \frac{\alpha_i^{(k)}}{\sqrt{n_k}} \right)'$$

where $\alpha_i^{(j)}$ is an indicator function that is equal to 1 if the individual belongs to group G_j and 0 otherwise.

A permutational linear test statistic is of the form:

$$L_0 = \sum_{i=1}^{n} \mathbf{z}_i c_i, \tag{10}$$

where c_i is a scalar score associated to the i^{th} observation which is independent of the covariates. The idea behind the permutational test is that, if the null hypothesis is true and the censoring mechanism does not depend on the grouping, the labels on the scores are exchangeable. Thus, the permutational distribution of L_0 is obtained by permuting the labels and recomputing the test statistic for all the possible rearranged labels. The main key for these procedures is to use scores that are sensitive to the alternative hypothesis and, in that case, the null hypothesis will be rejected if L_0 is an extreme value for the permutational distribution. This permutational distribution can be computed exactly when the sample size is small. When n is large, a version of the Central Limit theorem for exchangeable random variables allow us to rely on a normal asymptotic approximation for the permutational distribution of L_0 where $E(L_0) = n\bar{c}\bar{\mathbf{z}}'$ ($\bar{c} = 0$ in our examples) and variance

$$\text{Var}(L_0) = \frac{\left(\sum_{i=1}^{n} c_i^2 - n\bar{c}^2\right)\left(\sum_{i=1}^{n} (\mathbf{z}_i \mathbf{z}_i' - \bar{\mathbf{z}}\bar{\mathbf{z}}')\right)}{(n-1)}.$$

The Wilcoxon–Gehan (WG) score for each observation is the difference between the number of time observations that are clearly to its left and the number of time observations that are clearly to its right. Intervals which overlap with the i^{th} interval don't contribute in the computation of the i^{th} score. That is,

$$WGc_i = \sum_{j=1}^{n} 1\{R_j < L_i\} - \sum_{j=1}^{n} 1\{L_j > R_i\}.$$

Gehan (1965) proposes these scores in order to extend the two sample Wilcoxon test for right–censored data. The proposal is reviewed by Mantel (1967) to allow the use of interval–censored data.

The Wilcoxon–Peto (WP) score for each observation is the difference between Turnbull's estimated proportion of time observations that are to the left and Turbull's estimated proportion of time observations that are to the right, that is,

$$WPc_i = \hat{W}(L_i^-) - (1 - \hat{W}(R_i)) = \hat{W}(L_i^-) + \hat{W}(R_i) - 1.$$

Note that $\hat{W}$ is Turnbull's estimator for the pooled sample given in (4). This proposal is introduced by Peto and Peto (1972) and it is asymptotically efficient for time distributions in the logistic family.

In the same article Peto and Peto extend the Savage or Log-rank (LR) test to interval–censored data. The Log-rank scores are,

$$LRc_i = \frac{(1 - \hat{W}(R_i))\log(1 - \hat{W}(R_i)) - (1 - \hat{W}(L_i^-))\log(1 - \hat{W}(L_i^-))}{\hat{W}(R_i) - \hat{W}(L_i^-)},$$

where again $\hat{W}$ is given in (4). This proposal is asymptotically efficient for time distributions with Lehmann–type alternatives.

Fay and Shih (1998) introduce what they call distribution permutation tests, which provides another interesting approach to the k–sample problem. These are permutational tests with scalar scores obtained as follows: an estimate of the distribution function for each observation is compared to the overall Turnbull's estimate of the distribution function. The use of the self–consistent equations allow them to define these empirical estimates of the distribution function for each observation. For particular ways of comparing these estimated distributions Fay and Shih obtain the Wilcoxon–Peto test, the Log–rank test and a new test called the difference in means (DiM) test. In what follows we describe the difference in means test as an extension of the permutational t–test. In order to calculate the total mean of the distribution induced by $\hat{W}$, they identify each Turnbull's interval $[q_j, p_j]$ with the right endpoint p_j and assign all the probability of $[q_j, p_j]$, $\hat{w}_j$, to p_j. When $p_m = \infty$, they let $p_m = q_m$. The use of this distribution allows, as well, to compute for the i^{th} individual the imputed mean value of its interval, that is, the conditional expectation of T given that $T \in [L_i, R_i]$. Because of the self–consistent property of Turnbull's estimate, the mean of these imputed means is equal to the total mean of the distribution. The scalar score they propose for each individual is the difference between the above

imputed mean value and the total mean, that is,

$$DiMc_i = \frac{\sum_{j=1}^{m} p_j \hat{w}_j \alpha_j^i}{\hat{W}(R_i) - \hat{W}(L_i^-)} - \sum_{j=1}^{m} p_j \hat{w}_j.$$

Example: We use the fifth interval observation, $[2,4]$, in the example in Section 3.1 to illustrate the computation of the different scores. The only interval observation that is to the left to $[2,4]$ is the interval $[0,1]$, while to its right is $[5,7]$. Thus, the Wilcoxon–Gehan score value is, $WGsc_5 = 1 - 1 = 0$. Wilcoxon–Peto score value is $WPsc_5 = \frac{1}{4} - \frac{3}{8} = -0.125$, because the probability mass assigned by Turnbull's distribution function to the interval $[0,1]$ is $\hat{w}_1 = \frac{1}{4}$ and to the interval $[5,6]$ is $\hat{w}_4 = \frac{3}{8}$. The Log–rank score value is given by,

$$LRsc_5 = \frac{(1 - \hat{W}(4))\log(1 - \hat{W}(4)) - (1 - \hat{W}(2^-))\log(1 - \hat{W}(2^-))}{\hat{W}(4) - \hat{W}(2^-)} = -0.4055.$$

Since the interval $[2,4]$ contains Turnbull's intervals $[2,3]$ and $[4,4]$, with respective probability mass $\hat{w}_2 = 0.25$ and $\hat{w}_3 = 0.125$, the imputed mean is, $(3{\cdot}0.25{+}4{\cdot}0.125)/(0.25{+}0.125) = 3.33$. Furthermore, the total mean using Turnbull's estimate of the distribution function is, $1{\cdot}0.25{+}3{\cdot}0.25{+}4{\cdot}0.125{+}6{\cdot}0.375 = 3.75$. Therefore, the score value is given by, $DiMsc_5 = 3.33 - 3.75 = -0.4267$.

4.2 Illustration 3

Another instance of interval-censored data is found in an AIDS Clinical Trial designed to study the benefits of zidovudine therapy in patients in the early stages of the human immunodeficiency virus (HIV) infection (Volberding *et al.*, 1995). The design compares three groups. The first group, G_1, corresponds to those patients who started zidovudine monotherapy after their CD4 cell count fell below 500 per cubic millimeter. In the second and third groups, G_2 and G_3, two different dosages of zidovudine were given immediately after randomization. Among the 1607 subjects who could be evaluated, 541 were in the deferred-therapy group, 538 in the 500–mg group and 528 in the 1500–mg group. Subjects were followed prospectively until the development of AIDS or death. As a measure of the clinical progression of the disease, CD4 cell counts were periodically determined. The reported data included the times of the first count below 500 cells per cubic millimeter, as well as below 400 and below 300. We will focus on the time T, measured in months from randomization, until the CD4 count first reaches 400 cells per cubic millimeter. The random variable T is interval-censored, that is, for each individual i, we know that T_i is between L_i and R_i where R_i is the time of the first visit when CD4 was observed to be below 400 cells per cubic millimeter and L_i is defined to be the time of the preceding visit.

We illustrate now the above permutational methodology with the comparison of the survival of these three groups ($k = 3$). The choice for the $\mathbf{z}_i$ covariates is

the following,

$$\mathbf{z'}_i = \left(\frac{\alpha_i^{(1)}}{\sqrt{n_1}}, \frac{\alpha_i^{(2)}}{\sqrt{n_2}}, \frac{\alpha_i^{(3)}}{\sqrt{n_3}}\right) = \left(\frac{\alpha_i^{(1)}}{23.2594}, \frac{\alpha_i^{(2)}}{23.1948}, \frac{\alpha_i^{(3)}}{22.9783}\right)$$

where $\alpha_i^{(j)}$ is an indicator function that is equal to 1 if the individual belongs to group G_j and 0 otherwise. Then the linear permutational statistic form simplifies to the expression,

$$L_0 = \sum_{i=1}^{n} \mathbf{z}_i c_i = \begin{pmatrix} \sqrt{n_1}\ \bar{c}_{(1)} \\ \sqrt{n_2}\ \bar{c}_{(2)} \\ \sqrt{n_3}\ \bar{c}_{(3)} \end{pmatrix} = \begin{pmatrix} 23.2594\ \bar{c}_{(1)} \\ 23.1948\ \bar{c}_{(2)} \\ 22.9783\ \bar{c}_{(3)} \end{pmatrix},$$

where $\bar{c}_{(j)} = \frac{1}{n_j}\sum_{i=1}^{n} c_i \alpha_i^{(j)}$. The permutational distribution of L_0 is asymptotically distributed as a k–dimensional normal and we can use the Mahalanobis distance (Md) to obtain a $\chi^2_{k-1} = \chi^2_2$ distribution:

$$Md = L_0' V^{-} L_0 = \frac{n-1}{\sum_{i=1}^{n} c_i^2} \sum_{j=1}^{k} n_j \bar{c}^2_{(j)} = \frac{1606}{\sum_{i=1}^{n} c_i^2}(541\ \bar{c}^2_{(1)} + 538\ \bar{c}^2_{(2)} + 528\ \bar{c}^2_{(3)}),$$

where V^{-} is the generalized inverse of $\mathrm{Var}(L_0)$. The results using each of the permutational tests (see Table 1) show significant evidence of the differences between the survival curves. In this paper, the data are again analyzed in Subsection 5.2 to illustrate the nonparametric Bayesian method for interval–censored data.

Taula 1: Permutational test statistic (L_0) for different score choices, the related Mahalanobis distance (Md) and p–values for the null hypothesis of equal distributions: $H_0 : W_1 = W_2 = W_3$ *versus* the alternative of some differences between the distributions $H_a : W_i \neq W_j$ for some i,j

	Wilcoxon–Gehan	*Wilcoxon–Peto*	*Log–Rank*	*Difference in Means*
L_0	$\begin{pmatrix} -1804.732 \\ 337.9202 \\ 1485.709 \end{pmatrix}$	$\begin{pmatrix} -1.5351 \\ 0.2687 \\ 1.2826 \end{pmatrix}$	$\begin{pmatrix} -2.2098 \\ 0.3449 \\ 1.8887 \end{pmatrix}$	$\begin{pmatrix} -84.0323 \\ 16.4337 \\ 68.4719 \end{pmatrix}$
Md	16.3978	16.6800	17.6607	17.8151
p-value	0.000275	0.000239	0.000146	0.000135

4.3 Likelihood approaches

In this section we review different papers that introduce test statistics derived directly from the likelihood function. The first two papers derive equivalent forms of the Wilcoxon–Peto and the Log–rank tests from regression models. Finkelstein (1986) proposes an extension to interval–censored data of the proportional hazards model. Finkelstein assumes a discrete interval–censored time distribution and derives, from the likelihood function, the score vector that results for testing the hypothesis of a null regression coefficient. This statistic has the form $\sum(O - E)$ and it can be seen as the Log–rank test proposed by Peto and Peto. Because of the discrete nature of the data, Finkelstein uses the Fisher information matrix to derive the asymptotic distribution of the statistic instead of the permutational distribution. Their approach, however, produces numerical problems when applied to a large group of patients. Fay (1996) extends Finkelstein's work to the grouped continuous model. The score vector for testing the null hypothesis that the failure times are unrelated to the covariates, reduces to the Wilcoxon-Peto or the Log–rank tests as special cases. Fay (1999) shows the equivalence between the weighted Log–rank form of these score vectors given by $\sum w \cdot (O - E)$ and the permutational linear form (equation 10).

The approach by Petroni and Wolfe (1994) is different from all the above methods. Their proposal is a class of two sample tests based on Turnbull's estimated survival function from each group and requires a finite pre–specified number of intervals. These tests are based on the integrated weighted difference in Turnbull's estimators and extend the weighted Kaplan–Meier class developed by Pepe and Fleming (1989) for right–censored data. Under the null hypothesis of no difference between the distributions, the distribution of these tests is asymptotically normal and the variance is obtained via information matrices. This approach is specially indicated under crossing hazard alternatives.

4.4 Computational aspects

The following four S–Plus routines: `WGsc(·,·)`, `WPsc(·,·,·)`, `LRsc(·,·,·)` and `DiMsc(·,·,·)` implement, respectively, the Wilcoxon–Gehan scores, the Wilcoxon–Peto scores, the Log–Rank scores, and the Difference in Means scores. The test statistic can be computed from each set of scores using either the two sample methodology (`w2test(·,·)`), or the k–sample methodology, (`wktest(·,·)`). We illustrate these routines with the k–sample Wilcoxon–Peto test. First, we estimate the survival function from the pooled sample using Turnbull's method,

```
surv.est<-kaplanMeier(censor(lower,upper,censor.codes)~1,
data=int.data)
```

Then, we compute the Wilcoxon–Peto scores,

```
scores<-WPsc(lower,upper,surv.est)[[6]]
```

Afterwards we create a vector of covariates, `covar`, that assigns the value 1 for individuals in the first group, the value 2 for individuals in the second group and likewise until the k^{th} group. The `wktest(·,·)` routine would transform each covariate value s in a k–vector whose s–component is $1/\sqrt{n_s}$ and the rest of the components are 0. At last, we compute the permutational test statistic and the

corresponding Mahalanobis distance, `wktest(scores,covar)`

5 Bayesian Approach

The Bayesian approach is tempting in survival analysis because of the direct probabilistic interpretation of the posterior distribution and because many problems can be formulated in terms of integrals with respect to the posterior distribution. Furthermore, this framework allows the incorporation of prior beliefs about the distribution function. The reason why Bayesian methods had not been widely used in survival analysis until the last few years is because, for realistic models, the posterior distribution under censoring is extremely difficult to obtain directly. The development of new numerical algorithms, such as Markov chain Monte Carlo algorithms, which allow to obtain a sample from the posterior of interest has open the door to the use of Bayesian methods to survival analysis.

In this section we discuss both parametric and nonparametric approaches to interval–censored data. The review paper by Sinha and Dey (1997) and the recent book by Ibrahim, Chen and Sinha (2001) give details on semiparametric Bayesian models.

5.1 Bayesian Parametric Approach

As usual, let $T_1, \ldots, T_n$ be the potential times for the n individuals and denote by $\mathcal{D} = \{[L_i, R_i], \quad 1 \leq i \leq n\}$ the observed censoring intervals. We assume that $T_1, \ldots, T_n$ are independent and identically distributed with density function $w(t;\theta)$. As in section 3.3, the likelihood function $L(\theta|\mathcal{D})$ is given by (9) if we assume that the censoring occurs noninformatively. By means of Bayes theorem and after assuming a prior distribution $p(\theta)$ for θ, the posterior distribution of θ is given by:

$$p(\theta|\mathcal{D}) = \frac{L(\theta|\mathcal{D}) \cdot p(\theta)}{\int L(\theta|\mathcal{D}) \cdot p(\theta)\, d\theta}.$$

Usually the integral in the denominator does not admit an explicit solution and numerical methods are needed to obtain the posterior distribution function. As suggested in Smith and Roberts (1993), the Gibbs sampler is a very useful method in problems involving incomplete or censored data. The unobserved data are reintroduced in the model as further unknowns and this leads in general to more tractable situations. This strategy of introducing additional or latent variables in the model is also called the *data augmentation algorithm* (Tanner and Wong, 1987).

5.1.1 Data augmentation method

The basic idea behind the data augmentation algorithm is the following: Let $p(x)$ be the distribution of interest which does not have an explicit form and is difficult to sample from. Let y be an additional variable, which is referred to as *latent variable*, so that we can calculate or sample from $p(x|y)$ and also from

$p(y|x)$. The data augmentation algorithm consists on sampling iteratively from these two conditional distributions. That is, given an initial value $x^{(0)}$, draw a value $y^{(1)}$ from $p(y|x^{(0)})$ and then draw a value $x^{(1)}$ from $p(x|y^{(1)})$. Tanner and Wong (1987) proved that performing iteratively these two steps provides pairs $(X^{(i)}, Y^{(i)})$ such that the sequence $X^{(i)}$ converges in distribution to a variable X with distribution $p(x)$ and the sequence $Y^{(i)}$ converges in distribution to a variable Y with distribution $p(y)$.

In our setting the distribution of interest is $p(\theta|\mathcal{D})$ and the latent variables which are introduced in the model as additional parameters are the censored times $T_1, \ldots, T_n$. Then, the Gibbs sampler consists in sampling iteratively from $p(T_i|\theta, \mathcal{D})$, for each i and from $p(\theta|T_1, \ldots, T_n, \mathcal{D})$. In the first step each censored time is imputed; this produces as a result a complete data set. In the second step, since the noninformative condition implies that $p(\theta|T_1, \ldots, T_n, \mathcal{D}) = p(\theta|T_1, \ldots, T_n)$, the parameter θ is updated based on the complete imputed sample. The successive implementation of these two steps provides a sample of the parameter θ which, under weak conditions (Gelfand and Smith, 1990), converges to the posterior distribution of θ. Averages from these samples are used to estimate posterior quantities.

This data augmentation scheme also applies to the analysis of regression models where the parameter θ in the parametric distribution is related to some covariates $x_1, \ldots, x_k$ through a link function $\theta = g(x_i, \beta)$. The goal in this case is the estimation of the regression parameter β. The Gibbs sampling algorithm to obtain the posterior distribution of β is given by the the successive iteration of the following steps:

1. Impute a value T_i sampled from $w(t;\theta)$ truncated in the interval $[L_i, R_i]$.
2. Update the value of β sampling from the posterior distribution $p(\beta|T_1, \ldots, T_n)$ where

$$p(\beta|T_1, \ldots, T_n) = \frac{\prod_{i=1}^{n} w(T_i; \theta = g(x_i, \beta)) \cdot p(\beta)}{\int \prod_{i=1}^{n} w(T_i; \theta = g(x_i, \beta)) \cdot p(\beta)\, d\beta}$$

   and $p(\beta)$ is the prior density for the regression parameter.
3. Update the value of $\theta = g(x_i, \beta)$.

5.1.2 Computational aspects

The program BUGS, which stands as an acronym for Bayesian inference Using Gibbs Sampling, is a very useful tool for the implementation of this algorithm. This program provides a language for specifying complex Bayesian models and performs the Gibbs sampler by simulating from the full conditional distributions. Further details of the program are given in Spiegelhalter *et al.*(1996). The software is freely available at http://www.mrc-bsu.cam.ac.uk/bugs/ .

5.1.3 Illustration 2

We reanalyze the data from a cohort of hemophiliacs described in Illustration 2 assuming a log-normal model for the time to HIV infection. For each individual T_i denotes its infection time which is interval censored in $[L_i, R_i]$. The covariate x_i indicates the treatment group: $x_i = 0$ for the heavily-treated group and $x_i = 1$ for the lightly-treated group. The model assumptions and prior specifications can be expressed through the following hierarchical model:

$$
\begin{array}{lrcl}
\text{[Stage1]} & T_i & \sim & \log N(\mu_i, \sigma^2) \text{ truncated in } [L_i, R_i] \\
 & \mu_i & = & \beta_0 + \beta_1 \cdot x_i \\
 & & & \\
\text{[Stage2]} & \beta_0 & \sim & N(\alpha_0, \sigma_0^2) \\
 & \beta_1 & \sim & N(\alpha_1, \sigma_1^2) \\
 & \sigma^2 & \sim & IG(0.001, 0.001) \\
 & & & \\
\text{[Stage3]} & \alpha_0 & \sim & N(0, 1.10^{-6}) \\
 & \sigma_0^2 & \sim & IG(0.001, 0.001) \\
 & \alpha_1 & \sim & N(0, 1.10^{-6}) \\
 & \sigma_1^2 & \sim & IG(0.001, 0.001)
\end{array}
$$

In stage 1 we specify the observational model: for each individual we assume a log-normal model truncated in the corresponding censoring interval. The mean μ_i is assumed to be equal to β_0 for the heavily–treated group and equal to $\beta_0 + \beta_1$ for the lightly–treated group. The normal prior distributions for these parameters are specified in stage 2 and an inverse gamma distribution for the variance. In stage 3 we specify vague priors for the hyperparameters.

The analysis was performed using BUGS. We implemented 2000 iterations of the algorithm and the results are in Table 2. For illustrative purposes we give in Figure 4 the posterior distribution for β_0 and β_1.

Taula 2: Posterior means and 95% credible intervals for Illustration 1

Parameter		
β_0	mean	2.422
	95% credible interval	(2.345, 2.494)
β_1	mean	0.2401
	95% credible interval	(0.1383, 0.3411)
σ	mean	0.3635
	95% credible interval	(0.3176, 0.4151)

Using these results and the expression of the mean of a lognormal distribution ($\mu_T = exp(\mu + 0.5 \cdot \sigma^2)$), we obtain that the mean infection time for the heavily–

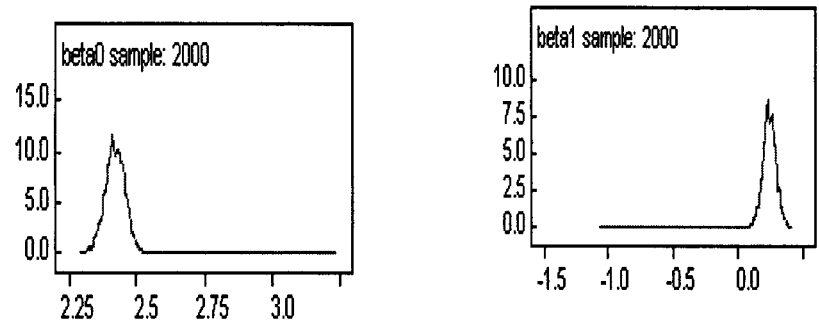

Figura 4: Posterior distribution for β_0 and β_1

treated group is 12.03 (which corresponds to 6 years) while for the lightly–treated group is 15.3 (approximately 7.6 years). In Figure 5 we have plotted the distribution functions of infection time for both groups. We can observe that the lightly–treated group has larger infection times than the heavily–treated group.

5.2 Nonparametric Bayesian Approach

Here we describe the analysis of interval–censored data in the Bayesian paradigm without the assumption of any parametric model. Susarla and Van Ryzin (1976) were the first to derive a nonparametric Bayesian estimator (NPBE) of the survival function for right-censored data. Their estimator is based on the class of Dirichlet processes *a priori* introduced by Ferguson (1973). They proved that the nonparametric Bayesian estimator includes the Kaplan-Meier estimator as a special case, both estimators are asymptotically equivalent and that NPBE has better small sample properties than the Kaplan-Meier estimator (Rai *et al*, 1980). The extension of this approach to more complex censoring situations and, in particular, to interval censoring is not straightforward. Nonparametric Bayesian estimators of the survival curve have only been obtained in an explicit way for special cases of interval–censored data. For instance, Johnson and Christensen (1986) obtain explicit formulas for the survival curve estimator using a Dirichlet process prior for the special case of nested interval data. What is meant by nested interval is that the intervals do not overlap, that is, given two censoring intervals, either one interval is contained into the other or they both are disjoint.

Since for a more general situation the estimation of the survival curve cannot be achieved explicitly, computing intensive methods can provide a solution. Doss (1994) propose a Gibbs sampling algorithm to deal with interval censoring based on the simulation of samples from the Dirichlet process. In what follows we propose an alternative approach (Calle and Gómez, 2001) which, by means of the use of latent count variables, only require simulation from a Dirichlet

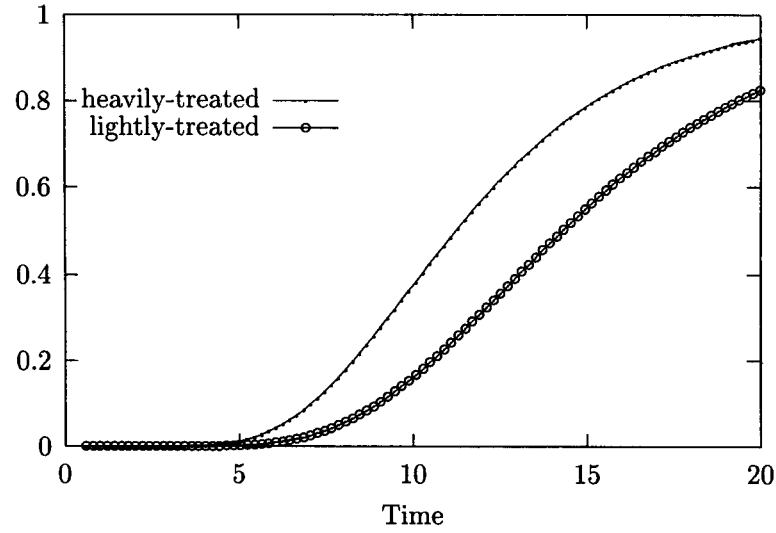

Figura 5: Estimated cumulative distributions of times to HIV infection

distribution.

5.2.1 Calle and Gómez estimator

Let $T_1, \ldots, T_n$ be the sample of the potential times for the n individuals and denote by $\mathcal{D} = \{[L_i, R_i], \quad 1 \leq i \leq n\}$ the observable data. Let $\{0, t_1, \ldots, t_{r-1}, t_r = \infty\}$ denote the unique ordered elements of the lower and upper limits of the censoring intervals $\{[L_i, R_i], \ i = 1, \ldots, n\}$ and define w_j as the probability of T being between t_{j-1} and t_j. Assume that there is some prior belief in the shape of the distribution function that can be summarized by a parametric model W_0 for the distribution function W. The uncertainty on this parametric form W_0 is modeled through a prior Dirichlet distribution for $(w_1, \ldots, w_r)$ with parameters $\alpha_j = \beta(W_0(t_j) - W_0(t_{j-1}))$, $j = 1, \ldots, r$ where β is a positive real number that represents the precision or the measure of faith in the prior guess W_0.

Since the posterior distribution of the vector $\mathbf{w}$ given a sample from a Dirichlet process, only depends on the number of events, n_j, that fall in the interval $(t_{j-1}, t_j]$, and not on where they fall exactly, the posterior distribution of $\mathbf{w}$ can be derived by introducing the vector $\mathbf{n} = (n_1, \ldots, n_r)$ in the model as a latent variable. If $\delta_j^i = \mathbf{1}\{T_i \in (t_{j-1}, t_j]\}$ is an indicator for the i^{th} individual that represents whether or not the event has occurred in the j^{th} interval, then every component n_j can be expressed as $n_j = \sum_{i=1}^{n} \delta_j^i$. As a matter of fact $\boldsymbol{\delta}^i = (\delta_1^i, \ldots, \delta_r^i)$ is a vector such that every component equals zero, except one. We assume that the prior distribution of $\boldsymbol{\delta}^i$ conditioned to $\mathbf{w}$ and $\mathcal{D}$ is a multinomial distribution of sample size 1.

To obtain the posterior distribution of the vector $\mathbf{w}$, given the data $\mathcal{D}$, a

multiple sequence Gibbs sampling method is proposed. The following algorithm iterates alternatively from the posterior conditional distribution of $\mathbf{n}$ given $\mathbf{w}$ and from the posterior conditional distribution of $\mathbf{w}$ given $\mathbf{n}$.

For each sequence $m = 1, \ldots, M$ we perform the following steps:

A. Initial values: Define the initial probabilities $\mathbf{w_m^{(0)}} = (w_{m1}^{(0)}, \ldots, w_{mr}^{(0)})$.

B. Updated $\mathbf{n}$: For each individual $i = 1, \ldots, n$, generate $\delta^{\mathbf{i}}$ from a truncated multinomial of sample size 1 and parameters $\mathbf{w_m^{(0)}}$. Compute $n_j^{(0)} = \sum_{i=1}^{n} \delta_j^i$, the number of events in each interval $(t_{j-1}, t_j]$.

C. Updated $\mathbf{w}$: Generate $\mathbf{w_m^{(1)}} = (w_{m1}^{(1)}, \ldots, w_{mr}^{(1)})$ from a Dirichlet distribution of parameter vector $(\alpha_1 + n_1^{(0)}, \ldots, \alpha_r + n_r^{(0)})$.

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

Repeat steps B, C and D until convergence.

It can be shown under rather weak conditions (Gelfand and Smith, 1990) that the Markovian sequence $(\mathbf{w}^{(\ell+1)}, \mathbf{n}^{(\ell)})$ converges to an equilibrium distribution that is the joint distribution of $(\mathbf{w}, \mathbf{n})$. After generating M samples from Gibbs sampling chains one can approximate the marginal posterior distribution of $\mathbf{w}$ by the empirical sampling distribution, or by using the average of the posterior conditional distributions of $\mathbf{w}$ given $\mathbf{n}$. Since the distribution function at time t_j can be expressed as $W(t_j) = \sum_{s \leq j} w_s$, a sample from the posterior distribution of $W(t_j)$ can as well be derived.

Calle and Gómez (2001) illustrate the effect of different prior weights (β) of the Dirichlet process. The Bayesian estimator is shown to be close to the nonparametric maximum likelihood estimator as the prior weight of the Dirichlet process approaches zero. On the other hand, as the prior weight increases, the Bayesian estimator approaches the parametric prior guess W_0. To illustrate this behaviour, Doss and Narasimhan (1998) describe an importance sampling scheme which allows the dynamic display of the changing estimated survival curves for different prior hyperparameters. This can be used to show how the nonparametric Bayesian estimator based on a Dirichlet process prior is a compromise between purely parametric and purely nonparametric estimators.

5.2.2 Computational aspects

The methodology has been implemented in a C-language program, ICGIBBS.C. The data file has the same structure that the one required for the program MODGL.C in Subsection 3.2, that is, 2 columns and $n + 1$ rows. The first row includes the sample size n and the number that plays the role of infinity. The following n rows include the left and the right endpoint of the censoring interval for each individual.

5.2.3 Illustration 3

The above methodology is illustrated in Calle and Gómez (2001) with data from the AIDS Clinical Trials Group protocol 019 described in Subsection 4.2. The variable of interest is the time T, measured in months from randomization, until the CD4 count first reaches 400 cells per cubic millimeter and it is interval-censored.

Taula 3: Mean and posterior 95% credible interval (in parentheses) of the survival function by treatment group

Month	Treatment group		
	Deferred therapy	500 mg ZDV	1500 mg ZDV
12	0.63 (0.61,0.65)	0.68 (0.65,0.70)	0.71 (0.69,0.73)
24	0.46 (0.42,0.48)	0.55 (0.53,0.57)	0.59 (0.56,0.61)
36	0.36 (0.33,0.38)	0.46 (0.43,0.48)	0.49 (0.46,0.51)
48	0.29 (0.26,0.31)	0.38 (0.35,0.40)	0.43 (0.40,0.45)
60	0.25 (0.22,0.27)	0.32 (0.29,0.34)	0.39 (0.36,0.42)
72	0.19 (0.17,0.25)	0.27 (0.26,0.32)	0.32 (0.30,0.38)

The Bayesian estimators were obtained through the implementation of the Gibbs sampling scheme described above taking $M = 5$ independent sequences and $i = 2000$ iterations in each sequence and discarding the first 500 iterations. Convergence of the Gibbs sampler was established both graphically and numerically using the program CODA (Best *et al.*, 1995). Figure 6 shows the estimated survival function for T according to treatment group and using $\beta = \sqrt{n}$. The survival curves suggest differences between the deferred–therapy group and the immediate–therapy groups (500-mg and 1500-mg). In particular, the median time to a CD4 cell count equal 400 in the immediate–therapy groups is approximately 32 months while the median time is 20 months in the deferred–therapy group. Table 3 gives the mean and the posterior pointwise 95% credible interval of the survival function by treatment group, every 12 months. We observe that the mean survival time is always smaller in the deferred therapy group than in the immediate therapy groups. For instance, if we focus on 48 months, we see that the probability that the time to reach 400 cells be larger than 4 years is 29% in the deferred therapy group while it is around 40% in the immediate therapy groups. We can also see that, after 60 months of randomization, only 25% of the patients have a CD4 cell count above 400 in the deferred therapy group while this percentage is 32% and 39% in the other groups. This behaviour remains the same at any time in the study. Furthermore, while the credible intervals for the immediate–therapy groups overlap, the corresponding credible interval for the deferred–therapy group lies always below the other two, which indicates that the observed differences in the survival times between these groups are significant. Therefore, the CD4 cell counts in the immediate–therapy groups declined significatively more slowly than those of the deferred–therapy group.

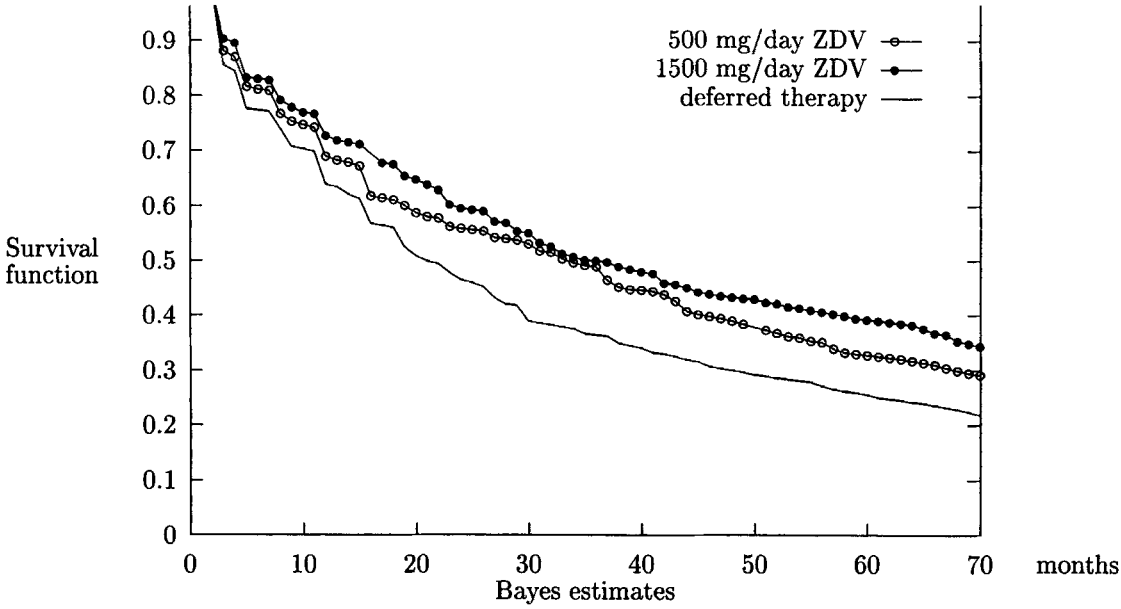

Figura 6: Estimated survival function for the elapsed time to CD4= 400, according to treatment group.

Acknowledgements

This survey paper is the result of two conferences and a seminar on interval-censored data. The authors are grateful to the many suggestions given by the audience and in particular to the GRASS group for their fruitful discussions. The authors are grateful to the doctors from Hospital Universitari Germans Trias i Pujol and to the AIDS Clinical Trial Group from Harvard University for providing the data used in illustrations 1 and 3, respectively. This research was partially supported by the Dirección General de Enseñanza Superior e Investigación Científica Grant PB98–0919.

References

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

Böhning, D., Schlattmann, P. and Dietz, E. (1996) Interval–censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika* **83**, 462–466.

Brookmeyer, R. and Goedert, J.J. (1989) Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* **45**, 325–335.

Calle, M.L. and Gómez, G. (2001) Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference* **98**, 73–87.

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.

Courgeau, D. and Najim, J. (1996) Interval–censored event history analysis. *Population: An English Selection*, **8**, 191–298.

De Gruttola, V. and Lagakos, S.W. (1989) Analysis of doubly censored survival data, with application to AIDS. *Biometrics* **45**, 1–11.

Doksum, K. (1974) Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

Doss, H. (1994) Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22**, 1763–1786.

Doss, H. and Narasimhan (1998) Dynamic diplay of changing posterior in Bayesian survival analysis, in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. Dey, D. Müller, P. and Sinha, D.), New York: Springer-Verlag, 63–87.

Fay, M.P. (1996) Rank invariant tests for interval-censored data under the grouped continuous model. *Biometrics* **52**, 811–822.

Fay, M.P. (1999) Comparing several score tests for interval-censored data. *Statistics in Medicine* **18**, 273–285.

Fay, M.P. and Shih, J.H. (1998) Permutation tests using estimated distribution functions. *Journal of the American Statistical Association* **93**, 387–396.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Finkelstein, D.M. (1986) A proportional hazards models for interval-censored failure time data. *Biometrics* **42**, 845–854.

Fleming, T.R. and Harrington, D.P. (1991) *Counting processes and survival analysis.* New York: John Wiley and Sons.

Gehan, E.A. (1965) A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* **52** 203–223.

Gelfand, A.E. and Smith, F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gentleman, R. and Geyer, C.J. (1994). Maximum-likelihood for interval-censored data: Consistency and computation. *Biometrika* **81**, 618–623.

Goetghebeur, E. and Ryan, L. (2000) Semiparametric regression analysis of interval–censored data. *Biometrics* **56**, 1139–1144.

Goggins, W. B. and Finkelstein, D.M. (2000) A proportional hazards models for multivariate interval-censored failure time data. *Biometrics* **56**, 940–943.

Gómez, G. and Calle, M.L. (1999) Nonparametric estimation with doubly censored data. *Journal of Applied Statistics* **26(1)**, 45–58.

Gómez, G., Calle, M.L., Muga, R. and Egea, J.M. (2000) Estimation of the risk of HIV Infection as a function of the length of intravenous drug use. A nonparametric Bayesian approach. *Statistics in Medicine.* **19**, 2641–2656

Gómez, G.and Julià, O. (1990) Estimation and asymptotic properties of the distribution of time-to-tumour in carcinogenesis experiments. *IMA Journal of Mathematics Applied in Medicine and Biology* **7**, 109–123.

Gómez, G., Julià, O. and Utzet, F. (1992) Survival Analysis for Left Censored Data. *Survival Analysis: State of the Art.* Editors: J.P. Klein and P.K. Goel. Kluwer Academic Publishers. ISBN 0-7923-1634-7.

Gómez, G., Julià, O. and Utzet, F. (1994) Asymptotic properties of the left Kaplan-Meier estimator. *Communications in Statistics: Theory and Methods* **23**, 123–135.

Gómez, G. and Lagakos, S. (1994) Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics* **50**, 204–212.

Gómez, G and van Ryzin, J. (1992) Estimation of the subsurvival function for time-to-tumor in survival/sacrifice experiments. *Statistics and Probability Letters* **13**, 5–13.

Groeneboom, P. and Wellner, J.A. (1992) *Information bounds and nonparametric maximum likelihood estimation* Basel: Birkhäuser Verlag.

Ibrahim, J.G., Chen, M.H and Sinha, D. (2001) *Bayesian Survival analysis.* New York: Springer–Verlag.

Johnson, W. and Christensen, R. (1986) Bayesian nonparametric survival analysis for grouped data. *The Canadian Journal of Statistics* **14**, 307–314.

Kooperberg, C. and Clarkson, D.B. (1997) Hazard regression with interval-censored data. *Biometrics* **53**, 1485–1494.

Lindsey, J.C. (1998) A study of interval censoring in parametric regression models. *Lifetime Data Analysis* **4**, 329–354.

Lindsey, J.C. and Ryan, L.M. (1998) Tutorial in Biostatistics. Methods for interval–censored data. *Statistics in Medicine* **17**, 219–238

Mantel, N. (1967) Ranking procedures for arbitrarily restricted observation. *Biometrics* **23**, 65–78.

Muñoz, A. and Xu, F. (1996) Models for the incubation of AIDS and variations according to age and period. *Statistics in Medicine* **15**, 2459–2473.

Ng, M.P (2002) A modification of Peto's nonparametric estimation of survival curves for interval-censored data. *Biometrics* **58**, 439–442.

Pan, W. and Chappell, R. (2002) Estimation in the Cox proportional hazards models with left–truncated and interval-censored data. *Biometrics* **58**, 64–70.

Pepe, M.S. and Fleming, T.R. (1989) Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497–507.

Peto, R. and Peto, J. (1972) Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A, General* **135**, 185–207.

Peto, R. (1973) Experimental survival curves for interval-censored Data. *Journal of the Royal Statistical Society, Series C* **22**, 86–91.

Petroni, G.R. and Wolfe, A. (1994) A two sample test for stochastic ordering with interval-censored data. *Biometrics* **50**, 77–87.

Rai, K., Susarla, V. and van Ryzin, J. (1980) Shrinkage estimation in nonparametric Bayesian survival analysis: A simulation study. *Communications in Statistics: Simulation and Computation* **3**, 271–298.

Schick, A. and Yu, Q. (2000) Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics* **27**, 45–55.

Sinha, D. and Dey, D.K. (1997) Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195–1212.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B,* **55**, 3–23.

Smith, P.J., Thompson, T.J. and Jereb, J.A. (1997) A model for interval-censored tuberculosis outbreak data. *Statistics in Medicine* **16**, 485–496.

Spiegelhalter, D. *et al.* (1996) Bayesian Inference Using Gibbs Sampling, Version 0.5, (version ii). *MRC Biostatistics Unit, Cambridge.*

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.

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.

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.

Volberding, P.A., Lagakos, S.W., Grimes, J.M., *et al.* (1995) A comparison of immediate with deferred zidovudine therapy for asymptomatic HIV-infected adults with CD4 cell counts of 500 or more per cubic millimeter. *New England Journal of Medicine* **333**, 401–451.

Younes, N. and Lachin, J. (1997) Link-based models for survival data with interval and continous time censoring. *Biometrics* **53**, 1199–1211.

Yu, Q., Schick, A., Li, L. and Wong, G.Y.C. (1998) Asymptotic properties of the GMLE with case 2 interval-censored data. *Statistics and Probability Letters* **37**, 223–228.

Yu, Q., Li, L. and Wong, G.Y.C. (2000) On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics* **27**, 35–44.