Nonparametric Bayesian Survival Analysis

by

Jean-Marie Rolin*

April 1997

Abstract

This article is a review of nonparametric Bayesian Survival Analysis and is a part of a contribution to the Encyclopedia of Biostatistics to be published by John Wiley and Sons Ltd. Most of the prior specifications proposed in the literature since Ferguson's seminal paper are presented with a special emphasis on the class of neutral to the right processes and on its subclass introduced by Hjort under the denomination of Beta processes. Berliner-Hill proposal of robust predictive distributions is also discussed. In the simple model with censored observations and without explanatory variables or heterogeneity factors, a complete description of the posterior distribution of the "predictive" hazard function is provided in the conditional model and in the joint model.

AMS MOS primary: 62N05
secundary: 62F15 62699

*Institut de Statistique and CORE, Université Catholique de Louvain, Belgium.
Financial support from the contract "Projet d'Actions de Recherche Concertées" no. 93/98-164 of the Belgian Government are gratefully acknowledged.

1 Introduction

In sampling theory, a statistical model is decribed by a measurable space $(S, \mathcal{S})$, a state space for the observation called the sample space, and a family of probability measures on $(S, \mathcal{S})$ indexed by a parameter θ belonging to a parameter set Θ, i.e., $\{P_\theta : \theta \in \Theta\}$, known as the **sampling distributions** of the observation. In this set-up, the parameter θ is totally unknown and any one of its possible values plays the same role. On the contrary, in Bayesian statistics, one considers that some prior information on the parameter is available and has to be incorporated in the statistical model. This must be done in the following way: the parameter set Θ is endowed by a σ-field $\mathcal{T}$ ensuring that P_θ is a probability transition, and $(\Theta, \mathcal{T})$ is endowed by a probability measure μ, called the **prior distribution** of the parameter θ expressing the available information. The sampling distributions are therefore naturally considered as (regular) conditional probabilities on the sample space given the parameter. A Bayesian statistical model is then a unique probability P on the product space parameter $\times$ observation, i.e., $(\Theta \times S, \mathcal{T} \otimes \mathcal{S})$ obtained by extending

$$P[C \times B] = \int_C P_\theta(B)\mu(d\theta) \quad \forall\, C \in \mathcal{T}, \quad \forall\, B \in \mathcal{S}.$$

From this joint probability measure, Bayesian inference aims at evaluating and analysing the revision of the prior distribution by the observation, i.e., the conditional probabilities on the parameter given the observation, to be called the **posterior distributions**. The marginal distribution of the observation is called the **predictive distribution** and is obtained as an average of the sampling probabilities weighted by the prior probability.

In the simplest case of **nonparametric duration statistical model**, the data form an i.i.d. sample of size n, $\{T_i : 1 \leq i \leq n\}$ of a positive random variable T defined on an abstract probability space $(\Omega, \mathcal{M})$ with values in $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+)$ ($\overline{\mathbb{R}}^+ = [0, \infty]$ and $\overline{\mathcal{B}}^+$ are the Borel sets of $\overline{\mathbb{R}}^+$) and distribution F, i.e. there exists a probability P_F on $(\Omega, \mathcal{M})$ such that

$$\begin{aligned} &(i) \qquad \underset{1\leq i\leq n}{\|} T_i \,;P_F \\ &(ii) \qquad P_F \circ T_i^{-1} = F \end{aligned} \tag{1.1}$$

The parameter F is therefore a probability distribution on $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+)$ and the Bayesian analysis consists in specifying a **prior distribution** on the

parameter space. This specification permits to consider a unique probability measure P on $(\Omega, \mathcal{M})$ (see e.g. Florens, Mouchart, Rolin (1990)) and with respect to this probability the parameter becomes a random probability measure, i.e.,

$$\begin{aligned} & F(B)(\omega) : \overline{\mathcal{B}}^+ \times \Omega \to [0,1] \text{ is such that} \qquad (1.2) \\ (i) \quad & F(B)(\cdot) \text{ is a random variable } \forall\, B \in \mathcal{B}^+ \\ (ii) \quad & F(\cdot)(\omega) \text{ is a probability measure on } (\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+) \quad \forall\, \omega \in \Omega. \end{aligned}$$

In this Bayesian set-up, (1.1) may be rewritten as:

$$\begin{aligned} (i) \quad & \underset{1 \le i \le n}{\|} T_i \mid F\,; P \qquad (1.3) \\ (ii) \quad & P[T_i \in B \mid F] = F(B), \quad \forall\, B \in \overline{\mathcal{B}}^+ \\ & \text{or, } T_i \mid F \sim F. \end{aligned}$$

The Bayesian inference consists in determining the **posterior distribution** of F, i.e., the distribution of F conditionally on the observation $T_1^n = \{T_i : 1 \le i \le n\}$ and the marginal distribution of the observation, i.e., the **predictive distribution**.

In his seminal paper, Ferguson (1973) shows that, to specify a distribution on F, a probability measure (or more generally a σ-finite measure) on a measurable space $(S, \mathcal{S})$, it suffices, in view of the Kolmogorov extension theorem, to define coherently the (finite dimensional) distributions of $\{F(B_\ell) : 1 \le \ell \le k\}$ for any measurable (non trivial) partition of S, $\{B_\ell : 1 \le \ell \le k\} \subset \mathcal{S}$. As noted later on by Doksum (1974) in the real case, we can more generally suppose that (almost) all realizations of F are probability measures (σ-finite measures) on $(S, \mathcal{S})$ if we furthermore assume that

$$(i) \quad (S, \mathcal{S}) \text{ is a standard Borel space (see, e.g. Breiman (1968))}$$

and that

$$\begin{aligned} (ii) \quad & F_0(B) = E[F(B)] \quad \forall\, B \in \mathcal{S} \text{ is a probability measure} \qquad (1.4) \\ & (\sigma-\text{finite measure}) \text{ on } (S, \mathcal{S}). \end{aligned}$$

In "classical" nonparametric statistics, F is usually assumed to be diffuse (i.e. $F(\{x\}) \doteq F(x) = 0 \quad \forall\, x \in S$) while the empirical distribution function $F_n = \dfrac{1}{n} \sum\limits_{1 \le i \le n} \varepsilon_{T_i}$ is purely discrete. In Bayesian nonparametric statistics,

F is very often a mixture of these two kinds of distributions. But their properties are quite different and, very often, F is decomposed into its diffuse part and its purely discrete part as follows (we will always use "mutatis mutandis" the same notation)

$$\begin{aligned} I_F &= \{x \in S : F(x) > 0\} \\ F &= F_c + F_d \\ F_c &= F(\cdot \cap I_F^c) \\ F_d &= \sum_{x \in I_F} F(x)\varepsilon_x \end{aligned} \tag{1.5}$$

(ε_x is the unit mass at point x also called the Dirac measure at point x)

As also noted by Ferguson (1973,1974) the prior specification must possess two desirable properties in point of view of a Bayesian nonparametric analysis: (1) the set of all realizations of F must be dense with respect to some suitable topology (pointwise or weak convergence) in the space of all probability measures on $(S, \mathcal{S})$ and (2) the posterior distribution of F should be analytically tractable.

We first review the proposed prior distributions for this simple set-up.

2 Uncensored Observations

2.1 The Natural Conjugate Model

For an i.i.d. sample of size n, $\{T_i : 1 \leq i \leq n\}$, the empirical counting process, defined as

$$N_n = nF_n = \sum_{1 \leq i \leq n} \varepsilon_{T_i} \tag{2.1.1}$$

is a sufficient statistic, i.e., $T_1^n \perp\!\!\!\perp F \mid N_n$ and N_n is a multinomial process in the sense that for any measurable (non trivial) partition $\{B_\ell : 1 \leq \ell \leq k\}$ of S, $\{N_n(B_\ell) : 1 \leq \ell \leq k\}$ conditionally on F has a multinomial distribution of parameters n and $\{F(B_\ell) : 1 \leq \ell \leq k\}$. The natural conjugate prior of the multinomial distribution is the Dirichlet distribution (see Wilks (1962)). According to these facts, Ferguson (1973) defined the Dirichlet measure as follows:

Definition 2.1.1 A random probability F is a Dirichlet measure of parameter n_0F_0 where $n_0 \in \mathbb{R}_0^+$ and F_0 is a probability measure on $(S, \mathcal{S})$, to

be written as $F \sim \mathcal{D}i(n_0 F_0)$, if for any measurable (non trivial) partition, $\{B_\ell : 1 \leq \ell \leq k\}$ of S, $\{F(B_\ell) : 1 \leq \ell \leq k\}$ has the Dirichlet distribution $Di[\{n_0 F_0(B_\ell) : 1 \leq \ell \leq k\}]$ which may be characterized as follows:

$$\forall\ a_\ell > -n_0 F_0(B_\ell)\ , 1 \leq \ell \leq k \tag{2.1.2}$$

$$E[\prod_{1 \leq \ell \leq k} F(B_\ell)^{a_\ell}] = \frac{\Gamma(n_0)}{\Gamma(n_0 + a)} \prod_{1 \leq \ell \leq k} \frac{\Gamma[n_0 F_0(B_\ell) + a_\ell]}{\Gamma[n_0 F_0(B_\ell)]} \mathbb{1}_{\{F_0(B_\ell) > 0\}}$$

where $a = \sum_{1 \leq \ell \leq k} a_\ell$.

■

Clearly this implies that $\forall\ B \in \mathcal{S}$, $F(B)$ has a Beta distribution; more precisely

$$F(B) \sim Be[n_0 F_0(B), n_0 F_0(B^c)] \tag{2.1.3}$$

This implies that

$$E[F(B)] = F_0(B) \tag{2.1.4}$$

and

$$V[F(B)] = \frac{F_0(B) F_0(B^c)}{n_0 + 1}. \tag{2.1.5}$$

Therefore F_0 may be seen as a prior guess of the true distribution of T, F_* let us say, and n_0 as a measure of the confidence we have in this prior guess.

As may easily be seen in the finite case by the natural conjugate property, Ferguson showed that the posterior distribution of F is again a Dirichlet measure, more precisely:

Proposition 2.1.2 If $F \sim \mathcal{D}i(n_0 F_0)$ and $\{T_i : 1 \leq i \leq n\}$ is a random sample of F (see 1.3) then $F \mid T_1^n \sim \mathcal{D}i(n_* F_{n_*})$ where

$$n_* = n_0 + n \tag{2.1.6}$$

$$n_* F_{n_*} = n_0 F_0 + n F_n \tag{2.1.7}$$

■

In survival analysis, the prime concern is the inference on the survival function,

$$S_t = F[(t, \infty]] = P(T > t \mid F] \tag{2.1.8}$$

In this context, we obtain that

$$S_t \mid T_1^n \sim Be[n_* S_{n_* t}, n_* - n_* S_{n_* t}] \tag{2.1.9}$$

and, for instance, the Bayesian estimator of S_t for a quadratic loss function is given by

$$E[S_t \mid T_1^n] = S_{n_* t} = \frac{n_0}{n_0 + n} S_{0t} + \frac{n}{n_0 + n} S_{nt} \tag{2.1.10}$$

i.e., a convex combination between the prior survival function and the sampling survival function.

It is interesting to note that no matter what is the true survival function S_{*t}, the Bayesian estimator converges to it uniformly a.s. since it has the same behaviour as the sampling survival function.

To complete the Bayesian analysis, it remains to describe the predictive distribution. The sequential description is the easiest one.

$$\begin{aligned} &(i) \qquad T_1 \sim F_0 \\ &(ii) \qquad T_{i+1} \mid T_1^i \sim \frac{n_0}{n_0 + i} F_0 + \frac{1}{n_0 + i} \sum_{1 \le j \le i} \varepsilon_{T_j} \quad \forall\ 1 \le i \le n-1 \end{aligned} \tag{2.1.11}$$

Note that even if F_0 is diffuse ($F_{0d} = 0$), then $P(T_2 = T_1) = \dfrac{1}{n_0 + 1}$.

Therefore, this Bayesian model generates ties in the sample with positive probability and the joint distribution of the sample is analytically complex. But it can be characterized in an alternative way when F_0 is diffuse. Let $\{T_{(j)} : 1 \le j \le M\}$ be the set of distinct values of the sample in order of their appearance and let $I_j = \{i : T_i = T_{(j)}\}$, $D_j = |I_j|$ (the cardinality of I_j) so that

$$F_n = \frac{1}{n} \sum_{1 \le j \le M} D_j \varepsilon_{T_{(j)}} \tag{2.1.12}$$

Then conditionally on M, $T_{(1)}^{(M)} = \{T_{(j)} : 1 \le j \le M\}$ and $I_1^M = \{I_j : 1 \le j \le M\}$ are independent and $T_{(j)}$, $1 \le j \le M$, are i.i.d. F_0. The distribution of the configuration of ties (I_1^M) is more involved but depends only on n_0 and may be described as in a Polya's urn scheme (see Blackwell and MacQueen (1973), Antoniak (1974), Yamato (1984) and Rolin (1992)).

The fact that this Bayesian model generates ties in the sample with positive probability has been considered as troublesome by many statisticians even if this discreteness is no more troublesome in many problems than the

discreteness of the empirical distribution function. However, if n is large, Korwar, Hollander (1973) have shown that M the number of distinct values is of order $n_0 \ln n$ and this may be contradicted by the sample. Some solutions to cope with this problem will be presented in section 2.3.

Note also that when n is large and when the statistician wants to incorporate the prior information that F_*, the true distribution, is diffuse, i.e. $M = n$, he can choose $n_0 = n/\ln n$. Such a choice preserves the consistency of the Bayesian estimator since $n_0/n \to 0$.

2.2 The Moment Problem

In the natural conjugate model, the inference on $P[T \in B|F] = F(B)$ is entirely characterized by (2.1.3) and proposition 2.1.2. But if the interest lies into a more general moment, i.e., $E[f(T)|F] = \int_S f dF \doteq F(f)$, for $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$, like for instance the expected lifetime or the expected residual lifetime, inference becomes much more difficult. It is known that $F(f)$ admits a density w.r.t. the Lebesgue measure (see, e.g. Florens, Rolin (1994)) but only in exceptional cases an analytical expression for this density is available (see Hannum, Hollander, Langberg (1981), Yamato (1984b), Cifarelli, Regazzini (1990), Diaconis, Kemperman (1995)). If these results have a great interest in their own, they are less important for a practical statistician who wants to make inference on $F(f)$ whatever f and F_0 must be. Two solutions for this general problem are available: simulation techniques and asymptotic approximations. Simulation requires a complete description of the trajectories of the Dirichlet process.

2.2.1 Trajectories of the Dirichlet Process

As noted by Ferguson (1973) and Blackwell (1973), (almost) all realizations of F are purely discrete, i.e., $F_c = 0$. Using the relation between Gamma processes and Dirichlet processes and results on the distribution of the size of the jumps of a Gamma process (see, e.g., Ferguson, Klass (1972)), Ferguson (1973) gave a complete description of the distributions of the locations and sizes of the jumps of a Dirichlet measure. However, his description is not very easy to implement and Sethuraman (1994) and Rolin (1992,1993) independently found another description in the general case which is much easier to simulate (see also Florens, Rolin (1994)). We will provide details for this alternative description for the posterior distribution (given in proposition 2.1.2) under the assumption that F_0 is diffuse. In such a case,

$I_{F_{n*}} = I_{F_n} = \{T_i : 1 \le i \le n\}$ and if we define $p = F(I_{F_n})$, $F^a = F(\ |I^c_{F_n})$ and $F^s = F(\ |I_{F_n})$ so that

$$F = (1-p)F^a + pF^s \tag{2.2.1}$$

Then,

$$(i) \qquad p \perp\!\!\!\perp F^a \perp\!\!\!\perp F^s | T_1^n$$

$$(ii) \qquad p|T_1^n \sim Be(n, n_0)$$

$$(iii) \qquad F^a|T_1^n \sim \mathcal{D}i(n_0 F_0)$$

$$(iv) \qquad F^s|T_1^n \sim \mathcal{D}i(nF_n).$$

Now

$$F^a = \sum_{1 \le k < \infty} \alpha_k \varepsilon_{\tau_k} \tag{2.2.2}$$

where

$$(i) \qquad \alpha_k = \delta_k \prod_{1 \le \ell \le k-1} (1 - \delta_\ell)$$
$$(ii) \qquad \delta_1^\infty \perp\!\!\!\perp \tau_1^\infty$$
$$(iii) \qquad \tau_i,\ 1 \le i < \infty, \text{ are i.i.d. } F_0$$
$$(iv) \qquad \delta_i,\ 1 \le i < \infty, \text{ are i.i.d. } Be(1, n_0)$$

and

$$F^s|T_1^n = \sum_{1 \le j \le M} \beta_j \varepsilon_{T_{(j)}} \tag{2.2.3}$$

where

$$\{\beta_j : 1 \le j \le M\}|T_1^n \sim Di[\{D_j : 1 \le j \le M\}]$$

Representations (2.2.2) and (2.2.3) are almost sure, but as far as only distributional properties are relevant, it may be shown, using properties of Dirichlet measures, that

$$F^s|T_1^n \sim \sum_{1 \le i \le n} \hat{\beta}_i \varepsilon_{T_i} \tag{2.2.4}$$

where

$$\{\hat{\beta}_i : 1 \leq i \leq n\}|T_1^n \sim Di[\{1 : 1 \leq i \leq n\}]$$

Now, simulation of $F(f)$ is performed as follows:

$$F(f)|T_1^n \sim (1-p)F_K^a(f) + pF^s(f)$$

where F_K^a is the truncation of F^a, i.e.,

$$F_K^a = (\sum_{1\leq k\leq K} \alpha_k)^{-1} \sum_{1\leq k\leq K} \alpha_k \varepsilon_{\tau_k} \tag{2.2.5}$$

The error induced by this truncation is completely described in Florens, Rolin (1994) (see also Muliere, Tardella (1996)).

2.2.2 Asymptotic Normality

By general properties of Dirichlet measures (Ferguson (1973)),

$$E[F(f)|T_1^n] = F_{n_*}(f) \tag{2.2.6}$$

and

$$V[F(f)|T_1^n] = \frac{1}{n_* + 1} V_{F_{n_*}}[f(T)] \tag{2.2.7}$$

where

$$V_{F_{n_*}}[f(T)] = \frac{n_0}{n_*} V_{F_0}[f(T)] + \frac{n}{n_*} V_{F_n}[f(T)] + \frac{n_0 n}{n_*^2}\{F_0(f) - F_n(f)\}^2 \tag{2.2.8}$$

$(V_F[f(T)] = F(f^2) - F(f)^2)$.

Therefore, by the strong law of large numbers, if F_* is the true distribution of T and if $F_0(f^2) < \infty$ and $F_*(f^2) < \infty$, then

$$F_{n_*}(f) \to F_*(f) \text{ as. } P_{F_*} \tag{2.2.9}$$

and

$$V_{F_{n_*}}[f(T)] \to V_{F_*}[f(T)] \text{ as. } P_{F_*}.$$

Moreover, it may be deduced from the functional central limit theorem for Dirichlet processes given by Lo (1983,1986,1987), that
$\sqrt{n_*}\{F(f) - F_{n_*}(f)\}|T_1^n$ converges in law to a $N[0, V_{F_*}[f(T)]]$ as. P_{F_*}.

Therefore under the assumptions that $F_0(f^2) < \infty$ and $F_*(f^2) < \infty$, as n is sufficiently large,

$$F(f)|T_1^n \overset{a}{\sim} N[E[F(f)|T_1^n], V[F(f)|T_1^n]] \tag{2.2.10}$$

Remark: The author can give a direct proof of this result which does not require the assumptions of Lo namely $F_* \ll F_0$ and F_* diffuse.

2.3 Other Prior Specifications

2.3.1 Mixtures of Dirichlet Measures

To enlarge the family of possible prior specifications, Antoniak (1974) introduced and studied mixtures of Dirichlet measures. This is done by introducing a random variable $\tilde{\theta}$ defined on $(\Omega, \mathcal{M})$ with values in a measurable space $(\Theta, \mathcal{T})$ (generally a Euclidean space) and specifying besides (1.3) that

$$\begin{aligned} (i) \quad & T_1^n \underline{\parallel} \tilde{\theta} | F \\ (ii) \quad & F|\tilde{\theta} \sim \mathcal{D}i(n_0 F_{0\tilde{\theta}}) \\ (iii) \quad & \tilde{\theta} \sim G \end{aligned}$$

where G is a probability measure on $(\Theta, \mathcal{T})$ and $F_{0\theta}(B) : \Theta \times \mathcal{S} \to [0,1]$ is a probability transition.

From (1.3), (i) and (ii), it is easily seen that

$$(iv) \quad T|\tilde{\theta} \sim F_{0\tilde{\theta}}$$

Now, a posteriori,

$$(v) \quad F|\tilde{\theta}, T_1^n \sim \mathcal{D}i(n_0 F_{0\tilde{\theta}} + nF_n)$$

and therefore $F|T_1^n$ is also a mixture of Dirichlet measures with

$$(vi) \quad \tilde{\theta}|T_1^n \sim G_n$$

Given the complexity of the predictive distribution, $T_1^n|\tilde{\theta}$, G_n seems to be difficult to compute. However, Antoniak who also noticed this, did not emphasize that, when $F_{0\theta}$ is diffuse $\forall\ \theta \in \Theta$, this problem is very simple using the alternative characterization of the predictive distribution. Indeed, the configuration of ties being independent of $\tilde{\theta}$, it can be seen that

$$(vii) \quad \tilde{\theta}|T_1^n = \tilde{\theta}|M, T_{(1)}^{(M)},$$

and since $T_{(1)}^{(M)}$ conditionally on M and $\tilde{\theta}$ is an i.i.d. sample of $F_{0\tilde{\theta}}$, the computation of G_n is done as in a usual parametric Bayesian model using only the distinct values in the sample. This is somewhat disappointing but mixtures of Dirichlet measures have been very successful in models with heterogeneity factors or exogenous variables. Such models are out of the scope of this article.

2.3.2 Tailfree and Neutral Random Measures

Very large families of prior specifications have been introduced by Doksum (1974) under the names of tailfree, neutral and F-neutral random probabilities. They are defined by considering a sequence of nested partitions $\mathcal{P}_n = \{B_{n,\ell} : 1 \leq \ell \leq k_n\}$ generating $\mathcal{S}$, i.e., $\mathcal{S} = \sigma(\bigcup_n \mathcal{P}_n)$ (this happens if and only if $\mathcal{S}$ is separable) and by stipulating independence relations between $F(B_{n,\ell})$, $1 \leq \ell \leq k_n$, $1 \leq n < \infty$ without any other assumptions on the distributions. Doksum shows that the posterior distribution shares the same properties. Some of these random probabilities are diffuse a.s. or even absolutely continuous w.r.t. Lebesgue measure on the real line (see Ferguson (1974)). However, they heavily depend on the chosen sequence of nested partitions. If the defining properties must hold for every sequence of nested partitions generating $\mathcal{S}$, then the random probabilities are Dirichlet measures or are degenerate random probabilities. This explains why such models have not been considered very often.
In fact, it may be shown (see Rolin (1992)) that if

$$F(C|B) = \frac{F(C \cap B)}{F(B)} \mathbb{1}_{\{F(B)>0\}} + \mathbb{1}_{\{F(B)=0\}}$$

and

$$\mathcal{F}_B = \sigma\{F(D) : D \subset B \quad D \in \mathcal{S}\}$$

then

Proposition 2.3.1

$F \sim \mathcal{D}i(n_0 F_0)$ if and only if $\forall\ C, B \in \mathcal{S}$ with $0 < F_0(B) < 1$,

$$(i) \qquad F(C|B) \perp\!\!\!\perp \mathcal{F}_{B^c}$$

$$(ii) \qquad F(B) \sim Be[n_0 F_0(B), n_0 F_0(B^c)]$$

■

In fact (ii) is almost irrelevant, because: If F is a random probability of expectation F_0 satisfying (i), if there exists $B_0 \in \mathcal{S}$ with $0 < F_0(B_0) < 1$ such that $P[F(B_0) \neq F_0(B_0)] > 0$ and if there exists a measurable partition $\{B_1, B_2, B_3\}$ such that $F_0(B_i) > 0 \quad \forall\ 1 \leq i \leq 3$, then there exists $n_0 \in (0, \infty)$ such that $F \sim \mathcal{D}i(n_0 F_0)$.

From this, it is natural to define a class of prior specifications containing the Dirichlet measures by imposing property (i) to hold for a (sufficiently large) subclass of $\mathcal{S}$.

There is a very natural choice of this subclass for survival data for which $S = \overline{\mathbb{R}}^+$ and $\mathcal{S} = \overline{\mathcal{B}}^+$, namely $\mathcal{S}_0 = \{(t, \infty] : t \in \overline{\mathbb{R}}^+\}$ and a random probability measure satisfying (i) $\quad \forall\, C, B \in \mathcal{S}_0$ will be called a neutral to the right process. This will be presented in section 2.4.

2.3.3 Berliner-Hill Predictive Distributions

The starting idea of Berliner-Hill (1988) is the following: if the prior specification on F is such that almost all realizations of F are diffuse probability measures, the order statistics of the sample, $T_{n,i} : 1 \leq i \leq n$, with $0 < T_{n,1} < T_{n,2} < \ldots < T_{n,n} < \infty$ are well defined for all n and the marginal distribution of the sample is exchangeable for all n. Therefore a simple computation yields

$$P[T_{n+1} \in (T_{n,i}, T_{n,i+1})] = \frac{1}{n+1} \tag{2.3.1}$$

$\forall\ 0 \leq i \leq n$, where $T_{n,0} = 0$ and $T_{n,n+1} = \infty$. They claimed that this "posterior predictive distribution" is more sensible than the one which could be deduced from the empirical distribution function. Indeed

$$P(T_{n+1} \in B | F, T_1^n) = P(T_{n+1} \in B | F) = F(B)$$

and replacing F by its pseudo-maximum likelihood estimator F_n will yield.

$$\hat{P}[T_{n+1} = T_{n,i} | F_n, T_1^n] = \frac{1}{n} \tag{2.3.2}$$

Such a "posterior predictive distribution" is relatively unreasonable since it gives probability 0 to a future observation larger than the largest.

They also claimed that estimators of F especially in small samples are not suitable for use as predictive distributions. However in a pure Bayesian set up,

$$P[T_{n+1} \in B | T_1^n] = E[F(B) | T_1^n] \tag{2.3.3}$$

i.e. the real posterior predictive distribution is the estimator of F.

Therefore, their approach has nothing to do with Bayesian methodology and formula (2.3.1) is not a posterior predictive distribution but is simply a marginal probability. As the authors said this model may nevertheless be

used to represent extremely vague subjective knowledge about the form of the true distribution.

This model may be extended to compute probabilities concerning m future observations. Using de Finetti's theorem and ergodic theory, the authors provide a good argument for using this model. Indeed, conditionally on T_1^n,

$$\frac{1}{m} \sum_{1 \leq k \leq m} \mathbb{1}_{\{T_{n,i} < T_{n+k} < T_{n,i+1}\}} \to \theta_i^n$$

a.s. as $m \to \infty$ and

$$\theta_i^n = F[(T_{n,i}, T_{n,i+1})] \tag{2.3.4}$$

Moreover,

$$\{\theta_i^n : 0 \leq i \leq n\} \sim Di(1, 1, 1, \ldots, 1) \tag{2.3.5}$$

implying that

$$E[\theta_i^n] = \frac{1}{n+1} \tag{2.3.6}$$

2.4 Neutral to the Right Processes. Hazard Functions.

As mentioned at the end of section 2.3.2. this large class of prior specification for F on the (positive) real line including the Dirichlet measures has been introduced by Doksum (1974) who established the main properties of such processes and derived, up to a Radon-Nikodym derivative, their posterior distributions. Ferguson (1974) gave an alternative description of this posterior distribution (see also Rolin (1983)) for another description). Further descriptions have been recently used, as for instance in Hjort (1990). We will therefore present a hopefully up-to-date version of these processes.

Let F be a random probability measure on $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+)$ with expectation F_0, i.e.

$$E[F(B)] = F_0(B) \quad \forall\ B \in \overline{\mathcal{B}}^+ \tag{2.4.1}$$

We will frequently use the following notations

$$\begin{aligned} F_t &= F([0, t]) \\ S_t &= F((t, \infty]) = 1 - F_t \\ F(t) &= F(\{t\}) \end{aligned} \tag{2.4.2}$$

Let us define

$$t_0 = \inf\{t : F_{0t} > 0\} \tag{2.4.3}$$

$$t_1 = \inf\{t : F_{0t} = 1\}$$
$$I = \{t : F_0(t) > 0\}$$

Note that besides very special experiments, it is natural to choose $t_0 = 0$, $t_1 = \infty$ and $I = \phi$ since F_0 is a prior guess of the true distribution F_*.

Definition 2.4.1. F is neutral to the right if
$\forall\ t \in [t_0, t_1)$ and $\forall\ s \in (0, t_1 - t)$,

$$\frac{S_{t+s}}{S_t} \underline{\parallel}\ \mathcal{F}_t$$

where $\mathcal{F}_t = \sigma\{F_u : 0 \le u \le t\}$. ∎

This definition may be translated in terms of the hazard function

$$L_t = -\ln S_t \tag{2.4.4}$$

Note that L_t is increasing right-continuous and we may suppose without loss of generality that $L_t = 0 \quad \forall\ t < t_0$ and $L_t = \infty \quad \forall\ t \ge t_1$. Now the definition of F to be neutral to the right becomes

$$\forall\ t \in [t_0, t_1) \ \text{ and } \ \forall\ s \in (0, t_1 - t) \tag{2.4.5}$$
$$L_{t+s} - L_t \underline{\parallel}\ \mathcal{L}_t$$

where $\mathcal{L}_t = \sigma\{L_u : 0 \le u \le t\}$.

Therefore, L_t is a Levy process, i.e., an increasing right-continuous process with independent increments. Such a process may be viewed as a cumulative distribution function of a "**purely**" random σ-finite measure on $[t_0, t_1)$ defined by the Lebesgue-Stieltjes integral

$$L(B) = \int_B dL_t \quad \forall\ B \in \overline{\mathcal{B}}^+ \tag{2.4.6}$$

Such a measure is called purely random because

$$\underset{1 \le \ell \le k}{\underline{\parallel}}\ L(B_\ell),\ \forall\ \{B_\ell : 1 \le \ell \le k\} \subset \overline{\mathcal{B}}^+ \ \text{ partition of } \ \mathbb{R}^+ \tag{2.4.7}$$

The distribution of such a purely random measure is entirely characterized by

$$\psi(p, B) = -\ln E[e^{-pL(B)}] \tag{2.4.8}$$

a transition measure on $(\overline{\mathbb{R}}^+, \overline{\mathcal{B}}^+)$, σ-finite on $[t_0, t_1)$ and concave in $p \in \mathbb{R}^+$.

We will briefly sketch the representation and path properties of such purely random measures according to Ferguson, Klass (1972) (see also Karr (1986) and Daley, Vere-Jones (1988)). Since these measures have to be used as prior specifications of a Bayesian nonparametric survival analysis, we will introduce some reasonable simplifying assumptions.

First of all, in the decomposition of L into its diffuse part and its purely discrete part, the diffuse part is deterministic. It is difficult to imagine a statistical experiment in which it will be interesting to introduce such a deterministic component. Hence we will suppose that L is purely discrete. This will be achieved if

$$\lim_{p \to \infty} \frac{1}{p} \psi(p, B) = 0, \ \forall \ B \in \mathcal{B}^+ \cap [t_0, t_1) \tag{2.4.9}$$

Secondly, L has fixed points of discontinuities corresponding to I since

$$I = \{t : P[L(t) > 0] > 0\} \tag{2.4.10}$$

and we may write, since I is countable,

$$L = \tilde{L} + \sum_{t \in I} L(t) \varepsilon_t \tag{2.4.11}$$

(Recall that $L(t) \doteq L(\{t\}) = L_t - L_{t-}$)

Now, we have

$$\underset{t \in I}{\parallel} L(t) \underline{\parallel} \tilde{L} \tag{2.4.12}$$

The distributions of $L(t)$, $t \in I$, are arbitrary and $\tilde{L}$ is a Levy process without fixed point of discontinuity.

Here again, as mentioned before, it will rarely happen that in a statistical experiment we have to consider times at which there is a positive probability to die. Nevertheless, such a phenomenon will appear in the posterior distribution.

Finally, the counting measure on $(\mathbb{R}^+ \times \mathbb{R}_0^+, \mathcal{B}^+ \otimes \mathcal{B}_0^+)$ defined by

$$N(D) = |\{t : (t, \tilde{L}(t)) \in D\}| \tag{2.4.13}$$

is a Poisson process with expectation

$$N_0(D) = E[N(D)] \tag{2.4.14}$$

which is called the Levy measure and is σ-finite on $[t_0, t_1) \times (0, \infty)$. (Indeed, for $t_0 \leq t < t_1$, $\tilde{L}_t < \infty$ a.s.)

Therefore,

$$\begin{aligned} \tilde{\psi}(p, B) &= -\ln E[e^{-p\tilde{L}(B)}] \\ &= \int_{B \times \mathbb{R}_0^+} (1 - e^{-pv}) N_0(dt, dv) \end{aligned} \tag{2.4.15}$$

More generally, $\forall\ f : (\mathbb{R}^{+2}, \mathcal{B}^{+2}) \to (\mathbb{R}, \mathcal{B})$ such that $f(t, 0) = 0\ \ \forall\ t \in \mathbb{R}^+$ it can be shown that

$$E[\sum_{t\ \in\ \mathbb{R}^+} f(t, \tilde{L}(t))] = \int\limits_{\mathbb{R}^+ \times \mathbb{R}_0^+} f(t, v) N_0(dt, dv) \tag{2.4.16}$$

and

$$-\ln E\left[e^{-\sum\limits_{t\ \in\ \mathbb{R}^+} f[t, \tilde{L}(t)]}\right] = \int\limits_{\mathbb{R}^+ \times \mathbb{R}_0^+} \{1 - e^{-f(t,v)}\} N_0(dt, dv) \tag{2.4.17}$$

In particular, formally because it may be infinite,

$$\begin{aligned} E[\tilde{L}(B)] &= E[\sum_{t \in B} \tilde{L}(t)] \\ &= \int_{B \times \mathbb{R}_0^+} v N_0(dt, dv) \end{aligned} \tag{2.4.18}$$

We may obtain a more useful expression by using the "predictive" hazard measure which is defined as

$$L^-(B) = \int_B \frac{F(dt)}{S_{t-}} = \int_B \frac{F(dt)}{F([t, \infty])} \tag{2.4.19}$$

In general, the continuous part of L^- is equal to the continuous part of L and the jumps are in bijection (note that the jumps of L^- are in $(0, 1)$), i.e.

$$L_c^- = L_c \tag{2.4.20}$$

and

$$L(t) = -\ln[1 - L^-(t)] \tag{2.4.21}$$

This implies that L^- is also a Levy process and the whole description given for L may be translated for L^-. But, in particular, since

$$\tilde{L}^-(B) = \sum_{t\in B} \tilde{L}^-(t) = \sum_{t\in B} \{1 - e^{-\tilde{L}(t)}\}$$

we have by (2.4.16) and (2.4.18)

$$\begin{aligned} \tilde{L}_0^-(B) &= E[\tilde{L}^-(B)] \\ &= \int_{B\times(0,1)} u N_0^-(dt, du) \\ &= \int_{B\times(0,\infty)} (1 - e^{-v}) N_0(dt, dv) \\ &= \tilde{\psi}(1, B) \end{aligned} \tag{2.4.22}$$

Therefore, $\tilde{L}_0^-$ is a σ-finite measure on $[t_0, t_1)$ and we may decompose N_0^- as

$$N_0^-(dt, du) = \frac{1}{u} \tilde{L}_0^-(dt) R_0^-(t, du) \tag{2.4.23}$$

where R_0^- is a probability transition on $[t_0, t_1) \times (0, 1)$.

Therefore, we obtain successively,

$$\begin{aligned} \tilde{\psi}^-(p, B) &= -\ln E[e^{-p\tilde{L}^-(B)}] \\ &= \int_B r_0^-(p, t) \tilde{L}_0^-(dt) \end{aligned}$$

where

$$r_0^-(p, t) = \int_{(0,1)} \frac{1 - e^{-pu}}{u} R_0^-(t, du) \tag{2.4.24}$$

Note that

$$\frac{d}{dp} r_0^-(p, t) = \int_{(0,1)} e^{-pu} R_0^-(t, du) \tag{2.4.25}$$

is the Laplace transform of the probability transition.

Since $\tilde{L}(B) = \sum_{t\in B} -\ln[1 - \tilde{L}^-(t)]$ using (2.4.17), we obtain

$$\tilde{\psi}(p, B) = -\ln E[e^{-p\tilde{L}(B)}] = \int_B r_0(p, t) \tilde{L}_0^-(dt) \tag{2.4.26}$$

where

$$r_0(p,t) = \int_{(0,1)} \frac{1-(1-u)^p}{u} R_0^-(t,du)$$

Returning to the purely discrete neutral to the right probability measure, we obtain

$$\begin{aligned} S_t &= e^{-L_t} \\ &= \prod_{0\le s\le t} [1-L^-(s)] \end{aligned} \tag{2.4.27}$$

$$\begin{aligned} E[S_t^p] &= \prod_{\substack{0\le s\le t \\ s\in\tilde{I}}} E[e^{-pL(s)}]\, E[e^{-p\tilde{L}_t}] \\ &= \prod_{\substack{0\le s\le t \\ s\in\tilde{I}}} E[\{1-L^-(s)\}^p]\, E[e^{-p\tilde{L}_t}] \end{aligned} \tag{2.4.28}$$

where

$$E[e^{-p\tilde{L}_t}] = e^{-\int_{(t_0,t)} r_0(p,s)\tilde{L}_0^-(ds)} \tag{2.4.29}$$

In particular, for $p=1$, the prior hazard function is given by

$$\begin{aligned} L_{0t} &= -\ln S_{0t} \\ &= \tilde{L}_{0t}^- + \sum_{0\le s\le t} -\ln[1-L_0^-(s)] \end{aligned} \tag{2.4.30}$$

This shows that $\tilde{L}_0^- = L_{0c}^- = L_{0c}$ and that L_0^- is the prior predictive hazard measure i.e.

$$L_0^-(B) = \int_B \frac{F_0(dt)}{F_0([t,\infty])} \tag{2.4.31}$$

Note that $E[L(B)] \neq L_0(B)$ while $E[L^-(B)] = L_0^-(B)$.
Note also that expression (2.4.28) is generally known as the product integral formula.

Hence, in order to specify a prior purely discrete neutral to the right process, we have to specify a prior expectation F_0 or equivalently a prior hazard function L_0 or a prior predictive hazard function L_0^-. This gives t_0, t_1 and I. For each $t \in I$ we have to specify a prior distribution of $L^-(t)$ or of $L(t)$ and we have to specify R_0^-, a probability transition on $[t_0,t_1)\times(0,1)$.

As mentioned before, $F \sim \mathcal{D}i(n_0F_0)$ is neutral to the right and from proposition 2.3.1, it may be deduced that

$$(i) \qquad L^-(t) = \frac{F(t)}{S_{t-}} \sim Be[n_0F_0(t), n_0S_{0t}] \quad \forall\ t \in I$$

$$(ii) \qquad R_0^-(t, du) = n_0S_{0t-}(1-u)^{n_0S_{0t-}-1}du$$

i.e. is a $Be(1, n_0S_{0t-})$ distribution.

Some special neutral to the right processes other than Dirichlet processes have been used by several authors. A very often encountered special case, named homogeneous neutral to the right process is a process with a probability transition independent of t, i.e., $R_0^-(t, du) = R_0^-(du)$. This entails that

$$E[e^{-p\tilde{L}_t}] = e^{-r_0(p)\tilde{L}_{0t}^-} \tag{2.4.32}$$

For instance, Kalbfleisch (1978) specifies L to be a Gamma measure, denoted as $L \sim \mathcal{G}a(c_0, M_0)$, where $c_0 \in \mathbb{R}_0^+$ and M_0 is a σ-finite measure on $(\mathbb{R}^+, \mathcal{B}^+)$, and defined as follows: L is a purely random measure such that

$$L(B) \sim \Gamma(c_0, M_0(B)) \quad \forall\ B \in \mathcal{B}^+, \tag{2.4.33}$$

i.e., is Gamma distributed with scale parameter c_0 and shape parameter $M_0(B)$.

Note that a Gamma measure is homogeneous since

$$\psi(p, B) = \ln(1 + \frac{p}{c_0})M_0(B).$$

In particular $E[L(B)] = \frac{1}{c_0}M_0(B)$ and $L_{0t} = -\ln S_{0t} = \ln(1 + \frac{1}{c_0})M_{0t}$. Hence, $\forall\ t \in I,\ \ L(t) \sim \Gamma[c_0, M_0(t)]$ and $\tilde{L} \sim \mathcal{G}a(c_0, M_{0c})$.

Simple computations show that $N_0(dt, dv) = M_{0,c}(dt)\frac{e^{-c_0v}}{v}dv$ while $R_0^-(du) = \{\ln(1 + \frac{1}{c_0})\ln(\frac{1}{1-u})\}^{-1}u(1-u)^{c_0-1}du$.

Ferguson and Phadia (1979) use a so called simple homogeneous Levy process, i.e., a Levy process without fixed point of discontinuity such that $N_0(dt, dv) = N_0(dt)\frac{1-e^{-pv}}{1-e^{-v}}e^{-\tau v}dv$ where $\tau \in \mathbb{R}_0^+$ and N_0 is a σ-finite diffuse measure on $(\mathbb{R}^+, \mathcal{B}^+)$. Simple computations show that $\tilde{L}_0^- = \frac{1}{\tau}N_0$ and that $R_0^-(du) = \tau(1-u)^{\tau-1}du$, i.e. is a $Be(1, \tau)$ distribution.

Finally, Hjort (1990) generalizes the Dirichlet process by defining L^- to be a Beta process.

Definition 2.4.2. L^- is a Beta process of parameter c_0 a positive Borel function defined on $\mathbb{R}^+$ and $L_0^- = E(L^-)$, to be written as $L^- \sim \mathcal{B}e(c_0, L_0^-)$ if

$$\begin{array}{ll} (i) & L^-(t) \sim Be[c_0(t)L_0^-(t), c_0(t)[1 - L_0^-(t)]] \quad \forall\ t \in I \\ (ii) & R_0^-(t, du) = c_0(t)(1-u)^{c_0(t)-1}du \\ & \text{i.e. is a } Be(1, c_0(t)) \text{ distribution.} \end{array}$$

■

Clearly the Dirichlet process arises from a Beta process for which $c_0(t) = n_0 S_{0t-}$ and simple homogeneous Levy processes correspond to Beta processes with constant c_0 and diffuse expectations.

We delay the description of the posterior distributions because they will be a special case of the posterior distributions under right censoring. The latter will be the subject of the next section.

3 Censored Observations

In most applications of survival analysis, the real lifetimes, T_i, $1 \leq i \leq n$, are not completely observed because they are censored on the right by censoring times C_i, $1 \leq i \leq n$, that are random variables defined on $(\Omega, \mathcal{M})$ with values in $(\mathbb{R}^+, \mathcal{B}^+)$. The available observations are

$$\begin{aligned} Y_i &= T_i \wedge C_i \doteq \min(T_i, C_i) \\ A_i &= \mathbb{1}_{\{T_i \leq C_i\}} \end{aligned} \tag{3.0.1}$$

If $A_i = 0$ (i.e. $T_i > C_i$), Y_i is called a loss and if $A_i = 1$ (i.e. $T_i \leq C_i$), Y_i is called a death. Let us remark that if we allow T_i to be infinite, we suppose that C_i is finite and so is Y_i, $\forall\ 1 \leq i \leq n$.

In this set-up, we will often encounter the following empirical random (sub)-probabilities

$$F_{un} = \frac{1}{n} \sum_{1 \leq i \leq n} A_i \varepsilon_{Y_i} \tag{3.0.2}$$

$$F_{cn} = \frac{1}{n} \sum_{1 \le i \le n} (1 - A_i) \varepsilon_{Y_i}$$
$$F_n = \frac{1}{n} \sum_{1 \le i \le n} \varepsilon_{Y_i} = F_{un} + F_{cn}$$

and we will use, "mutatis mutandis", the same notations as in formula (2.4.2).

Several models are available according to assumptions made on the distribution of the censoring times.

3.1 Conditional Model

Like in regression, the analysis is made conditionally on the censoring times and the assumptions are

$$(i) \qquad \underset{1 \le i \le n}{\perp\!\!\!\perp} T_i \mid F, C_1^n \tag{3.1.1}$$
$$(ii) \qquad T_i \underline{\perp\!\!\!\perp} C_1^n \mid F$$
$$(iii) \qquad F \underline{\perp\!\!\!\perp} C_1^n \tag{3.1.2}$$
$$(iv) \qquad T_i \mid F \sim F$$

Therefore, no assumptions are made on the distribution of the censoring times C_1^n.

In this model, the first results were obtained by Susarla-Van Ryzin (1976). Using a Dirichlet prior, i.e. $F \sim \mathcal{D}i(n_0 F_0)$, they derive the Bayesian estimator of the survival function S_t under a square error lost, i.e., the posterior expectation of S_t. In fact, they made the assumption that the C_i's were independent random variables but the proof goes through without this assumption. The expression they gave is not correct because if they define A_i as we did, in the proof they use $\tilde{A}_i = \mathbb{1}_{\{T_i < C_i\}}$ instead. So, if $\{Z_j : 1 \le j \le M\}$ are the distinct values of the sample $\{Y_i, 1 \le i \le n\}$, they obtained

$$\begin{aligned} \tilde{S}_t &= E[S_t \mid Y_1^n, \tilde{A}_1^n] \qquad (3.1.3) \\ &= S_{n_* t} \cdot \prod_{\{j : Z_j \le t\}} \{1 + \frac{L_j}{n_* S_{n_* Z_{j^-}} - L_j}\} \end{aligned}$$

where, as usual,

$$n_* F_{n_*} = n_0 F_0 + n F_n \tag{3.1.4}$$

and

$$L_j = nF_{cn}(Z_j) \tag{3.1.5}$$

is the number of losses at Z_j.
The correct estimator is given by

$$\begin{aligned} \hat{S}_t &= E[S_t \mid Y_1^n, A_1^n] \\ &= S_{n_* t} \cdot \prod_{\{j: Z_j \leq t\}} \{1 + \frac{L_j}{n_* S_{n_* Z_j}}\} \end{aligned} \tag{3.1.6}$$

Taking a pseudo "non informative" prior distribution, i.e., $n_0 = 0$, this entails

$$\tilde{S}_t = \prod_{\{j: Z_j \leq t\}} \{1 - \frac{D_j}{N_j - L_j}\} \tag{3.1.7}$$

and

$$\hat{S}_t = \prod_{\{j: Z_j \leq t\}} \{1 - \frac{D_j}{N_j}\} \tag{3.1.8}$$

where

$$D_j = nF_{un}(Z_j) \tag{3.1.9}$$

is the number of deaths at Z_j and

$$N_j = nS_{nZ_{j^-}} = \sum_{1 \leq i \leq n} \mathbb{1}_{\{Y_i \geq Z_j\}} \tag{3.1.10}$$

is the number of individuals at risk at time Z_j.

The last estimator (3.1.8) is therefore nothing else than the celebrated Kaplan-Meyer estimator.

They also gave an expression of the posterior expectation of S_t^p where p is an integer showing that the posterior distribution of S_t is a mixture of Beta distributions and they conjectured that the posterior distribution of F was a mixture of Dirichlet measures. This was proved in a subsequent paper (Blum, Susarla (1977)). However, their formula actually proves that the posterior distribution of F is a neutral to the right process. In a subsequent paper, Susarla, Van Ryzin (1978) showed that their estimator converges weakly to a Gaussian process under the assumptions that the C_i's are i.i.d. and that the true distributions of T_1 and C_1 are continuous and strictly smaller than one on the real line.

Ferguson and Phadia (1979) generalize these results as follows. If F is a priori a neutral to the right process then F is a posteriori a neutral to the

right process even if some observations are "exclusively censored" (A_i) and if the other observations are "inclusively censored" ($\tilde{A}_i$). They describe the posterior distribution with a little bit of trouble for the posterior distribution of the jump at a death observation outside the fixed points of discontinuity. This trouble disappears for an homogeneous Levy process and also for the Dirichlet process. They compare the estimators obtained with different special priors: Gamma processes, Dirichlet measures and simple homogeneous Levy processes. Finally, Hjort (1990) proves that Beta processes is a natural conjugate family of prior distributions for right censored observations (smaller that the natural conjugate family of neutral to the right processes since it replaces the specification of a transition probability $R_0^-(t, du)$ by the specification of a function $c_0(t)$). Indeed

Proposition 3.1.1. If $L^- \sim \mathcal{B}e(c_0, L_0^-)$, then $L^- \mid Y_1^n, A_1^n \sim \mathcal{B}e(c_{n_*}, L_{n_*}^-)$ where

$$(i) \qquad c_{n_*}(t) = c_0(t) + nS_{nt-}$$

$$(ii) \qquad L_{n_*}^-(dt) = \frac{c_0(t)L_0^-(dt) + nF_{un}(dt)}{c_{n_*}(t)}$$

■

This shows that $c_0(t)$ may be interpreted as a prior guess of the population at risk at time t and it would be natural to choose $c_0(t)$ of the form $c_0(t) = n_1 S_{1t-}$.

Note also that if the observations are inclusively censored, then proposition 3.1.1 holds with $c_{n_*}(t) = c_0(t) + nS_{unt-} + nS_{cnt}$ and so to include both types of censoring we can choose $c_0(t) = n_1 S_{1t-} + n_2 S_{2t}$.

All the results described in this section are easily deduced from the following description of the posterior distribution of a neutral to the right process (a translation in our set-up of theorem 4 of Ferguson, Phadia (1979)). It will also show that Beta processes are quite natural because if the prior is any neutral to the right process, then the posterior distribution tends to be close to a Beta process.

Let $J = \{Y_i : 1 \leq i \leq n\} = \{Z_j : 1 \leq j \leq M\}$ where Z_j are the distinct observed values. Then, a posteriori, we may write

$$L = \tilde{L} + \sum_{t \in I \cap J^c} L(t)\varepsilon_t + \sum_{1 \leq j \leq M} L(Z_j)\varepsilon_{Z_j}$$

and

$$L^- = \tilde{L}^- + \sum_{t \in I \cap J^c} L^-(t)\varepsilon_t + \sum_{1 \le j \le M} L^-(Z_j)\varepsilon_{Z_j} \tag{3.1.11}$$

Now,

$$(i) \quad -\ln E[e^{-p\tilde{L}_t} \mid Y_1^n, A_1^n] \tag{3.1.12}$$
$$= \int_0^t L_{0c}^-(ds) \int_{(0,1)} \frac{1-(1-u)^p}{u}(1-u)^{nS_{ns-}} R_0^-(s, du)$$

$$(ii) \quad \forall\ t \in I \cap J^c \tag{3.1.13}$$
$$P[L^-(t) \in du \mid Y_1^n, A_1^n] = \frac{(1-u)^{nS_{nt-}} P[L^-(t) \in du]}{\int_{(0,1)} (1-u)^{nS_{nt-}} P[L^-(t) \in du]}$$

$$(iii) \quad \text{If } \ Z_j \in I \tag{3.1.14}$$
$$P[L^-(Z_j) \in du \mid Y_1^n, A_1^n]$$
$$= \frac{u^{D_j}(1-u)^{nS_{nZ_j-}-D_j} P[L^-(Z_j) \in du]}{\int_{(0,1)} u^{D_j}(1-u)^{nS_{nZ_j-}-D_j} P[L^-(Z_j) \in du]}$$

$$(iv) \quad \text{If } \ Z_j \in I^c \ \text{ and } \ D_j = 0 \tag{3.1.15}$$
$$L^-(Z_j) = 0$$

$$(v) \quad \text{If } \ Z_j \in I^c \ \text{ and } \ D_j > 0 \tag{3.1.16}$$
$$P[L^-(Z_j) \in du \mid Y_1^n, A_1^n]$$
$$= \frac{u^{D_j-1}(1-u)^{nS_{nZ_j-}-D_j} R_0^-(Z_j, du)}{\int_{(0,1)} u^{D_j-1}(1-u)^{nS_{nZ_j-}-D_j} R_0^-(Z_j, du)}$$

If A_i is replaced by $\tilde{A}_i$, it suffices to replace $nS_{nZ_j-} - D_j$ by nS_{nZ_j} in formulae (3.1.14) and (3.1.16).

From this, the result of Hjort becomes clear and the posterior expectation of the survival function is given in that case by

$$E[S_t | Y_1^n, A_1^n] \tag{3.1.17}$$
$$= \exp\{-\int_{[0,t]} \frac{c_0(s)}{c_0(s) + nS_{ns-}} L_{0c}(ds)\}$$
$$\times \prod_{\substack{0 \le s \le t \\ s \in I \cap J^c}} \left\{1 - \frac{c_0(s) L_0^-(s)}{c_0(s) + nS_{ns-}}\right\}$$

$$\times \prod_{\{j:Z_j \le t\}} \left\{ 1 - \frac{c_0(Z_j) L_0^-(Z_j) + D_j}{c_0(Z_j) + n S_{n Z_{j-}}} \right\}$$

Finally note that if $c_0(t) = n_1 S_{1t-}$ and if we let n_1 tend to 0, then we obtain the Kaplan-Meyer estimator.

3.2 Berliner-Hill Predictive Distributions under Noninformative Censoring

Under censoring, Berliner-Hill (1988) assumed that the censoring scheme is non informative, i.e., for a censored observation $A_i = 0$ and Y_i the only available information is that $T_i > C_i$. Hence if $\mathcal{L} = \{i : A_i = 0\}$, the "predictive" distribution will be given by

$$P[T_{n+1} \in (T_{m,i}, T_{m,i+1}) \mid \bigcap_{k \in \mathcal{L}} \{T_k > C_k\}] \tag{3.2.1}$$

where $T_{m,i}$, $1 \le i \le m$, are the order statistics of the uncensored observations. Unfortunately, in their very general set-up (3.2.1) is not computable. However, if $C_k \in (T_{m,i_k}, T_{m,i_{k+1}})$, then $T_k > T_{m,i_k}$ and so they approximate (3.2.1) by

$$P[T_{n+1} \in (T_{m,i}, T_{m,i+1}) \mid \bigcap_{k \in \mathcal{L}} \{T_k > T_{m,i_k}\}] \tag{3.2.2}$$

They first give an expression for this probability requiring extensive algebraic and computational work. But another argument leads to a much nicer expression, i.e.,

$$P[T_{n+1} > T_{m,i} \mid \bigcap_{k \in \mathcal{L}} \{T_k > T_{m,i_k}] = \prod_{1 \le j \le i} \left(1 - \frac{1}{1 + N_j} \right) \tag{3.2.3}$$

where N_j is the population at risk at $T_{m,j}$ i.e.

$$N_j = \sum_{1 \le i \le n} \mathbb{1}_{\{Y_i \ge T_{m,j}\}} \tag{3.2.4}$$

Note that in this set-up $D_j = 1$, and hence this "predictive" distribution is very close to the Kaplan-Meyer estimator given by (3.1.7). However, as remarked in section 2.3.3., the probability given by (3.2.3) has nothing to do with the real posterior predictive distribution, i.e., $P[T_{n+1} > T_{m,i} \mid Y_1^n, A_1^n]$.

3.3 Joint Model

In this model, the conditional assumptions (3.1.1) are completed by supposing that the censoring times are an i.i.d. sample of the distribution G or equivalently

$$\begin{array}{ll} (i) & \underset{1\le i\le n}{\perp\!\!\!\perp} (T_i, C_i) \mid K \\ (ii) & T_i \perp\!\!\!\perp C_i \mid K \quad \forall\ 1 \le i \le n \\ (iii) & T_i \mid K \sim F \\ (iv) & C_i \mid K \sim G \end{array} \tag{3.3.1}$$

where K is the distribution of (T_i, C_i), i.e., under (ii) (iii) and (iv) $K = F \otimes G$.

This entails the following assumptions for the observations (Y_i, A_i), $1 \le i \le n$ (see (3.0.1))

$$\begin{array}{ll} (i) & \underset{1\le i\le n}{\perp\!\!\!\perp} (Y_i, A_i) \mid H \\ (ii) & (Y_i, A_i) \mid H \sim H \end{array} \tag{3.3.2}$$

This model is identified because H and K are in bijection provided that $G_t < 1\ \forall\ t \in \mathbb{R}^+$ (see, e.g., Mouchart, Rolin (1995)). Indeed, if

$$\begin{aligned} p &= P[A_i = 1 \mid H] \\ F_u(B) &= P[Y_i \in B \mid H, A_i = 1] \\ F_c(B) &= P[Y_i \in B \mid H, A_i = 0] \end{aligned} \tag{3.3.3}$$

then, the "predictive" hazard measures of F and G are respectively given by

$$L^-(B) \doteq \int_B \frac{F(dt)}{F([t,\infty])} = \int_B \frac{pF_u(dt)}{pF_u([t,\infty]) + (1-p)F_c([t,\infty])} \tag{3.3.4}$$

$$M^-(B) \doteq \int_B \frac{G(dt)}{G([t,\infty])} = \int_B \frac{(1-p)F_c(dt)}{pF_u((t,\infty]) + (1-p)F_c([t,\infty])} \tag{3.3.5}$$

These relations are in close connection with the self-consistency properties as introduced and studied by Tsai (1986) and Lo (1993). Their results and some extensions may be summarized as follows:

If we specify for H a natural conjugate prior, i.e., $H \sim \mathcal{D}i(n_0 H_0)$, then

$$\begin{array}{ll} (i) & p \underline{\parallel} F_u \underline{\parallel} F_c \\ (ii) & p \sim Be(n_0 p_0, n_0(1-p_0)) \\ (iii) & F_u \sim \mathcal{D}i(n_0 p_0 F_{0u}) \\ (iv) & F_c \sim \mathcal{D}i(n_0(1-p_0)F_{0c}) \end{array} \tag{3.3.6}$$

These are usual properties of Dirichlet measures we have already met. However, the following results are much more striking. Indeed,

$$\begin{array}{ll} (i) & L^- \underline{\parallel} M^- \quad \text{or equivalently} \\ & F \underline{\parallel} G \\ (ii) & L^- \sim \mathcal{B}e(d_0, L_0^-) \\ (iii) & M^- \sim \mathcal{B}e(c_0, M_0^-) \end{array} \tag{3.3.7}$$

where

$$\begin{aligned} d_0(t) &= n_0 p_0 F_{0u}([t,\infty]) + n_0(1-p_0)F_{0c}([t,\infty]) \\ &= n_0 P(Y_i \geq t \mid H_0) \end{aligned} \tag{3.3.8}$$

and

$$\begin{aligned} c_0(t) &= n_0 p_0 F_{0u}((t,\infty]) + n_0(1-p_0)F_{0c}([t,\infty]) \\ &= n_0 P[Y_i > t, A_i = 1 \mid H_0] + n_0 P[Y_i \geq t, A_i = 0 \mid H_0] \end{aligned} \tag{3.3.9}$$

Note that c_0 and d_0 have the form suggested by the results of Hjort (1990) in case of both types of censoring.

All these results hold a posteriori, since

$$H \mid Y_1^n, A_1^n \sim \mathcal{D}i(n_* H_{n_*}) \tag{3.3.10}$$

with

$$n_* H_{n_*} = n_0 H_0 + n H_n \tag{3.3.11}$$

and

$$H_n = \frac{1}{n} \sum_{1 \leq i \leq n} \varepsilon_{(Y_i, A_i)} \tag{3.3.12}$$

This allows us to rediscover the posterior distribution obtained by Hjort (1990) but also to obtain much more useful properties. Thanks to (3.3.6), p, F_u and F_c are easy to simulate and have a posteriori asymptotic normal

approximations of the distributions of their functionals. Using (3.3.4), we may obtain

$$F((t,\infty]) = \prod_{0\le s\le t} \frac{pF_u((s,\infty]) + (1-p)F_c([s,\infty])}{pF_u([s,\infty]) + (1-p)F_c([s,\infty])} \tag{3.3.13}$$

and

$$G((t,\infty]) = \prod_{0\le s\le t} \frac{pF_u((s,\infty]) + (1-p)F_c((s,\infty])}{pF_u((s,\infty]) + (1-p)F_c([s,\infty])} \tag{3.3.14}$$

Therefore F and G are also easy to simulate and it is possible to deduce asymptotic normal approximations for the posterior distributions of their functionals.

Finally, if $p_0 = 1$, then F is a priori a Dirichlet measure, i.e., $F \sim \mathcal{D}i(n_0 F_{0u})$, and this gives a way to simulate the posterior distribution of F the expectation of which is the Susarla-Van Ryzin estimator.

3.4 Dependent Censoring

The independence between the real life-times and the censoring times (assumption (ii) of (3.3.1)) is subject to criticism in many applications. However, without this assumption the joint model is not identified in the sense that several choices of K generate the same H which is the maximal exactly estimable parameter, i.e., for an infinite sample, H and nothing else is known.

Two major attempts have been made to cope with this problem.

At first Tsai (1986) introduced identifying restrictions he called generalized self-consistent property that may be written as

$$P(T_i > t \mid C_i = u) = \frac{P(T_i > t)}{P(T_i > u)} P(T_i > u \mid C_i = u) \tag{3.4.1}$$

a.s. G on $[0,t)$ $\forall\ t \in \mathbb{R}^+$ and he showed that under this restriction, (3.3.4) and (3.3.13) hold. Note that (3.4.1) is trivially satisfied under assumption (ii) of (3.3.1).

At last, Phadia, Susarla (1983) specified K to be a Dirichlet measure, i.e. $K \sim \mathcal{D}i(n_0 K_0)$ and gave a formula for the posterior expectation of K. However, their work is a little bit disappointing because it may be schematically described as follows. Clearly H is a function of K and therefore $H \sim \mathcal{D}i(n_0 H_0)$. Now it is possible to "complete" H by R, another function of K, such that K is in bijection with (H,R) and such that a priori

$H \underline{\parallel} R$. Therefore, as in (3.3.10), we will have $H \mid Y_1^n, A_1^n \sim \mathcal{D}i(n_* H_{n_*})$ but R will not be revised by the sample and will keep its prior specification i.e. $R \mid Y_1^n, A_1^n \sim R$.

4 Concluding Remarks

This presentation of nonparametric Bayesian survival analysis is far to be exhaustive and represents a very personal view on this subject. The author surely missed a lot of important contributions to this field by ignorance or by lack of understanding and he apologizes for these omissions.

References

[1] Antoniak C., Mixtures of Dirichlet process with applications to Bayesian nonparametric problems, *Annals of Statistics* 2, 1152-1174, (1974).

[2] Berliner L.M., B.M. Hill, Bayesian nonparametric survival analysis, *Journal of the American Statistical Association* 83, 772-779, (1988).

[3] Blackwell D., Discreteness of Ferguson selection, *Annals of Statistics* 1, 356-358, (1973).

[4] Blackwell D., J.B. MacQueen, Ferguson distributions via Polya urn schemes, *Annals of Statistics* 1, 353-355, (1973).

[5] Blum K., V. Susarla, On the posterior distribution of a Dirichlet process given randomly right censored data, *Stochastic Process. Appl.* 5, 207-211, (1977).

[6] Breiman L., *Probability*, Addison-Wesley Publishing Company Inc., Reading, Menlo Park and London, (1974).

[7] Cifarelli D.M., E. Regazzini,Distributions functions of means of a Dirichlet process, *Annals of Statistics* 18, 429-442, (1990).

[8] Daley D.J., D. Vere-Jones, *An introduction to the Theory of Point Processes*, Springer-Verlag, New York (1988).

[9] Diaconis P., J.B.K. Kemperman, Some new tools for Dirichlet priors, in *Bayesian Statistics V: Fifth Valencia International Meeting on Bayesian Statistics*, Oxford Sciences Publications, 589-597, (1995).

[10] Doksum K., Tailfree and neutral random probabilities and their posterior distributions, *Annals of Probability* 2, 183-201, (1974).

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

[12] Ferguson T.S., Prior distributions on spaces of probability measures, *Annals of Statistics* 2, 615-629, (1974).

[13] Ferguson T.S., M.J. Klass, A representation of independent increment processes without Gaussian components, *Annals of Mathematical Statistics* 43, 1634-1643, (1972).

[14] Ferguson T.S., E. Phadia, Bayesian nonparametric estimation based on censored data, *Annals of Statistics* 7, 163-186, (1979).

[15] Florens J.P., M. Mouchart, J.M. Rolin, *Elements of Bayesian Statistics*. Marcel Dekker, New York, (1990).

[16] Florens J.P., J.M. Rolin, Bayes, Bootstrap, Moments, Discussion Paper 9413, Institut de Statistique, Université catholique de Louvain, Belgium, (1994).

[17] Hannum R.C., M. Hollander, N.A. Langberg, Distributional results for random functionals of a Dirichlet process, *Annals of Probability* 7, 665-670, (1981).

[18] Hjort N.L., Nonparametric Bayes estimators based on beta process in models for life history data, *Annals of Statistics* 18, 1259-1294, (1990).

[19] Kalbfleisch J.D., Nonparametric Bayesian analysis of survival time data, *Journal of the Royal Statistical Society* B, 40, 214-221, (1978).

[20] Karr A.F., *Point Processes and their Statistical Inference*, Marcel Dekker, New York, (1986).

[21] Korwar R.M., M. Hollander, Contributions to the theory of Dirichlet processes, *The Annals of Probability* 1, 705-711, (1973).

[22] Lo A.Y., Weak convergence for Dirichlet processes, *Sankhya*, Ser. A, 45, 105-111, (1983).

[23] Lo A.Y., A remark on the limiting posterior distribution of the multiparameter Dirichlet process, *Sankhya*, Ser. A, 48, 247-249, (1986).

[24] Lo A.Y., A large sample study for the Bayesian bootstrap, *Annals of Statistics* 15, 360-375, (1987).

[25] Lo A.Y., A Bayesian bootstrap for censored data, *Annals of Statistics* 21, 100-123, (1993).

[26] Mouchart M., J.M. Rolin, Competing risks models: problems of modelling and of identification, Discussion Paper 9507, Institut de Statistique, Université catholique de Louvain, Belgium, (1995).

[27] Muliere P., L. Tardella, Approximating distributions of random functionals of Ferguson-Dirichlet priors, technical report, Duke University, (1995).

[28] Phadia E.G., V. Susarla, Nonparametric Bayesian estimation of a survival curve with dependent censoring mechanism, *Ann. Inst. Statist. Math.*, 35, 389-400, (1983).

[29] Rolin J.M., Nonparametric Bayesian Statistic: a stochastic process approach, in *Specifying Statistical Models, from Parametric to Nonparametric using Bayesian or Non-Bayesian Approaches*. Lecture Notes in Statistics, vol 16, Springer, Berlin, 108-133, (1983).

[30] Rolin J.M., Some useful properties of the Dirichlet process, Discussion Paper 9202, Institut de Statistique, Université catholique de Louvain, Belgium, (1992).

[31] Rolin J.M., On the distribution of jumps of the Dirichlet process, Discussion Paper 9302, Institut de Statistique, Université catholique de Louvain, Belgium, (1993).

[32] Sethuraman J., A constructive definition of the Dirichlet prior, *Statistica Sinica* 2, 639-650, (1994).

[33] Susarla V., J. Van Ryzin, Nonparametric Bayesian estimation of survival curves from incomplete observations, *Journal of the American Statistical Association* 71, 897-902, (1976).

[34] Susarla V., J. Van Ryzin, Empirical Bayes estimation of a distribution (survival) function from right censored observations, *Annals of Statistics* 6, 740-754, (1978a).

[35] Susarla V., J. Van Ryzin, Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples, *Annals of Statistics* 6, 755-768, (1978b).

[36] Tsai W.Y., Estimation of survival curves from dependent censorship models via a generalized self-consistent property with nonparametric Bayesian estimation application, *Annals of Statistics* 14, 238-249, (1986).

[37] Wilks S.S, *Mathematical Statistics*, Wiley, New York, (1962).

[38] Yamato H., Properties of samples from distributions chosen from a Dirichlet process, *Bulletin of Informatics and Cybernetics* 21, 77-83, (1984a).

[39] Yamato H., Characteristic functions of means of distributions chosen from a Dirichlet process, *Annals of Probability* 12, 262-267, (1984b).