THE ANALYSIS OF INTERVAL CENSORING AND DOUBLE CENSORING VIA MARKOV CHAIN MONTE CARLO METHODS

M. LUZ CALLE

Grup de Recerca en Modelització de Sistemes Biològics, Departament d'Informàtica i Matemàtica, Escola Politècnica Superior, Universitat de Vic, Carrer de la Sagrada Família, 7 – 08500 Vic, e-mail: callem@uvic.es

Data de recepció: 18/05/02
Data de publicació: 31/05/02

ABSTRACT

Survival analysis is used in different fields to analyze the elapsed time between two events. What distinguishes survival analysis from other areas in statistics is that data are usually censored. Interval censoring arises when the occurrence of the final event of interest cannot be exactly observed and the failure time is only known to lie in an interval. A more complex censoring scheme is found when both initial and final times are interval–censored. This situation is referred as double censoring. In this paper we provide a formal description of a parametric Bayesian method for the analysis of interval–censored and doubly–censored data and clear guidelines for its practical use. The proposed methodology is illustrated with data from a cohort of hemophilia patients who were infected with HIV in the early 1980's.

RESUM

L'Anàlisi de la supervivència s'utilitza en diferents camps per analitzar el temps transcorregut entre dos esdeveniments. El que distingeix l'anàlisi de la supervivència d'altres àrees de l'estadística és que les dades normalment estan censurades. La censura en un interval apareix quan l'esdeveniment final d'interès no és directament observable i només se sap que el temps de fallada està en un interval concret. Un esquema de censura més complex encara apareix quan tant el temps inicial com el temps final estan censurats en un interval. Aquesta situació s'anomena doble censura. En aquest article donem una descripció formal d'un mètode bayesià paramètric per a l'anàlisi de dades censurades en un interval i dades doblement censurades així com unes indicacions clares de la seva utilització pràctica. La metodologia proposada s'ilustra amb dades d'una cohort de pacients hemofílics que es varen infectar amb el virus VIH a principis dels anys 1980's.

RESUMEN

El análisis de la supervivencia se utiliza en diferentes campos para analizar el tiempo transcurrido entre dos sucesos. Lo que distingue el análisis de la supervivencia de otras áreas de la estadística es que los datos normalmente están censurados. La censura en un intervalo aparece cuando el suceso final de interés no es directamente observable y sólo se sabe que el tiempo de fallo está en un intervalo concreto. Un esquema de censura más complejo todavía aparece cuando tanto el tiempo inicial como el tiempo final están censurados en un intervalo. Esta situación se denomina doble censura. En este articulo

damos una descripción formal de un método bayesiano paramétrico para el análisis de datos censurados en un intervalo y datos doblemente censurados así como unas indicaciones claras de su utilización práctica. La metodología propuesta se ilustra con datos de una cohorte de pacientes hemofílicos que se infectaron con el virus VIH a principios de los años 1980.

1 Introduction

Survival or time to event analysis is the term used to describe the methodologies for analyzing duration times between two events. To determine the survival times it is necessary to define two time points: the origin time corresponding to the time at which an original event occurs and the failure time corresponding to the time at which the final event occurs. A common problem in many time-to-event studies is that the occurrence of the final event of interest cannot be exactly observed and the failure time is only known to lie in an interval. For each individual i we observe an interval $[X_L^i, X_R^i]$ that contains the survival time X^i which is said to be interval–censored. This happens, for instance, in longitudinal studies where patients are monitored periodically and the event of interest is detectable only at specific times of observation, for example, at the time of a medical examination.
A more complex censoring scheme is found when both initial and final times are interval–censored. We refer to this situation as double censoring. Let X denote the initial time, Y the final time and $T = Y - X$ the elapsed time of interest. For an individual i we observe the vector $(X_L^i, X_R^i, Y_L^i, Y_R^i)$ which means that $P(X^i \in [X_L^i, X_R^i], Y^i \in [Y_L^i, Y_R^i]) = 1$. The elapsed time T^i is doubly–censored, in the origin and at the end. Figure 1 illustrates this kind of censoring.

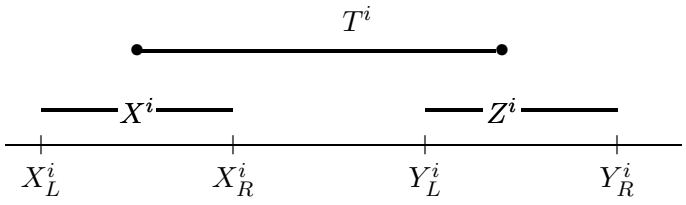

Figure 1: Double censoring

In the context of HIV–AIDS studies X^i is usually taken as the infection time of a patient which is only known to lie between the time, X_L^i, of the last negative antibody test and the time, X_R^i, of the first positive antibody test. Y^i is the time of the AIDS diagnosis which can be

exactly observed, that is $Y_L^i = Y_R^i$, or can be right–censored, $Y_R^i = +\infty$, for those patients which at the end of the study have not developed the disease. The elapsed time between X^i and Y^i, that is $T^i = Y^i - X^i$, is the AIDS latency time.

The analysis of interval–censored and doubly–censored data has been mainly approached through nonparametric frequentist methods. One of the first papers approaching the interval–censored situation is due to Peto (1973) who 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. The paper by Finkelstein (1986) proposes a test for covariate effects. A more recent approach to nonparametric estimation under interval censoring can be found in Gentleman and Geyer (1994). For the nonparametric analysis of doubly–censored data we find DeGruttola and Lagakos (1989), Gómez and Lagakos (1994) and Gómez and Calle (1999), between others, which extend Turnbull's algorithm to double censoring.

Interval–censoring has also been approached nonparametrically from a Bayesian perspective. See, for example, Doss (1994), Sinha and Dey (1997), Gómez *et alt.* (2000), Calle and Gómez (2001a) and the book on Bayesian survival analysis by Ibrahim, Chen and Sinha (2001). The Bayesian approach provides a direct probabilistic interpretation of the posterior distribution and 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 opened the door to the use of Bayesian methods to survival analysis.

Frequentist parametric methods have not been widely used in survival analysis, mainly because this approach depends on the model assumptions which are difficult to check under censoring. However, sometimes their use is indicated by the nature of the problem in study or suggested by a previous similar situation. Lindsey (1998) justifies the benefits of parametric models for analyzing interval–censored data. Lindsey and Ryan (1998) provide a useful tutorial of both parametric and nonparametric methods. On the contrary, Bayesian parametric methods through Markov Chain Monte Carlo methods have become a very used approach for the analysis of complex hierarchical models, see for instance, Stang and Huerta (2000). However, most of the applications involve right censoring and there is a need for a more general formulation of the methology under interval censoring. As we will illustrate in the paper, this approach is specially appropriate to deal with doubly–censored data. The goal of this paper is to provide a formal description of a sampling–based method for the analysis of interval–censored and doubly–censored data and to give clear guidelines for its practical use. We hope that this will contribute to make the parametric Bayesian approach an interesting alternative for the analysis

of this kind of censoring.
The rest of the paper is organized as follows: In section 2 we introduce the notation for interval–censored data and propose a methodology to analyze this kind of censoring. In section 3 we extend the former approach to deal with doubly–censored data. The proposed method is illustrated in section 4 with data from De Gruttola and Lagakos (1989) corresponding to a cohort of hemophilia patients who were infected with HIV in the early 1980's.

2 Inference from interval–censored data

Let X be the random variable of interest. In our setting X is a positive random variable representing the time until the occurrence of a certain event $\mathcal{E}$ with right-continuous distribution function $W(x;\theta_X) = \text{Prob}\{X \leq x\}$ and density function $w(x;\theta_X)$, with unknown θ_X. In a study of n items or individuals, their potential times to $\mathcal{E}$, namely, $X_1, \ldots, X_n$, are unknown and instead we observe intervals that contain the unobserved values of $X_1, \ldots, X_n$. Let $\mathcal{D} = \{[X_L^i, X_R^i], \quad 1 \leq i \leq n\}$ be the interval–censored survival data where X_L^i is the last observed time for the i^{th} individual before the event $\mathcal{E}$ has occurred and X_R^i indicates the first time the event $\mathcal{E}$ has been observed. We are in fact formally observing random censoring vectors (X_L^i, X_R^i), $i = 1, \ldots, n$, coming from a joint density function, $f_{[X_L, X_R]}(l, r; \gamma)$, such that $X_L^i \leq X^i \leq X_R^i$ with probability 1.
We suppose that censoring occurs noninformatively in the sense that for any x, l, r such that $l \leq x \leq r$, the conditional density of X given X_L and X_R, $f_{[X|X_L,X_R]}(x|l, r; \theta_X, \gamma)$, satisfies

$$f_{[X|X_L,X_R]}(x \mid l, r; \theta_X, \gamma) \quad = \quad \frac{w(x;\theta_X)}{W(r;\theta_X) - W(l-;\theta_X)}, \tag{1}$$

where we define $W(t-) = \lim_{\Delta \to 0^+} W(t-\Delta)$. This noninformative censoring condition means that the only information provided by the censoring interval $[X_L^i, X_R^i]$ of an individual about the distribution of X^i is that the interval contains X^i.
It can be proved (Gómez *et alt.*, 2001) that if censoring occurs noninformatively and if the law governing X_L and X_R does not involve any of the parameters of interest, we can base our inferences on the likelihood function $L(\theta_X|\mathcal{D})$ given by

$$L(\theta_X|\mathcal{D}) = \prod_{i=1}^{n} \int_{X_L^i}^{X_R^i} w(u;\theta_X)\ du.$$

By means of Bayes theorem and after assuming a prior distribution $p(\theta_X)$ for θ_X, the posterior distribution of θ_X is given by:

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

Usually the integral in the denominator is analytically intractable and does not admit an explicit solution. As an alternative we propose sampling–based method, in particular, the Gibbs sampler (Gelfand an Smith, 1990) to obtain a sample from the posterior distribution of interest, $p(\theta_X|\mathcal{D})$. As suggested by Smith and Roberts (1993), the Gibbs sampler is a very useful method in problems involving incomplete or censored data. The unobserved data $X^1, \ldots, X^n$ 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). The vector of interest is now $(X^1, \ldots, X^n, \theta_X)$ and its posterior distribution can be obtained by performing the Gibbs algorithm. This method consists in sampling iteratively from the full conditional distributions, that is the conditional distribution of each variable given all the rest. In this case we have:

1. The conditional distribution of each censored time given the other survival times, the parameter vector and the observed censoring intervals:
   $p(X^i|X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^n, \theta_X, \mathcal{D})$, for each $i = 1, \ldots, n$, and
2. the conditional distribution of the parameter vector given the survival times and the observed censoring intervals:
   $p(\theta_X|X^1, \ldots, X^n, \mathcal{D})$.

In the first step each censored observation X^i is imputed from its full conditional distribution. In the second step the parameter θ_X is updated based on the complete imputed sample.
In the following two propositions we state how these conditional distributions can be simplified by using the noninformative censoring condition (1).

Proposition 1 *The full conditional distribution for X^i, that is $p(X^i|X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^n, \theta_X, \mathcal{D})$, is the prior distribution for X, $w(x; \theta_X)$, truncated in the interval $[X_L^i, X_R^i]$.*

Proof. Using the fact that $X^1, \ldots, X^n$ are i.i.d., the full conditional distribution for X^i reduces to $p(X^i \mid \theta_X, X_L^i, X_R^i)$. From the noninformative condition (1) this conditional distribution is given by

$$p(X^i = x \mid \theta_X, X_L^i, X_R^i) \quad = \quad \frac{w(x; \theta_X)}{W(X_R^i; \theta_X) - W(X_R^i -; \theta_X)} \cdot \mathbf{1}\{X_L^i \leq x \leq X_R^i\},$$

which is the prior distribution for X, $w(x; \theta_X)$, truncated in the interval $[X_L^i, X_R^i]$. □

Proposition 2 *The full conditional distribution for θ_X, that is $p(\theta_X|X^1, \ldots, X^n, \mathcal{D})$, is equal to $p(\theta_X|X^1, \ldots, X^n)$*

Proof. This result follows directly from the noninformative condition which implies that θ_X is conditionally independent of the censoring intervals given the complete sample $X^1, \ldots, X^n$. □

This scheme can be extended to a regression model with covariates $z_1, \ldots, z_k$ related to θ_X through the link function $\theta_X = g(z^i, \beta_X)$. We assume a prior distribution $p(\beta_X|\theta_0)$ for β_X and $p(\theta_0)$ for the hyperparameter θ_0.
The Gibbs sampling algorithm to obtain the posterior distribution of β_X is then given by the successive iteration of the following steps:

Gibbs sampling algorithm for interval censoring

1. Impute a value X^i sampled from $w(x;\theta_X)$ truncated in the interval $[X_L^i, X_R^i]$.
2. Sample a new value of β_X from its full conditional distribution $p(\beta_X|X^1, \ldots, X^n, \theta_0)$ and update the value of $\theta_X = g(z^i, \beta_X)$.
3. Sample a new value of θ_0 from its full conditional distribution $p(\theta_0|\beta_X)$.

The successive implementation of these steps provides a sample of the vector of unknowns $(X^1, \ldots, X^n, \beta_X, \theta_0)$ which, under weak conditions (Gelfand and Smith, 1990), converges to its posterior distribution. Averages from these samples are used to estimate posterior quantities.

3 Inference from doubly–censored data

Let X and Y be the random variables corresponding to the chronological times of the initial and final events, respectively. Define the duration time to be $T = Z - X$. We wish to estimate the parameters of the density functions, $w(x;\theta_X)$ and $f(t;\theta_T)$, of X and T under the assumption that X and T are independent random variables. We assume that X and Y are interval–censored in $[X_L, X_R]$ and $[Y_L, Y_R]$, respectively. For each subject i of a random sample of size n the observable data are of the form $\mathcal{D} = \{(X_L^i, X_R^i, Y_L^i, Y_R^i),\ 1 \leq i \leq n\}$. Under the assumption of a noninformative censoring (1), the joint likelihood is given by:

$$L(\theta_X, \theta_T|\mathcal{D}) = \prod_{i=1}^{n} \int_{X_L^i}^{X_R^i} \int_{Y_L^i - x}^{Y_R^i - x} f(t;\theta_T)\ w(x;\theta_X)\ dt\ dx.$$

We assume in addition that there is a set of covariates $z_1, \ldots, z_k$ related to θ_X and to θ_T through the link function $\theta_X = g(z^i, \beta_X)$ and $\theta_T = h(z^i, \beta_T)$, respectively. The prior distribution for the regression parameters are $p(\beta_X|\theta_0)$ and $p(\beta_T|\theta_1)$ and $p(\theta_0)$ and $p(\theta_1)$ for

the corresponding hyperparameters. As in the case of interval censoring, we introduce the censoring times X^i and T^i, for i from 1 to n, as further latent variables.
The vector of interest is then $(X^1, \ldots, X^n, T^1, \ldots, T^n, \theta_X, \theta_T)$. The Gibbs algorithm to sample from its posterior distribution consists on sampling iteratively from the full conditional distributions:

1. The conditional distribution of each censored initial time:
   $p(X^i|X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^n, T^1, \ldots, T^n, \theta_X, \theta_T, \mathcal{D})$, for each $i = 1, \ldots, n$;

2. The conditional distribution of each censored latency time:
   $p(T^i|X^1, \ldots, X^n, T^1, \ldots, T^{i-1}, T^{i+1}, \ldots, T^n, \theta_X, \theta_T, \mathcal{D})$, for each $i = 1, \ldots, n$;

3. the conditional distribution of θ_X:
   $$p(\theta_X|X^1, \ldots, X^n, T^1, \ldots, T^n, \theta_T, \mathcal{D}) \text{ and}$$

4. the conditional distribution of θ_T:
   $$p(\theta_T|X^1, \ldots, X^n, T^1, \ldots, T^n, \theta_X, \mathcal{D}).$$

In the first step each censored observation X^i is imputed from its full conditional distribution. In the second step the parameter θ_X is updated based on the complete imputed sample.
Using the assumption that X and T are independent and the same reasoning as in proposition (1) it follows that:

Proposition 3 *The full conditional distribution for X^i is the prior distribution for X, $w(x; \theta_X)$, truncated in the interval $[X_L^i, X_R^i]$.*

To obtain the full conditional distribution of the doubly–censored latency time T^i we use the fact that, given X^i, the variable T^i is interval censored in $[Y_L^i - X^i, Y_R^i - X^i]$. Thus, as in the previous result, it follows that:

Proposition 4 *The full conditional distribution for T^i is the prior distribution for T, $f(t; \theta_T)$, truncated in the interval $[Y_L^i - X^i, Y_R^i - X^i]$.*

It is also easy to prove that:

Proposition 5 *The full conditional distributions for the parameter vectors θ_X and θ_T are equal to $p(\theta_X|X^1, \ldots, X^n)$ and $p(\theta_T|T^1, \ldots, T^n)$, respectively.*

The Gibbs sampler to obtain the posterior distribution of interest is then given by the successive simulation from the following steps:

Gibbs sampling algorithm for double censoring

1. Impute a value X^i sampled from $w(x;\theta_X)$ truncated in the interval $[X_L^i, X_R^i]$.
2. Impute a value T^i sampled from $f(t;\theta_T)$ truncated in the interval $[Y_L^i - X^i, Y_R^i - X^i]$.
3. Sample a new value of β_X from its full conditional distribution $p(\beta_X|X^1,\ldots,X^n,\theta_0)$ and update the value of $\theta_X = g(z^i,\beta_X)$.
4. Sample a new value of β_T from its full conditional distribution $p(\beta_T|T^1,\ldots,T^n,\theta_1)$ and update the value of $\theta_T = h(z^i,\beta_T)$.
5. Sample a new value of θ_0 from its full conditional distribution $p(\theta_0|\beta_X)$.
6. Sample a new value of θ_1 from its full conditional distribution $p(\theta_1|\beta_T)$.

4 Illustration

4.1 Data description and notation

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. 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. 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. A complete description of this data set is given in De Gruttola and Lagakos (1989). 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 $(z^i, X_L^i, X_R^i, d^i, Y_L^i, c^i)$. Covariate z indicates the treatment group. The value $z^i = 0$

corresponds to the heavily-treated grup and $z^i = 1$ to the lightly-treated group. X_L^i and X_R^i are the chronological times of the patient's last negative and first positive antibody test, respectively, d^i stands for the infection indicator. For those individuals who developed AIDS, $c^i = 1$ and Y_L^i denotes the chronological time of first clinical symptom of AIDS. For those individuals who had not developed AIDS at the end of the study, $c^i = 0$ and Y_L^i is the time of the last blood sample tested. The observed data can be divided into three groups according to their censoring patterns.

1. The first group corresponds to those individuals with a right-censored infection time.
2. The second group corresponds to those individuals with an interval-censored infection time and an observed AIDS diagnosis.
3. The last group corresponds to those individuals with an interval-censored infection time and a right-censored AIDS diagnosis time.

These censoring schemes are outlined in the following diagram (Figure 2), where X^i denotes infection time and Y^i AIDS diagnosis time.

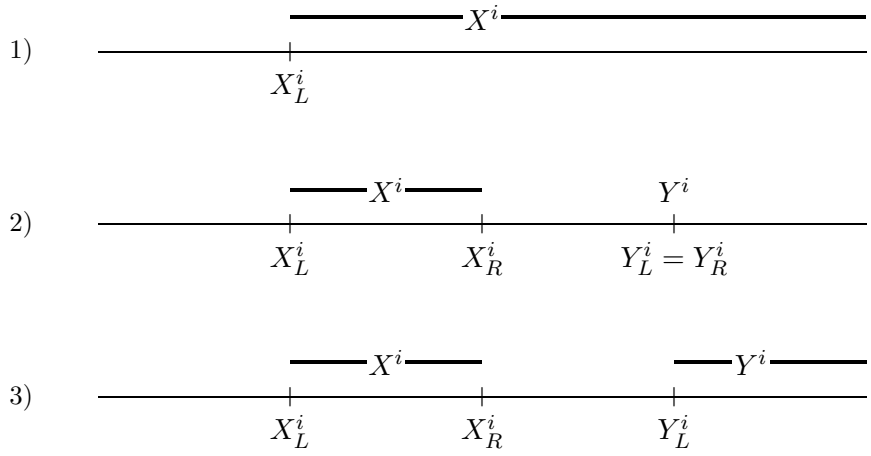

Figure 2: Different censoring schemes

4.2 Joint analysis of infection and latency times

We analyze the data assuming a log-normal model for both the time to HIV infection and the latency time of AIDS. We have chosen the log-normal distribution because it is known to

provide a good fit for long term survival times. Taking into account that T^i given X^i is interval–censored and that only individuals with an observed infection time ($d^i = 1$) contribute to the latency inference process, the model assumptions and prior specifications can be expressed through the following hierarchical model:

$$
\begin{aligned}
[\text{Stage1}] \quad & \text{for } (i \text{ in } 1:N)\{ \\
& \quad X^i \sim \log N(\mu_X^i, \sigma_X^2) \text{ truncated in } [X_L^i, X_R^i] \\
& \quad X_R^i = +\infty \text{ if } d^i = 0 \\
& \quad \mu_X^i = \beta_0 + \beta_1 \cdot z^i \\
& \quad \text{if } d^i = 1\{ \\
& \qquad T^i | X^i \sim \log N(\mu_T^i, \sigma_T^2) \text{ truncated in } [Y_L^i - X^i, +\infty) \text{ if } c^i = 0 \\
& \qquad T^i = Y_L^i - X^i \text{ if } c^i = 1 \\
& \qquad \mu_T^i = \beta_2 + \beta_3 \cdot z^i \\
& \quad \} \\
& \} \\
[\text{Stage2}] \quad & \beta_k \sim N(\alpha_k, \sigma_k^2) \text{ for } k = 0,1,2,3 \\
& \sigma_X^2 \sim IG(0.001, 0.001) \\
& \sigma_T^2 \sim IG(0.001, 0.001) \\
[\text{Stage3}] \quad & \alpha_k \sim N(0, 1.10^{-6}) \text{ for } k = 0,1,2,3 \\
& \sigma_k^2 \sim IG(0.001, 0.001) \text{ for } k = 0,1,2,3
\end{aligned}
$$

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.
Now, to implement the proposed algorithm in section 3 we have to derive all the full conditional distribution and perform the successive simulations. Alternatively, we have used the program BUGS which stands as an acronym for Bayesian inference Using Gibbs Sampling and is a very useful tool for the implementation of this algorithm. Given the model assumptions, this program performs the Gibbs sampler by simulating from the full conditional distributions. Further details of the program are given in Spiegelhalter *et al.*(1996). The code to specify this model and to obtain the posterior distributions of the parameters is in the appendix.

The Bayesian estimators were obtained through the implementation of the Gibbs sampling scheme described above. We implemented 2000 iterations of the algorithm and discarded the first 500 iterations. Convergence of the Gibbs sampler was established both graphically and numerically using the program CODA (Best *et al.*, 1995).
We have computed the sample mean and the 95% credible interval for each parameter in the model. The results are given in Table 1. Figure 3 gives the posterior distribution of each parameter.

Table 1: Posterior means together with the 95% credible intervals for parameters of interest

Parameter	mean	95% credible interval
β_0	2.426	$(2.348, 2.502)$
β_1	0.231	$(0.134, 0.334)$
β_2	2.787	$(2.517, 3.109)$
β_3	0.468	$(0.114, 0.845)$
σ_X	0.363	$(0.321, 0.413)$
σ_T	0.916	$(0.711, 1.172)$

Using these results and the expression of the mean of a lognormal distribution ($E(X) = exp(\mu_X + 0.5 \cdot \sigma_X^2)$), we obtain that the mean infection time for the heavily–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 4 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. The difference between the two groups becomes clear after the first 3 years.
The results for the latency times are as follows. Using as before the expression of the mean of a lognormal distribution ($E(T) = exp(\mu_T + 0.5 \cdot \sigma_T^2)$), we obtain that the mean latency time for the heavily–treated group is 24.70 (which corresponds to 12 years) while for the lightly–treated group is 39.45 (approximately 19 years). The estimated distribution curves of the latency times for the two groups are plotted in Figure 5. From this plot 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.

5 Discussion

We have detailed the methodology for a Bayesian analysis of interval–censored and doubly–censored data. The use of Markov Chain Monte Carlo methods, such as the Gibbs sampler, is shown to be very appropriate for these kind of censoring. Though much emphasis has been placed on nonparametric or semiparametric models for censored data, parametric models provides a useful framework for the analysis of complex models.

We have analyzed the data corresponding to the cohort of heamophiliacs using a log-normal model for both the infection times and the latency times. The purpose of this analysis was illustrative of the methology. For a more realistic analysis of the data it would be necessary to check the model assumptions. The problem is that, as far as we know, model fitting test for interval censoring or double censoring are not available in statistical packages. One possibility is to use the Bayesian model selection method proposed by Sinha *et alt.* (1999) as a model fitting test. Their methodology compares two alternative models. It could be used as a model fitting test by comparing the parametric model with the nonparametric estimate given for instance by Turnbull's algorithm.

An alternative to complete parametric methods for the analysis of interval–censored data is the Mixture of Dirichlet process model. This model allow a hierarchical model structure where some components are treated parametrically while others are analyzed nonparametrically. The paper by Calle and Gómez (2001b) follows this approach in the context of a linear regression model where one covariate is interval–censored. Further research is needed in developing the methodology to other regression models, such us the logistic regression model, or to allow the response variable to be also censored.

6 Acknowledgements

The author would like to thank Prof. G. Gómez for her advice and suggestions on this paper. This research was partially supported by the Dirección General de Investigación Científica y Técnica 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.

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

Calle, M.L. and Gómez, G. (2001b) Semiparametric Bayesian Analysis of Regression Models with an Interval-censored Covariate; Technical Report, DR2001/04, Dept. Statistics and Operations Research. Universitat Politècnica de Catalunya.

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

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

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

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.

Gómez, G.; Calle, M.L. and Oller, R. (2001) A walk through Interval-Censored Data. Technical Report, 2001/16, Dept. Statistics and Operations Research. Universitat Politècnica de Catalunya.

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

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

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

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

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

Sinha, D., Chen, M. and Ghosh, S. (1999) Bayesian Analysis and Model Selection for Interval–Censored Survival Data. *Biometrics* **55**, 585–590.

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.

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

Stang, D. and Huerta, G. (2000) Assessing the impact of Managed–Care on the Distribution of Lenght–of–Stay Using Bayesian Hierarchical Models. *Life Time Data Analysis* **6**,

123–139.

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.

A Program Code

Here we give the program code to analyzed the data described in section (4.2) with the program BUGS. The program includes the log-normal distribution but we were not able to implement it under interval censoring. For that reason we first transform the data with the logarithm function and then use the normal distribution.

```
model log-normal; # name of the  program
    {
        for(i in 1:149){    # Patients with interval--censored
                            # infection time and  who have not
                            # developed AIDS at the end of the study.
            logxl[i]<- log(xl[i]); # log transformation
            logxr[i]<- log(xr[i]); # of the data
                muX[i]<-beta0+beta1*z[i];
                muT[i]<-beta2+beta3*z[i];
                logX[i] ~ dnorm(muX[i],tauX) I(logxl[i],logxr[i]);
                                    #  truncated normal distribution
                                    # in the interval [logxl, logxr]

                X[i]<-exp(logX[i]);
                tl[i]<-(yl[i]-X[i]);
                logtl[i]<-log(tl[i]);
                logT[i] ~ dnorm(muT[i],tauT) I(logtl[i],);
                                    #  truncated normal distribution
                                    # in the interval [logtl, infinity)

        }
        for(i in 150:192){  # Patients with interval--censored
                            # infection time and who have developed
                            # AIDS at the end of the study.
            logxl[i]<- log(xl[i]);
            logxr[i]<- log(xr[i]);
                muX[i]<-beta0+beta1*z[i];
```

```
        muT[i]<-beta2+beta3*z[i];
        logX[i] ~ dnorm(muX[i],tauX) I(logxl[i],logxr[i]);
        X[i]<-exp(logX[i]);
        tl[i]<-(yl[i]-X[i]);
        tr[i]<-(yl[i]-X[i])+1;
        logtl[i]<-log(tl[i]);
        logtr[i]<-log(tr[i]);
        logT[i] ~ dnorm(muT[i],tauT) I(logtl[i],logtr[i]);

  }
for(i in 193:262){# Patients with right--censored
                  # infection time
    logxl[i]<- log(xl[i]);
        muX[i]<-beta0+beta1*z[i];
        logX[i] ~ dnorm(muX[i],tauX) I(logxl[i],);

}

beta0 ~ dnorm(alpha0,tau0); # Prior distributions
beta1~ dnorm(alpha1,tau1);  # of the parameters of interest
beta2 ~ dnorm(alpha2,tau2);
beta3~ dnorm(alpha3,tau3);
sigmaX <- 1/sqrt(tauX);
tauX  ~ dgamma(1.0E-3, 1.0E-3);
sigmaT <- 1/sqrt(tauT);
tauT ~ dgamma(1.0E-3, 1.0E-3);
alpha0 ~ dnorm(0, 1.0E-6);     # Prior distributions
tau0~ dgamma(1.0E-3, 1.0E-3);  # of the hyperparameters
alpha1 ~ dnorm(0, 1.0E-6);
tau1~ dgamma(1.0E-3, 1.0E-3);
alpha2 ~ dnorm(0, 1.0E-6);
tau2~ dgamma(1.0E-3, 1.0E-3);
alpha3 ~ dnorm(0, 1.0E-6);
tau3~ dgamma(1.0E-3, 1.0E-3);


}
```

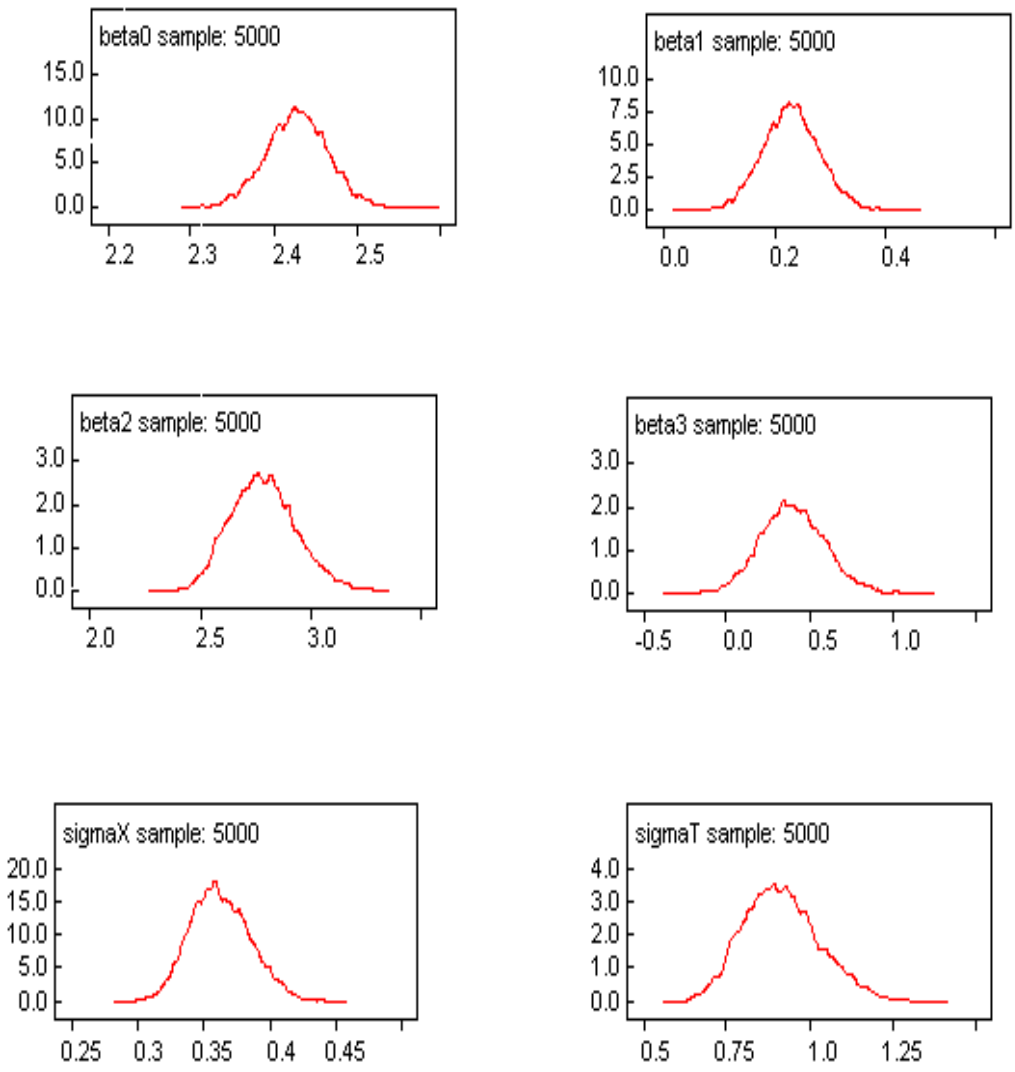

Figure 3: Posterior distribution of the model parameters: $\beta 0, \beta 1, \beta 2, \beta 3, \sigma_X$ and σ_T

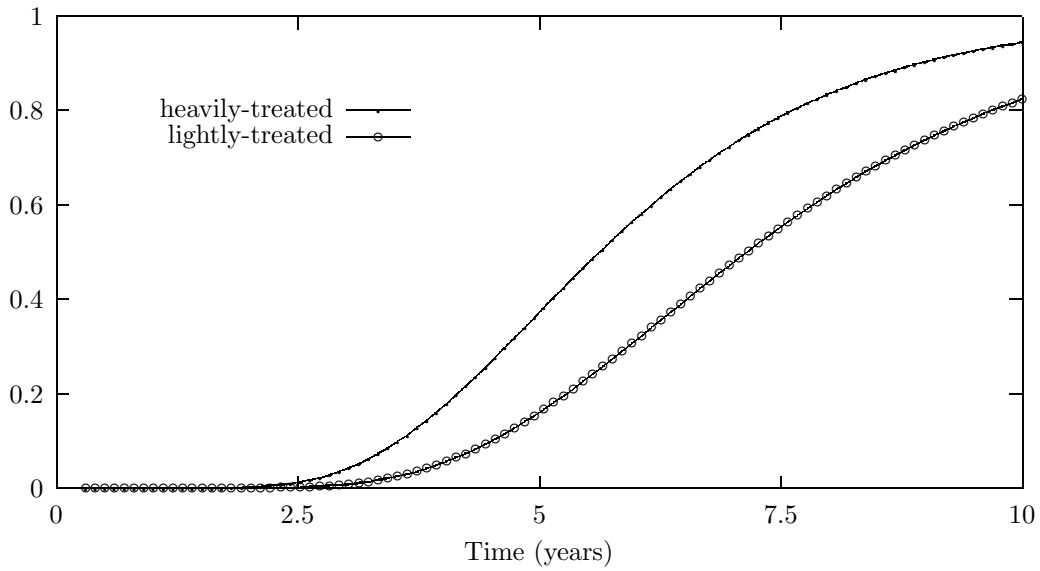

Figure 4: Estimated cumulative distributions of times to HIV infection

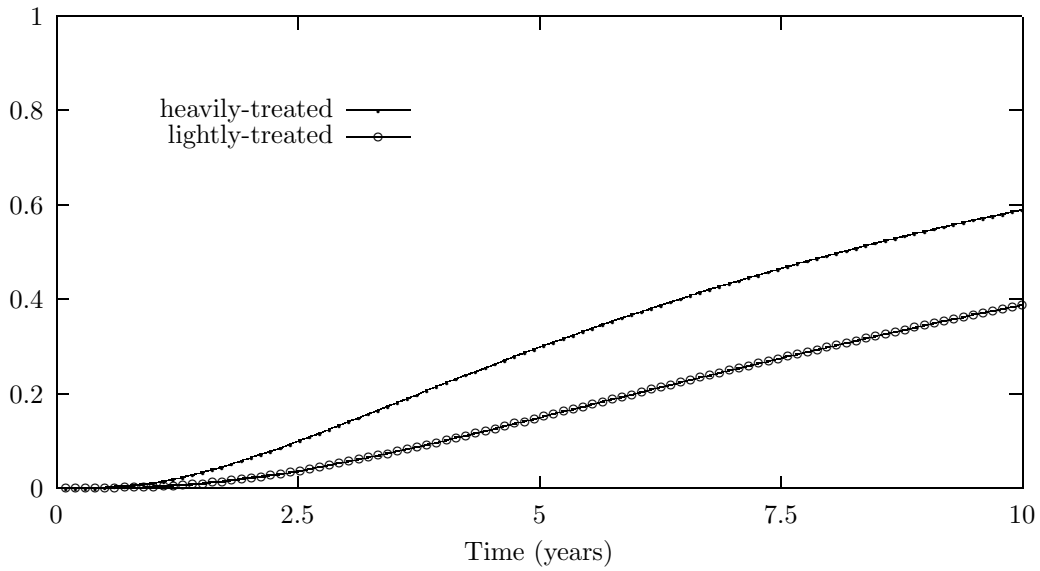

Figure 5: Estimated cumulative distributions of latency times to AIDS