Analysis of Competing Risks by Using Bayesian Smoothing

DARIO GASBARRA and S. R. KARIA

University of Helsinki

ABSTRACT. We consider the competing risks set-up. In many practical situations, the conditional probability of the cause of failure given the failure time is of direct interest. We propose to model the competing risks by the overall hazard rate and the conditional probabilities rather than the cause-specific hazards. We adopt a Bayesian smoothing approach for both quantities of interest. Illustrations are given at the end.

Key words: censoring, data augmentation, gamma process prior, Dirichlet distribution, kernel smoothing, Markov chain Monte Carlo

1. Introduction

We shall deal with the following problem: we have i.i.d. random pairs (U_j, ζ_j), where $U_j \geqslant 0$ can be interpreted as failure time and $\zeta_j \in \{1, \ldots, d\}$ as the specific cause of failure. We observe a right censored sample $(T_j, \delta_j)_{j=1,\ldots,m}$, where T_j is the last time the individual was seen functioning, and the mark δ_j is 0 if the individual was censored, and otherwise $\delta_j = \zeta_j$. Under the independent censoring assumption, the question is how to estimate the joint distribution of (U, ζ).

Examples arise in many fields, such as pharmaceutical studies, in psychology and education, and in sociology. The general approach in modelling competing risks data is via the cause-specific hazards or subdistribution functions when the so called "latent lifetimes" under various risks are dependent. Our approach is to estimate separately the conditional distribution of the cause of failure given the failure time, and the overall hazard rate. Also, the non-parametric Bayesian approach is adopted to incorporate prior information which is not so with the usual non-parametric approach in which cause-specific hazards are estimated using Nelson–Aalen estimators. The proposed method enables one to model the competing risks via the conditional probability and the overall hazard rate and use the Bayesian technique to estimate the parameters involved in the model.

The cause-specific conditional probability can be viewed as a non-linear regression of ζ on U. Classical non-parametric regression techniques (see Prakasa Rao, 1983) motivate our specification of the prior. We take a (random) partition of the time interval and on each segment an independent d-dimensional Dirichlet distributed random vector is assigned. Then we define the cause-specific conditional distribution at time t by smoothing this piecewise constant d-dimensional process. Given the observations, posterior computations are carried out by a data-augmented Markov chain Monte Carlo.

The prior for the overall hazard rate is specified in a similar fashion. A number of authors have considered non-parametric Bayesian estimation of the cumulative hazard function $\Lambda(t)$ assuming it to be a stochastic process with independent increments (see Hjort, 1990 and references therein).

A conceptual problem is that the resulting random distribution is discrete with probability one, and a corresponding hazard rate does not exist. In order to obtain a random hazard rate, at some stage one should apply a convolution procedure to the random cumulative hazard. For

example, one could first compute the posterior expectations $\hat{\Lambda}_t = E(\Lambda_t|\text{data})$ and then estimate the hazard rate by

$$\hat{\lambda}_t = \int k(t, s)\, d\hat{\Lambda}_s, \tag{1}$$

where k is a continuous stochastic kernel.

Here we "smooth the prior first", and model the overall hazard rate $h(t)$ as the convolution of a gamma process $\Gamma(s)$ with a kernel $\hat{k}_X(t, s)$, where X is a random parameter. Such construction of a hazard rate was introduced by Lo & Weng (1989). For any choice of the parameter x in the convolution kernel, the posterior distribution of Γ will be a mixture of gamma processes, with an awkward mixing distribution over all possible data sets with the given sample size (th. 4.1. in Lo and Weng, 1989). For posterior computations, they propose Monte Carlo sampling from the posterior mixing distribution by using Aldous' "Chinese restaurant process", cf. Aldous (1985).

Wolpert & Ickstadt (1998) have the same construction in the context of spatial point processes and use Markov chain Monte Carlo (MCMC) to sample from the same data-augmented posterior. In their paper there is also a novel technique for sampling a general Lévy process (inverse Lévy measure (ILM) method).

We follow the same data-augmented MCMC approach, but because of the structure of the kernel $\tilde{k}_x(t, s)$, we sample only the increments of Γ over a coarse grid, without going to the fine resolution which is possible to achieve by the ILM-method. In section 5, we also compare the proposed hazard rate estimator with the estimator (1) based on Hjort's posterior beta process and the Nelson–Aalen estimator..

2. Preliminaries

2.1. Counting processes and competing risks

In this section we discuss some basic ideas about counting processes and competing risks. We assume that the ultimate failure is due to only one risk.

We define the individual counting processes as

$$N_{ij}(t) = 1_{\{U_j \leq t, \zeta_j = 1\}}, \quad i = 1, 2, \ldots, d,$$

$$N_j(t) = \sum_{i=1}^{d} N_{ij}(t),$$

$$N(t) = \sum_{j=1}^{n} N_j(t).$$

Let $\mathcal{F} = (\mathcal{F}_t, t \geq 0)$ be the filtration where $\mathcal{F}_t = \sigma(U_j, \zeta_j: U_j \leq t, j = 1, \ldots, m)$. The cause-specific and overall hazards are, respectively

$$d\Lambda_i(t) = \frac{dF_i(t)}{1 - F(t-)}, \quad i = 1, 2, \ldots, d,$$

and

$$d\Lambda(t) = \frac{dF(t)}{1 - F(t-)}$$

where $F_i(t) = P[U \leq t, \zeta = i]$ and $F(t) = P[U \leq t]$.

We assume that $\Lambda_i(t)$ are absolutely continuous with respect to Lebesgue measure. The cause-specific hazards and the conditional probability of interest are

$$h_i(t) = \frac{d\Lambda_i}{dt}(t), \tag{2}$$

$$h(t) = \frac{d\Lambda}{dt}(t) = \sum_{i=1}^{d} h_i(t) \tag{3}$$

and

$$\pi_i(t) = \frac{d\Lambda_i}{d\Lambda}(t). \tag{4}$$

By definition $h_i(t) = \pi_i(t)h(t)$, and it follows that $\pi_i(t) = P[\zeta = i | U = t]$, for $i = 1, 2, \ldots, d$ (see Brémaud, 1981, th 15 ch. II). Note that $\sum_{i=1}^{d} \pi_i(t) = 1$. Denote the vector $(\pi_1(t), \ldots, \pi_d(t))$ by $\pi(t)$. We concentrate on π and h to carry out the analysis. Comparative study of risks can be based on π. For example, proportionality of cause-specific hazards is equivalent to saying that the failure time and the cause of failure are independent, i.e. $\pi(t)$ is constant in t.

As usual, we assume independent censoring so that there is no need to model the censoring process. We refer to Andersen *et al*. (1993, ch. III).

The data is a right censored sample $(T_j, \delta_j)_{j=1,\ldots,m}$, where T_j is the last time the jth individual was under observation, and if a failure was observed at T_j we have $U_j = T_j$ and $\delta_j = \zeta_j$, otherwise $U_j > T_j$ and $\delta_j = 0$.

In the frequentist approach, $\{h(t),\ t > 0,\ i = 1, 2, \ldots, d\}$ are unknown determinsitic functions subject to estimation. One could either estimate the cause-specific hazard rates $\{h_i(t),\ t > 0,\ i = 1, 2, \ldots, d\}$ separately, or estimate the overall hazard rate $\{h(t),\ t > 0\}$ together with the time-dependent probability distribution $\{\pi_i(t),\ t > 0,\ i = 1, 2, \ldots, d\}$. To develop a Bayesian framework, we shall treat these unknown functions as random processes with a prior distribution to be specified. As mentioned earlier, one way is to specify a prior distribution for the cause-specific hazard rates $\{h_i(t),\ i = 1, 2, \ldots, d\}$ directly. Therefore, in the Bayesian framework, it makes perfect sense to say, for example, that the cause-specific hazard rate processes are stochastically independent under an assigned prior. Since the object of interest is

$$E(\pi_i(t)|\text{data}) = E\left(\frac{h_i(t)}{\sum_j h_j(t)}|\text{data}\right) \neq \frac{E(h_i(t)|\text{data})}{\sum_j E(h_j(t)|\text{data})},$$

we adopt the alternative way, and specify independent priors for the vector valued process $\{\pi_i(t),\ i = 1, \ldots, d\}$ and the overall hazard rate process $\{h(t)\}$. Note that the resulting cause-specific hazard rate processes $h_i(t) = h(t)\pi_i(t),\ i = 1, \ldots, d$ are not necessarily independent under such a prior specification.

2.2. The likelihood function

The likelihood function based on right censored survival data (t_j, δ_j) corresponding to a sample of m identical units, is

$$L(\pi, h) = \left(\prod_{j:\delta_j \neq 0} \pi_{\delta_j}(t_j)\right)\left(\exp\left\{-\int_0^\infty Y(u)h(u)du\right\}\prod_{j:\delta_j \neq 0} h(t_j)\right) = L(\pi)L(h), \tag{5}$$

where $Y(u) = \#\{\text{individuals at risk at time } u\}$.

Since the likelihood function factorizes into two parts, the estimation of $\{\pi_i(t)\}$ and $\{h(t)\}$ can be carried out separately. In particular, assuming that π and h are independent under the

prior distribution implies that they are also independent under the posterior. The posterior expectation of the cause specific hazards will be simply $E(h_i(t)|\text{data}) = E(\pi_i(t)|\text{data})E(h(t)|\text{data})$.

3. Prior specifications

For smoothing the priors we use an absolutely continuous kernel $K(t, ds) = k(t, s)ds$ on $[a, b] \times [a, b]$. It is assumed to have the following truncated convolution form hereafter:

$$k(t, s) = \frac{\phi\left(\frac{t-s}{\eta}\right)1_{[a,b]}(s)}{\int_a^b \phi\left(\frac{t-u}{\eta}\right)du} \tag{6}$$

where $1_I(s)$ is the indicator function of the interval I, ϕ is the standard normal density and the bandwidth η is fixed. Smoothing by a convolution procedure is mathematically equivalent to accounting for independent errors in the location of the data points. It is quite common to model such errors with the normal distribution. Also the bandwidth η will be fixed according to our prior beliefs about the smoothness (or the roughness) of the process we are modelling.

3.1. Prior distribution for the process $\{\pi(t), t \in [a, b]\}$

We start from a Poisson process $X = \{X_1, \ldots, X_N\}$ with a given intensity μ on the interval $[a, b]$. This defines a Voronoi partition of $[a, b]$ into subintervals $I_1, \ldots, I_N$. Let $Q_1, \ldots, Q_N$ be independent d-dimensional Dirichlet distributed random vectors with the same parameter $\alpha \in (\Re^+)^d$. We define a piecewise constant vector valued function on $[a, b]$

$$q(s) = \sum_{l=1}^{N} Q_l 1_{I_l}(s), \tag{7}$$

and finally we smooth this function by the kernel (5), getting the conditional probability function

$$\pi(t) = \int_a^b k(t, s)q(s)ds = \frac{\sum_{l=1}^{N} Q_l \Phi\left(\frac{I_l - t}{\eta}\right)}{\sum_{l=1}^{N} \Phi\left(\frac{I_l - t}{\eta}\right)}, \tag{8}$$

where Φ denotes the probability measure under the standard normal distribution, and $(I - t)/\eta$ is a shifted and rescaled interval.

Prior knowledge could be represented by a suitable choice of the vector α. The parameter μ, as well as η, control the smoothness of the vector valued process $\{\pi(t)\}$.

We rewrite the likelihood of the process $\{\pi(t)\}$ as a function of $\{q(s)\}$,

$$L(\pi) = \prod_{j:\delta_j \neq 0} \frac{\sum_{l=1}^{N} Q_l^{(\delta_j)} \Phi\left(\frac{I_l - T_j}{\eta}\right)}{\Phi\left(\left[\frac{a - T_j}{\eta}, \frac{b - T_j}{\eta}\right]\right)}, \tag{9}$$

where $Q_l^{(i)}$ is the ith coordinate of the vector Q_l. Note that the likelihood does not depend on the censored observations.

3.2. Data augmentation

The smoothed parameter function $\{\pi(t)\}$ is a convolution of the function $\{q(s)\}$ with a (truncated) normal density. This is equivalent to the following construction: for each observation T_j such that $\delta_j \neq 0$, sample $S_j \simeq \mathcal{N}(T_j, \eta^2)1_{[a,b]}$ and take δ_j discrete with distribution $q(S_j)$. Then the marginal conditional distribution of δ_j given T_j is just $\pi(T_j)$. S_j can be interpreted as the latent "exact" location of the observation which was then contaminated by a normal error. In the sequel we shall use the latent variables S_j in order to simplify the computations.

3.3. The posterior distribution

Once we have introduced the latent random variables S_j for each observation (T_j, δ_j) with $\delta_j \neq 0$, the augmented likelihood of $\{\pi(t)\}$ is

$$\tilde{L}(\pi) = \prod_{j:\delta_j \neq 0} 1_{[a,b]}(S_j) \exp\left(-\frac{1}{2}\left(\frac{T_j - S_j}{\eta}\right)^2\right) q^{(\delta_j)}(S_j). \tag{10}$$

Note that, instead of the normal density, any other distribution could have been used.

The posterior measure is then proportional to

$$\text{Poisson}_\mu(dx) \prod_{j:\delta_j \neq 0} 1_{[a,b]}(s_j) \exp\left(-\frac{1}{2}\left(\frac{T_j - s_j}{\eta}\right)^2\right) ds_j \prod_{l=1}^{N(x)} C_l(x, S)\text{Dirichlet}(dq_l | \alpha + M_l), \tag{11}$$

where the factors C_l are given by

$$C_l(x, S) = \frac{\alpha_1(\alpha_1+1)\cdots(\alpha_1+M_l(1)-1)\cdots\alpha_d(\alpha_d+1)\cdots(\alpha_d+M_l(d)-1)}{(\alpha_1+\cdots+\alpha_d)(\alpha_1+\cdots+\alpha_d+1)\cdots(\alpha_1+\cdots+\alpha_d+M_l(1)+\cdots+M_l(d)-1)}, \tag{12}$$

I_l is the interval in the partition corresponding to the marked point (X_l, Q_l), and the $M_l(i) = M_l(i; S, X) = \#\{j: S_j \in I_l \text{ and } \delta_j = i\}$ depend on the data and both the point configuration X and the vector of latent variables S.

Note that under the posterior distribution (i.e. conditionally on the r.v.s T_j, δ_j) the probability vectors Q_l follow a mixed Dirichlet distribution: conditionally on the latent variables S_j and on the point process X, the Q_ls are independent Dirichlet vectors. Here X and (S_j) are the mixing parameters.

3.4. A Markov chain Monte Carlo procedure

We describe a cycle of the MCMC algorithm, where in the three different steps the variables (S_j), the random grid X, and the random variables (Q_l) are updated in turn given the current values of the remaining parameters. When possible we update by sampling a parameter from a conditional distribution, otherwise we use a Metropolis–Hastings step.

In this way we construct an ergodic Markov chain on the parameter space, such that the

posterior distribution is the invariant measure of the chain. Therefore posterior expectations coincide with ergodic averages, and are approximated by generating, on a computer, one realization of the Markov chain (see Besag *et al*., 1995).

We resume a complete updating cycle into three updates:

Update 1. Sample independently the random variables S_j from the full conditionals given the current values of the marked point process $\{(X_l, Q_l),\ l = 1, \ldots, N(X)\}$ and the data (T_j). By (9) it is straightforward to see that S_j become independent with respective distributions

$$\frac{k(T_j, s)q^{(\delta_j)}(s)ds}{\int_0^\infty k(T_j, u)q^{(\delta_j)}(u)du}. \tag{13}$$

It is easy to sample S_j from (13) by inverting the distribution function or by rejection sampling.

Update 2. Update with Metropolis–Hastings' step the grid configuration x from the marginal conditional of X given the current values of (S_j) and the data (T_j), for this purpose, we use the birth and death Metropolis algorithm proposed by Geyer & Møller (1994) for point processes which have a density with respect to the Poisson process. Starting from $X = \{X_1, \ldots, X_{N(X)}\}$, with probability $1/2$ we propose either to add a point to the configuration (upstep) or to delete a point unless $N(X) = 1$ (downstep).

upstep. Sample Y uniformly in the interval $[a, b]$, and accept the proposal $X^* = X \cup \{Y\}$ with probability

$$\min\left(1, \frac{\mu(b-a)}{(N(X)+1)} \frac{\prod_{l=1}^{N(X^*)} C_l(X^*, S)}{\prod_{l=1}^{N(X)} C_l(X, S)}\right) \tag{14}$$

downstep. Choose uniformly at a random point Y within $\{X_1, \ldots, X_{N(X)}\}$ and accept the proposal $X^* = X \backslash \{Y\}$ with probability

$$\min\left(1, \frac{N(X)}{\mu(b-a)} \frac{\prod_{l=1}^{N(X^*)} C_l(X^*, S)}{\prod_{l=1}^{N(X)} C_l(X, S)}\right). \tag{15}$$

So in each cycle either we add one point, we delete one point or we remain with the same point configuration.

Update 3. To complete the cycle, given the correct values of (X_l), (S_j), and the data (T_j), sample independently the random vectors Q_l from the full conditionals

$$Q_l \simeq \text{Dirichlet}(\alpha + M_l(X, S)), \quad l = 1, \ldots, N(X). \tag{16}$$

Note that just after update 2 only (X_l) and (S_l) are in equilibrium with the posterior distribution. This is allowed since we are able to restore the equilibrium with step 3, sampling (Q_l) from the full conditionals (see Besag *et al*., 1995).

3.5. Prior distributions for the process $\{h(t) \in [a, b]\}$

Here we denote by $|I|$ the length of the interval I.

We construct the hazard rate process as a special case of the model in Lo & Weng (1989), with additional randomness resulting from the random choice of the convolution kernel. We take a Poisson process $X' = \{X'_1, \ldots, X'_{N(X')}\}$ with constant intensity μ', and we define another random Voronoi partition $I'_1, \ldots, I'_{N(X')}$ of $[a, b]$, and given independent random variables G_l, $l = 1, \ldots, N(X')$ with respective distributions Gamma$(\alpha|I'_l|, \beta|I'_l|)$, we define the piecewise constant process

$$g(s) = \sum_{l=1}^{N(X')} G_l 1_{I'_l}(s). \tag{17}$$

This is then smoothed, yielding the overall hazard rate

$$h(t) = \int_a^b k'(t, s) g(s) ds = \frac{\sum_{l=1}^{N(X')} G_l \Phi\left(\frac{I'_l - t}{\eta}\right)}{\sum_{l=1}^{N(X')} \Phi\left(\frac{I'_l - t}{\eta}\right)}. \tag{18}$$

To simplify the numerical calculations, here we omit the normalizing constant from the kernel (5), and instead use the substochastic kernel

$$k'(t, s) = \frac{1}{\eta} \phi\left(\frac{t - s}{\eta}\right) 1_{[a,b]}(s) \tag{19}$$

for the hazard. We also introduce the notation

$$K'(t, I) = \int_I k'(t, s) ds.$$

A careful reader will note that we are within the framework of Lo & Weng (1989), since the process g and h can be given also in the form

$$g(s) = \int_0^\infty \prod_{l=1}^{N(X')} 1_{I'_l}(s) 1_{I'_l}(u) \Gamma(du) \tag{20}$$

$$h(t) = \int_0^\infty \prod_{l=1}^{N(X')} K'(t, I'_l) 1_{I'_l}(u) \Gamma(du) = \int_0^\infty \tilde{k}_{X'}(t, u) \Gamma(du), \tag{21}$$

where $\Gamma(u)$ is a homogeneous gamma process with driving measure $d\alpha = \alpha dt$ and constant scale function β. The random variates $G_l|I'_l|$ are the increments of $\Gamma(u)$ on the interval I'_l. In fact the point process X' is a random parameter which controls the form of the kernel $\tilde{k}'_X(t, u) = \sum_{l=1}^{N(X')} K'(t, I'_l) 1_{I'_l}(u)$ in (21).

This two step Bayesian convolution procedure with a randomly chosen kernel, is an analogue to frequentist adaptive kernel smoothing procedures: under the posterior distribution, the random choice of the kernel $\tilde{k}'_X(t, u)$ will depend on the global pattern of the data.

Prior knowledge can be represented by assigning the driving measure $d\alpha$ and the scaling function β of the underlying gamma process $\{\Gamma(u)\}$.

The parameter μ' driving the point process X' as well as η, control the smoothness of the process $\{h(t)\}$.

4. Posterior distribution and MCMC for the hazard rate

Recall that the likelihood for the hazard rate process has the form

$$L(h) = \prod_{j:\delta_j \neq 0} \left(\int_0^\infty k'(T_j, s) g(s) ds \right) \exp\left\{ - \int_0^\infty \left(\int_0^\infty Y(t) k'(t, s) dt \right) g(s) ds \right\} \tag{22}$$

and the prior is

$$\text{Poisson}_{\mu'}(dx') \prod_{l=1}^{N(x')} \text{Gamma}(G_l; \alpha|I'_l|, \beta|I'_l|) \tag{23}$$

To perform Bayesian computations we need to sample the process $\{g(s)\}$ from the posterior distribution $\Pr(dg|T_j; j = 1, \ldots, n) \propto \Pr(dg)L(h(g))$.

We introduce for j such that $\delta_j \neq 0$ random variables S'_j, which are conditionally independent given $\{g(s)\}$, (T_j), with distributions

$$\frac{k'(T_j, s) g(s) ds}{\int_0^\infty k'(T_j, u) g(u) du}. \tag{24}$$

Note that the joint density of the augmented data (T_j, S'_j), conditionally on $\{g(s)\}$, is proportional to

$$\prod_{j:\delta_j \neq 0} k'(T_j, S'_j) g(S'_j) \exp\left\{ - \int_0^\infty \left(\int_0^\infty Y(t) k'(t, s) dt \right) g(s) ds \right\}. \tag{25}$$

Then we construct a Markov chain Monte Carlo sampler for the posterior distribution of $\{g(s)\}$ with one cycle consisting of three updates.

(*Update 1*) Sample independently the random variables S'_j from the full conditionals (24) given the current vaues of (X'_l), (G_l) and the data (T_j, δ_j). These are univariate distributions and are easy to sample by inverting the distribution function or by rejection sampling. This is much simpler than the procedure of Lo & Weng (1989) where the same random variables (S'_j) are sampled jointly conditioning only on the data (T_j, δ_j), losing conditional independence.

(*Update 2*) Update with a Metropolis–Hastings' step the grid configuration X' from the marginal conditional given the current values of (S'_j) and the data (T_j, δ_j). Note that the marginal likelihood of the point process X' is

$$\prod_{l=1}^{N(X')} B_l(X', S', T), \tag{26}$$

where

$$B_l(X', S', T) = \frac{\Gamma(M'_l + \alpha|I'_l|)}{\Gamma(\alpha|I'_l|)} \frac{(\beta|I'_l|)^{\alpha|I'_j|}}{\left(\beta|I'_l| + \int_0^\infty Y(t) K'(t, I'_l) dt \right)^{M'_l + \alpha|I'_l|}} \tag{27}$$

with

$$M'_l = \#\{j: S'_j \in I'_l\} \tag{28}$$

We again use a birth and death proposal kernel, where with probability $1/2$ we add to the

configuration a new point sampled from the uniform distribution over the interval $[a, b]$, and otherwise we propose to delete from the configuration a uniformly chosen point.

Then the acceptance probabilities for the transition $X' \to X^*$ are

upstep:

$$\min\left(1, \frac{\mu'(b-a)}{(N(X')+1)} \frac{\prod_{l=1}^{N(X^*)} B_l(X^*, S, T)}{\prod_{l=1}^{N(X')} B_l(X', S', T)}\right) \tag{29}$$

downstep:

$$\min\left(1, \frac{N(X')}{\mu'(b-a)} \frac{\prod_{l=1}^{N(X')} B_l(X', S', T)}{\prod_{l=1}^{N(X^*)} B_l(X^*, S', T)}\right) \tag{30}$$

(*Update 3*) Sample the random variables (G_l) from the full conditionals given the current values of (X'_l), (S'_j) and the data (T_j, δ_j).

It is easy to see that, given the augmented sample (T_j, S'_j), the conditional distribution of the random variables $\{G_l, l = 1, \ldots, N(X')\}$ becomes conjugate, meaning that they are conditionally independent and gamma distributed with shape and scale parameters $(\alpha|I'_l| + M'_l)$ and $(\beta|I'_l| + \int_0^\infty Y(t)\, K'(t, I'_l)dt)$ respectively, where M'_l was defined in (28).

5. Illustration and discussion

As an illustration, we analysed contraceptive failure data. The data consist of eventually right censored failure times of 547 hormonal intra-uterine devices, giving a total of 185 observed failures, in a 5-year study period.

The possible causes of failure were categorized into five risk classes: (1) pregnancy, (2) expulsion, (3) amenorrhea (suppression of menstruation), (4) bleeding and pain, and (5) hormonal disturbances. For the analysis of the 5-dimensional process $\pi(t)$ prior parameters were given values $\alpha = (1, 1, 1, 1, 1)$, $\mu = 10$ and $\eta = 0.27$, where our time unit is one year. We did not have any prior information about $\pi(t)$ and this prior is a tentative way to express our ignorance. The pointwise posterior mean of $\pi(t)$, resulting from the Bayesian computations, which is also the posterior predictive distribution of ζ given $U = t$, is shown in Fig. 1. The posterior predictive probability of termination due to pregnancy increases with time after the insertion of an IUD. The posterior predictive probability of termination due to expulsion assumes its highest value right after the insertion of an IUD, and then decreases with time. The posterior predictive probability of termination due to amenorrhea increases with time since such changes in the bleeding pattern may not be acceptable to a woman. Special counselling is needed in order to prevent terminations due to amenorrhea. The posterior predictive probability of termination due to hormonal reasons increases with time during the initial months of usage. The reason for this is that the device used is a hormonal contraceptive method. Once the body accepts the hormonal changes, the probability of termination decreases. The posterior predictive probability of termination due to bleeding and pain fluctuates around 0.3 over the study period. For comparison, the classical Nadaraya–Watson non-parametric regression estimator based on

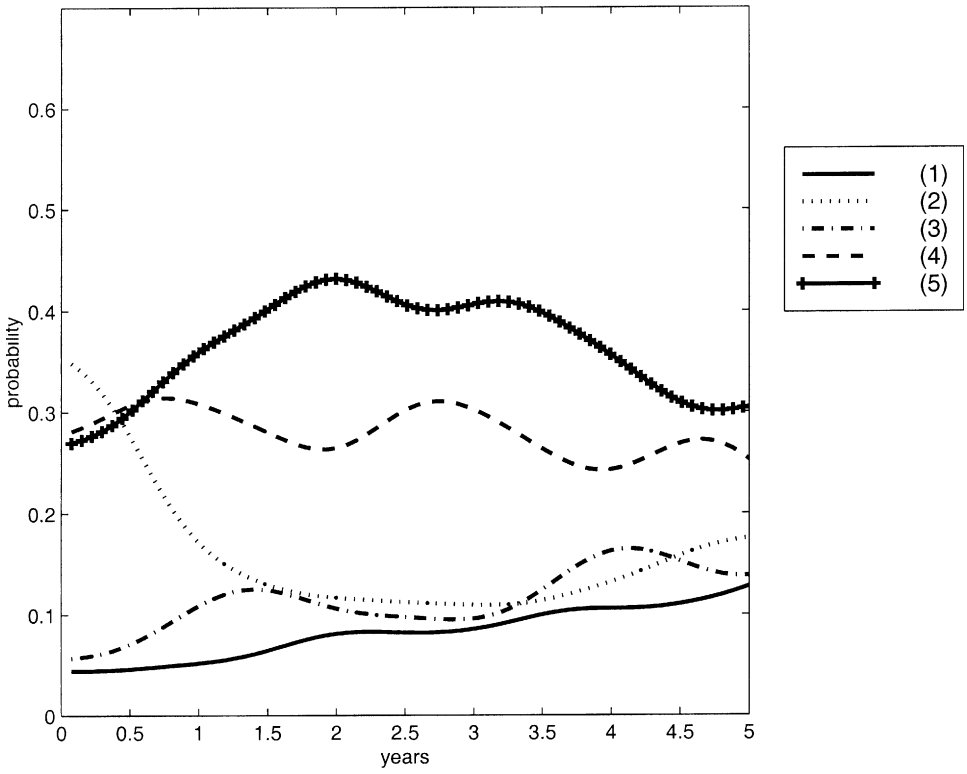

Fig. 1. Posterior means of the 5-dimensional probability process $\pi(t)$. The curves are labelled by the corresponding coordinates: (1) pregnancy, (2) expulsion, (3) amenorrhea, (4) bleeding and pain, and (5) hormonal disturbances. Prior parameters were given values $\alpha = (1, 1, 1, 1, 1)$, $\mu = 10$ and $\eta = 0.27$.

the same kernel and bandwidth is shown in Fig. 2. Note that the ordering of the estimators of $\pi(t)$ is the same in Figs 1 and 2.

For the overall hazard rate $h(t)$ we used our smooth prior with parameters $\alpha(dt) = dt$, $\beta = 1$, $\mu' = 10$, $\eta = 0.27$. This prior was not based on actual consultation with experts, but simply reflects our guesses about the order of magnitude of the hazard rate.

For a comparison, here we consider also the "smooth the posterior afterwards" approach. We model the cumulative hazard $\Lambda(t)$ with a beta process prior. It means that the increments $\Lambda(dt)$ over infinitesimal intervals dt are independent and with distributions Beta$(b(t)a(dt), b(t)(1 - a(dt)))$, where $b(t)$ is a non-negative function and da is a non-negative Radon measure with $a(\{t\}) \leqslant 1$ for all t. This family of processes was introduced in Hjort (1990).

It follows that the posterior expectation given the counting process' observations is

$$\hat{\Lambda}(t) = \int_0^t \frac{1}{b(s) + Y(s-)}(a(ds) + dN(s)),$$

where $N(t)$ is defined in section 2. For $a = b = 0$ this is the Nelson–Aalen estimator. The parameters were chosen in order to match with our prior beliefs for the smooth hazard rate. We had $a(dt) = a\,dt$, $b(t) = b$ with $a = b = 1$. In Fig. 3 we compare the estimators of the cumulative hazard: we have $\int_0^t \hat{h}(s)ds$ where $\hat{h}(t) = E(h(t)|\text{data})$, the posterior beta

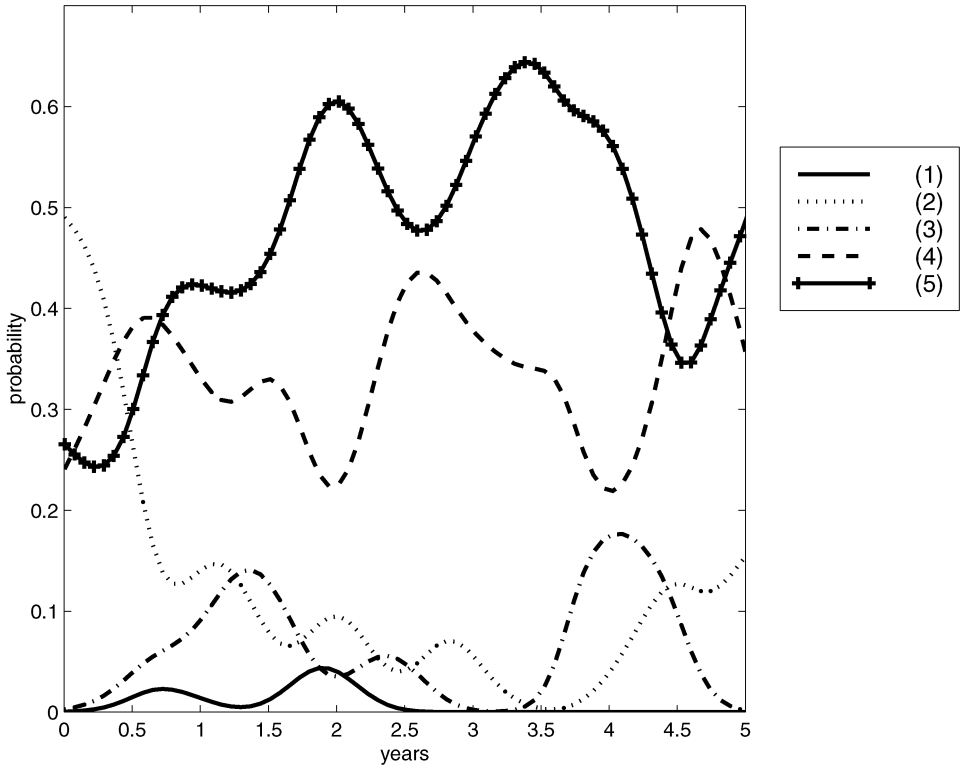

Fig. 2. Kernel smoothed non-parametric regression estimator of $\pi(t)$, based on a truncated normal kernel with bandwidth 0.27. The curves are labelled by the corresponding coordinates: (1) pregnancy, (2) expulsion, (3) amenorrhea, (4) bleeding and pain, and (5) hormonal disturbances.

process $\hat{\Lambda}(t)$, and the Nelson–Aalen estimator. To show the uncertainty in the posterior about the parameter curves, we give the pointwise posterior mean $\pm$ posterior pointwise standard deviations in Figs 3 and 4. Our cumulative hazard estimator is close to Hjort's posterior beta process. The deviation from the Nelson–Aalen estimator is due to the prior.

By convolving the latter two processes with the normalized kernel (6) as in formula (1), we obtain estimators of the hazard rate process, which are compared with our estimator $\hat{h}(t)$ in Fig. 4.

Looking at these pictures, one may argue that in practice it does not really matter whether the smoothing is built-in at the prior level or it is simply performed afterwards on the posterior expectation of Hjort's beta process. This is not a surprise, and it reassures us on the issue of convergence of the MCMC algorithm. Although a few thousands of iterations of the algorithm were enough to obtain the results, we also had longer runs of 10^5 iterations discarding a burn-in period of 1000 iterations. We monitored the acceptance rate in the Metropolis step which was around 80%. Because of these facts and the heavy use of conditional independence in the algorithm, we are quite confident on the numerical results.

This Bayesian cumulative hazard estimator can be compared with other Bayesian non-parametric methods based on piecewise constant processes. In Arjas & Gasbarra (1994), and in Arjas & Heikkinen (1996), the hazard rate function is supposed to be piecewise constant over a

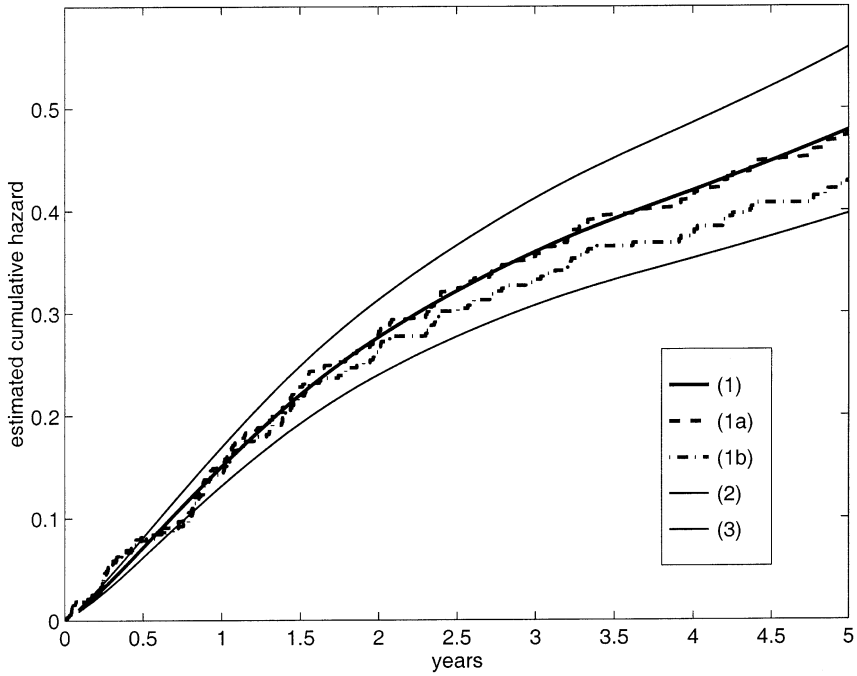

Fig. 3. Cumulative hazard estimates: (1) the integral of the posterior expectation of our smooth hazard rate process (2, 3) the curve (1) $\pm$ the corresponding pointwise posterior standard deviations, (1a) the posterior expectation of Hjort's beta process, and (1b) Nelson–Aalen's estimator. The smooth hazard rate process had prior parameters $\alpha(dt) = dt$, $\beta = 1$, $\mu = 10$, $\eta = 0.27$. The beta process had prior parameters $a(dt) = dt$ and $b(t) = 1$.

random grid, and a prior dependence structure is assigned for the values over neighbouring intervals.

When a Bayesian statistican models the cumulative hazard rate as an increasing process with independent increments and wants to say something about the hazard rate function, it is clear that a smoothing procedure has to apply at some stage. Whether this is built-in at the prior level, or it is done only at the end of the analysis by convolving the posterior expectation with a kernel, is a matter of "Bayesian" taste. To smooth the posterior Lévy process according to (1), is not a problem and it gives a perfectly reasonable approach. However, if the prior hazard rate function is a kernel smoothed Lévy process, then under the posterior the latter process will be a mixture of Lévy processes, where the mixing parameter is the vector of latent variables (S_j). In the prior specification it is also possible to choose the kernel randomly. Under the posterior distribution, the kernel will adapt itself to the data-pattern, according to the marginal likelihood (26).

Acknowledgements

We are thankful to Professor Elja Arjas for making this joint work possible, and for the constructive discussions we had with him. We acknowledge Leiras OY, which provided the data. The work of the first author was supported by the Department of Computer Science, University of Helsinki, and by the Academy of Finland.

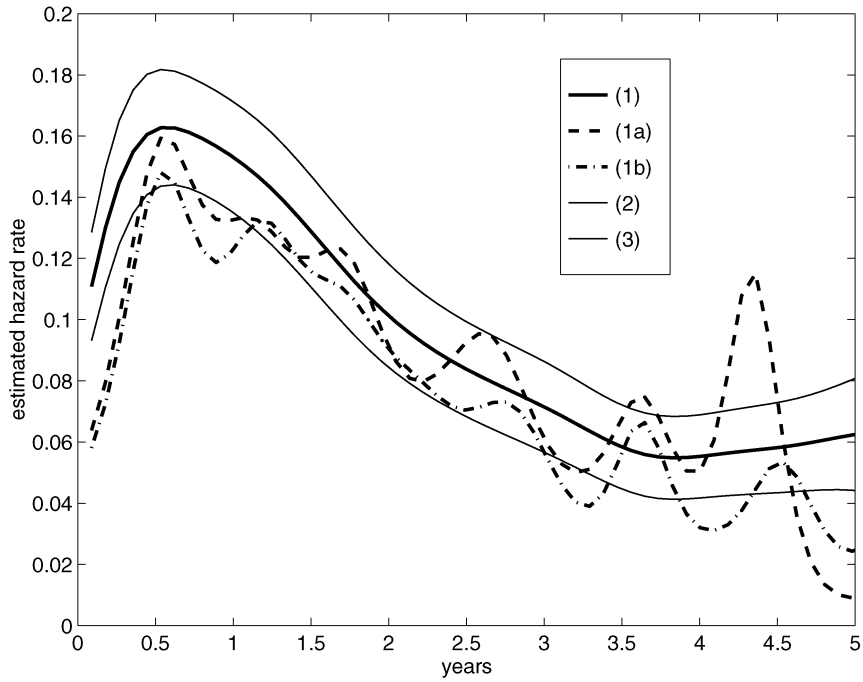

Fig. 4. Hazard rate estimates: these are given by the posterior expectation of our smoothed hazard rate process (1), the curve (1) $\pm$ the corresponding pointwise posterior standard deviations (2, 3), the convolution of the posterior expectation of the beta process (1a) and the convolution of the Nelson–Aalen estimator (1b), both with a truncated normal kernel with bandwidth $\eta = 0.27$.

References

Aldous, D. J. (1986). Exchangeability and related topics. In *Ecole d'eté de probabilités de St. Flour 1983*. Spinger Lecture Notes in Mathematics 1117, Spinger-Verlag, New York.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Spinger-Verlag, New York.

Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Arjas, E. & Heikkinen, J. (1996). An algorithm for nonparametric estimation of a Poisson intensity. *Comput. Statist*. **12**, 385–402.

Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci*. **10**, 3–66.

Brémaud, P. (1981). *Point processes and queues* Springer-Verlag, New York.

Geyer, C. J. & Møller, J. (1994). Simulation and likelihood inference for spatial point processes. *Scand. J. Statist*. **21**, 359–373.

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

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

Prakasa Rao, B. L. S. (1983). *Nonparametric functional estimation.* Academic Press, New York.

Wolpert, R. L. & Ickstadt, K. (1998). Gamma/Poisson random field models for spatial statistics. *Biometrika* **85**, 251–267.

Received September 1997, in final form January 2000

Dario Gasbarra, Rolf Nevanlinna Institute, University of Helsinki, P. L. 4, 00014 Helsinki Finland.