Statistica Sinica 4(1994), 505-524

NONPARAMETRIC BAYESIAN INFERENCE FROM RIGHT CENSORED SURVIVAL DATA, USING THE GIBBS SAMPLER

Elja Arjas and Dario Gasbarra

University of Oulu

Abstract: Consider simple right censored survival data with a common unknown hazard rate. The hazard rate is here modelled nonparametrically, as a jump process having a martingale structure with respect to the prior distribution. For an evaluation of posterior probabilities, given the data, sample paths of the hazard rate are generated from the posterior distribution by using a dynamic version of the Gibbs sampler. The algorithm is described in detail. It is also shown how, by slightly modifying the algorithm, the procedure can be altered to correspond to a constrained estimation problem where the hazard rate is known to be increasing (or decreasing). The methods are illustrated by simulation examples.

Key words and phrases: Hazard rate, Markov chain Monte Carlo integration, posterior distribution, predictive distribution, Bayesian smoothing, constrained estimation.

1. Introduction

It is often felt that statistical models based on conventional parametric distributions are not flexible enough to provide a realistic description of survival data. This has led to a widespread use of nonparametric estimators, notably the Kaplan-Meier estimator for the survival function and the Nelson-Aalen estimator for the cumulative hazard. The corresponding distributions are discrete, with point masses at data points, and therefore some form of kernel smoothing is often applied afterwards to arrive at a suitable density estimator. The idea of nonparametric "baseline" hazard rate is also present in the well known Cox regression model, where a parametrically defined relative risk function is then used for describing covariate effects.

One can make a compromise between conventional parametric and nonparametric estimators, and assume that the hazard rate is some simple function, such as piecewise constant, piecewise linear, or spline, involving, if necessary, a fairly large number of points supporting the curve and acting as its parameters. But such models can create other problems in conventional statistical inference. While asymptotic theory often provides a justification of appropriate confidence statements concerning single coefficients, their combination, for example, for obtaining confidence bands for survival predictions is far from being a clear-cut procedure.

Bayesian inference is in many ways a tempting approach to problems in survival analysis. In particular, the direct probabilistic interpretation of the posterior distribution is pleasant, and many concrete problems are formulated naturally in terms of integrals with respect to the posterior. The predictive survival distribution seems particularly important in this respect. A number of authors, e.g. Ferguson (1973) and Susarla and van Ryzin (1976), have considered nonparametric Bayesian estimation of survival (or distribution) functions, using the Dirichlet process as a prior. This has the advantage that the posterior distribution has again the Dirichlet structure, making the problem analytically tractable. An extension to so called "neutral to the right" family of distributions was considered by Doksum (1974) and by Ferguson and Phadia (1979). For a recent contribution, and for an extensive list of references, see Hjort (1990). In these papers the cumulative hazard is a stochastic process with independent increments, according to the prior distribution. A somewhat different approach was chosen by Dykstra and Laud (1981), who modelled the hazard rate as a gamma process, therefore postulating an IHR property of the model. This work was later generalized, in different directions, by Lo (1984), Ammann (1985), Lo and Weng (1989) and Thompson and Thavaneswaran (1992).

On the other hand, apart from statisticians' general unwillingness to specify priors, the Bayesian approach has suffered from the difficulty of choosing 'between two evils': restricting the model distributions to conjugate families, because of their analytic tractability, or facing the sizable computational difficulties in cases where the dimension of the parameter space is large. Recently the application of Monte Carlo integration methods, in particular the Gibbs sampler, has changed this picture dramatically (Gelfand and Smith (1990); Smith (1991); Roberts and Smith (1993) and, in survival modeling context, Clayton (1991)).

In this paper, the specification of the prior is done in terms of four hyperparameters, by adopting a hierarchical model structure. Of the above papers on nonparametric Bayesian estimation we come closest to Dykstra and Laud (1981) and its extensions, in that our model is based on the hazard rate. According to the prior, the cumulative hazard, therefore, does not have the, perhaps, often unrealistic independent increments property. As the prototype model, we assume that the hazard rate has a simple martingale jump process structure. In case the hazard is known to be monotone, we can change this assumption and consider only processes with positive (negative) jumps.

The plan of this paper is as follows: In Section 2 we introduce the statistical model. In Section 3 we describe the Gibbs sampler algorithm and illustrate the method by examples based on simulated and real data. Section 4 shows how the method can be adapted in an instance of constrained estimation, where we assume that the hazard rate is monotonically increasing. The paper concludes with a discussion.

2. The Model

We consider simple right censored survival data $\{(X_j, \delta_j);\ 1 \le j \le n\}$ arising in a study of n individuals; here X_j is the time the jth individual was last seen and δ_j is the indicator of an observed failure at X_j (i.e. $\delta_j = 1$ refers to an observed failure and $\delta_j = 0$ to censoring).

All the individuals are assumed to have a common hazard rate for failure, $\{\lambda_t;\ t \ge 0\}$. This function becomes the parameter of the model. Since the hazard rate cannot be observed, we shall treat it as a random function, or stochastic process, also assigning probabilities to its sample paths.

Remark. Since the hazard rate is a functional of the distribution (or survival) function, it may at first seem problematic to assign (subjective) probabilities to functions which themselves have a probabilistic interpretation. This is not a real difficulty, however. It seems most natural to think about the distribution functions as limits of the corresponding empirical quantities from a hypothetical sample of n "similar" (exchangeable) individuals, as $n \to \infty$. An alternative point of view is that the survival times of the individuals are conditionally independent, given their common hazard (or survival) function as a parameter.

In this paper we make the convention that $\{\lambda_t; t \ge 0\}$ has a simple jump process structure:

$$\lambda(t) = \sum_{i \ge 0} 1_{\{T_i < t \le T_{i+1}\}} \lambda_i,$$

where $1_{\{\ \}}$ is the indicator function, $0 = T_0 < T_1 < T_2 < \cdots$ is an increasing sequence of jump times, and $\lambda_i > 0$ are the corresponding levels of the piecewise constant hazard rate. The specification of our "prototype" prior distribution of the hazard rate sample paths is completed by assuming that

(i) the jump times $T_1, T_2, \ldots$ form a time-homogeneous Poisson process with parameter μ;
(ii) the initial level λ_0 has gamma distribution $\gamma(\cdot; \alpha_0, \beta_0)$ (with α_0 as the shape parameter and β_0 as the scale parameter);
(iii) given $\lambda_0, \ldots, \lambda_{i-1}$, and independently of the jump times T_k, λ_i has distribution $\gamma(\cdot; \alpha, \beta_i)$, where α is the shape parameter and $\beta_i = \alpha/\lambda_{i-1}$.

Here μ, α_0, α and β_0 are given hyperparameters controlling the "constancy" and initial level of the hazard rate. The prior mean number of jump points in an interval $(0,t]$ is μt. Therefore, if μ is small the hazard rate is likely to remain constant over long time intervals, and conversely. The initial level of hazard has prior mean $E_{prior}(\lambda_0) = \frac{\alpha_0}{\beta_0}$ and coefficient of variation $\frac{\sqrt{\mathrm{Var}_{prior}(\lambda_0)}}{E_{prior}(\lambda_0)} = \frac{1}{\sqrt{\alpha_0}}$. Note also that the conditional expected values and standard deviations are given by

$$E_{prior}(\lambda_i|\lambda_0,\ldots,\lambda_{i-1}) = \lambda_{i-1}$$
$$\sqrt{\mathrm{Var}_{prior}(\lambda_i|\lambda_0,\ldots,\lambda_{i-1})} = \frac{\lambda_{i-1}}{\sqrt{\alpha}}.$$

Therefore, we are here assuming that the hazard rate has a martingale structure with respect to the prior distribution and the internal filtration.

A brief description of this "prototype" prior could therefore be given as follows:

(i) there is no built-in prior assumption of trend of the hazard rate;
(ii) the level of hazard at time $t = 0$ is controlled by the mean $\frac{\alpha_0}{\beta_0}$ and the "tightness" parameter $\sqrt{\alpha_0}$. This prior is flat (tight) if α_0 is small (large);
(iii) The variability of the hazard rate over time is controlled by the intensity μ and tightness $\sqrt{\alpha}$, the former controlling the number of jumps and the latter their size. Thus, for example, a small value of μ and a large value of α correspond to a prior assumption of few but significant change points in the level of the hazard, whereas the opposite choice allows for frequent small changes. Letting either $\mu \to 0$ or $\alpha \to \infty$ we have, in the limit, a prior assumption of constant hazard rate, corresponding to an exponential model.

As always, the choice of the prior should reflect the honest understanding which the analyst has about the problem at hand. If the choice of the four hyperparameters does not offer enough room for this, for example, because of an assumed monotonicity property of the hazard, one should of course modify the prior. An example of this is discussed in Section 4.

Remark. Apart from this Bayesian interpretation, the hyperparameters could be interpreted in terms of costs in penalized ML-estimation, incurred by large and/or frequent fluctuations in the hazard rate.

Given the hazard rate $\lambda(t)$, $t \geq 0$, the individuals are assumed to live and die independently. Assuming that the censoring mechanism is non-informative with regard to the hazard rate, the likelihood will be proportional to the usual product form

$$\prod_{j=1}^{n}\Big[\lambda(X_j)^{\delta_j}\exp\Big\{-\int_0^{X_j}\lambda(s)ds\Big\}\Big] = \prod_{j=1}^{n}(\lambda(X_j))^{\delta_j}\cdot\exp\Big\{-\int_0^{T_{\max}}Y(s)\lambda(s)ds\Big\},$$

where $T_{\max} = \max_{1\leq j\leq n} X_j$ is the largest observation time, and $Y(t) = n - \sum_{j=1}^{n} 1_{\{X_j<t\}}$ is the number of individuals at risk at time t. The posterior density

is then proportional to the product of the prior density and the likelihood. The prior density of a "segment" $\{\lambda(t); 0 \leq t \leq T_{\max}\}$ of hazard rate can now be written as

$$\mu^m e^{-\mu T_{\max}} \gamma(\lambda_0; \alpha_0, \beta_0) \prod_{i=1}^{m} \gamma\Big(\lambda_i; \alpha, \frac{\alpha}{\lambda_{i-1}}\Big),$$

where $m = \sum_i 1_{\{T_i \leq T_{\max}\}}$ is the number of jumps the hazard rate makes in the interval $(0, T_{\max}]$.

3. The Gibbs Sampler Algorithm

3.1 Description of the algorithm

We now describe the algorithm which is used in the numerical calculations to generate samples of the hazard rate process from the posterior distribution. We begin by generating a parameter process history (see Arjas (1989), Arjas et al. (1992))

$$H^0_{T_{\max}} = \{(T^0_k, \lambda^0_k); 0 \leq k \leq m_0\}$$

from the prior distribution, to be used as a starting value for the Gibbs sampler algorithm. Here $m_0 = \sum_{k \geq 1} 1_{\{T^0_k \leq T_{\max}\}}$ is the number of points T^0_k in the interval $(0, T_{\max}]$.

In a general step of the Gibbs sampler, a history

$$H^i_{T_{\max}} = \{(T^i_0, \lambda^i_0), \ldots, (T^i_{m_i}, \lambda^i_{m_i})\}$$

is replaced by a history

$$H^{i+1}_{T_{\max}} = \{(T^{i+1}_0, \lambda^{i+1}_0), \ldots, (T^{i+1}_{m_{i+1}}, \lambda^{i+1}_{m_{i+1}})\},$$

where $m_i = \sum_{k \geq 1} 1_{\{T^i_k \leq T_{\max}\}}$ and $m_{i+1} = \sum_{k \geq 1} 1_{\{T^{i+1}_k \leq T_{\max}\}}$. As a result of the chosen algorithm we have always $m_{i+1} \geq m_i - 1$.

The general step can be decomposed into $2m_{i+1}$ substeps, each corresponding to a new value of T_k or λ_k. The new value is sampled from a conditional distribution, where the conditioning is on the data (which stays always fixed during the algorithm) and on the "current" values of the other sampled parameters.

Denoting densities generically by the letter p and using an obvious shorthand, we sample from the distributions

$$T^{i+1}_k \simeq p(T_k \,|\, T^{i+1}_0, \lambda^{i+1}_0, \ldots, T^{i+1}_{k-1}, \lambda^{i+1}_{k-1}, \lambda^i_k, T^i_{k+1}, \lambda^i_{k+1}, \ldots, T^i_{m_i}, \lambda^i_{m_i}, \text{data}),$$

$$\lambda^{i+1}_k \simeq p(\lambda_k \,|\, T^{i+1}_0, \lambda^{i+1}_0, \ldots, T^{i+1}_{k-1}, \lambda^{i+1}_{k-1}, T^{i+1}_k, T^i_{k+1}, \lambda^i_{k+1}, \ldots, T^i_{m_i}, \lambda^i_{m_i}, \text{data}).$$

These conditional densities of the parameters T_k and λ_k are proportional respectively to the "full" probability densities $p(H^{i,i+1}_{T_k}, \text{data})$ and $p(H^{i,i+1}_{\lambda_k}, \text{data})$, where we use the following notation for the current histories:

$$H^{i,i+1}_{T_k} = \{(T_0^{i+1}, \lambda_0^{i+1}), \ldots, (T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}), (T_k, \lambda_k^i), (T_{k+1}^i, \lambda_{k+1}^i), \ldots, (T_{m_i}^i, \lambda_{m_i}^i)\},$$

$$H^{i,i+1}_{\lambda_k} = \{(T_0^{i+1}, \lambda_0^{i+1}), \ldots, (T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}), (T_k^{i+1}, \lambda_k), (T_{k+1}^i, \lambda_{k+1}^i), \ldots, (T_{m_i}^i, \lambda_{m_i}^i)\}.$$

In other words, we can sample each parameter from the joint distribution of the current history and the data when all the remaining parameters are fixed.

In a typical step we first generate a new value T_k^{i+1} of T_k to the interval $(T_{k-1}^{i+1}, T_{k+1}^i)$. Because of the Markovian structure of the model the density is proportional in T_k to the joint density of the parameter, the data points within the interval $(T_{k-1}^{i+1}, T_{k+1}^i]$ and $Y(T_{k+1}^i)$, i.e. the number of individuals at risk after the interval. So we have

$$\begin{aligned}
&p(T_k \mid T_0^{i+1}, \lambda_0^{i+1}, \ldots, T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}, \lambda_k^i, T_{k+1}^i, \lambda_{k+1}^i, \ldots, T_{m_i}^i, \lambda_{m_i}^i, \text{data}) \\
&= p(T_k \mid T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}, \lambda_k^i, T_{k+1}^i, \{(X_j, \delta_j); X_j \in (T_{k-1}^{i+1}, T_{k+1}^i]\}, Y(T_{k+1}^i)) \\
&\propto p(T_k, T_{k+1}^i, \{(X_j, \delta_j); X_j \in (T_{k-1}^{i+1}, T_{k+1}^i]\}, Y(T_{k+1}^i) \mid T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}, \lambda_k^i) \\
&= \exp\Big\{-Y(T_{k+1}^i)\Big(\int_{T_{k-1}^{i+1}}^{T_k} \lambda_{k-1}^{i+1} ds + \int_{T_k}^{T_{k+1}^i} \lambda_k^i ds\Big) \\
&\quad - \sum_{X_j \in (T_{k-1}^{i+1}, T_k]} \int_{T_{k-1}^{i+1}}^{X_j} \lambda_{k-1}^{i+1} ds - \sum_{X_j \in (T_k, T_{k+1}^i]} \Big(\int_{T_{k-1}^{i+1}}^{T_k} \lambda_{k-1}^{i+1} ds + \int_{T_k}^{X_j} \lambda_k^i ds\Big) \\
&\quad - \int_{T_{k-1}^{i+1}}^{T_k} \mu ds - \int_{T_k}^{T_{k+1}^i} \mu ds\Big\} \cdot (\lambda_{k-1}^{i+1})^{\sharp\{X_j \in (T_{k-1}^{i+1}, T_k], \delta_j = 1\}} \cdot (\lambda_k^i)^{\sharp\{X_j \in (T_k, T_{k+1}^i], \delta_j = 1\}} \cdot \mu^2,
\end{aligned}$$

where $\sharp A$ denotes the cardinality of set A. The position of T_k is determined by considering the partition of $(T_{k-1}^{i+1}, T_{k+1}^i]$ induced by the ordered observations X_j in that interval. Suppose that there are n_k such points, and denote the $n_k + 1$ intervals of the partition by $I_1, \ldots, I_{n_k+1}$. Then in each interval I_j the conditional probability density of T_k has a form proportional to $\exp(a_j T_k) c_j$ and it can be normalized by dividing by the constant

$$C_k = \sum_{j=1}^{n_k+1} c_j \int_{I_j} \exp(a_j s) ds.$$

T_k^{i+1} can now be sampled from this piecewise continuous density.

In a second substep of the algorithm we have to generate the parameter λ_k^{i+1} conditionally on all the remaining parameter values, including the position T_k^{i+1}

and the data. Again, because of the Markovian apriori structure of the model, only the parameter values $\lambda_{k-1}^{i+1}, \lambda_{k+1}^{i}, T_k^{i+1}, T_{k+1}^{i}$, the observations in the interval $(T_k^{i+1}, T_{k+1}^{i}]$, and $Y(T_{k+1}^{i})$ are relevant (up to proportionality in λ_k): Denoting by

$$r_k = \sum_{T_k^{i+1} < X_j \leq T_{k+1}^{i}} \delta_j$$

the number of observed failures in that interval, we have

$$\begin{aligned}
&p(\lambda_k \mid T_0^{i+1}, \lambda_0^{i+1}, \ldots, T_{k-1}^{i+1}, \lambda_{k-1}^{i+1}, T_k^{i+1}, T_{k+1}^{i}, \lambda_{k+1}^{i}, \ldots, T_{m_i}^{i}, \lambda_{m_i}^{i}, \text{data})\\
&\propto p(H_{\lambda_k}^{i,i+1}, \text{data})\\
&\propto p(\lambda_k, \lambda_{k+1}^{i}, \{(X_j, \delta_j); X_j \in (T_k^{i+1}, T_{k+1}^{i}]\}, Y(T_{k+1}^{i}) \mid \lambda_{k-1}^{i+1}, T_k^{i+1}, T_{k+1}^{i})\\
&= \gamma(\lambda_k; \alpha, \beta_k^{i+1}) \gamma\Big(\lambda_{k+1}^{i}; \alpha, \frac{\alpha}{\lambda_k}\Big) (\lambda_k)^{r_k}\\
&\quad \cdot \exp\Big\{-Y(T_{k+1}^{i}) \int_{T_k^{i+1}}^{T_{k+1}^{i}} \lambda_k ds - \sum_{X_j \in (T_k^{i+1}, T_{k+1}^{i}]} \int_{T_k^{i+1}}^{X_j} \lambda_k ds\Big\},
\end{aligned}$$

where $\beta_1^{i+1} = \beta_0$ and $\beta_k^{i+1} = \alpha / \lambda_{k-1}^{i+1}, k > 1$. This is proportional in λ_k to

$$\lambda_k{}^{(r_k - 1)} \exp\Big\{-\lambda_k \Big(\beta_k^{i+1} + \int_{T_k^{i+1}}^{T_{k+1}^{i}} Y(s) ds\Big)\Big\} \cdot \exp\Big\{-\frac{\alpha \lambda_{k+1}^{i}}{\lambda_k}\Big\} = f_\zeta(\lambda_k) \cdot g_\zeta(\lambda_k),$$

where for $\zeta \geq 0$ we define

$$f_\zeta(\lambda) = \lambda^{(\zeta + r_k - 1)} \exp\Big\{-\lambda \Big(\beta_k^{i+1} + \int_{T_k^{i+1}}^{T_{k+1}^{i}} Y(s) ds\Big)\Big\}$$

and

$$g_\zeta(\lambda) = \Big(\frac{1}{\lambda}\Big)^\zeta \exp\Big\{-\frac{\alpha \lambda_{k+1}^{i}}{\lambda}\Big\}.$$

We note that $f_\zeta(\cdot)$ is proportional to the gamma density

$$\gamma\Big(\cdot; \zeta + r_k, \beta_k^{i+1} + \int_{T_k^{i+1}}^{T_{k+1}^{i}} Y(s) ds\Big).$$

(Note also that, for $\zeta > 1$, $g_\zeta(\cdot)$ is proportional to an inverse gamma density).

Solving the equation

$$\frac{\zeta + r_k - 1}{\beta_k^{i+1} + \int_{T_k^{i+1}}^{T_{k+1}^{i}} Y(s) ds} = \frac{\alpha \lambda_{k+1}^{i}}{\zeta}$$

one can find a value $\zeta^\star$ such that $f_{\zeta^\star}(\cdot)$ and $g_{\zeta^\star}(\cdot)$ have their modes at the same point. Generating λ_k^{i+1} from the density proportional to $f_\zeta(\cdot)g_\zeta(\cdot)$ can then be accomplished by the rejection sampling technique (Ripley (1987), Gilks (1992), Gilks and Wild (1992)):

First compute

$$M_k = \max_\lambda \; g_{\zeta^\star}(\lambda) = g_{\zeta^\star}\Big(\frac{\alpha\lambda^i_{k+1}}{\zeta^\star}\Big) = \Big(\frac{\zeta^\star}{\alpha\lambda^i_{k+1}}\Big)^{\zeta^\star}\exp(-\zeta^\star),$$

then do:

Step1: generate $\lambda^\star$ from the density $\gamma(\cdot;\zeta^\star + r_k, \beta_k^{i+1} + \int_{T_k^{i+1}}^{T^i_{k+1}} Y(s)ds)$;
Step2: generate U uniform in $[0,1]$;
Step3: if

$$g_{\zeta^\star}(\lambda^\star) > U \cdot M_k$$

then accept $\lambda_k^{i+1} = \lambda^\star$, **otherwise** reject this $\lambda^\star$ and go back to **Step1** to generate a new $\lambda^\star$.

Note that the choice $\zeta = \zeta^\star$ is optimal for the rejection sampling procedure, minimizing the probability of rejection.

In this way we can update successively the old values $\{(T_k^i, \lambda_k^i);\ 1 \le k \le m_i - 1\}$ to a set of new ones $\{(T_k^{i+1}, \lambda_k^{i+1});\ 1 \le k \le m_i - 1\}$. To complete the algorithm, we only have to determine its behaviour after $(T^{i+1}_{m_i-1}, \lambda^{i+1}_{m_i-1})$ has already been generated. When replacing the last jump time $T^i_{m_i}$ and generating a new one, i.e., $T^{i+1}_{m_i}$ from the distribution

$$p(\cdot \,|\, T^{i+1}_{m_i-1}, \lambda^{i+1}_{m_i-1}, \lambda^i_{m_i}, \{(X_j,\delta_j); X_j \in (T^{i+1}_{m_i-1}, T_{\max}]\}),$$

it may happen that the algorithm produces a point which falls outside the observation interval, i.e., $T^{i+1}_{m_i} > T_{\max}$. In that case $T^{i+1}_{m_i}$ is discarded, so that the $(i+1)$st generated hazard function has one jump point less than the previous one and we set $m_{i+1} = m_i - 1$. The probability of this event is proportional to

$$\exp\Big\{-Y(T_{\max})\int_{T^{i+1}_{m_i-1}}^{T_{\max}} \lambda^{i+1}_{m_i-1}ds - \sum_{X_j\in(T^{i+1}_{m_i-1},T_{\max}]} \int_{T^{i+1}_{m_i-1}}^{X_j} \lambda^{i+1}_{m_i-1}ds - \int_{T^{i+1}_{m_i-1}}^{T_{\max}} \mu ds\Big\}$$
$$\cdot\, (\lambda^{i+1}_{m_i-1})^{\sharp\{failures \in (T^{i+1}_{m_i-1}, T_{\max}]\}} = p_{tail}.$$

The normalizing constant becomes $(m = m_i)$

$$C_m = \sum_{j=1}^{n_m+1} c_j \int_{I_j} \exp(a_j s)ds + p_{tail}$$

so that the actual probability is p_{tail}/C_m. In the complementary case where $T^{i+1}_{m_i} < T_{\max}$ we first generate a new parameter $\lambda^i_{m_i+1} \simeq \gamma(\cdot; \alpha, \beta_{m_i+1})$, where $\beta_{m_i+1} = \alpha/\lambda^i_{m_i}$, representing the next level of the hazard after $T_{\max}$. At this stage $\lambda^i_{m_i+1}$ is a "fictitious" parameter, because it doesn't enter the likelihood. However, it is needed for the next steps of the Gibbs sampler. By updating the jump level

$$\lambda_{m_i} = \lambda^{i+1}_{m_i} \simeq p(\cdot \mid \lambda^{i+1}_{m_i-1}, T^{i+1}_{m_i}, \lambda^i_{m_i+1}, \{(X_j, \delta_j); X_j \in (T^{i+1}_{m_i}, T_{\max}]\}),$$

we get a new point $(T^{i+1}_{m_i}, \lambda^{i+1}_{m_i})$.

It is possible that the $(i+1)$st generated hazard rate will have even more than m_i jump points. To see this, we generate a new jump time

$$T^{i+1}_{m_i+1} \simeq p(\cdot \mid T^{i+1}_{m_i}, \lambda^{i+1}_{m_i}, \lambda^i_{m_i+1}, \{(X_j, \delta_j); X_j \in (T^{i+1}_{m_i}, T_{\max}]\});$$

again, if $T^{i+1}_{m_i+1} > T_{\max}$ we discard it, and this iteration of the Gibbs sampler ends giving the history $H_{T^{i+1}_{\max}}$ with $m_{i+1} = m_i + 1$ points. Otherwise we generate a new parameter λ_{m_i+2} (the next hazard level after $T_{\max}$) from the distribution $\gamma(\cdot; \alpha, \beta_{m_i+2})$, where $\beta_{m_i+2} = \alpha/\lambda^i_{m_i+1}$, and then update the last jump level

$$\lambda_{m_i+1} = \lambda^{i+1}_{m_i+1} \simeq p(\cdot \mid \lambda^{i+1}_{m_i}, T^{i+1}_{m_i+1}, \lambda^i_{m_i+2}, \{(X_j, \delta_j); X_j \in (T^{i+1}_{m_i+1}, T_{\max}]\}).$$

The algorithm continues by generating and updating points until, for the first time, some point goes beyond $T_{\max}$ during the updating step.

We can resume an algorithm's cycle as follows:

(1): update the initial level λ_0 at time $T_0 = 0$;
(2): update successively all the marked points (T_k, λ_k) from the first to the last but one;
(3): generate the last jump time T_m;
if this goes beyond $T_{\max}$ **then** discard it and go back to **(1)** to start a new updating cycle;
else generate $\lambda_{m+1} \simeq \gamma(\cdot; \alpha, \alpha/\lambda_m)$, update the last jump level λ_m and go back to the updating step **(3)** with m replaced by $m + 1$.

So in each cycle any number of marked points can be added, or one point can be erased.

3.2. Two illustrations

As a result of the algorithm we obtain a Markov chain $\{H^i_{T_{\max}}\}_{i \geq 0}$, which is ergodic under mild regularity conditions (Roberts and Smith (1993, 1994)). It is a straightforward matter to verify, by direct calculation, that the posterior distribution specified at the end of Section 2 and viewed as a measure on the space $\mathcal{H}$ of such histories, is invariant for that chain. As a consequence, for each real function $\phi : \mathcal{H} \rightarrow \Re$ integrable with respect to the posterior distribution on the space of parameter histories, we have almost surely that

$$\lim_{n\rightarrow\infty} \frac{1}{n}\sum_{i=1}^{n} \phi(H_T^i) = \int_{\mathcal{H}} \phi(H_T)dP(H_T|\text{data}).$$

Therefore, by using the simulated histories, we can approximate the posterior expectation of any integrable function ϕ. Perhaps the most important special case here is the predictive survival function obtained by considering, for each time point t, $\phi(H_T) = \exp\{-\int_0^t \lambda(s)ds\}$ (the hazard rate $\lambda(\cdot)$ is chosen to correspond to the history H_T). We denote this survival function by $\bar{F}_{pred}(t)$. Its density $f_{pre\bar{d}}(t)$ is approximated most conveniently in the simulation by considering, for each fixed t, $\phi(H_T) = \lambda(t)\exp\{-\int_0^t \lambda(s)ds\}$. The natural notion of hazard, called here the *predictive hazard*, is defined as $\lambda_{pred}(t) = f_{pred}(t)/\bar{F}_{pred}(t)$, and corresponds to the hazard, according to the posterior distribution, of some individual (not in the data but "similar") still alive at t.

As a first illustration, we used data generated from the model, plugging in a known "true hazard" function $\lambda_{true}(t)$, which in turn was sampled from the prior parameter distribution. The data were then censored in a noninformative way, at independent exponential random times and truncated at a fixed terminal time $T = 10$.

We present the numerical results arising from two sets of simulated data, with respectively 50 and 200 censored survival times (Figures 1(a,b,c) and 2). The dashed lines correspond to the hazard $\lambda_{true}(t)$ used in generating the data and the "true" survival function $\exp\{-\int_0^t \lambda_{true}(s)ds\}$. The curves referred to as *predictive survival probability*, and *predictive hazard*, were obtained as explained above.

In Figure 1(a) we used the hyperparameter values $(\mu, \alpha_0, \beta_0, \alpha) = (2, 5, 25, 5)$ corresponding to $\mu \cdot 10 = 20$ change points on the interval $(0, 10]$ as the prior mean, initial level mean $E_{prior}(\lambda_0) = \frac{\alpha_0}{\beta_0} = 0.2$, and "tightness" parameter $\sqrt{\alpha_0} = \sqrt{\alpha} = \sqrt{5}$. With such loose control on the number of jump points and the jump sizes, the predictive survival probabilities are seen to follow closely the Kaplan-Meier curve. The corresponding predictive hazard rate oscillates considerably over time. Some of those oscillations are caused by (compared to the true hazard) random clusters and gaps in the observations. Such sensitivity of the estimation can be conveniently tuned down by choosing a smaller value of μ and/or a larger value of α. In Figure 1(b) we have chosen $\mu = 0.5$, and in Figure 1(c) $\alpha = 20$, leaving the other hyperparameters unchanged compared to Figure 1(a). The changes stabilize the behaviour of the predictive hazard considerably. On the other hand, the predictive survival probabilities are almost identical in all three figures. This was to be expected, as the survival probabilities depend on the cumulative hazard rather than on the local behaviour of the hazard rate.

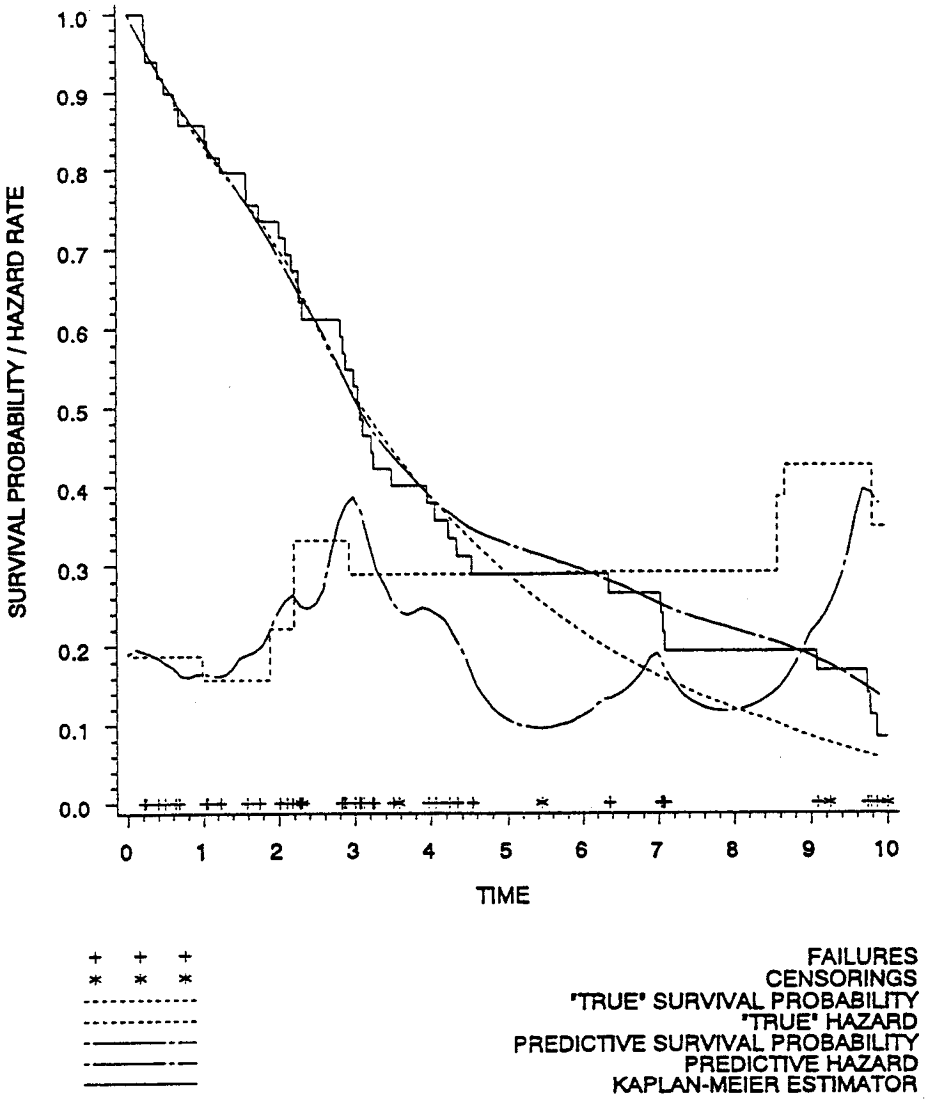

Figure 1(a). Results from an analysis of simulated data with 50 individuals under observation in the time interval $[0, 10]$; from those 2 were censored independently within the time interval and 5 at the end of the study. The hyperparameters had the values $\mu = 2$, $\alpha_0 = \alpha = 5$ and $\beta_0 = 25$. The Gibbs sampler ran for 5000 iterations and the first 500 generated histories were discarded.

It's always a delicate point assessing the convergence of the Gibbs sampler algorithm; however, observing independent runs of the algorithm with different lenghts and different initial values, and comparing naively the approximations of the predicitive hazard rate, we observed very soon (after a few hundred iterations) essentially the same behaviour. Further evidence of good convergence properties of the algorithm was obtained from experiments with large data sets where the posterior survival probabilities seemed to follow, irrespective of the prior, closely the Kaplan-Meier curves (Figure 2).

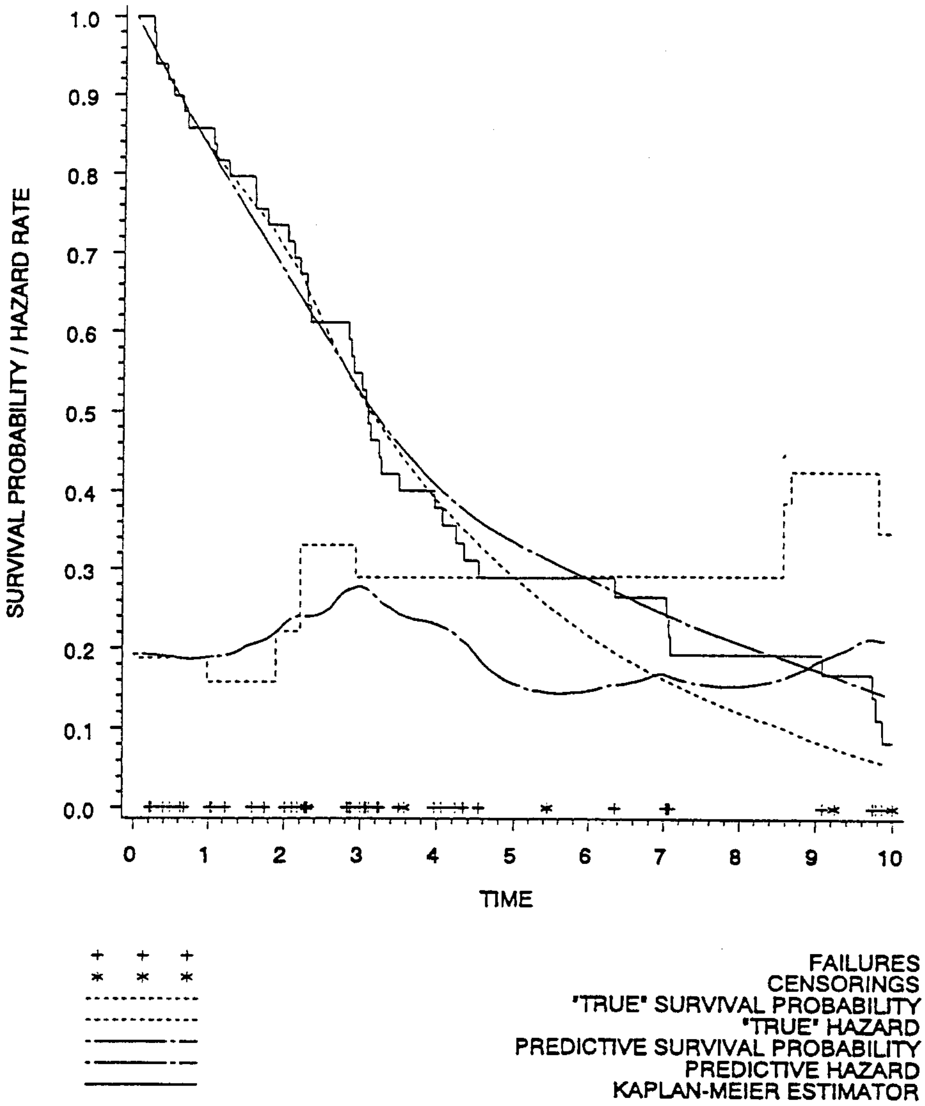

Figure 1(b). The data set is the same as in Figure 1(a); only the prior distribution was slightly changed setting $\mu = 0.5$. The Gibbs sampler ran for 5000 iterations and the first 500 generated histories were discarded.

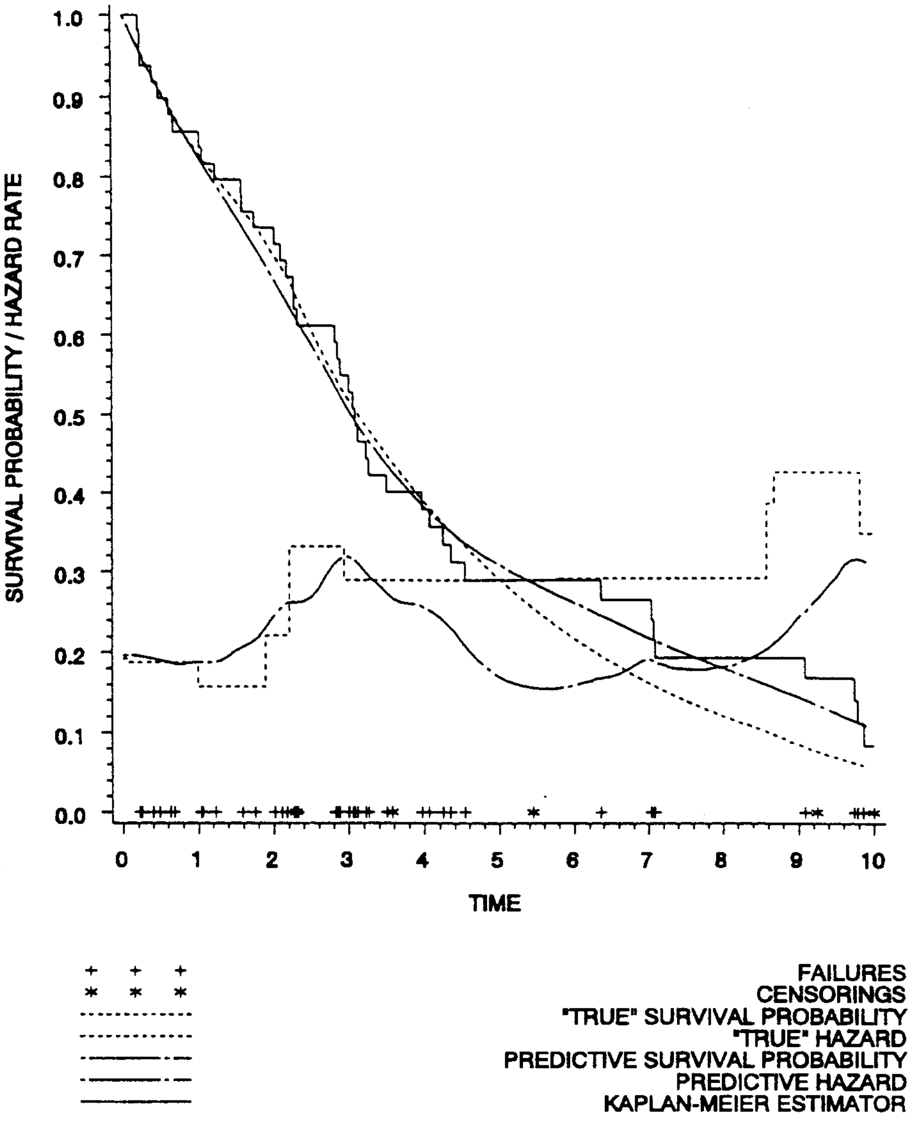

Figure 1(c). The data set is the same as in Figure 1(a); only the prior distribution was slightly changed setting $\alpha = 20$. The Gibbs sampler ran for 5000 iterations and the first 500 generated histories were discarded.

As a second illustration, we considered the data set given in Nair (1984), consisting of the failure times of 40 randomly selected mechanical switches. Three of these observations were right censored because of the termination of the test. The Kaplan-Meier curve and the 90% Greenwood confidence bands for the survival function (see e.g. Kalbfleisch and Prentice (1980)) are displayed in Figure 3, where the time and hazard where rescaled so that the test was terminated at time $T = 10$. We have drawn into this same figure the predictive survival probabilities

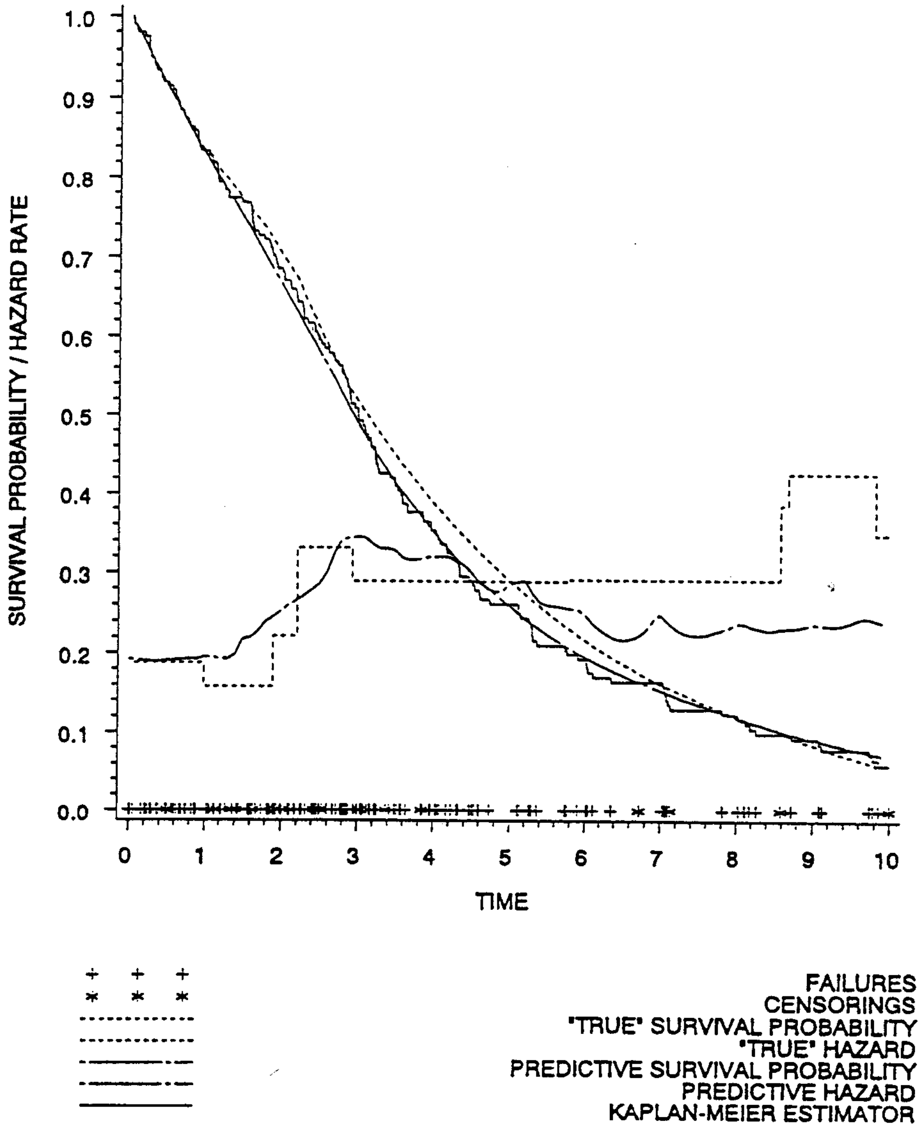

Figure 2. Continuing the analysis we increased the data set in Figures 1(a)-1(b) to 200 individuals; 11 of them were censored in the time interval and 11 at the end of the study. The hyperparameters had the values $\mu = 0.6; \alpha_0 = \alpha = 5$ and $\beta_0 = 25$. The Gibbs sampler ran for 5000 iterations and the first 500 were discarded.

and the corresponding predictive hazards using two rather different hyperparameters values $(\mu, \alpha_0, \beta_0, \alpha) = (0.2, 3, 200, 15)$ and $(0.7, 1, 270, 4)$. The predictive survival curves are again in good agreement with the Kaplan-Meier, and stay between the approximate confidence bands for times after the first failure. Note, however, that for times until that failure the entire band collapses into a single curve corresponding to estimated survival probability one. We would therefore interpret this discrepancy between the predictive curves and the "band" in this area to be a consequence of the band being deficient, and not the curves.

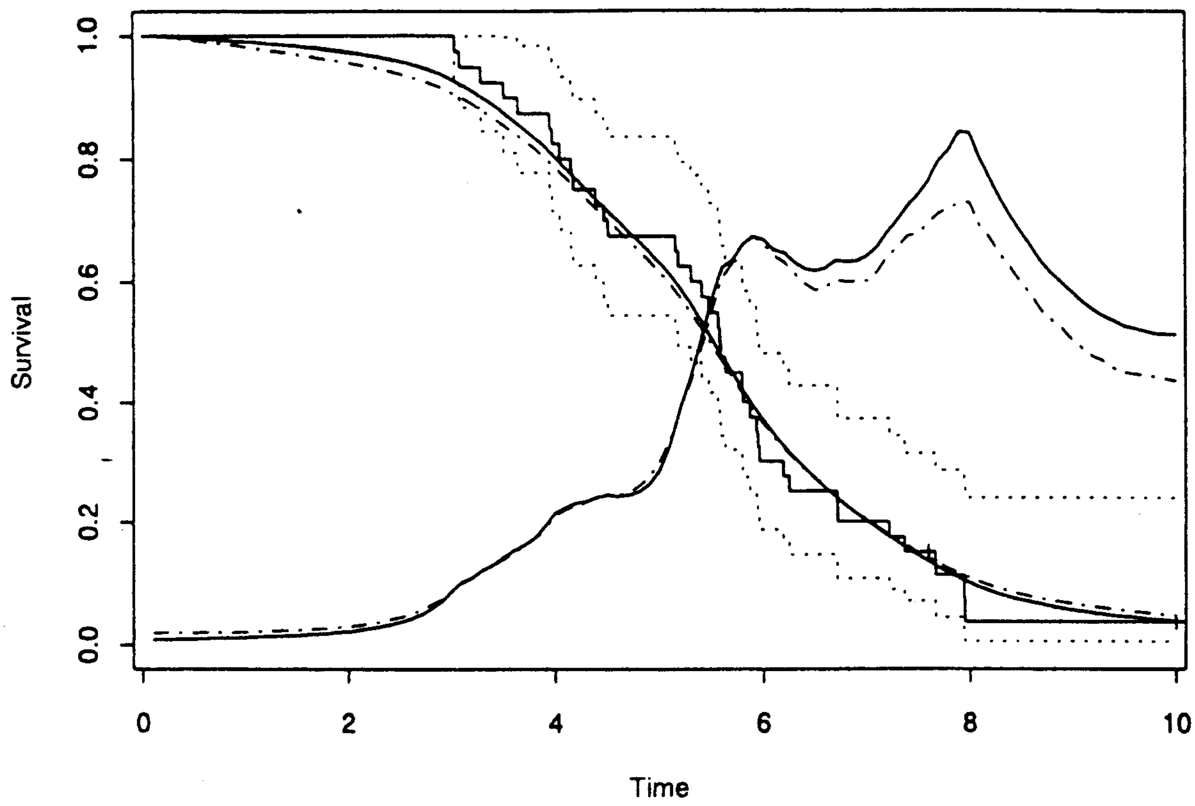

Figure 3. Analysis of Nair (1984) data. From 40 individuals, two were censored in the time interval and one at the end of the study. Predictive survival and hazard curves from two different priors are shown, with respective hyperparameters were $\alpha_0 = 3, \alpha = 15, \beta_0 = 200, \mu = 0.2$ (dashed line) and $\alpha_0 = 1, \alpha = 4, \beta_0 = 270, \mu = 0.7$ (solid line). The step functions are the Kaplan-Meier estimator and the 90% Greenwood confidence bands. The Gibbs sampler ran for 5000 iterations and the first 500 were discarded.

4. A Modification: Estimation of Increasing Hazard Rate

In the above algorithm we made the neutral assumption that, according to the prior distribution, the hazard rate has no trend up or down. Sometimes, however, we may have qualitative prior information and know, for example, with certainty that the hazard rate has to be IHR, i.e., a non-decreasing function of time. The background of such postulate can be aging resulting from physical wear, or some other similar argument. For cumulative hazard this means convexity.

It is not immediately obvious how conventional nonparametric hazard estimators, such as Nelson-Aalen, should be adjusted to take into account such monotonicity. A possible solution, studied recently in Huang and Wellner (1994), is to consider the derivative of the greatest convex minorant (GCM) of the Nelson Aalen estimator. The statistical properties of such an estimator are much more involved than those of Nelson-Aalen, however, since they cannot be linked directly with counting process martingales.

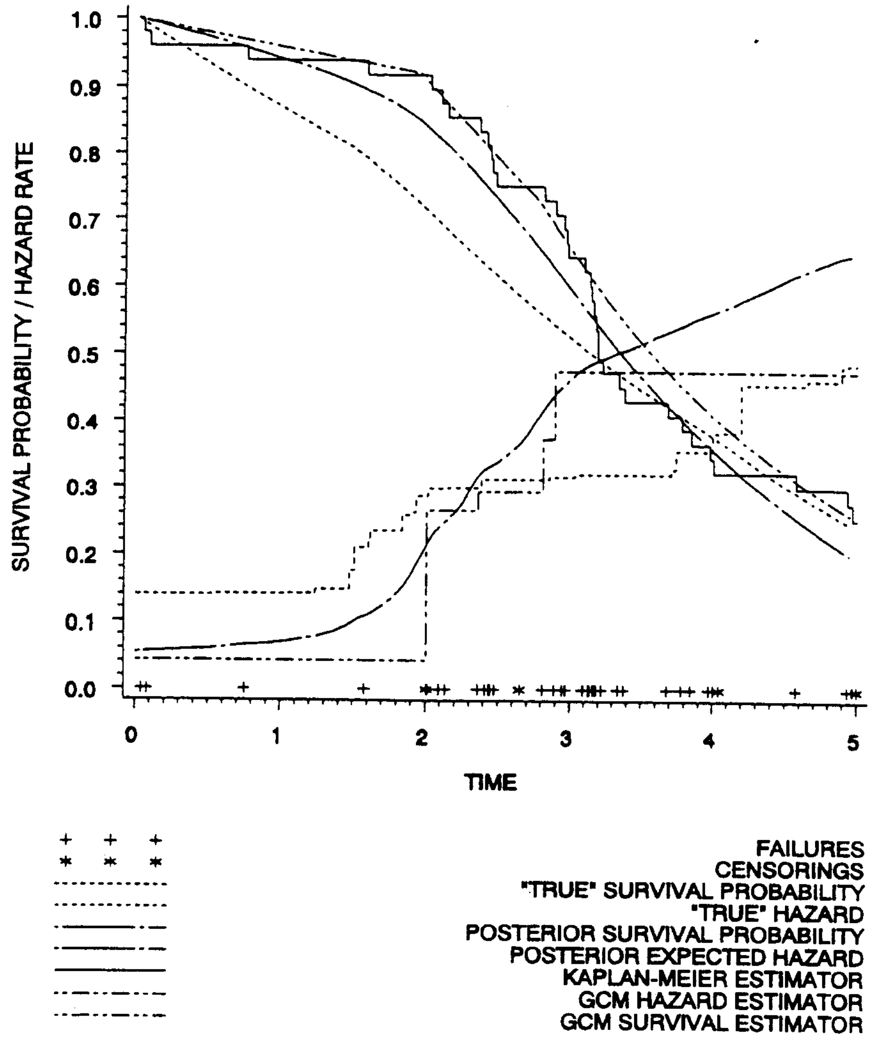

Figure 4. Results from fitting an IHR model to a simulated data set of 50 individuals, of which 3 were censored in the time interval $[0, 5]$ and 12 at the end of the study. The hyperparameters were given the values $\alpha_0 = 1.5$, $\beta_0 = 4$, $\mu = 1.5$, $\nu = 4$. The Gibbs sampler ran for 5000 iterations and the first 500 were discarded.

On the other hand, it turns out that the corresponding constrained Bayes estimation problem is solved very easily, by a slight adjustment on the model. This variant, which resembles closely that in Dykstra and Laud (1981), can be described as follows. We specify the prior probability by again assuming that the sequence of jump times $0 = T_0 < T_1 < T_2 < \cdots$ forms a Poisson process with fixed intensity μ and that the initial level of the hazard rate λ_0 has density $\gamma(\cdot;\alpha_0,\beta_0)$, with μ,α_0, and β_0 as fixed hyperparameters. Now, however, we assume that the (then positive) increments $\lambda_i-\lambda_{i-1}$, $i=1,2,\ldots$ of the hazard process are independent ν-exponential random variables, with ν a given hyperparameter. Within this prior model specification, the updating of the jump times T_i of the hazard process is accomplished just as in the previous trend-free model, whereas the updating of hazard level λ_i is from the distribution proportional to the gamma density $\gamma(\cdot;r_i+1,\int_{T_i}^{T_{i+1}}Y(s)ds)$ restricted to the subinterval $(\lambda_{i-1},\lambda_{i+1})$.

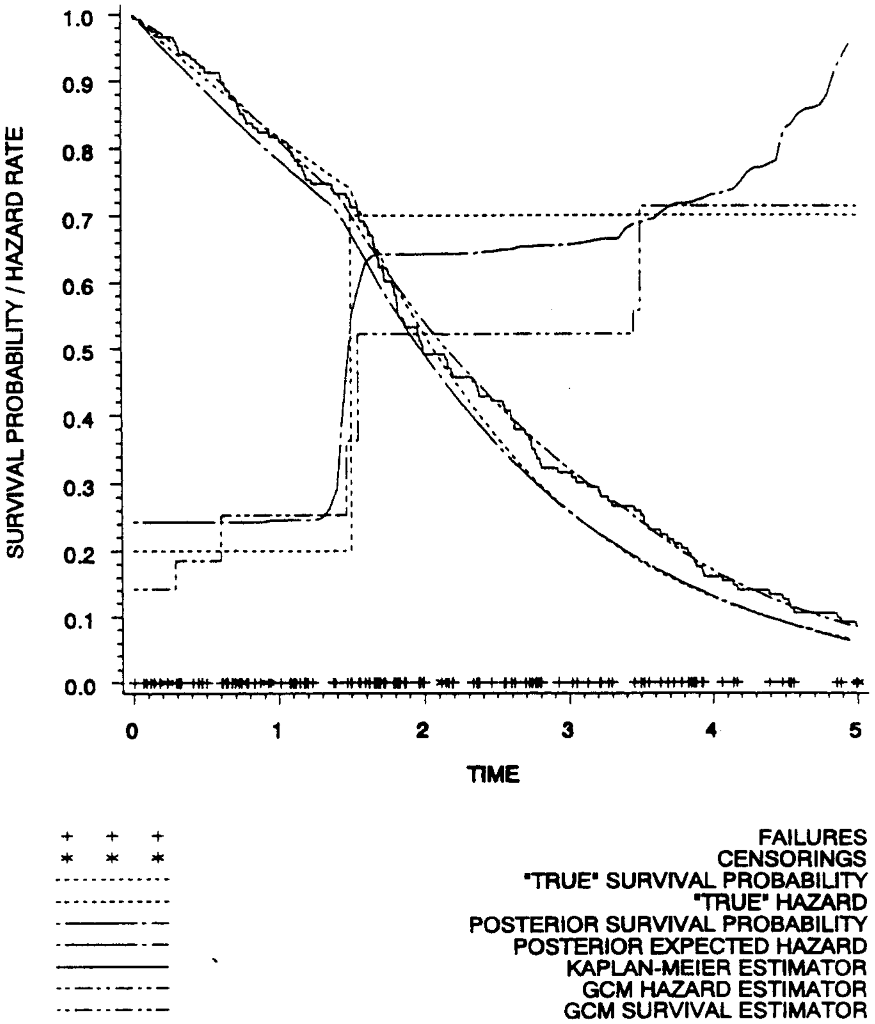

Figure 5. Results from fitting an IHR model to a simulated data set of 150 individuals; 5 were censored during the interval of study and 13 at the end. To check the robustness of the estimation algorithm, the true hazard process had only a single but very large jump (from 0.2 to 0.7). The hyperparameters were given the values $\alpha_0=1$, $\beta_0=2$, $\mu=0.5$, $\nu=2$. The Gibbs sampler ran for 5000 iterations and the first 500 were discarded.

As an illustration of the method, in Figures 4 and 5 are shown the results from the analysis of two simulated data sets. In addition to the plots used earlier in Figures 1 and 2, we have added the Huang and Wellner (1994) GCM-estimator of the hazard rate, which is piecewise constant, and the corresponding exponential plug-in survival probability estimator.

Note, finally, that the assumed IHR property concerns the hazard rate process $\{\lambda(t); t \geq 0\}$ which is here an unknown model parameter. The corresponding predictive hazard can be viewed as arising from a mixture of distributions, where the mixing is according to the posterior distribution. It is well known that IHR property is not necessarily preserved in mixing, and therefore the resulting hazard $\lambda_{pred}(t)$ may not be increasing even if the individual $\lambda(t)$-functions are. This is simply a consequence of the fact that a long survival is better explained by a low hazard rate, and therefore Bayes' rule gives it relatively more weight as the observed survival time of an individual becomes longer. On the other hand, in a setting where a large data set has been generated according to $\lambda_{true}(t)$, consistency of the estimation implies that $\lambda_{true}(t)$ can be "almost recovered". In practice this means that if $\lambda_{true}(t)$ is IHR and n is large, then also $\lambda_{pred}(t)$ should be IHR.

5. Final Remarks

Our implementation is different from other applications of the Gibbs sampler in the context of multiple change point models (Carlin et al. (1992), Stephens (1992)) in that here the number of change points is not specified in advance; in each iteration the algorithm updates dynamically the dimension of the problem, handling in this way the posterior distribution on an infinite-dimensional parameter space. From this point of view our method stands in a natural way between parametric (finite dimensional) and nonparametric (infinite dimensional) approaches to Bayesian inference from survival data. The particular structure of the considered model, i.e. piecewise constant hazard rate and prior distribution based on Poisson and gamma distributions, should be primarily viewed as a convenient way of arriving at a simple model formulation. In Bayesian estimation, where typically the main concern is in integrals with respect to the posterior distribution, such as predictive survival probabilities, it seems that the precise functional form of the hazard rate is less crucial than in the frequentist approach where obtaining a "good looking" point estimate function is important. Moreover, four hyperparameters with fairly clear-cut interpretations in their relation to the behaviour of the hazard rate seem to offer enough flexibility in actual statistical estimation. The choice of their values has a similar role in hazard estimation as the choice of the form and the bandwidth of the kernel in kernel smoothing. An important difference, however, is that here the hyperparameters are chosen first, and that the smoothing becomes a part of the algorithm itself. Should one prefer to think that also the hyperparameters are unknown, and therefore viewed as a random variables, it is a simple matter to add an additional level of hierarchy to the model and then extend the sampling algorithm to these variables.

Acknowledgement

This research was supported by a grant from the Academy of Finland. We are also grateful to Jian Huang and Jon Wellner for making their paper available to us prior to its publication.

References

Ammann, L. P. (1985). Conditional Laplace transforms for Bayesian nonparametric inference in reliability theory. *Stochastic Process. Appl.* **20**, 197-212.

Arjas, E. (1989). Survival models and martingale dynamics. *Scand. J. Statist.* **16**, 177-225.

Arjas, E., Haara, P. and Norros, I. (1992). Filtering the histories of a partially observed marked point process. *Stochastic Process. Appl.* **40**, 225-250.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.* **41**, 389-405.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467-485.

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

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356-367.

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

Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-186.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gilks, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4* (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 641-649, Oxford University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.

Huang, Y. and Wellner, J. (1994). Estimation of a monotone density or monotone hazard under random censoring.

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

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351-357.

Lo, A. Y. and Weng, C. S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227-245.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics* **26**, 265-275.

Ripley, B. D. (1987). *Stochastic Simulation.* John Wiley, New York.

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

Roberts, G. O. and Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Process. Appl.* **49**, 207-216.

Smith, A. F. M. (1991). Bayesian computational methods. *Philos. Trans. Roy. Soc. London Ser. A* **337**, 369-386.

Stephens, D. A. (1992). Bayesian retrospective multiple changepoint identification. Preprint.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897-902.

Thompson, M. E. and Thavaneswaran, A. (1992). On Bayesian nonparametric estimation for stochastic processes. *J. Statist. Plann. Inference* **33**, 131-141.

Department of Applied Mathematics and Statistics, University of Oulu, Linnanmaa 90570, Oulu, Finland.

(Received April 1993; accepted February 1994)