Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data

Nils Lid Hjort

The Annals of Statistics, Vol. 18, No. 3. (Sep., 1990), pp. 1259-1294.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199009%2918%3A3%3C1259%3ANBEBOB%3E2.0.CO%3B2-U

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1990, Vol. 18, No. 3, 1259–1294

NONPARAMETRIC BAYES ESTIMATORS BASED ON BETA PROCESSES IN MODELS FOR LIFE HISTORY DATA

BY NILS LID HJORT

Norwegian Computing Centre and University of Oslo

Several authors have constructed nonparametric Bayes estimators for a cumulative distribution function based on (possibly right-censored) data. The prior distributions have, for example, been Dirichlet processes or, more generally, processes neutral to the right. The present article studies the related problem of finding Bayes estimators for cumulative hazard rates and related quantities, w.r.t. prior distributions that correspond to cumulative hazard rate processes with nonnegative independent increments. A particular class of prior processes, termed beta processes, is introduced and is shown to constitute a conjugate class. To arrive at these, a nonparametric time-discrete framework for survival data, which has some independent interest, is studied first.

An important bonus of the approach based on cumulative hazards is that more complicated models for life history data than the simple life table situation can be treated, for example, time-inhomogeneous Markov chains. We find posterior distributions and derive Bayes estimators in such models and also present a semiparametric Bayesian analysis of the Cox regression model. The Bayes estimators are easy to interpret and easy to compute. In the limiting case of a vague prior the Bayes solution for a cumulative hazard is the Nelson–Aalen estimator and the Bayes solution for a survival probability is the Kaplan–Meier estimator.

1. Introduction and summary. Let $X_1, \dots, X_n$ be independent and identically distributed (iid) with an unknown cumulative distribution function (cdf) F on $[0, \infty)$ and suppose the data may be subject to right censoring. The problem of constructing nonparametric Bayes estimators for F involves placing a probability distribution on the space $\mathscr{F}$ of cdf's, i.e., viewing F as a stochastic process. Ferguson (1973) introduced the class of Dirichlet processes and used these as priors for F. Later Doksum (1974), Ferguson (1974), Susarla and Van Ryzin (1976), Ferguson and Phadia (1979), Dykstra and Laud (1981) and Padgett and Wei (1981) contributed new versions of prior processes and corresponding Bayes estimators for F.

The approach of Ferguson and Phadia (1979) was to write $F(t) = 1 - \exp\{-B(t)\}$ and let B be a Lévy process, i.e., one having independent nonnegative increments. They showed that B is still of the Lévy type given the data and they provided formulae for the posterior expectation of $\exp\{-B(t)\}$. One may also take interest in the *hazard rate* (or intensity, or force of transition) $\alpha(t) = F'(t)/F[t, \infty)$ in the model above. α is as basic as F when it comes to understanding the survival phenomenon under study, and the hazard rate

Received October 1985; revised September 1989.

AMS 1980 *subject classifications*. Primary 62C10; secondary 60G57.

Key words and phrases. Beta processes, censoring, Cox regression, cumulative hazard, Lévy process, nonparametric Bayes, time-discrete, time-inhomogeneous.

concept is more easily and understandably generalized to more complicated models like Markov chains with several possible states for each individual. Now $\alpha(t)$ is as difficult to estimate with good precision as is the density $F'(t)$, so we prefer working with the *cumulative hazard* $A(t) = \int_0^t \alpha(s)\,ds$. When F is continuous, A is just the $B = -\log(1 - F)$ mentioned above. The cumulative hazard can be defined even when F has no density, and such a general definition is necessary since most of the cdf's and cumulative hazards encountered in the present article happen to be discrete with probability 1. General correspondence formulae are

$$A(t) = \int_{[0,t]} \frac{dF(s)}{F(s,\infty)}, \qquad F(t) = 1 - \prod_{[0,t]} \{1 - dA(s)\},$$

as discussed in Section 3. One consequence of the discreteness is that A is not equal to $B = -\log(1 - F)$ anymore. We view the hazard rate, canonically defined as above, as a quantity of central importance in models for life history data and choose to concentrate on A instead of B. Thus a natural task to undertake is the nonparametric Bayesian estimation of A in the simple survival model above and later on in more complex models, like the competing risks framework, time-inhomogeneous Markov chains and regression models of the Cox variety.

It turns out that a particular class of prior distributions for A is particularly well suited to this task. The class is rich, each member has large support, its parameters are easily interpreted and explicit Bayes estimators can be derived, both for A and for related quantities. These processes produce cumulative hazard rates whose increments are independent and approximately beta distributed and are termed *beta processes*. A particular transformation of a given Dirichlet process produces a special case of the beta process, but the beta processes form a much larger and more flexible class.

The beta processes are discussed in Section 3. They are constructed and perhaps best understood as fine limits of time-discrete processes. The time-discrete case, reaching in generality to time-inhomogeneous Markov chains, is treated separately and briefly in Section 2 and has some independent interest. Section 4 considers posterior distributions and Bayes estimators in the simple time-continuous survival model mentioned first. The principal result is that if A is a beta process a priori, then it still is a posteriori. Section 5 extends the results to time-inhomogeneous Markov chains. The classical Nelson–Aalen nonparametric estimator of a cumulative hazard rate emerges as the Bayes solution in the limiting vague prior case and, correspondingly, Kaplan–Meier type estimators are limits of Bayes estimators for survival probabilities. Next, Section 6 gives a brief description of a Bayesian semiparametric treatment of Cox's regression model, in which the baseline hazard is given a beta process prior and the β coefficients a general prior density. This extension from homogeneous models to regression models is important and demonstrates the wider applicability of beta processes. Finally, some problems for future research are briefly discussed in Section 7, along with some complementing remarks.

2. Nonparametric time-discrete survival analysis.

2.1. *A time-discrete model with censoring.* Let X be a variable taking values in $\mathscr{X} = \{0, b, 2b, \dots\}$ and let

$$
\begin{aligned}
&f(jb) = \Pr\{X = jb\}, \qquad F(jb) = \Pr\{X \le jb\} = \sum_{i=0}^{jb} f(ib),\\
&\alpha(jb) = \Pr\{X = jb | X \ge jb\} = f(jb)/F[jb, \infty),\\
&A(jb) = \sum_{i=0}^{j} \alpha(ib),
\end{aligned}
\tag{2.1}
$$

for $j \ge 0$. α is the *hazard rate*, while A will be called the *cumulative hazard rate*. Note that F and f can be recovered from knowledge of A:

$$
\begin{aligned}
&F(jb) = 1 - \prod_{i=0}^{j} \{1 - \alpha(ib)\},\\
&f(jb) = \left[\prod_{i=0}^{j-1} \{1 - \alpha(ib)\}\right]\alpha(jb), \qquad j \ge 0.
\end{aligned}
\tag{2.2}
$$

Let $X_1, \dots, X_n$ be iid from this distribution, but assume that the observations are subject to right censoring. Thus what one observes is $(T_1, \delta_1), \dots, (T_n, \delta_n)$, where $T_i = \min(X_i, c_i)$, c_i being the censoring time for individual number i, and $\delta_i = I\{X_i \le c_i\}$. Consider the counting process N and the number-at-risk process Y, given by

$$
\begin{aligned}
&N(jb) = \sum_{i=1}^{n} I\{T_i \le jb \text{ and } \delta_i = 1\},\\
&Y(jb) = \sum_{i=1}^{n} I\{T_i \ge jb\}, \qquad j \ge 0.
\end{aligned}
\tag{2.3}
$$

We will sometimes write $dN(jb)$ for the increment $N(jb) - N((j - 1b)$ at time jb.

The likelihood of what is observed can be written

$$
\begin{aligned}
L(\text{data}) &= \left[\prod_{i:\, \delta_i = 1} f(t_i b)\right]\left[\prod_{i:\, \delta_i = 0} F(t_i b, \infty)\right]\\
&= \prod_{i=1}^{n} \prod_{j=0}^{\infty} \{1 - \alpha(jb)\}^{I\{j < t_i \text{ or } (j = t_i \text{ and } \delta_i = 0)\}} \alpha(jb)^{I\{j = t_i \text{ and } \delta_i = 1\}}.\\
&= \prod_{j=0}^{\infty} \left[\{1 - \alpha(jb)\}^{Y(jb) - dN(jb)} \alpha(jb)^{dN(jb)}\right],
\end{aligned}
\tag{2.4}
$$

utilizing (2.2). Of course, the product is really a finite one, every factor after the largest T_i being 1.

REMARK 2A. Let us comment upon the assumptions implicit in the derivation above. First, if we imagine a small time interval $[jb, jb + \varepsilon)$ during which two events are possible for an object still at risk at jb, viz. that it "dies" or "is censored", then it is assumed that the force of transition $\alpha(jb)$ gets its chance before the censoring mechanism. Alternatively, if the model is the result of a discretization of an essentially time-continuous process and if an object is censored in $[jb, (j + 1)b)$ and then dies before $(j + 1)b$, then it is assumed that this information will be available after all. Second, very general censoring mechanisms are allowed; (2.4) remains true as long as the censoring only depends upon the past and outside random variation [(Aalen (1978a); Gill (1980)]. Sometimes it is useful to model the behavior of the censoring variables c_i themselves. For the present Bayesian purpose it is, however, simpler to allow ourselves to condition on their observed values.

We may express the content of these conditions in another useful way. Let

$$\mathscr{F}_{jb} = \sigma\{N(ib), Y(ib); i \le j\}$$

be the processes' history up to and including time point jb. Then Y is *predictable* (or *pre-visible*) w.r.t. this increasing sequence of σ-algebras, i.e., $Y(jb)$ is $\mathscr{F}_{(j-1)b}$-measurable (known at time $(j - 1)b$) and

$$dN(jb) \mid \mathscr{F}_{(j-1)b} \sim \text{bin}\{Y(jb), \alpha(jb)\}. \tag{2.5}$$

It is immediate from (2.4) that the nonparametric maximum likelihood (ML) estimator of $\alpha(\cdot)$ is given by $\alpha^*(jb) = dN(jb)/Y(jb)$. These are the familiar occurrence/exposure rates. The ML property transfers, by (2.1) and (2.2), to

$$A^*(jb) = \sum_0^{jb} \frac{dN}{Y} \quad \text{and} \quad F^*(jb) = 1 - \prod_0^{jb}\left[1 - \frac{dN}{Y}\right], \qquad j \ge 0, \tag{2.6}$$

in obvious notation. These are the proper time-discrete analogues of, respectively, the Nelson–Aalen estimator $\int_0^t dN/Y$ and the Kaplan–Meier estimator $1 - \prod_0^t[1 - dN/Y]$; cf., for example, Andersen and Borgan (1985). One can indeed go on to the derivation of large-sample and other properties of A^* and F^*, paralleling those known in the continuous case, but this will not be pursued here.

2.2. *Nonparametric Bayes estimators.* The present aim is to construct a class of nonparametric Bayes estimators for the cumulative hazard A (and for α and F).

Let $\alpha(jb)$ be independent for $j = 0, 1, \ldots,$ with $\alpha(jb)$ having prior density $h_{jb}(s)\,ds$ on $[0, 1]$. This defines a probability distribution $\mathscr{P}$, with expectation operator $\mathscr{E}$, on the space $\{[0,1]^\infty, (\mathscr{B}_{[0,1]})^\infty\}$ in which the hazards $\alpha = \{\alpha(jb); j \ge 0\}$ live. Now $X_1, \ldots, X_n$ are iid with distribution (2.2) given α. This defines a simultaneous probability distribution $\mathscr{P}'$ on $\{\mathscr{X}^n \times [0,1]^\infty, \Gamma^n \times (\mathscr{B}_{[0,1]})^\infty\}$, in which Γ is the σ-algebra consisting of all subsets of $\mathscr{X}$, by

$$\begin{aligned}\mathscr{P}'\{X_1 \le j_1 b, \dots, X_n \le j_n b, \alpha \in G\} &= \int_G F(j_1 b) \cdots F(j_n b) \mathscr{P}(d\alpha) \\ &= \mathscr{E} F(j_1 b) \cdots F(j_n b) I\{\alpha \in G\},\end{aligned} \tag{2.7}$$

for integers j_i and $G \in (\mathscr{B}_{[0,1]})^\infty$. The proof of the following result utilizes (2.4) and the formula above.

Proposition 2.1. *Let $\alpha = \{\alpha(jb);\ j \ge 0\}$ have the prior distribution defined above. Then, given the censored data set $(T_1, \delta_1), \dots, (T_n, \delta_n)$, the $\alpha(jb)$'s are still independent, with posterior densities*

$$h^*_{jb}(s) = \text{const.}\, s^{dN(jb)} (1-s)^{Y(jb) - dN(jb)} h_{jb}(s).$$

The class of beta distributions now suggests itself. Let us write

$$A \sim \text{beta}\{c, A_0\} \tag{2.8}$$

to indicate that the cumulative hazard A, viewed as a stochastic process, has independent summands

$$\alpha(jb) \sim \text{beta}\{c(jb)\alpha_0(jb), c(jb)(1 - \alpha_0(jb))\}.$$

Then $\mathscr{E}\alpha(jb) = \alpha_0(jb) = dA_0(jb)$ is the prior guess and $\mathscr{E}A(jb) = A_0(jb)$, whereas $\operatorname{Var}\alpha(jb) = \alpha_0(jb)(1 - \alpha_0(jb))/(c(jb) + 1)$ is the prior uncertainty. Our previous considerations imply

$$A|\text{data} \sim \text{beta}\left\{c + Y, \sum_0^{\cdot} \frac{c\,dA_0 + dN}{c + Y}\right\}, \tag{2.9}$$

that is, the beta processes (2.8) constitute a natural class of conjugate prior distributions. (*Time-continuous* beta processes are studied in Section 3.) Furthermore, the nonparametric Bayes estimator of A becomes

$$\begin{aligned}\hat{A}(jb) = \mathscr{E}\{A(jb)|\text{data}\} &= \sum_{i=0}^{j} \frac{c(ib)\alpha_0(ib) + dN(ib)}{c(ib) + Y(ib)} \\ &= \sum_0^{jb} \frac{c\,dA_0 + dN}{c + Y},\end{aligned} \tag{2.10}$$

in obvious notation. Also, the conditional variance of $A(jb)$ is $\sum_0^{jb} d\hat{A}(1 - d\hat{A})/(c + Y + 1)$, and is useful, for example, when constructing (Bayesian) confidence bands for A.

Remark 2B. It is interesting to note that the Nelson–Aalen estimator A^* emerges as $c(\cdot)$ tends to zero and that $\hat{A}$ becomes the simple prior guess A_0 when $c(\cdot)$ grows large. Thus $c(jb)$, $j = 0, 1, \dots,$ are parameters that measure *strength of belief* in the prior guess. Extending the informal Bayesian notion of prior sample size a bit, we see that $c(jb)$ can be interpreted as the number at risk at jb in an imagined prior sample with intensity α_0: The combined sample would then consist of $c + Y$ at risk, c of them having hazard α_0 and Y of them having hazard dN/Y. Thus $\Pr\{\text{die at } jb | \text{at risk at } jb\} = (c/(c+Y))\alpha_0 + (Y(c+Y))\,dN/Y$, giving the Bayes solution (2.10).

Before going on to Markov chains, let us point out that a class of nonparametric Bayes estimators of the cdf F is an easy spin-off of the considerations above. For, let $A \sim \text{beta}\{c, A_0\}$ be the prior (which transforms into a prior for F, but this is not really needed). Then by (2.2) and (2.9),

$$\hat{F}(jb) = \mathscr{E}'\{F(jb)|\text{data}\}$$

$$= 1 - \prod_{i=0}^{j}[1 - \mathscr{E}'\{\alpha(ib)|\text{data}\}] = 1 - \prod_{0}^{jb}\left[1 - \frac{c\,dA_0 + dN}{c+Y}\right]. \tag{2.11}$$

The Kaplan–Meier estimator and the prior guess $F_0 = 1 - \prod_0^{\cdot}(1 - \alpha_0)$ come forward in the limiting cases $c(\cdot) \to 0$ and $c(\cdot) \to \infty$, respectively.

2.3. *Time-inhomogeneous Markov chains.* The methods and results of the previous subsection will now be generalized considerably. Let $X = \{X(ub);\ u = 0, 1, 2, \dots\}$ be a Markov chain in the state space $\{1, \dots, k\}$ with transition probabilities

$$P_{ij}(ub, vb) = \Pr\{X(vb) = j|X(ub) = i\}, \qquad 0 \le u \le v, \qquad i, j = 1, \dots k.$$

The situation of Section 2.2 corresponds to state space $\{1, 2\}$ and $1 \to 2$ being the only possible transition. The natural analogues of the hazards $\alpha(jb)$ in (2.1) are simply the one-step probabilities $\alpha_{ij}(vb) = P_{ij}((v-1)b, vb)$. The corresponding cumulative hazard rate from i to j is

$$A_{ij}(vb) = \sum_{u=1}^{v} \alpha_{ij}(ub), \qquad v \ge 1. \tag{2.12}$$

These are of fundamental interest in many applications, for example, in demography, quantitative sociology and actuarial statistics. Plots of $A_{ij}(\cdot)$ for perhaps different j's, give information about when individuals tend to leave i and for which destinations. Studying these hazard rates if often judged more informative than studying transition probabilities.

One cannot expect to be able to observe the Markov chain X for all time points $0, b, 2b, \dots,$ at least not when no absorbing states are present. Assume that X is observed up to and including time wb and put $X_{wb} = \{X(ub);\ u = 0, 1, \dots, w\}$. We are to estimate the A_{ij}'s using data of this type collected for n individuals (or objects), moving around in the state space independently of each other, each with transition probabilities $P_{ij}(\cdot, \cdot)$.

To do this we need a workable expression for the likelihood, as in (2.4). Let the data be

$$X^{(a)}_{w(a)b} = \{X^{(a)}(ub);\, u = 0, 1, \dots, w(a)\}, \qquad a = 1, \dots, n. \tag{2.13}$$

Define first

$$\begin{aligned} dN_{ij}(ub) &= \sum_{a=1}^{n} I\{X^{(a)}((u-1)b) = i, X^{(a)}(ub) = j\}, \\ Y_i(ub) &= \sum_{a=1}^{n} I\{X^{(a)}((u-1)b) = i, u \le w(a)\}, \qquad u \ge 1. \end{aligned} \tag{2.14}$$

So $Y_i(ub)$ is the number of individuals at risk in state i just before time ub; these are exposed to the $k-1$ forces of transition $\alpha_{ij}(ub)$, $j \ne i$ [and may remain in i, with probability $\alpha_{ii}(ub) = 1 - \sum_{j \ne i} \alpha_{ij}(ub)$]. Not counted in $Y_i(ub)$ are those that had $X^{(a)}((u-1)b) = i$, but were censored before ub. The increments $dN_{ij}(ub)$ add up to counting processes N_{ij}; $N_{ij}(vb)$ counts the number of transitions $i \to j$ seen in $(0, vb]$.

Using the Markov property and independence between individuals, one arrives at the likelihood

$$L(\text{data}) = \left[\prod_{a=1}^{n} \pi^{(a)}(X^{(a)}(0))\right] \prod_{u=1}^{\infty} \prod_{i,j} \alpha_{ij}(ub)^{dN_{ij}(ub)},$$

where $\pi^{(a)}$ is the start distribution for $X^{(a)}$. Assumptions about the censoring mechanisms that ensure the above likelihood are as in Remark 2A; see in particular (2.5), the parallel of which in the present situation is

$$\begin{aligned} &\{dN_{ij}(ub);\, j = 1, \dots, k\} \,|\, \mathscr{F}_{(u-1)b} \\ &\qquad \sim \text{multinomial}\{Y_i(ub);\, \alpha_{i1}(ub), \dots, \alpha_{ik}(ub)\}. \end{aligned} \tag{2.15}$$

Here $\mathscr{F}_{ub}$ is the complete history of Y_i's and N_{ij}'s up to and including time ub. We shall also assume that the start distributions contain no information about the hazard rates.

To parallel the theory of Section 2.2, define first

$$\alpha(ub) = \{\alpha_{ij}(ub);\, i, j = 1, \dots, k\} = \mathbf{P}((u-1)b, ub), \qquad u \ge 1.$$

Let $\mathscr{M}_k$ be the space of such size k Markov matrices. Delete the diagonal to get $\alpha^0(ub) = \{\alpha_{ij}(ub);\, i = 1, \dots, k, j \ne i\}$. If a prior distribution $\mathscr{P}$ for the sequence of matrices $\alpha(ub)$ is established, in $\mathscr{M}_k^{\infty}$, a simultaneous distribution $\mathscr{P}'$ can be defined, along with expectation operator $\mathscr{E}'$, for both α and Markov chain data (2.13), in analogy to (2.7). In the following proposition, which generalizes Proposition 2.1, $(Z_1, \dots, Z_k) \sim \text{Dir}\{\beta_1, \dots, \beta_k\}$ is used to indicate that $\sum_{j=1}^{k} Z_j = 1$ and that $(Z_1, \dots, Z_{k-1})$ has Dirichlet density

$$\frac{\Gamma(\sum_{j=1}^{k} \beta_j)}{\Gamma(\beta_1) \cdots \Gamma(\beta_k)} z_1^{\beta_1 - 1} \cdots z_{k-1}^{\beta_{k-1}-1} \left(1 - \sum_{j=1}^{k-1} z_j\right)^{\beta_k - 1},$$

in the simplex where $z_1, \dots, z_{k-1}$ are nonnegative and have a sum not exceeding 1. The proof of the proposition requires more formalism than ingenuity and is left out here. Details are available in Hjort (1984).

PROPOSITION 2.2. *Let the hazard rate matrices $\alpha(ub)$ be independently distributed in $\mathscr{M}_k$.*

(i) *Then, given the n abridged sample paths* (2.13), *the $\alpha(ub)$'s are still independent.*

(ii) *If the distribution of $\alpha(ub)$ is given by a prior density $h_{ub}(\mathbf{s}) = h_{ub}(\{s_{ij}; i \neq j\})$ for $\alpha^0(ub)$, then $\alpha^0(ub)$ admits a posterior density*

$$h^*_{ub}(\mathbf{s}) = \text{const.} \prod_{i=1}^{k} \left[\left\{ \prod_{j \neq i} s_{ij}^{dN_{ij}(ub)} \right\} (1 - s_{i\cdot})^{Y_i(ub) - dN_{i\cdot}(ub)} \right] h_{ub}(\mathbf{s}),$$

where $s_{i\cdot} = \sum_{j \neq i} s_{ij}$ and $N_{i\cdot} = \sum_{j \neq i} N_{ij}$.

(iii) *If in particular the rows of $\alpha(ub)$ are independent and Dirichlet distributed, then this is true also a posteriori. Specifically, if the ith row $\alpha_i(ub) \sim \mathrm{Dir}\{c_i(ub)\alpha_{0,i1}(ub), \dots, c_i(ub)\alpha_{0,ik}(ub)\}$, then the updated parameter vector is*

$$\{c_i(ub)\alpha_{0,i1}(ub) + dN_{i1}(ub), \dots, c_i(ub)\alpha_{0,ik}(ub) + dN_{ik}(ub)\}.$$

It is now easy to match (2.8) and (2.10). Assume that a prior distribution for the $k(k-1)$ cumulative hazard rates A_{ij} is chosen which specifies that its summands are independent and that the k rows of $\alpha(ub)$ are distributed as in (iii) above. Note that $\mathscr{E}\alpha_{ij}(ub) = \alpha_{0,ij}(ub)$, so $A_{0,ij}(vb) = \sum_{u=1}^{v} \alpha_{0,ij}(ub)$ is the prior guess for A_{ij}. If the A_{ij}'s are to be estimated under a loss function which is quadratic in $\hat{A}_{ij}(ub) - A_{ij}(ub)$, or equivalently quadratic in the terms $\hat{\alpha}_{ij}(ub) - \alpha_{ij}(ub)$, then the Bayes solution has

$$\mathscr{E}'\{\alpha_{ij}(ub) | \text{data}\} = \frac{c_i(ub)\alpha_{0,ij}(ub) + dN_{ij}(ub)}{c_i(ub) + Y_i(ub)}.$$

This leads to

$$\hat{A}_{ij}(vb) = \sum_{b}^{vb} \frac{c_i \, dA_{0,ij} + dN_{ij}}{c_i + Y_i}, \qquad v \geq 1, \tag{2.16}$$

in natural notation. Similarly, if a Bayes estimate is required for the waiting time distribution F_i for state i, i.e., $F_i(vb) = \Pr\{X \text{ leaves } i \text{ before time } vb\}$, then the answer is

$$\hat{F}_i(vb) = 1 - \prod_{b}^{vb} \left[1 - \frac{c_i \, dA_{0,i\cdot} + dN_{i\cdot}}{c_i + Y_i} \right].$$

REMARK 2C. The interpretation of the strength of belief parameter $c(\cdot)$ in Remark 2B can be invoked and yields qualitative meaning to the parameters $c_1(\cdot), \dots, c_k(\cdot)$ entering the present situation.

REMARK 2D. Let $c_i(\cdot)$ tend to zero in (2.16). Then the non-Bayesian estimator $A^*_{ij}(vb) = \Sigma_b^{vb}\, dN_{ij}/Y_i$ emerges. This is the time-discrete analogue of Aalen's nonparametric estimator $\int dN_{ij}/Y_i$ [Aalen (1978a, b)]. One can prove that these sums of occurence/exposure rates are ML estimators for A_{ij} by maximizing each term of the likelihood while using that α_{ij}'s and dN_{ij}'s sum to 1 over j for fixed i. This is a modest generalization of a result by Fleming and Harrington (1978), who assumed that all chains were observed over the same time span.

REMARK 2E. To estimate the hazard rate $i \to j$, using either the Aalen analogue A^*_{ij} or the Bayes solution (2.16), one needs only a portion of the data, viz. the risk set for state i (the process Y_i) and the times at which transitions $i \to j$ occur (the N_{ij} process). Thus the original, intended data collection (2.13) need not be complete, i.e., it may be censored in some respects, as far as the calculation of a specific A^*_{ij} or $\hat{A}_{ij}$ is concerned.

3. The time-continuous case.

3.1. *Cumulative hazard rates and the product-integral.* Let T be a random variable with cdf $F(t) = \Pr\{T \le t\}$ on $[0, \infty)$ and $F(0) = 0$. The *cumulative hazard rate* for F or T is a nonnegative, nondecreasing, right continuous function A on $[0, \infty)$ which we should like to satisfy

$$dA(s) = A[s, s + ds) = \Pr\{T \in [s, s + ds) | T \ge s\} = dF(s)/F[s, \infty),$$

in analogy with the discrete counterpart (2.1). So we *define*

$$A[a, b) = \int_{[a,b)} \frac{dF(s)}{F[s, \infty)} \tag{3.1}$$

and also have

$$F[a, b) = \int_{[a,b)} F[s, \infty)\, dA(s), \qquad 0 \le a \le b < \infty. \tag{3.2}$$

If F is absolutely continuous, then it is easily seen that $A(t) = -\log\{1 - F(t)\}$ or $F(t) = 1 - \exp\{-A(t)\}$. We shall, however, encounter cdf's having jumps, in which case this classic correspondence no longer holds and we prefer (3.1) as the starting point for interpretational reasons.

We need to know that F is uniquely determined by A. F is indeed restorable from equation (3.2), whose solution, for our purposes, is most easily given using the *product integral*:

$$F(t) = 1 - \prod_{[0,t]} \{1 - dA(s)\}, \qquad t \ge 0; \tag{3.3}$$

compare (2.2). See, e.g., Gill [(1980), Lemma 3.2.1]. One can show that $\prod_{[a,b]}\{1 - dA(s)\} = \exp\{-A[a,b]\}$ if and only if A is continuous, so that $A = -\log(1 - F)$ holds under this assumption. In general, we infer from $1 - x \le \exp(-x)$ that $F(t) \ge 1 - \exp\{-A(t)\}$ or $-\log\{1 - F(t)\} \ge A(t)$. Gill and Johansen (1987) give a good account of the properties of the product integral.

3.2. *Lévy processes.* We now turn to the construction of prior distributions for A.

Let $\mathscr{F}$ be the set of all cdf's F on $[0, \infty)$ having $F(0) = 0$ and let $\mathscr{B}$ be the set of all nondecreasing, right continuous functions B on $[0, \infty)$ having $B(0) = 0$. We can always write $F = 1 - \exp(-B)$ with a B in $\mathscr{B}$ [and we allow $B(\infty) < \infty$ or $F(\infty) < 1$]. We also need to define

$$\mathscr{A} = \{\text{those } A \text{ in } \mathscr{B} \text{ for which (3.3) leads to an } F \text{ in } \mathscr{F}\}. \tag{3.4}$$

This is the space of cumulative hazard rates. We are to place a probability distribution on $\{\mathscr{A}, \Sigma_{\mathscr{A}}\}$, where $\Sigma_{\mathscr{A}}$ is the σ-algebra generated by the Borel cylinder sets. Such a probability distribution, say $\mathscr{P}$, is determined if the distribution of every finite set of increments $A[a_{j-1}, a_j)$ is specified, as long as this is done in a Kolmogorov-consistent way. In other words, we demand $A[a, c) =_d A[a, b) + A[b, c)$ when $a < b < c$, in which $=_d$ means equality in distribution; see Ferguson (1973, 1974) and Doksum (1974).

The natural analogues of the prior distributions considered in Section 2 are the nonnegative, nondecreasing processes on $[0, \infty)$ that start at zero and have independent increments. Term these *Lévy processes*. Such processes were studied extensively by Lévy (1936) and have been utilized in nonparametric Bayesian analysis by Doksum (1974), Ferguson (1974), Ferguson and Phadia (1979), Kalbfleisch and Prentice [(1980), Chapter 8] and Wild and Kalbfleisch (1981). Ferguson's Dirichlet processes (1973) are the most widely used ones in this context. These are related to Lévy processes in that $-\log(1 - F)$ is Lévy when F is Dirichlet.

These authors use Lévy processes as priors for $B = -\log(1 - F)$, the reason for this being a fundamental result due to Doksum (1974) and Ferguson and Phadia (1979): If B is a Lévy process, then given a set of possibly censored observations from F, the B process is still Lévy. We shall deviate slightly, but significantly, from this approach, starting with A instead of B. The reasons for this are the desire to parallel the construction and results of Section 2, A being more easily interpreted as cumulative hazard than $B = -\log(1 - F)$, and the fact that A is easier to generalize to more general models like the competing risks framework or time-inhomogeneous Markov chains. Other authors have mainly been interested in F, while we prefer the hazard rate and the cumulative hazard rate as the fundamental concepts with which to construct, interpret and analyze models for life history data.

The A approach entails some mathematical difficulties, but they can be managed and the gain will be substantial. The first inconvenience we encounter is the observation that not every Lévy process can be used as a prior for A. The gamma process, for example, which has independent increments of the form $G[s, s+\varepsilon) \sim \text{gamma}\{cG_0[s, s+\varepsilon), c\}$, does *not* have paths that a.s. produce proper cdf's $F = 1 - \Pi_{[0,\cdot]}\{1 - dG(s)\}$, i.e., the subset $\mathcal{A}$ of $\mathcal{B}$ does not have outer measure 1. Remark 3A gives a characterization of the subclass of Lévy priors for A that yield paths in $\mathcal{A}$ with probability 1. We shall first confine ourselves to the construction of a particular rich class that resembles the beta process priors used in the time-discrete case [see (2.8)] and whose paths certainly lie in $\mathcal{A}$.

3.3. *A beta process with independent increments.* What is needed, then, is a beta process on $[0, \infty)$, with paths in $\mathcal{A}$, which has independent increments of the type

$$(3.5) \qquad dA(s) \sim \text{beta}\{c(s)\,dA_0(s), c(s)(1 - dA_0(s))\},$$

infinitesimally speaking. The existence of such a process is not at all obvious, since beta distributed variables have cumbersome convolution properties. (The Dirichlet process has marginal distributions of this type, but with *dependent* increments.)

Let A be any Lévy process. There exists a separable version with right-continuous paths [Breiman (1968), page 299], i.e., $\mathcal{P}(\mathcal{B}) = 1$, where $\mathcal{P}$ is the probability measure governing A. Let $\mathcal{E}$ be the expectation operator associated with $\mathcal{P}$ and let $t_1, t_2, \ldots$ be the times at which A a.s. is discontinuous, say with jumps $S_j = A\{t_j\} = A(t_j) - A(t_j -)$. Then A admits a *Lévy representation*

$$(3.6) \quad \mathcal{E}\exp\{-\theta A(t)\} = \left[\prod_{j:\, t_j \le t} \mathcal{E}\exp(-\theta S_j)\right] \exp\left\{-\int_0^\infty (1 - \varepsilon^{-\theta s})\, dL_t(s)\right\},$$
$$t \ge 0, \theta \ge 0,$$

where $\{L_t;\, t \ge 0\}$ is a continuous Lévy measure. This means that L_t for each t is a measure on $(0, \infty)$, $L_t(D)$ is nondecreasing and continuous in t for each Borel set D in $(0, \infty)$, and $L_0(D) = 0$. It holds that $A(t)$ is finite a.s. whenever $\int_0^\infty s/(1+s)\, dL_t(s)$ is finite. In the Lévy formula (3.6), which follows from Ferguson [(1974) page 623], it is assumed that A contains no nonrandom part. The distribution of such a $\mathcal{P}$ is specified by $\{t_1, t_2, \ldots\}$, the distributions of $S_1, S_2, \ldots$ and $\{L_t;\, t \ge 0\}$.

Remark 3A. The Lévy measures $\{L_t\}$ can of course have full support $[0, \infty)$. The gamma process mentioned above, for example, has $\mathcal{E}\exp\{-\theta G(t)\} = (c/(c+\theta))^{cG_0(t)} = \exp[\{-c\log(1 + \theta/c)\}G_0(t)]$, which can be written in the form (3.6) with $dL_t(s) = cs^{-1}\exp(-cs)\, ds G_0(t)$, valid for s in $[0, \infty)$. Theorem 3.1 gives an explicit construction of a time-continuous beta process aiming at (3.5) and developed as a fine limit of time-discrete ones, and one conspicuous feature of the result is the fact that the accompanying Lévy measures are concentrated on $[0, 1]$. It is indeed the case that the condition $L_t(1, \infty) = 0$ for all t is necessary and sufficient for a Lévy process to be a.s. a proper cumulative hazard. This can be proved using a representation theorem for such processes by Ferguson and Klass (1972). The general A process of (3.6), after removing the S_j jumps at the nonrandom jump times, admits a representation as a countable sum of random jumps at a countable collection of random sites. Using methods of proof from Ferguson and Klass one can show that the largest of all the jumps occurring in the interval $[0, T]$, say, has probability distribution $\mathscr{P}\{J \le x\} = \exp\{-L_T[x, \infty)\}$. The characterization follows since this largest jump must be in $[0, 1]$ a.s.

THEOREM 3.1. *Let A_0 in $\mathscr{A}$ be continuous, and let $c(\cdot)$ be a piecewise continuous, nonnegative function. Then there exists a Lévy process $A(\cdot)$, whose paths a.s. fall in $\mathscr{A}$ and whose Lévy representation is*

$$\begin{aligned}\mathscr{E}\exp\{-\theta A(t)\} &= \exp\left\{-\int_0^1 (1 - e^{-\theta s})\, dL_t(s)\right\} \\ &= \exp\left[\sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \int_0^t \frac{\Gamma(m)\Gamma(c(z))}{\Gamma(m + c(z))} c(z)\, dA_0(z)\right],\end{aligned} \tag{3.7}$$

where

$$dL_t(s) = \left\{\int_0^t c(z) s^{-1} (1 - s)^{c(z)-1}\, dA_0(z)\right\} ds, \qquad t \ge 0, 0 < s < 1. \tag{3.8}$$

PROOF AND CONSTRUCTION. For each n, define independent variables $X_{n,i} \sim \text{beta}\,\{a_{n,i}, b_{n,i}\}$ for $i = 1, 2, \dots$, where

$$a_{n,i} = c_{n,i} A_0\left(\frac{i-1}{n}, \frac{i}{n}\right],$$

$$b_{n,i} = c_{n,i}\left(1 - A_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\right), \qquad c_{n,i} = c\left(\frac{i - 1/2}{n}\right).$$

Let

$$A_n(0) = 0 \quad \text{and} \quad A_n(t) = \sum_{i/n \le t} X_{n,i} \quad \text{for } t \ge 0. \tag{3.9}$$

Then A_n has independent beta increments and its jumps become smaller [expected size is $A_0((i-1)/n, i/n]$ at i/n], but occur more often, as n increases. We have

$$\mathscr{E}A_n(t) = \sum_{i/n \le t} A_0\left(\frac{i-1}{n}, \frac{i}{n}\right] \to A_o(t),$$

$$\begin{aligned}\operatorname{Var} A_n(t) &= \sum_{i/n \le t} A_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\left(1 - A_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\right) \Big/ (c_{n,i} + 1) \\ &\to \int_0^t \frac{dA_0(s)}{c(s) + 1},\end{aligned} \tag{3.10}$$

and shall see that $\{A_n\}$ converges in distribution in each of the spaces $D[0, R]$, $R > 0$, to a Lévy process A having the required properties.

Fix t for the moment. We show now that

$$\mathscr{E}\exp\{-\theta A_n(t)\} \to \exp\left\{-\int_0^1 (1 - e^{-\theta s})\, dL_t(s)\right\}, \tag{3.11}$$

for each θ, where L_t is as in (3.8), and intend to employ Lemma A.1 in the Appendix. The right-hand side involves

$$-\int_0^1 (1 - e^{-\theta s})\, dL_t(s) = \sum_{m=1}^{\infty} (-1)^m (\theta^m/m!) \int_0^1 s^m\, dL_t(s),$$

where

$$\int_0^1 s^m\, dL_t(s) = \int_0^t \int_0^1 c(z) s^{m-1}(1 - s)^{c(z)-1}\, ds\, dA_0(z)$$
$$= \int_0^t \frac{\Gamma(m)\Gamma(c(z) + 1)}{\Gamma(m + c(z))}\, dA_0(z),$$

by Fubini's theorem; this verifies the second equality in (3.7). The left-hand side can be written $\mathscr{E}\prod_{i/n \le t} \exp(-\theta X_{n,i}) = \prod_{i/n \le t}(1 + z_{n,i})$, where

$$1 + z_{n,i} = \int_0^1 e^{-\theta x} \frac{\Gamma(c_{n,i})}{\Gamma(a_{n,i})\Gamma(b_{n,i})} x^{a_{n,i}-1}(1 - x)^{b_{n,i}-1}\, dx$$
$$= 1 + \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} y_{n,i}(m),$$

in which

$$y_{n,i}(m) = A_0\left(\frac{i-1}{n}, \frac{i}{n}\right] \frac{(a_{n,i} + 1) \cdots (a_{n,i} + m - 1)}{(c_{n,i} + 1) \cdots (c_{n,i} + m - 1)}.$$

It can now be seen that

$$\sum_{i/n \le t} y_{n,i}(m) \to \int_0^t \frac{(m-1)!\, dA_0(z)}{(c(z) + 1) \cdots (c(z) + m - 1)},$$

implying by Lebesgue's theorem on dominated convergence that

$$\sum_{i/n \le t} z_{n,i} = \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \sum_{i/n \le t} y_{n,i}(m) \to -\int_0^1 (1 - e^{-\theta s})\, dL_t(s).$$

It is not difficult to show that the rest of the assumptions of Lemma A.1 are satisfied, proving (3.11) as required.

It may be shown by similar arguments that

$$\mathscr{E}\exp\left\{-\sum_{j=1}^{k}\theta_j A_n(a_{j-1},a_j]\right\}\to\exp\left\{-\sum_{j=1}^{k}\int_0^1(1-e^{-\theta_j s})\,dL_{(a_{j-1},a_j]}(s)\right\},$$

i.e., the finite-dimensional distributions of $\{A_n\}$ converge properly. Furthermore, $\{A_n\}$ is *tight* in the space $D[0,R]$ of all right-continuous functions on [0, R] with left-hand limits, equipped with the Skorohod topology [see Billingsley (1968), Chapter 3]. To see this, note that $\mathscr{E}A_n(s,t]A_n(t,u]=A_0(s_n,t_n]A_n(t_n,u_n]\le\{A_0(u_n)-A_0(s_n)\}^2$, writing $s_n=[ns]/n$, $t_n=[nt]/n$, $u_n=[nu]/n$. This can be seen to imply tightness in $D[0,R]$ by the arguments used in the proof of Billingsley's Theorem 15.6 (but not quite by Theorem 15.6 itself).

This entails that for each R, $\{A_n\}$ converges in distribution to a random element of $D[0,R]$, say $A_{(R)}$. This process has finite-dimensional distributions arising as limits to those of A_n. Thus we may take $A_{(R)}$ to be simply the $[0,R]$ restriction of a Lévy process A on $[0,\infty)$ with Lévy representation as given in the theorem.

It remains only to show that A is in the space $\mathscr{A}$ of cumulative hazard rates with probability 1. But $\mathscr{A}_R$, the space $\mathscr{A}$ restricted to $[0,R]$, is closed in $D[0,R]$ w.r.t. the Skorohod topology, according to Lemma A.2 in the Appendix, and A_n certainly lies in $\mathscr{A}_R$ for all large n. That $\mathscr{P}\{A\in\mathscr{A}\}=1$ follows now from Billingsley's Theorem 2.1. Another but less direct argument can appeal to Remark 3A. □

Agree to say that the Lévy process A constructed in the theorem, i.e., satisfying (3.7) and (3.8), is a *beta process* with parameters $c(\cdot)$ and $A_0(\cdot)$ and write this as $A\sim\text{beta}\{c(\cdot),A_0(\cdot)\}$. The construction in the proof of Theorem 3.1 indeed aimed at the fulfillment of (3.5) in some sense. We check below that some of the beta characteristics are preserved. Although (3.5) is a useful guide and provides interpretation for the process A, one should bear in mind that $A(s,s+\varepsilon]$, say, is generally not *exactly* beta distributed.

From (3.7) and the Lévy structure, it is clear that $\mathscr{E}\exp\{-\theta A(a,b]\}=\exp[-\int_0^1\{1-\exp(-\theta s)\}d(L_b-L_a)(s)]$. Differentiating w.r.t θ, putting θ equal to zero and using (3.8) gives

$$\begin{aligned}\mathscr{E}A(a,b]&=\int_0^1 s\,d(L_b-L_a)(s)\\&=\int_a^b\int_0^1 sc(z)s^{-1}(1-s)^{c(z)-1}\,ds\,dA_0(z)=A_0(a,b],\end{aligned}$$

earning A_0 the "prior guess" label again. Differentiating once more leads to $\operatorname{Var}A(a,b]=\int_{(a,b]}dA_0(s)/(c(s)+1)$, in harmony with (3.5) and (3.10), since $dA_0(s)(1-dA_0(s))=dA_0(s)$ in the continuous case.

We have assumed so far that the prior guess A_0 is continuous. We find in Section 4 that if $A \sim \text{beta}\{c, A_0\}$ and a sample is drawn from $F = F_A$, then the posterior distribution of A is such that the jump $A\{x\}$ is positive, whenever x is a point at which an observation occurs. Thus the posterior guess $\mathscr{E}\{A(\cdot)|\text{data}\}$ has fixed points of discontinuity. It is therefore necessary to extend the earlier definition by allowing A_0 a finite number of discontinuities.

DEFINITION. Let A_0 be a cumulative hazard with a finite number of jumps taking place at $t_1, t_2, \dots,$ and let $c(\cdot)$ be a piecewise continuous, nonnegative function on $[0, \infty)$. Say that the Lévy process A is a beta process with parameters $c(\cdot)$, $A_0(\cdot)$, and write again

$$(3.12) \qquad A \sim \text{beta}\{c(\cdot), A_0(\cdot)\}$$

to indicate this, if the following holds: A has Lévy representation (3.6), with

$$(3.13) \qquad S_j = A\{t_j\} \sim \text{beta}\big\{c(t_j)A_0\{t_j\}, c(t_j)\big(1 - A_0\{t_j\}\big)\big\}$$

and

$$(3.14) \qquad dL_t(s) = \int_0^t c(z)s^{-1}(1-s)^{c(z)-1}\,dA_{0,c}(z)\,ds$$

for $t \geq 0$ and $0 < s < 1$,

in which $A_{0,c}(t) = A_0(t) - \Sigma_{t_j \leq t} A_0\{t_j\}$ is A_0 with its jumps removed. This can be rephrased as follows:

$$(3.15) \qquad A(t) = \sum_{t_j \leq t} S_j + A_c(t),$$

where the jumps S_j are independent and distributed as above, and A_c is $\text{beta}\{c(\cdot), A_{0,c}(\cdot)\}$ of the earlier type.

Note that the existence of this more general beta process is guaranteed by Theorem 3.1 and that the definition is in harmony with (3.5). Furthermore,

$$(3.16) \qquad \mathscr{E}A(t) = \sum_{t_j \leq t} \mathscr{E}S_j + A_{0,c}(t) = A_0(t)$$

by previous efforts, i.e., A_0 is still the prior guess, and the generally valid formula for the variance becomes

$$(3.17) \quad \operatorname{Var} A(t) = \sum_{t_j \leq t} \operatorname{Var} S_j + \int_0^t \frac{dA_{0,c}(s)}{c(s)+1} = \int_0^t \frac{dA_0(s)(1 - dA_0(s))}{c(s)+1};$$

compare again with (3.5) and (3.10).

Consider finally the random distribution function $F(t) = 1 - \prod_{[0,t]}\{1 - dA(s)\}$ which has a beta process as its cumulative hazard. Let $F_1(t) =$

$\mathscr{E}F(t)$ be its expectation. Then by (3.2) and the property of independent increments, $F_1(a,b] = \int_{(a,b]} F_1(s,\infty)\, dA_0(s)$ for $0 \le a \le b$. But according to Lemma 3.2.1 in Gill (1980), this implies that F_1 is the cdf having A_0 as its cumulative hazard, i.e.,

$$\mathscr{E}F(t) = F_0(t) = 1 - \prod_{[0,t]} \{1 - dA_0(s)\}. \tag{3.18}$$

This proves in particular that the order of expectation and product can be interchanged. See also Remark 7A.

4. Posterior distributions and Bayes estimators.

4.1. *Posterior distribution of a general Lévy process.* The main aim in this subsection will be to find the distribution of a beta process A, given a set of possibly censored observations having A as a cumulative hazard, thus providing a time-continuous analogue of the basic result (2.9). We will, however, study more general Lévy processes as priors for A.

We show next that if A is Lévy and is a.s. a cumulative hazard, then A is still Lévy given the data, and provide formulae for its revised Lévy representation. Accordingly this subsection may be regarded as a parallel to the work by Doksum [(1974), Section 4] and Ferguson and Phadia [(1979), Section 2]. These authors use $B = -\log(1 - F)$ instead of A, and show that B Lévy a priori implies B Lévy a posteriori. There is indeed a connection between our result and theirs, since one can prove that A is Lévy if and only if B is Lévy, using the relations $dB(s) = -\log\{1 - dA(s)\}$, $dA = 1 - \exp\{-dB(s)\}$. Thus we already know that the property of being Lévy (i.e., having independent, nonnegative increments) is preserved for A, passing from prior to posterior distributions. However, the formulae involving $\mathscr{E}\{\exp(-\theta B(t))|\text{data}\}$ obtained by Ferguson and Phadia cannot easily be translated into formulae for $\mathscr{E}\{\exp(-\theta A(t))|\text{data}\}$, which are the ones we shall need. Therefore, a new proof is given.

We now describe the general class of possible prior distributions for A. Let A have fixed points of discontinuity $M = \{t_1, \dots, t_m\}$, with jumps $S_j = A\{t_j\}$ that have densities $f_j(s)\, ds$ in $[0,1]$. The process

$$A_c(t) = A(t) - \sum_{t_j \le t} S_j \tag{4.1}$$

is free of fixed discontinuity points and has Lévy formula

$$\mathscr{E} \exp\{-\theta A_c(t)\} = \exp\left\{-\int_0^1 (1 - e^{-\theta s})\, dL_t(s)\right\}, \tag{4.2}$$

where the continuous family of Lévy measures $\{L_t;\ t \ge 0\}$ is assumed to be of the form

$$\begin{aligned} dL_t(s) &= \int_0^t a(s,z)\, dH(z)\, ds, && t \ge 0, \quad \text{if } s \in (0,1), \\ &= 0, && \text{if } s \ge 1. \end{aligned} \tag{4.3}$$

Here H is a continuous nondecreasing function having $H(0) = 0$ and $a(s, z)$ is some nonnegative function, assumed to be continuous in (s, z) except possibly on the line segments where $z \in M$ and chosen such that $\int_0^1 s\, dL_t(s)$ is finite.

Observe that $A \sim \text{beta}\{c, A_0\}$, according to its definition in Section 3.3, corresponds to the case where the $f_j(s)$'s are beta densities and $a(s, z) = c(z)s^{-1}(1 - s)^{c(z)-1}$, $H = A_0$. Note further that the main restriction on A above is not the regularity conditions involving a and H, but the assumption that all Lévy measures L_t are concentrated in $(0, 1)$, as opposed to the general case (3.6). This is done to ensure that $\mathscr{P}(\mathscr{A}) = 1$; see Remark 3A.

Let $X_1, \dots, X_n$ be iid given A, i.e., with the common cdf $F = F_A$ defined in (3.3). Similarly to the discrete case (2.7), this defines a simultaneous probability distribution $\mathscr{P}'$ on $\{[0, \infty)^n \times \mathscr{B}, \mathscr{C}_n \times \Sigma_{\mathscr{B}}\}$, in which $\mathscr{C}_n$ denotes the Borel sets on $[0, \infty)^n$ and $\Sigma_{\mathscr{B}}$ is the σ-algebra generated by the Borel cylinder sets on $\mathscr{B}$, the function space defined in connection with (3.4), as follows:

$$\begin{aligned}\mathscr{P}'\{X_1 \in D_1, \dots, X_n \in D_n, A \in G\} &= \int_G F(D_1) \cdots F(D_n)\mathscr{P}(dA)\\ &= \mathscr{E}F(D_1) \cdots F(D_n)I\{A \in G\},\end{aligned} \tag{4.4}$$

for Borel sets D_i and $G \in \Sigma_{\mathscr{B}}$.

We will attack the case of $n = 1$ observation X first. The result can then be applied repeatedly to establish the posterior distribution given a full sample.

We are concerned with finite measures on $\{\mathscr{B}, \Sigma_{\mathscr{B}}\}$. Such a measure, say $\mathscr{P}_0$, with integral operator $\mathscr{E}_0$ defined by $\mathscr{E}_0\psi(A) = \int\psi(A)\mathscr{P}_0(dA)$, is known when all finite-dimensional $\mathscr{P}_0\{A[a_{j-1}, a_j) \in D_j, j = 1, \dots, k\}$ are specified. But this equivalent to giving the Laplace transforms $\mathscr{E}_0 \exp\{-\sum_{j=1}^k \theta_j A[a_{j-1}, a_j)\}$. It is convenient to put this in the following way: Knowledge of $\mathscr{E}_0 \exp\{-\int_0^\infty \theta(z)\, dA(z)\}$, for all $\theta(z)$ of the type $\theta(z) = \sum_{j=1}^k \theta_j I\{z \in [a_{j-1}, a_j)\}$, is sufficient to specify $\mathscr{P}_0$ on $\{\mathscr{B}, \Sigma_{\mathscr{B}}\}$ completely. For example, the general A defined in (4.1)–(4.3) is seen to have

$$\begin{aligned}&\mathscr{E} \exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\}\\ &\quad= \left[\prod_{t_j \in M} \mathscr{E} \exp\{-\theta(t_j)S_j\}\right] \mathscr{E} \exp\left\{-\int_0^\infty \theta(z)\, dA_c(z)\right\}\\ &\quad= \left[\prod_{t_j \in M} \mathscr{E} \exp\{-\theta(t_j)S_j\}\right] \exp\left\{-\int_0^1\int_0^\infty (1 - e^{-\theta(z)s})a(s, z)\, dH(z)\, ds\right\}.\end{aligned}$$

The number of times we encounter integrals like this excuses our using

$$R_{(b,c)}[g(s, z)] = \int_0^1\int_b^c g(s, z)a(s, z)\, dH(z)\, ds \tag{4.5}$$

as shorthand notation. Thus

$$\mathscr{E}\exp\left\{-\int_0^\infty \theta(z)\,dA(z)\right\} \tag{4.6}$$
$$= \left[\prod_{t_j\in M}\mathscr{E}\exp\{-\theta(t_j)S_j\}\right]\exp\{-R_{(0,\infty)}[1-e^{-\theta(z)s}]\}$$

defines the prior distribution $\mathscr{P}$ for A. Observe that $\mathscr{P}$ is specified by the finite set M, the densities $f_j(s)$ of the jumps and the functions $a(s,z)$, $H(z)$. It is convenient to term these quantities the parameters of $\mathscr{P}$, even though a and H are not quite uniquely determined by $\mathscr{P}$, since only the product $a(s,z)\,dH(z)$ matters.

THEOREM 4.1. *Let X given A be distributed according to F in* (3.3) *and let A be a Lévy process as defined in* (4.1)–(4.3), *or equivalently*, (4.6).

(i) *Given $X > x$, A is still a Lévy process with posterior parameters M^*, $\{f_j^*(s)\}$, $a^*(s,z)$ and $H^*(z)$ as follows*: $M^* = M$, $H^* = H$,

$$f_j^*(s) = \begin{cases}\text{const.}(1-s)f_j(s), & \textit{if } t_j \le x,\\ f_j(s), & \textit{if } t_j > x,\end{cases} \tag{4.7}$$

$$a^*(s,z) = \begin{cases}(1-s)a(s,z), & \textit{if } z \le x,\\ a(s,z), & \textit{if } z > x.\end{cases} \tag{4.8}$$

(ii) *Given $X = x$, where $x = t_i$ is among the prior jumps M, A is still Lévy with posterior parameters $M^* = M$, $H^* = H$ and $a^*(s,z)$ as in* (4.8) *and*

$$f_j^*(s) = \begin{cases}\text{const.}(1-s)f_j(s), & \textit{if } t_j < x,\\ \text{const.}\,sf_i(s), & \textit{if } t_j = x,\\ f_j(s), & \textit{if } t_j > x.\end{cases} \tag{4.9}$$

(iii) *Given $X = x$, where x is not among the prior jumps M, A is again Lévy with posterior parameters $M^* = M \cup \{x\}$, $a^*(s,z)$ as in* (4.8), *$H^* = H$ once more and*

$$f_j^*(s) = \begin{cases}\text{const.}(1-s)f_j(s), & \textit{if } t_j < x,\\ f_j(s), & \textit{if } t_j > x,\end{cases}$$

while the new jump $S = A\{x\}$ has density

$$f^*(s) = \text{const.}\,sa(s,x), \qquad s \in (0,1). \tag{4.10}$$

PROOF. The simultaneous distribution of (A, X) is defined by $\mathscr{P}'\{X \in D, A \in G\} = \mathscr{E}F(D)I\{A \in G\}$, for D in $[0,\infty)$ and $G \in \Sigma_{\mathscr{B}}$. In particular, the marginal distribution of X is

$$F_0(x) = \mathscr{E}F(x) = 1 - \left[\prod_{t_j \le x} \mathscr{E}(1 - S_j)\right]\exp\{-R_{(0,x)}[s]\},$$

using Lemma A.3 in the Appendix and with notation as in (4.5). Now define

$$\mathscr{P}_1(G) = \mathscr{P}'\{X > x, A \in G\} = \mathscr{E}\left[\prod_{[0,x]} \{1 - dA(z)\}\right] I\{A \in G\}.$$

This is a finite measure on $\{\mathscr{B}, \Sigma_{\mathscr{B}}\}$ with integral operator $\mathscr{E}_1$, say. By extension, $\mathscr{E}_1\psi(A) = \mathscr{E}[\prod_{[0,x]}\{1 - dA(s)\}]\psi(A)$ for every measurable bounded ψ. In particular,

$$\begin{aligned}
&\mathscr{E}_1 \exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\}\\
&\quad = \mathscr{E}\left[\prod_{[0,x]} \{1 - dA(z)\}\right]\exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\}\\
&\quad = \left[\prod_{t_j \le x} \mathscr{E}(1 - S_j)\exp\{-\theta(t_j)S_j\}\right]\left[\prod_{t_j > x} \mathscr{E}\exp\{-\theta(t_j)S_j\}\right]\\
&\qquad \times \exp\{-R_{(0,x)}[1 - e^{-\theta(z)s} + se^{-\theta(z)s}]\}\exp\{-R_{(x,\infty)}[1 - e^{-\theta(z)s}]\},
\end{aligned}$$

where Lemma A.3 is used again. $\theta(z) = 0$ gives $\mathscr{E}_1 1 = F_0(x, \infty)$, as it should, and the following analogue to (4.6) comes forward for the posterior distribution $\mathscr{P}'\{A \in G | X > x\} = \mathscr{P}_1(G)/F_0(x, \infty)$:

$$\begin{aligned}
&\mathscr{E}'\left[\exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\}\middle| X > x\right]\\
&\quad = \frac{\mathscr{E}_1\exp\{-\int_0^\infty\theta(z)\, dA(z)\}}{\mathscr{E}_1 1}\\
&\quad = \prod_{t_j \le x} \frac{\mathscr{E}(1 - S_j)\exp\{-\theta(t_j)S_j\}}{\mathscr{E}(1 - S_j)} \prod_{t_j > x} \mathscr{E}\exp\{-\theta(t_j)S_j\}\\
&\qquad \times \frac{\exp\{-R_{(0,x)}[1 - e^{-\theta(z)s} + se^{-\theta(z)s}]\}}{\exp\{-R_{(0,x)}[s]\}}\exp\{-R_{(x,\infty)}[1 - e^{-\theta(z)s}]\}.
\end{aligned}$$

Since the ratio in the middle equals $\exp\{-R_{(0,x)}[(1 - e^{-\theta(z)s})(1 - s)]\}$, this proves, using (4.5) and by analogy with (4.6), that statement (i) of the theorem is true.

Case (ii) is similar, but requires slightly lengthier arguments. The following is a sketch. Consider the finite measure

$$\mathscr{P}_2(G) = \mathscr{P}'\{X = x, A \in G\} = \mathscr{E}\left[\prod_{[0,x)} \{1 - dA(z)\}\right] A\{x\} I\{A \in G\}$$

on $\mathscr{B}$. It has integral operator satisfying

$$
\begin{aligned}
&\mathscr{E}_2 \exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\} \\
&\quad = \mathscr{E}\left[\prod_{[0,x)} \{1 - dA(z)\}\right]\exp\left\{-\int_{[0,x)} \theta(z)\, dA(z)\right\} \\
&\qquad \times \mathscr{E}A\{x\}\exp[-\theta(x)A\{x\}]\mathscr{E}\exp\left\{-\int_{(x,\infty)} \theta(z)\, dA(z)\right\} \\
&\quad = \left[\prod_{t_j<x} \mathscr{E}(1 - S_j)\exp\{-\theta(t_j)S_j\}\right] \\
&\qquad \times \mathscr{E}S_i\exp\{-\theta(x)S_i\}\left[\prod_{t_j>x} \mathscr{E}\exp\{-\theta(t_j)S_j\}\right] \\
&\qquad \times \exp\{-R_{(0,x)}[1 - e^{-\theta(z)s} + se^{-\theta(z)s}]\}\exp\{-R_{(x,\infty)}[1 - e^{-\theta(z)s}]\},
\end{aligned}
\tag{4.11}
$$

by elaborations similar to those in the proof of Lemma A.3. $\theta(z) = 0$ yields $F_0\{x\}$. Dividing these expressions leads to an expression for $\mathscr{E}'[\exp\{-\int_0^\infty \theta(z)\, dA(z)\}|X = x]$, which, by comparison with the general formula (4.6), is seen to imply (ii).

The most difficult case is (iii), conditioning on the event $\{X = x\}$ when this has zero probability. What needs to be proved is that

$$\mathscr{P}'\{X \in D, A \in G\} = \int_D \mathscr{P}_x^*\{A \in G\}F_0(dx),$$

where $\mathscr{P}_x^*$ is the candidate for $\mathscr{P}'\{\cdot|X = x\}$ given in the theorem, for all Borel sets D and every G in $\Sigma_{\mathscr{B}}$. In view of (ii) it suffices to show this for $D = (a, b]$, an arbitrary interval free of points from M. Define

$$\mathscr{P}_3\{X \in (a,b], A \in G\} = \mathscr{E}\left[\prod_{[0,a]} \{1 - dA(z)\} - \prod_{[0,b]} \{1 - dA(z)\}\right]I\{G \in A\}$$

and $\mathscr{E}_3\psi(A) = \int\psi(A)\mathscr{P}_3(dA)$. The problem is reduced to that of showing

$$\mathscr{E}_3 \exp\left\{-\int_0^\infty \theta\, dA\right\} = \int_{(a,b]}\left[\mathscr{E}_x^* \exp\left\{-\int_0^\infty \theta\, dA\right\}\right]F_0(dx)$$

for all right-continuous step functions ending in zero, where $\mathscr{E}_x^*$ is expectation evaluated according to $\mathscr{P}_x^*$. This is demonstrated via heroic integrations in Hjort (1984).

Another and more intuitive approach is based on evaluating

$$\mathscr{E}'\left\{\exp\left(-\int_0^\infty \theta\, dA\right)\middle| X \in [x, x + \varepsilon)\right\}$$

and its limit as ε goes to zero. This can be handled by arguments resembling those used to prove (i) and (ii) above. One arrives at an expression similar to (4.11), but instead of $\mathscr{E}S_i \exp\{-\theta(x)S_i\}$, it includes a factor which is close to

(respectively, converges to)

$$\frac{\mathscr{E}A[x, x+\epsilon)\exp\{-\theta(x)A[x, x+\varepsilon)\}}{\mathscr{E}A[x, x+\varepsilon)} \to \frac{\int_0^1 s\exp\{-\theta(x)s\}a(s,x)\,ds}{\int_0^1 sa(s,x)\,ds},$$

and this agrees with statement (iii). This line of reasoning is included here because it is more intuitive than the formal proof and leads in a simpler way to the correct answer; because similar arguments are helpful in the more complicated world of Markov chains discussed in Section 5; and finally because I believe even a formal proof can be constructed based on the $\varepsilon \to 0$ arguments, with the additional help of general results in Pfanzagl (1979) or Diaconis and Freedman [(1986b), Section 4]. □

Let us next consider a full sample. Let $X_1, \dots, X_n$ be iid given A and assume that $(T_1, \delta_1), \dots, (T_n, \delta_n)$ is observed, where $T_i = \min(X_i, c_i)$, $\delta_i = I\{X_i \le c_i\}$ and $c_1, \dots, c_n$ are the censoring times. Define, analogously to (2.5) and (2.6), the counting process N and the left-continuous at-risk process Y:

$$(4.12) \qquad N(t) = \sum_{i=1}^{n} I\{T_i \le t \text{ and } \delta_i = 1\}, \qquad Y(t) = \sum_{i=1}^{n} I\{T_i \ge t\}.$$

In particular, $dN(t) = N\{t\}$ is the number of observed X_i's at the exact spot t. The following theorem parallels Proposition 2.1 and also resembles Theorem 4 in Ferguson and Phadia (1979). It is proved by repeated application of Theorem 4.1, conditioning first on (T_1, δ_1), then on (T_2, δ_2), etc. The only difficulty lies in carefully sorting out the factors contributing to the density of a new jump.

THEOREM 4.2. *Let the Lévy process A have prior distribution as in (4.1)–(4.3), but with no fixed points of discontinuity, i.e., M is empty. Let $u_1, \dots, u_p$ be the distinct points at which noncensored observations occur. Then the posterior distribution of A is a Lévy process, with parameters $M^* = \{u_1, \dots, u_p\}$, $H^* = H$ and $a^*(s,z) = (1-s)^{Y(z)}a(s,z)$, and $A\{u_j\}$ has density*

$$f_j^*(s) = \text{const.}\, s^{dN(u_j)}(1-s)^{Y(u_j)-dN(u_j)}a(s, u_j).$$

The typical application will start out with a continuous prior guess A_0, making this theorem appropriate. However, a version also including a nonempty M of fixed points of discontinuity can be given. If $t \in M$ is such a point, with prior density $f_t(s)$ for the jump $S = A\{t\}$, then the posterior density of S can be shown to be

$$(4.13) \qquad f_t^*(s) = \text{const.}\, s^{dN(t)}(1-s)^{Y(t)-dN(t)}f_t(s).$$

The next result parallels the time-discrete result (2.9) and should become to the beta processes what Theorem 1 in Ferguson (1973) is to the Dirichlet processes.

COROLLARY 4.1. *Let* $A \sim \text{beta}\{c(\cdot), A_0(\cdot)\}$, *as defined in Section* 3.3. *Then, given the data* $(T_1, \delta_1), \dots, (T_n, \delta_n)$,

$$A \sim \text{beta}\left\{c(\cdot) + Y(\cdot), \int_0^{(\cdot)} \frac{c\, dA_0 + dN}{c + Y}\right\}.$$

PROOF. Considering jumps first, what must be shown is that $A\{t\}|\text{data}$ is distributed as a beta variable with parameters $c(t)A_0\{t\} + dN(t)$ and $c(t)(1 - A_0\{t\}) + Y(t) - dN(t)$; cf. (3.13). This follows from Theorem 4.2 when $A_0\{t\} = 0$, since $a(s,z) = c(z)s^{-1}(1-s)^{c(z)-1}$ and from its appendix (4.13) when $A\{t\} > 0$. Next, consider intervals between fixed jump sites. According to (3.14), one must demonstrate that the posterior Lévy formula can be written in the form

$$dL_t^*(s) = \int_0^t \{c(z) + Y(z)\}s^{-1}(1-s)^{c(z)+Y(z)-1} \frac{c(z)\, dA_{0,c}(z)}{c(z) + Y(z)}\, ds.$$

But this follows also from Theorem 4.2 □

REMARK 4A. By convention, a beta$\{0, b\}$ variable is equal to the constant 0 a.s. and a beta$\{a, 0\}$ is equal to 1 a.s., when a and b are positive parameters. In particular, for all but a finite number of t's it is the case that $\mathcal{P}'\{A\{t\} = 0|\text{data}\} = 1$. Nevertheless, in addition to the jumps that occur for A whenever $A_0\{t\}$ and/or $dN(t)$ are positive, it will with probability 1 have infinitely many tiny jumps at a random collection of sites. This is a fact following from the Lévy process property.

4.2. *Bayes estimators.* It is now easy to construct large classes of nonparametric Bayes estimators for A and for the accompanying cdf F. Consider the general A defined in (4.1)–(4.3). Then

$$\mathcal{E}A(t) = \sum_{t_j \le t} \mathcal{E}S_j + \int_0^1 s\, dL_t(s) = \sum_{t_j \le t} \mathcal{E}S_j + \int_0^t \int_0^1 sa(s,z)\, ds\, dH(z)$$

and

$$\begin{aligned}\mathcal{E}F(t) &= 1 - \mathcal{E}\prod_{[0,t]} \{1 - dA(z)\} \\ &= 1 - \prod_{t_j \le t} \mathcal{E}(1 - S_j) \prod_{[0,t]} \left\{1 - \int_0^1 sa(s,z)\, ds\, dH(z)\right\} \\ &= 1 - \prod_{t_j \le t} \mathcal{E}(1 - S_j) \exp\left\{-\int_0^t \int_0^1 sa(s,z)\, ds\, dH(z)\right\}.\end{aligned}$$

Bayes estimators for A and F (w.r.t. quadratic loss functions) are obtained by applying these formulae to the appropriate a posteriori situations. One can also compute a posteriori variances (and even higher moments) in reasonable generality.

Let A have the prior distribution described in Theorem 4.2, with prior guess $\mathscr{E}A(t) = \int_0^t \int_0^1 sa(s,z)\,ds\,dH(z)$. Define $K[z;\,b,c] = \int_0^1 s^b(1-s)^c a(s,z)\,ds$. Then, in the notation used in Theorem 4.2,

$$\begin{aligned} \hat{A}(t) &= \mathscr{E}'\{A(t)|\text{data}\} \\ &= \sum_{u_j \le t} \frac{K\left[u_j; dN(u_j) + 1, Y(u_j) - dN(u_j)\right]}{K\left[u_j; dN(u_j), Y(u_j) - dN(u_j)\right]} \qquad (4.14) \\ &\quad + \int_0^t K[z; 1, Y(z)]\, dH(z), \end{aligned}$$

$$\begin{aligned} \hat{F}(t) &= \mathscr{E}'\{F(t)|\text{data}\} \\ &= 1 - \prod_{u_j \le t} \frac{K\left[u_j; dN(u_j), Y(u_j) + 1 - dN(u_j)\right]}{K\left[u_j; dN(u_j), Y(u_j) - dN(u_j)\right]} \qquad (4.15) \\ &\quad \times \exp\left\{ - \int_0^t K[z; 1, Y(z)]\, dH(z) \right\}. \end{aligned}$$

If the prior has $a(s,z) = a(s)$ independent of z [e.g., $c(z) \equiv c$ in a beta process], then $\mathscr{E}A(t) = \int_0^1 sa(s)\,ds\,H(t)$ is the prior guess, $K[z;\,b,c] = K[b,c]$ is also independent of z and the formulae above simplify. This could be termed the *homogeneous case*. For example, if the prior has $a(s,z) \equiv 2$, so that H is in fact the prior expectation, then

$$(4.16) \qquad \hat{A}(t) = \sum_{u_j \le t} \frac{dN(u_j) + 1}{Y(u_j) + 2} + \int_0^t \frac{2\,dH(z)}{(Y(z)+1)(Y(z)+2)}.$$

In these equations, note that the typical time-continuous case would have every $dN(u_j)$ equal to 1. In (4.16) the second term is of smaller order than the first and the first term uses binomial estimates of the type $(x+1)/(n+2)$ instead of x/n.

The primary special case however, is the following, giving analogues to (2.10) and (2.11). It follows from Corollary 4.1 upon using equations (3.16) and (3.18).

THEOREM 4.3. *Let A be a beta process with parameters c and A_0, as defined in Section* 3.3. *Then the Bayes estimators of A and F, based on n possibly censored observations, are*

$$\hat{A}(t) = \int_0^t \frac{c\,dA_0 + dN}{c + Y}, \qquad \hat{F}(t) = 1 - \prod_{[0,t]} \left[1 - \frac{c\,dA_0 + dN}{c+Y}\right].$$

A couple of remarks are in order here. Firstly, as $c(\cdot)$ tends to zero, $\hat{A}$ and $\hat{F}$ tend to the usual nonparametric estimators of, respectively, Nelson and Aalen and Kaplan and Meier. These can accordingly be given a Bayesian justification under vague prior conditions. At the other extreme are the prior guesses A_0 and F_0, which arise as $c(\cdot)$ tends to infinity. In general, $c(\cdot)$ can be interpreted as the number at risk in an imagined prior sample, as in Remark 2B for the discrete case. Note finally that the precision of the Bayes estimate $\hat{A}$ is naturally measured by

$$\mathrm{Var}\{A(t)|\mathrm{data}\} = \int_0^t \frac{d\hat{A}(1 - d\hat{A})}{c + Y + 1}.$$

This can also be used to construct Bayesian confidence bands for A. For this purpose another perfectly feasible approach is to simulate realizations of $A(\cdot)$ from the posterior distribution.

5. Markov chains.

Suppose that $X_a = \{X_a(t); t \geq 0\}$ are Markov chains for $a = 1, \dots, n$, moving around in the state space $\{1, \dots, k\}$ independently of each other and each with the same set of cumulative hazard rates A_{ij}. Assume that X_a is followed over the interval $[0, w_a]$ (but see Remark 2E) and introduce

$$\begin{aligned} N_{ij}(t) &= \text{number of observed transitions } i \to j \text{ during } [0, t], \\ Y_i(t) &= \text{number at risk in state } i \text{ just before time } t. \end{aligned} \tag{5.1}$$

We were able to generalize from (2.10) to (2.16) in the time-discrete case and aim now at a similar generalization from results of Section 4 to the present setting. To establish such results for Markov chains, quite a bit of ground needs to be covered, partly parallelling the work done in Sections 3 and 4. A brief outline is given here, with most details left behind in Hjort (1984).

The beta process was obtained in Section 3.3 essentially as a fine limit of time-discrete ones. Thus a natural start of our program is a study of the limiting processes obtained by letting the time-interval length b tend to zero for the discrete processes of Proposition 2.2. The following theorem, whose proof is omitted, parallels Theorem 3.1.

THEOREM 5.1. *Let $A_{0,1}, \dots, A_{0,k-1}$ be continuous, nondecreasing functions on $[0, \infty)$, each starting with the value 0 at zero, and let $c(\cdot)$ be a piecewise continuous, nonnegative function. Define for each n independent vectors*

$$(X_{n,1,u}, \dots, X_{n,k,u}) \sim \mathrm{Dir}\{a_{n,1,u}, \dots, a_{n,k,u}\}, \qquad u = 1, 2, \dots,$$

where $a_{n,j,u} = c((u - \frac{1}{2})/n)A_{0,j}((u - 1)/n, u/n]$ for $j = 1, \dots, k - 1$ and $a_{n,k,u} = c((u - \frac{1}{2})/n)\{1 - \sum_{j=1}^{k-1} A_{0,j}((u - 1)/n, u/n]\}$, and construct the cumulative hazard rates

$$A_{n,j}(t) = \sum_{u/n \leq t} X_{n,j,u}, \qquad t \geq 0, j = 1, \dots, k - 1.$$

Then it holds that $(A_{n,1}, \dots, A_{n,k-1})$ converges in distribution to $(A_1, \dots, A_{k-1})$, in each of the spaces $D[0, R]^{k-1}$, where $A_1, \dots, A_{k-1}$ are independent beta processes, with $A_j \sim$ beta$\{c, A_{0,j}\}$.

This suggests that a natural Bayesian strategy is to start out specifying $k(k-1)$ independent beta processes as priors for A_{ij}. A rather long proof of Theorem 5.2 is available in Hjort (1984). It uses arguments similar to those given in Section 4, but necessarily becomes more cumbersome notationally.

THEOREM 5.2. *Let A_{ij} for $i \neq j$ be independent* beta$\{c_{ij}, A_{0,ij}\}$ *processes and assume that the prior guesses $A_{0,ij}$ and $A_{0,il}$ have disjoint sets of discontinuity points when $j \neq l$ (they will often be chosen continuous). Then, given data collected from the n individual Markov chains, the A_{ij}'s are still independent and*

$$A_{ij}|\text{data} \sim \text{beta}\left\{c_{ij} + Y_i, \int_0^{\cdot} \frac{c_{ij}\,dA_{0,ij} + dN_{ij}}{c_{ij} + Y_i}\right\}.$$

Bayes estimators for the cumulative hazard rates are

$$\hat{A}_{ij}(t) = \mathscr{E}\{A_{ij}(t)|\text{data}\} = \int_0^t \frac{c_{ij}\,dA_{0,ij} + dN_{ij}}{c_{ij} + Y_i}. \tag{5.2}$$

Further, the conditional variance of $A_{ij}(t)$ given data can be written $\int_0^t d\hat{A}_{ij}(1 - d\hat{A}_{ij})/(c_{ij} + Y_i + 1)$. We can also derive a Kaplan–Meier-resembling nonparametric Bayes estimator of the waiting time distribution in state i, i.e.,

$$G_i([s,t]) = \Pr\{X(z) \equiv i \text{ in } [s,t] | X(s) = i\} = \prod_{[s,t]} \{1 - dA_{i\cdot}(z)\} \quad \text{for } t \geq s.$$

The result is

$$\hat{G}_i([s,t]) = \prod_{[s,t]} \left[1 - \frac{c_i\,dA_{0,i\cdot} + dN_{i\cdot}}{c_i + Y_i}\right], \tag{5.3}$$

where $N_{i\cdot} = \sum_{j \neq i} N_{ij}$, $A_{0,i\cdot} = \sum_{j \neq i} A_{0,ij}$ and where for simplicity $c_i = c_{ij}$.

6. Semiparametric regression models. Assume that a covariate vector z_i is recorded for individual number i and that these measurements are believed to influence the individual's hazard function. Among several possible semiparametric models appropriate for such situations the most famous one is the Cox model, which postulates that

$$F_i(t, \infty) = F(t, \infty)^{\exp(\beta z_i)}, \qquad i = 1, \ldots, n,$$

for some parameter β, where F_i is the cdf for individual i and F is the cdf for an individual having covariate vector zero. This relation translates into $1 - dA_i(s) = \{1 - dA(s)\}^{\exp(\beta z_i)}$ in terms of hazard functions; cf. (3.3).

Suppose first that β is known and that A has a beta process prior with parameters c and A_0. We shall find the Bayes estimator for A, the baseline hazard for individuals with covariate vector zero. Data are in the form of $t_i = \min(x_i, c_i)$, $\delta_i = I\{x_i \leq c_i\}$ again. If we momentarily think of an observed life time t_i as the event $X_i \in [t_i, t_i + \varepsilon]$, then number i contributes

$$F_i[t_i, t_i + \varepsilon] = \prod_{[0, t_i)} \{1 - dA(s)\}^{\exp(\beta z_i)} \left[1 - \prod_{[t_i, t_i + \varepsilon]} \{1 - dA(s)\}^{\exp(\beta z_i)} \right]$$

to the likelihood if $\delta_i = 1$ and

$$F_i(t_i, \infty) = \prod_{[0, t_i]} \{1 - dA(s)\}^{\exp(\beta z_i)}$$

if $\delta_i = 0$. The total likelihood $L(\text{data}|A, \beta)$ can be written as

$$\prod_{\text{out}} \{1 - dA(s)\}^{R(s, \beta)},$$

taken over all s outside the $[t_i, t_i + \varepsilon]$ intervals where lifetimes are observed, multiplied by factors

$$\prod_{[t_i, t_i + \varepsilon]} \{1 - dA(s)\}^{R_{(i)}(s, \beta)} - \prod_{[t_i, t_i + \varepsilon]} \{1 - dA(s)\}^{R(s, \beta)},$$

for each i with $\delta_i = 1$, in which $R(s, \beta) = \sum_{j=1}^{n} \exp(\beta z_j) I\{t_j \geq s\}$ and $R_{(i)}(s, \beta)$ is the same quantity but evaluated without number i. It is assumed that the t_i's with $\delta_i = 1$ are distinct.

This can be used to work out the posterior distribution of A. We skip the somewhat laborious details here, but report that A given data is a Lévy process and is distributed like a beta process $\{c + R, cA_0/(c + R)\}$ between jumps. It has positive jumps $A\{s\}$ at observed lifetimes, i.e., where $dN(s) > 0$, where N is the total counting process for the data. The distribution of such an $A\{s\}$ is, however, more complex than a simple beta distribution. Some work yields the expected value as $\mathscr{E}(A\{s\}|\text{data}) = J(s, \beta)\, dN(s)/\{c(s) + R(s, \beta)\}$, where

$$J(s, \beta) = \frac{1}{c + R - \Delta} \frac{\Delta}{\psi(c + R) - \psi(c + R - \Delta)},$$

$\Delta(s, \beta) = \sum_{i=1}^{n} \exp(\beta z_i) I\{s = t_i, \delta_i = 1\}$ and $\psi(x)$ is the digamma function $\Gamma'(x)/\Gamma(x)$. We define $J(s, \beta)$ to be zero when $dN(s) = 0$. Since the observed lifetimes are distinct, $\Delta(s, \beta)$ is equal to $\exp(\beta z_i)$ at the precise point at which number i dies. Combining our efforts we arrive at

$$\hat{A}(t, \beta) = \mathscr{E}\{A(t)|\text{data}, \beta\} = \int_0^t \frac{c(s)\, dA_0(s) + J(s, \beta)\, dN(s)}{c(s) + R(s, \beta)}.$$

This can be compared with the traditional estimator $\int_0^t dN(s)/R(s, \beta)$. The expansion $\psi(z) = \log z - \frac{1}{2}/z - \frac{1}{12}/z^2 + \cdots$ can be used to show that $J(s, \beta)$ is reasonably close to 1 for most values of c and R and Δ (and is exactly equal to 1 when $\Delta = 1$).

This generalizes some of the results of Section 4 and can be viewed as parallelling work by Wild and Kalbfleisch (1981). These authors found it necessary to assume that covariates were constant in time, whereas they can be time-dependent in our reasoning above.

Of course β is seldom known in practice. Any reasonable estimate $\hat{\beta}$ based on the data produces upon insertion in $\hat{A}(t, \beta)$ an empirical Bayes type estimator for A. $\hat{\beta}$ could, for example, be of the Bayesian variety $\int \beta L(\beta)\pi(\beta)\,d\beta / \int L(\beta)\pi(\beta)\,d\beta$ studied in Hjort [(1986a) Section 4], where $L(\beta) = \prod_{i:\ \delta_i = 1}\{\exp(\beta z_i)/R(t_i, \beta)\}$ is Cox's familiar partial likelihood. But a more complete semiparametric Bayesian treatment is possible and works out as follows. Place a prior distribution $\pi(\beta)\,d\beta$ on β in addition to the beta process prior on A and take β and A to be independent. Lengthy arguments, involving $L(\text{data}|A, \beta)$ and reasoning similar to that of Section 4, in addition to technicalities similar to those spelled out in the proof of Lemma A.3 in the Appendix, yield the posterior density for β as

$$\begin{aligned}\pi(\beta|\text{data}) = \text{const.}\exp&\left[-\int_0^\infty \{\psi(c(s) + R(s,\beta)) - \psi(c(s))\}c(s)\,dA_0(s)\right]\\ &\times \prod_{i:\,\delta_i = 1}[\psi(c(t_i) + R(t_i,\beta))\\ &\qquad -\psi(c(t_i) + R(t_i,\beta) - \Delta(t_i,\beta))]\pi(\beta),\end{aligned}$$

provided that A_0 has a continuous hazard rate α_0 and that t_i's with $\delta_i = 1$ again are distinct. The semiparametric Bayes estimator for A becomes

$$\hat{A}(t) = \mathscr{E}\{\hat{A}(t,\beta)|\text{data}\} = \int \hat{A}(t,\beta)\pi(\beta|\text{data})\,d\beta.$$

It is perfectly possible to compute this posterior average by simulation, by some appropriate acceptance/rejection program.

It is similarly feasible to estimate more complex parameters. As an example, consider $\kappa = F(t, \infty)^{\exp(\beta z)}$, the survival probability for an individual with covariate vector z. The Bayes estimator is

$$\begin{aligned}\hat{\kappa} &= \mathscr{E}\mathscr{E}\{\kappa|\text{data},\beta\}\\ &= \int \prod_{[0,t]}\left[1 - \frac{c(s)\,dA_0(s) + J(s,\beta)\,dN(s)}{c(s) + R(s,\beta)}\right]^{\exp(\beta z)} \pi(\beta|\text{data})\,d\beta.\end{aligned}$$

The expression for the posterior density above represents an interesting spectrum of possible curves as $c(\cdot)$ varies. When c tends to zero it reduces to $\prod_{i:\ \delta_i = 1}[\psi(R(t_i,\beta)) - \psi(R(t_i,\beta) - \exp(\beta z_i))]$ times the prior and a multiplicative factor, and to a first order approximation this is equal to the ordinary partial likelihood $L(\beta)$. When c grows to infinity one can show that the posterior density becomes proportional to $L_0(\beta)\pi(\beta)$, where $L_0(\beta)$ is the likelihood under the assumption $A \equiv A_0$. This is similar to what Kalbfleisch and Prentice [(1980), Section 8.4] found starting out with a gamma process prior for $-\log(1 - F)$. In our view the estimators obtained here, particularly for the cumulative hazard, seem more natural and are easier to interpret, which suggests that the beta process approach is preferable to the gamma process approach.

Diaconis and Freedman (1986a, b), Hjort (1986b) and others have shown that semiparametric and nonparametric Bayesian schemes sometimes lead to estimators that are either plainly inconsistent or consistent for certain least false parameters that depend on the construction of the prior distribution. In this particular situation nothing dramatic happens, however, as one can show that $(1/n)\log\{\pi(\beta|\text{data})/\pi(0|\text{data})\}$ has exactly the same limit as that of the corresponding expression for Cox's partial likelihood $L(\beta)$, under reasonable regularity conditions. In particular, regardless of $c(\cdot)$ and $A_0(\cdot)$, as long as $c(\cdot)$ stays bounded as n grows, the Bayes solution $\hat{\beta} = \mathscr{E}\{\beta|\text{data}\}$ and the Cox estimator are consistent for the very same least false/best fitting parameter value β_0. This can be demonstrated using techniques of Hjort (1986a).

7. Further developments.

In this section some topics for possible future research are discussed briefly, along with some complementing remarks.

7A. *Generalized Dirichlet processes.* Let A be a beta process with parameters $c(\cdot)$ and $A_0(\cdot)$ and consider the random cdf $F(t) = 1 - \prod_{[0,t]}\{1 - dA_0(s)\}$. It has expectation $F_0(t) = 1 - \prod_{[0,t]}\{1 - dA_0(s)\}$ irrespective of $c(\cdot)$. It is interesting to note that F is simply a Dirichlet process with parameter $kF_0(\cdot)$, where k is a given positive constant, for the particular choice $c(s) = kF_0[s,\infty)$. One way of proving this is to use product integral techniques to evaluate the Lévy representation for $B = -\log(1 - F)$,

$$\mathscr{E}\exp\{-\theta B(t)\} = \mathscr{E}\{1 - F(t)\}^{\theta} = \prod_{[0,t]}\mathscr{E}\{1 - dA(s)\}^{\theta},$$

which becomes the required $\Gamma(k)\Gamma(kF_0[t,\infty) + \theta)/\Gamma(kF_0[t,\infty))\Gamma(k + \theta)$; see also the following formula. Accordingly, the class of Dirichlet processes has been extended. We may term F above a *generalized Dirichlet process* with *two* parameter functions $c(\cdot)$ and $F_0(\cdot)$.

The various probabilistic properties of these generalized Dirichlet processes may now be explored, in the tradition of Ferguson (1973, 1974), Doksum (1974), Antoniak (1974), Korwar and Hollander (1973), Hjort (1976) and others. Furthermore, the list of Bayes estimators based on Dirichlet processes in different situations published in the literature since and including Ferguson (1973) could be supplemented with accompanying Bayes solutions using generalized Dirichlet process priors.

Various authors have concentrated on $B = -\log(1 - F)$ instead of A, as mentioned earlier. It is not difficult to find Bayes estimators, etc., for B, using the following result. If A is a general beta process, as in (3.12)–(3.15), then B is Lévy with representation

$$\mathscr{E}\exp\{-\theta B(t)\} = \left[\prod_{t_j \le t}\mathscr{E}(1 - S_j)^{\theta}\right]\exp\left[-\int_0^t\{\psi(c(s) + \theta) - \psi(c(s))\}c(s)\,dA_{0,c}(s)\right].$$

7B. *Asymptotic distributions.* The Bayes estimators of Section 4 are interesting competitors to the traditional nonparametric estimators and one might explore their frequentist behavior. Thus let A denote the true, underlying cumulative hazard and introduce

$$M(t) = N(t) - \int_0^t Y(s)\, dA(s), \qquad t \geq 0. \tag{7.1}$$

This is a square integrable martingale w.r.t. the sigma-fields $\mathscr{F}_t = \sigma\{N(s), Y(s); s \leq t\}$; cf. Aalen (1978a) and Gill (1980). We have from

$$\hat{A}(t) = \int_0^t \frac{c(s)\, dA_0(s) + dN(s)}{c(s) + Y(s)} \tag{7.2}$$

that $\hat{A}(t) - A(t) = \int_0^t 1/(c + Y)(dM - c\, dK)$, in which $K(t) = A(t) - A_0(t)$. One consequence, among several, is that $E_A \hat{A}(t) = A(t) + \int_0^t E_A c/(c + Y)\, dK$. The bias is small when A is close to A_0 and/or when Y is large compared with c.

The limiting distribution of $\hat{A} - A$ may be derived in a framework in which n, the number of individuals under study, tends to infinity. A standard assumption often fulfilled is that $Y(s)/n \to y(s)$, say, uniformly on bounded sets, in probability. The martingale techniques of Aalen (1978a) may be used to show that $\sqrt{n}(\hat{A} - A)$ has the same limiting distribution, namely that of a certain Gaussian martingale, as $\sqrt{n}(A^* - A)$, where $A^*(t) = \int_0^t dN/Y$ is the maximum likelihood-like Nelson–Aalen estimator; cf. Andersen and Borgan (1985). The Bayes estimator $\hat{F}$ obtained in Theorem 4.3 can be studied similarly. Martingale techniques may be used to obtain results along the lines of Susarla and Van Ryzin (1978b).

7C. Dynamic Bayes estimators. The loss function under which (7.2) is the Bayes solution is

$$L(A, \tilde{A}) = \int_0^\infty \{\tilde{A}(t) - A(t)\}^2\, dW(t), \tag{7.3}$$

where W is any finite measure. (7.2) was derived when A was given a completely specified a priori distribution, namely the beta$\{c, A_0\}$. This is in the classical decision theoretic tradition of Wald and followers. It should be pointed out, however, that (7.2) makes perfect sense also when the $c(s)$ function is allowed to depend upon (portions of) the data. In particular most of the arguments pertaining to the behavior of $\hat{A}$ in the non-Bayesian frequentist framework of 7B go through when $c(s)$ is only assumed to be left-continuous and progressively measurable w.r.t. $\{\mathscr{F}_t; t \geq 0\}$, or more generally *predictable*; cf. Gill (1980). For example, the limit distribution result about $\sqrt{n}(\hat{A} - A)$ mentioned above holds true if only $c(s)/\sqrt{n} \to 0$ uniformly on bounded sets, in probability.

But (7.2) is not a proper Bayes solution when $c(\cdot)$ depends on the data, at least not in the traditional framework. In the present situation a full *function* $A(t)$ is to be estimated and one could, somewhat speculatively, allow a Bayesian to gradually adjust his beliefs about the future by also taking into account information relevant for earlier time points. Specifically, one might propose to judge an estimator by its dynamic risk function

$$R(A, \tilde{A}) = \int_0^\infty E_A\Big\{\big(d\tilde{A}(s) - dA(s)\big)^2\Big|\mathscr{F}_{s-}\Big\}\, dW(s).$$

The statistician is to estimate $dA(s) = A[s, s + ds]$, say, using prior information as well as data observed in $[0, s]$. The dynamic Bayes estimator becomes $d\hat{A}(s) = \mathscr{E}\{dA(s)|\mathscr{F}_{s-}\}$. If (3.5) is found appropriate for some predictable $c(s)$, then $\hat{A}$ becomes as in (7.2) again, but now with $c(s)$ more generally interpreted. $c(s)$ is, for example, allowed to depend on $Y(s)$.

7D. Empirical Bayes estimation. A statistical formalization sometimes placed between the pure Bayesian and the classical frequentist frameworks is the empirical Bayes setup in which data from previous experiments are available and modelled as being relevant for the present one. Assume that $A \sim \text{beta}\{c, A_0\}$ is one's prior process (dynamic or not), so that $\hat{A}$ in (7.2) is the Bayes estimator. Assume further that A_0 is unknown, but that m earlier independent experiments have resulted in observed processes $N_j(\cdot)$ and $Y_j(\cdot)$, defined as in (4.12), for $j = 1, \dots, m$. Suppose finally that the m cumulative hazard rates $A_1, \dots, A_m$ are each distributed as A. Then an estimator $\hat{A}_0$ may be constructed, for example,

$$\hat{A}_0(t) = \int_0^t \left\{\sum_{j=1}^m Y_j(s)\right\}^{-1} d\left\{\sum_{j=1}^m N_j(s)\right\}.$$

The insertion of this for A_0 in (7.2) defines an empirical Bayes estimator. Similarly, the parameter c can be estimated from earlier data, for example, when it is modelled as being constant. One may prove, using martingale techniques again [M in (7.1) is a martingale for given A and $A_j - A_0$ is also a martingale], that the procedure outlined here is asymptotically optimal as m increases. This yields results along the lines of Korwar and Hollander (1976) and Susarla and Van Ryzin (1978a).

7E. Semi-Markov processes. The present paper has introduced beta processes, and Bayes estimators under such, into models leading in generality to time-inhomogeneous finite-state Markov chains and to Cox-like regressions. The Markov assumption is sometimes too crude, however, and it appears useful to generalize some results to semi-Markov processes. This is at least possible for the case of forward-going semi-Markov processes, the class studied by Voelkel and Crowley (1984). See Phelan (1990) for a study of Markov renewal processes using the beta prior.

7F. Double censoring. It would also be useful to find Bayes estimators based on beta processes in situations where data can be censored also from the left, providing Bayesian competitors to the maximum likelihood-like solutions of Turnbull (1974) and Samuelsen (1989). This is a difficult problem to solve in any generality, although explicit formulae may be derived when the number of left-censored data is small. It appears that equations may be obtained from which Bayes estimators may be computed by numerical iteration procedures.

7G. General Aalen models. The estimator (7.2) can obviously be used as an estimator incorporating prior beliefs for the cumulative hazard in Aalen's (1978a) general multiplicative counting process model, of which the models considered in this paper are only special cases. I have not been able to formalize the Bayesian framework in this generality, however, due to difficulties caused by the fact that beta processes (and relatives) are discrete, with infinitely many jumps on each interval, with probability 1. Work by Jacobsen (1982) can possibly be exploited to arrive at (7.2) in the most general case.

7H. Admissibility and minimaxity. It is difficult to establish that (7.2) is admissible under loss function (7.3) in the unrestricted nonparametric sense, even though it is the almost unique Bayes solution. The (7.2) estimators probably are admissible; some relevant methods of proof are in Hjort (1976). Using slightly nonorthodox martingale techniques, in the simultaneous framework in which both A and the data N, Y are random, one can work out an expression for the minimum Bayes risk under a beta process prior, w.r.t. loss function (7.3), namely $\int_0^\infty \int_0^t \mathscr{E} c/\{(c+1)(c+Y)\}\, dA_0(s)(1 - dA_0(s))\, dW(t)$. The inner integral is maximal for $c(s) = Y(s)^{1/2}$. This hints at a minimax property for the estimator

$$\tilde{A}(t) = \int_0^t \frac{\frac{1}{2}\sqrt{Y(s)} + dN(s)}{\sqrt{Y(s)} + Y(s)}.$$

These considerations may be formalized, but necessarily involve the dynamic framework mentioned in 7C.

7I. More general prior processes. The beta processes have several good properties. Some characteristics may make them inappropriate in some situations, however, for example, the independent increment property (shared by all Lévy processes, of course). It would be useful to have in one's Bayesian toolkit prior processes with correlated increments. Similarly, in the competing risks and Markov chain situations of Section 5 we used independent priors for the cumulative hazards involved, whereas these perhaps should be negatively correlated, say, in some applications. The possibility offered by the dynamic beta processes with predictable parameter functions provides a partial answer to such needs. One could also try out linear combinations of independent Beta processes. More work is needed.

7J. Estimating the hazard rate itself. We have proposed (7.2) as a nonparametric Bayesian estimator of the cumulative hazard A. We may write

$$d\hat{A}(s) = \frac{c(s)}{c(s) + Y(s)} \alpha_0(s)\, ds + \frac{Y(s)}{c(s) + Y(s)}\, dA^*(s),$$

where again A^* is the non-Bayesian Nelson–Aalen solution, in the case of a prior guess $\alpha_0(s)$ for the hazard rate $\alpha(s)$ itself. Thus a reasonable nonparametric Bayesian estimator of α is $\hat{\alpha}(s) = c(s)/\{c(s) + Y(s)\}\alpha_0(s) + Y(s)/\{c(s) + Y(s)\}\alpha^*(s)$, in which $\alpha^*(s)$ is any of several possible non-Bayesian smoothers of $dA^*(s)$, for example the kernel-type in Ramlau-Hansen (1983) or the orthogonal expansion estimator in Hjort (1985).

APPENDIX

Lemmas A.1 and A.2 were needed in the proof of Theorem 3.1. Lemma A.3 was vigorously used in the proof of Theorem 4.1 and a more general version of it was used to obtain the posterior density $\pi(\beta|\text{data})$ in Section 6.

LEMMA A.1. *Let $z_{n,i}$ be real numbers, for $n \ge 1$ and $i \ge 1$. Assume that, as $n \to \infty$, (i) $\Sigma_{a<i/n\le b} z_{n,i} \to z$; (ii) $\max_{a<i/n\le b}|z_{n,i}| \to 0$; (iii) $\limsup \Sigma_{a<i/n\le b}|z_{n,i}| \le M < \infty$. Then $\Pi_{a<i/n\le b}(1 + z_{n,i}) \to \exp(z)$.*

PROOF. We have $\log(1 + z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \cdots = z + z^2K(z)$, say, where $K(z) \to -\frac{1}{2}$ as $z \to 0$ and $|K(z)| \le 1$ whenever $|z| \le \frac{1}{2}$. For n large enough, every $|z_{n,i} \le \frac{1}{2}$. It suffices to show $\Sigma_{a<i/n\le b} z_{n,i}^2 K(z_{n,i}) \to 0$. But the left-hand side is dominated by $\max_{a<i/n\le b}|z_{n,i}|\Sigma_{a<i/n\le b}|z_{n,i}|$. □

LEMMA A.2. *The space $\mathscr{A}_R$ of all cumulative hazard rates on $[0, R]$, i.e., $\mathscr{A}$ restricted to $[0, R]$, is closed in $D[0, R]$ w.r.t. the Skorohod topology.*

PROOF. Take for convenience $R = 1$. Let $\{A_n\}$ be a sequence of cumulative hazard rates, having by definition corresponding cdf's $\{F_n\}$ that satisfy

$$F_n(t) = 1 - \prod_{[0,t]}\{1 - dA_n(s)\}, \quad \text{for } t \ge 0;$$

see (3.3). Assume that $A_n \to A$ in $D[0, 1]$; we are to show that indeed also A is a cumulative hazard rate. In other words, if F defined by

$$F(t) = 1 - \prod_{[0,t]}\{1 - dA(s)\}, \qquad t \ge 0,$$

is well defined and lies in $\mathscr{F}_1$, the cdf's restricted to $[0, 1]$, then the lemma is proved.

Let ε be a given number in $(0, 1)$. There exist points $0 = t_0 < \cdots < t_m = 1$ having $\max_{i\le m} A(t_{i-1}, t_i) \le \varepsilon$; see Billingsley [(1968), page 110]. Let us write $F(t) = K(t) - \delta(t)$ as follows: For an arbitrary t, say in $(t_j, t_{j+1}]$, let

$$K(t) = 1 - \left[\prod_{i=1}^{j}\{1 - A(t_{i-1}, t_i]\}\right]\{1 - A(t_j, t]\}.$$

One has $0 \le \delta(t) \le \exp\{A(t)\}A(t)\max_{i\le m} A(t_{i-1}, t_i) \le \exp\{A(t)\}A(t)\varepsilon$, according to a version of Theorem 2.5 in Johansen (1987).

Now there is a sequence of strictly increasing, continuous functions λ_n, having $\lambda_n(0) = 0$ and $\lambda_n(1) = 1$, such that $\lambda_n(s) \to s$ and $A_n(\lambda_n s) \mapsto A(s)$, uniformly [see Billingsley (1968), page 112]. Write

$$F_n(\lambda_n t) = 1 - \prod_{[0,\lambda_n t]} \{1 - dA_n(s)\} = K_n(t) - \delta_n(t),$$

in which

$$K_n(t) = 1 - \left[\prod_{i=1}^{j} \{1 - A_n(\lambda_n t_{i-1}, \lambda_n t_i]\}\right]\left\{1 - A_n(\lambda_n t_j, \lambda_n t]\right\}$$

and

$$0 \le \delta_n(t) \le \exp\{A_n(\lambda_n t)\} A_n(\lambda_n t) \max_{i \le m} A_n(\lambda_n t_{i-1}, \lambda_n t_i).$$

Keeping m fixed, we get from $|F(t) - F_n(\lambda_n t)| \le |K(t) - K_n(t)| + \delta(t) + \delta_n(t)$ that $\limsup_{n\to\infty} |F(t) - F_n(\lambda_n t)| \le 2A(1)\exp\{A(1)\}\varepsilon$, showing us that $F_n \to F$ in the Skorohod topology in $D[0,1]$. But it is easy to prove that $\mathscr{F}_1$ is closed in $D[0,1]$, so that necessarily F is a cdf restricted to $[0,1]$. □

LEMMA A.3. *Let A be the general Lévy process defined in* (4.1)–(4.3), *or, equivalently*, (4.6). *Then*

$$\begin{aligned}
\mathscr{E}&\left[\prod_{[0,x]} \{1 - dA(z)\}\right] \exp\left\{-\int_0^\infty \theta(z)\, dA(z)\right\} \\
&= \prod_{t_j \le x} \mathscr{E}(1 - S_j)\exp\{-\theta(t_j)S_j\} \prod_{t_j > x} \mathscr{E}\exp\{-\theta(t_j)S_j\} \\
&\quad \times \exp\{-R_{(0,x)}[1 - e^{-\theta(z)s} + se^{-\theta(z)s}]\}\exp\{-R_{(x,\infty)}[1 - e^{-\theta(z)s}]\}.
\end{aligned}$$

PROOF. A_c, the part of A that has no fixed points of continuity, is independent of the jumps $S_j = A\{t_j\}$ and is itself a Lévy process [cf. Ferguson (1974), page 623] with Lévy formula

$$\mathscr{E}\exp\left\{-\int_0^\infty \theta(z)\, dA_c(z)\right\} = \exp\{-R_{(0,\infty)}[1 - e^{-\theta(z)s}]\}. \tag{A1}$$

By writing $\prod_{[0,x]}\{1 - dA(z)\}\exp\{-\int_0^\infty \theta\, dA\}$ as

$$\begin{aligned}
&\prod_{t_j \le x} (1 - S_j)\exp\{-\theta(t_j)S_j\} \prod_{t_j > x} \exp\{-\theta(t_j)S_j\} \\
&\quad \times \left[\prod_{[0,x]} \{1 - dA_c(z)\}\right] \exp\left\{-\int_0^x \theta(z)\, dA_c(z)\right\} \exp\left\{-\int_x^\infty \theta(z)\, dA_c(z)\right\},
\end{aligned}$$

and using independence and (A1), the problem is reduced to that of proving

$$\begin{aligned}
&\mathscr{E}\left[\prod_{[0,x]} \{1 - dA_c(z)\}\right] \exp\left\{-\int_0^x \theta(z)\, dA_c(z)\right\} \\
&\qquad = \exp\{-R_{(0,x)}[1 - e^{-\theta(z)s} + se^{-\theta(z)s}]\}.
\end{aligned} \tag{A2}$$

Before taking recourse to a rigorous proof we include some arguments that are heuristic but illustrative. If $E\exp(-\theta Y) = \exp(-h(\theta))$, then $E(1 - Y)\exp(-\theta Y) = (1 - h'(\theta))\exp(-h(\theta))$. Hence, since

$$\mathscr{E}\exp\{-\theta(z)\,dA_c(z)\} = \exp\left\{-\int_0^1 (1 - e^{-\theta(z)s})a(s,z)\,ds\,dH(z)\right\}$$

by (A1) and (4.5), we have

$$\begin{aligned}\mathscr{E}\{1 - dA_c(z)\}&\exp\{-\theta(z)\,dA_c(z)\}\\ &= \left\{1 - \int_0^1 se^{-\theta(z)s}a(s,z)\,ds\,dH(z)\right\}\\ &\quad\times\exp\left\{-\int_0^1 (1 - e^{-\theta(z)s})a(s,z)\,ds\,dH(z)\right\}.\end{aligned}$$

Multiplying together these infinitesimal factors gives (A2), since $\prod_{[0,x]}\{1 - f(z)\,dH(z)\} = \exp\{-\int_0^x f(z)\,dH(z)\}$ when the right-hand side is continuous in x.

These arguments may be formalized as follows. Note first that it suffices to demonstrate

$$\text{(A3)}\qquad \begin{aligned}\mathscr{E}\left[\prod_{[b,c)}\{1 - dA_c(z)\}\right]&\exp\{-\theta A_c[b,c)\}\\ &= \exp\{-R_{(b,c)}[1 - e^{-\theta s} + se^{-\theta s}]\}\end{aligned}$$

for a constant θ, since $\theta(z)$ is a step function. Write

$$\begin{aligned}J &= \left[\prod_{[b,c)}\{1 - dA_c(z)\}\right]\exp\{-\theta A_c[b,c)\},\\ J_i &= \{1 - A_c[x_{i-1}, x_i)\}\exp\{-\theta A_c[x_{x-1}, x_i)\}, \qquad i = 1,\dots,m,\end{aligned}$$

where $b = x_0 < \cdots < x_m = c$ is a fine partition. $\prod_{i=1}^m J_i$ provides an approximation to J; in fact it follows from Theorem 2.5 in Johansen (1987) that $\prod_{i=1}^m J_i \to J$ a.s. when $m \to \infty$ and the sequence of partitions $\{x_i\}_0^m$ is chosen such that the max mesh tends to zero. Since the product furthermore is bounded (by 1) it follows that $\mathscr{E}\prod_{i=1}^m J_i \to \mathscr{E}J$. But

$$\mathscr{E}J_i = \{1 - R_{(x_{i-1},x_i)}[se^{-\theta s}]\}\exp\{-R_{(x_{i-1},x_i)}[1 - e^{-\theta s}]\}$$

and

$$\begin{aligned}\mathscr{E}\prod_{i=1}^m J_i &= \exp\{-R_{(b,c)}[1 - e^{-\theta s}]\}\prod_{i=1}^m\{1 - R_{(x_{i-1},x_i)}[se^{-\theta s}]\}\\ &\to \exp\{-R_{(b,c)}[1 - e^{-\theta s}]\}\exp\{-R_{(b,c)}[se^{-\theta s}]\},\end{aligned}$$

by Lemma A.1. This proves (A3) and the present lemma. □

Acknowledgments. Conversations with Odd Aalen triggered some of the research that led to the present paper. I am grateful for the careful work of an Associate Editor and her/his helpers and for the persistent encouragement of Editor van Zwet, which finally caused me to carry out the necessary revision.

REFERENCES

AALEN, O. O. (1978a). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. O. (1978b). Nonparametric estimation of partial transition probabilites in multiple decrement models. *Ann. Statist.* **6** 534–545.

ANDERSEN, P. K. and BORGAN, Ø. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12** 97–158.

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BREIMAN, L. (1968). *Probability.* Addison-Wesley, Reading, Mass.

DIACONIS, P and FREEDMAN, D. (1986a). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

DIACONIS, P. and FREEDMAN, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87.

DOKSUM, K. A. (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. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

FERGUSON, T. S. and KLASS, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43** 1634–1643.

FERGUSON, T. S. and PHADIA, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7** 163–186.

FLEMING, T. R. and HARRINGTON, D. P. (1978). Estimation for discrete time nonhomogeneous Markov chains. *Stochastic Process. Appl.* **7** 131–139.

GILL, R. D. (1980). *Censoring and Stochastic Integrals.* Math. Centre Tracts **124**, Mathematical Centre, Amsterdam.

GILL, R. D. and JOHANSEN, S. (1987). Product integrals and counting processes. Research report, CWI, Amsterdam.

HJORT, N. L. (1976). Applications of the Dirichlet process to some nonparametric problems (in Norwegian). Graduate thesis, Univ. Tromsø. [Abstract in *Scand. J. Statist.* (1977) **4** 94.]

HJORT, N. L. (1984). Nonparametric Bayes estimators of cumulative intensities in models with censoring. Research Report No. 762, Norwegian Computing Centre, Oslo.

HJORT, N. L. (1985). Contribution to the discussion of Andersen and Borgan's "Counting process models for life history data: A review." *Scand. J. Statist.* **12** 141–150.

HJORT, N. L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scand. J. Statist.* **13** 63–85.

HJORT, N. L. (1986b). Contribution to the discussion of Diaconis and Freedman. *Ann. Statist.* **14** 49–55.

JACOBSEN, M. (1982). *Statistical Analysis of Counting Processes. Lecture Notes in Statist.* **12**. Springer, Berlin.

JOHANSEN, S. (1987). Product integrals and Markov processes. *CWI Newsletters* **12** 3–13.

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

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711.

KORWAR, R. M. and HOLLANDER, M. (1976). Empirical Bayes estimation of a distribution function. *Ann. Statist.* **4** 581–588.

LÉVY, P. (1936). *Théorie de l'Addition des Variables Aléatoire*.Gauthiers-Villars, Paris.

PADGETT, W. J. and WEI, L. J. (1981). A Bayesian nonparametric estimator of survival probability assuming increasing failure rate. *Comm. Statist. Theory Methods* **10** 49–63.

PFANZAGL, J. (1979). Conditional distributions as derivatives. *Ann. Probab.* **7** 1046–1050.

PHELAN, M. J. (1990). Bayes estimation from a Markov renewal process. *Ann. Statist.* **18** 603–616.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

SAMUELSEN, S. O. (1989). Asymptotic theory for nonparametric estimators from doubly censored data. *Scand. J. Statist.* **16** 1–21.

SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71** 897–902.

SUSARLA, V. and VAN RYZIN, J. (1978a). Empirical Bayes estimation of a distribution (survival) function from right-censored observations. *Ann. Statist.* **6** 740–754.

SUSARLA, V. and VAN RYZIN, J. (1978b). Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples. *Ann. Statist.* **6** 755–768.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69** 169–173.

VOELKEL, J. G. and CROWLEY, J. (1984). Nonparametric inference for a class of semi-Markov processes with censored observations. *Ann. Statist.* **12** 142–160.

WILD, C. J. and KALBFLEISCH, J. D. (1981). A note on a paper by Ferguson and Phadia. *Ann. Statist.* **9** 1061–1065.

NORWEGIAN COMPUTING CENTRE
GAUSTADALLÉEN 23
P.B. 114 BLINDERN
N-0314 OSLO 3
NORWAY