13

Survival Analysis

In survival analysis the data are times of occurrences of events, with the special feature that these times may be only partially observable. In this chapter we consider Bayesian nonparametric methods for survival data, allowing for right censoring of the survival times. After a general introduction, we consider the Dirichlet process prior for the survival distribution and the beta process prior for the cumulative hazard function, and study their properties. Next we introduce general "neutral to the right" priors on the survival distribution, or equivalently independent increment process priors for the cumulative hazard function, and derive their conjugacy, their asymptotic consistency, and Bernstein–von Mises theorems. We obtain similar results for the Cox proportional hazard model. We also discuss priors for the hazard function, and the Bayesian bootstrap for censored data.

13.1 Introduction

Survival analysis is concerned with inference on the distribution of times of occurrence of events of interest. These are referred to as *survival times*, as in most applications the event is associated with failure of an object, such as the death of a subject, the onset of a disease, or the end of functioning of equipment. Survival times are generally nonnegative, whence the analysis concerns inference on distributions on the positive half line $(0, \infty)$.

The distribution of a random variable X on $(0, \infty)$ is given by its cumulative distribution function F, but in survival analysis it is customary to make inference on the *survival function* $\bar{F}$, defined as $\bar{F}(x) = 1 - F(x)$, for $x \geq 0$. The main complicating factor is that the survival time X may not be directly observable, but be subject to censoring. Censoring blocks the actual value of the observation, and only reveals partial information in the form of a comparison with a *censoring variable* C. Various types of censoring are possible.

(i) *Right censoring*: X is observed if $X \leq C$; otherwise C is observed and $X \geq C$ is noted. In other words, we observe the pair $T = \min(X, C)$ and $\Delta = \mathbb{1}\{X \leq C\}$.

(ii) *Left censoring*: the pair $T = \max(X, C)$ and $\Delta = \mathbb{1}\{X \geq C\}$ is observed.

(iii) *Interval censoring, type I*: given two censoring variables $C_1 \leq C_2$, there are three possibilities: X is fully observed if $C_1 \leq X \leq C_2$; C_1 is observed and the fact that $X < C_1$; or C_2 is observed and the fact that $X \geq C_2$. In other words, we observe $C_1, C_2, \mathbb{1}\{X < C_1\}, X\mathbb{1}\{C_1 \leq X \leq C_2\}, \mathbb{1}\{X > C_2\}$.

(iv) *Interval censoring, type II*: two censoring variables $C_1 \leq C_2$, and the indicators $\mathbb{1}\{X < C_1\}, \mathbb{1}\{C_1 \leq X \leq C_2\}, \mathbb{1}\{X > C_2\}$ are observed.

(v) *Current status censoring*: a censoring variable C, and only the "current status" of the observation relative to C, i.e. whether $X \le C$ or not are observed.

In addition, the observation could be subject to *truncation*: not at all observed if the observation fails to cross a certain threshold. Whereas a censored observation gives partial information, a truncated observation is completely lost. In the latter case the distribution of the observed data may not be the quantity of interest: the sample is biased towards larger values.

Instead of modeling survival times through their survival function $\bar{F}$ it is often insightful to use the *cumulative hazard function*

$$H(t) = \int_{(0,t]} \frac{dF}{1 - F-}. \tag{13.1}$$

This is a nondecreasing right-continuous function, and hence can be thought of as a Lebesgue-Stieltjes measure. It is finite on finite intervals strictly within the support of F, but its total mass may exceed 1. In fact its total mass necessarily increases to infinity unless F has an atom at the right end of its support.[1] We shall use F and H interchangeably as functions and measures. If F possesses a density f relative to the Lebesgue measure, then H is absolutely continuous also, with density $h(t) = f(t)/(1 - F(t))$. This is referred to as the *hazard function* of X, and can be interpreted as the "instantaneous rate of failure at time t given survival up to time t," since

$$h(t) = \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathrm{P}(t \le X \le t + \delta |\, X \ge t) = \frac{f(t)}{1 - F(t)}. \tag{13.2}$$

Survival and cumulative hazard functions are in one-to-one correspondence. The survival function can be recovered uniquely from the cumulative hazard function through the relation

$$1 - F(t) = \prod_{(0,t]} (1 - dH) = e^{-H^c(t)} \prod_{u \in (0,t]} (1 - \Delta H(u)). \tag{13.3}$$

Here the expression in the middle is formal notation for the *product integral*, for which the expression on the far right is one definition, where H^c is the continuous part of H and $\Delta H(u) = H(u) - H(u-)$ the jump of H at u. An alternative definition of the product integral is

$$\prod_{(s,t]} (1 - dH) = \lim_{n \to \infty} \prod_{i=1}^{n} (1 - (H(u_i) - H(u_{i-1}))), \tag{13.4}$$

where the limit is taken over any sequence of meshes $s = u_0 < u_1 < \cdots < u_n = t$ with maximum mesh width tending to zero. If F (or equivalently H) is continuous, then $H = H^c$ and (13.3) reduces to the simpler relation $\bar{F} = e^{-H}$. It will be useful to define for general F the function $A = -\log \bar{F}$, so that $\bar{F} = e^{-A}$ always; and $A = H$ if F has no atoms.

In this chapter we mostly discuss *random right censoring*: right censoring by observation times C that are independent of the event times. In the i.i.d. model one has an independent random sample of survival times $X_i \overset{\text{iid}}{\sim} F$, independent censoring times $C_i \overset{\text{iid}}{\sim} G$, for $i = 1, \ldots, n$, and observes the n pairs of observations $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, where

[1] Proofs and further background to the results in this section can be found for instance in Andersen et al. (1993).

$T_i = \min(X_i, C_i)$ and $\Delta_i = \mathbb{1}\{X_i \le C_i\}$. We denote the data by D_n. The parameters in this model are identifiable, in that the relationship between the pair (F, G) and the distribution of the pair (T, Δ) is one-to-one if restricted to the interval where $(1-F)(1-G) > 0$.[2] The observations can be summarized through the two *counting processes* $N = (N(t)\colon t \ge 0)$ and $(Y(t)\colon t \ge 0)$ defined by

$$N(t) = \sum_{i=1}^{n} \mathbb{1}\{T_i \le t\}\Delta_i = \sum_{i=1}^{n} \mathbb{1}\{X_i \le t, X_i \le C_i\}, \tag{13.5}$$

$$Y(t) = \sum_{i=1}^{n} \mathbb{1}\{T_i \ge t\} = \sum_{i=1}^{n} \mathbb{1}\{X_i \ge t, C_i \ge t\}. \tag{13.6}$$

The processes N and Y give the numbers of observed failures and of subjects still alive (or *at risk*) at time t, respectively. It can be shown that the process $N - \int_0^{\cdot} Y\,dH$ is a martingale (the "predictable" process $\int_0^{\cdot} Y\,dH$ is the "compensator" of the counting process N), from which it can be concluded that the process $\int_0^{\cdot} \mathbb{1}\{Y > 0\}Y^{-1}\,dN - H$ is a (local) martingale as well. Since it is zero at zero, it has zero expectation, provided this exists. This may motivate the following estimator for the cumulative hazard function, known as the *Nelson-Aalen estimator*,

$$\hat{H}(t) = \int_{(0,t]} \mathbb{1}\{Y > 0\}\,\frac{dN}{Y}. \tag{13.7}$$

The classical estimator of the survival function $\bar{F}$, known as the *Kaplan-Meier estimator*, is the survival function corresponding to $\hat{H}$:

$$1 - \hat{F}(t) = \prod_{(0,t]} \Big(1 - \mathbb{1}\{Y > 0\}\frac{dN}{Y}\Big). \tag{13.8}$$

This estimator can also be derived as the *nonparametric maximum likelihood estimator*, defined as the maximizer of the *empirical likelihood*

$$F \mapsto \prod_{i=1}^{n} F\{X_i\}^{\Delta_i}(1 - F(C_i))^{1-\Delta_i}.$$

Alternatively, it can be motivated by a factorization of the survival probability $\bar{F}(t)$ as the product $\prod_i \mathrm{P}(X > t_i \,|\, X > t_{i-1})$ of conditional survival probabilities over a grid of time points $0 = t_0 < t_1 < \cdots < t_k = t$. A natural estimator of $\mathrm{P}(X > t_i \,|\, X > t_{i-1})$ is one minus the ratio of the number of deaths $N(t_{i-1}, t_i]$ in the interval $(t_{i-1}, t_i]$ and the number at risk $Y(t_{i-1})$ at the beginning of the interval. For a fine enough partition each interval will contain at most one distinct time of death (which may shared by several individuals), and each interval without a death contributes a factor 1 to the product. In the limit of finer and finer partitions the estimator thus takes the form

$$1 - \hat{F}(t) = \prod_{j:T_j^* \le t} \Big(1 - \frac{\sum_{i:X_i = T_j^*} \Delta_i}{\sum_{i:X_i \ge T_j^*} 1}\Big), \tag{13.9}$$

[2] See e.g. Lemma 25.74 in van der Vaart (1998).

where $T_1^* < T_2^* < \cdots < T_k^*$ are the distinct values of the uncensored observations. This is identical to (13.8), and for this reason the Kaplan-Meier estimator is also known as the *product limit estimator*. If the largest observation is censored, then the nonparametric maximum likelihood estimator is non-unique, as any placement of the mass beyond the last observation will give the same likelihood. The Kaplan-Meier estimator (13.8) will then be an improper distribution, in that $\hat{F}(t) = \hat{F}(t_{(n)}) < 1$ for $t > T_{(n)}$.

13.2 Dirichlet Process Prior

The random right censoring model has two parameters: the survival distribution F and the censoring distribution G. By the independence of the survival and censoring times, a prior on G has no role in the posterior computation for F if G is chosen a priori independent of F. The most obvious prior on the survival distribution F is the Dirichlet process $\mathrm{DP}(\alpha)$.

By Theorem 4.6 the posterior distribution of F given the complete set of survival times $X_1, \ldots, X_n$ is $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$. Adding the censoring times to the data does not change the posterior. Hence by the towering rule of conditioning, the posterior distribution of F given the observed data D_n is the $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$ distribution "conditioned on D_n." This is a mixture of Dirichlet processes with the mixing distribution equal to the conditional distribution of $X_1, \ldots, X_n$ given D_n, in the model $F \sim \mathrm{DP}(\alpha)$ and $X_1 \ldots, X_n | F \overset{\text{iid}}{\sim} F$.

In the latter setup the variables $X_1, \ldots, X_n$ are a "sample from the Dirichlet process" and their distribution is described through their predictive distributions in Section 4.1.4. The dependence between $X_1, \ldots, X_n$ makes the conditioning nontrivial. Denoting the distribution of $(X_1, \ldots, X_n)$ by m, and the conditional distributions of groups of variables within m by $m(\cdot|\,\cdot)$, we can write the posterior distribution of F given D_n as

$$\int \cdots \int \prod_{i:\Delta_i=0} \mathbb{1}_{x_i \geq C_i} \, \mathrm{DP}\Big(\alpha + \sum_{i=1}^n \delta_{X_i} \Delta_i + \sum_{i=1}^n \delta_{x_i}(1-\Delta_i)\Big) \\ dm((x_i\colon \Delta_i = 0)|\,(X_i\colon \Delta_i = 1)).$$

However, this representation is cumbersome, as m itself is a mixture of many components. It turns out that there is a tractable representation in terms of a Pólya tree processes.

We can view the censoring times $C_1, \ldots, C_n$ as constants. Fix any sequence $\mathcal{T}_m = \{A_\varepsilon\colon \varepsilon \in \mathcal{E}^m\}$ of successive, binary, measurable partitions of the sample space $\mathbb{R}^+$, as in Section 3.5, that includes the sets (C_i, ∞) for all observed censoring times C_i and generates the Borel sets. (Such a tree can be constructed by including the ordered times C_i as splitting points; see the proof below for an example.) The Dirichlet process prior is a Pólya tree process for this sequence of partitions, as it is for *any* sequence of partitions (see Section 4.1.2). The following theorem shows that it remains a Pólya tree process with respect to this special type of tree after updating it with censored data.

Theorem 13.1 *If D_n are random right censored data resulting from a sample of survival times $X_1, \ldots, X_n | F \overset{iid}{\sim} F$ with $F \sim \mathrm{DP}(\alpha)$, then the posterior distribution $F|\, D_n$ follows a Pólya tree prior* $\mathrm{PT}\,(\mathcal{T}_m, \alpha(A_\varepsilon) + N(A_\varepsilon) + M(A_\varepsilon))$ *on any sequence of nested partitions that includes the sets (C_i, ∞) for the observed censoring times C_i, where*

$$N(A)=\sum_{i=1}^n \Delta_i \mathbb{1}\{T_i\in A\},\qquad M(A)=\sum_{i=1}^n (1-\Delta_i)\mathbb{1}\{(T_i,\infty)\subset A\}.$$

In particular, for any partition $0=t_0<t_1<\cdot<t_k=t$ *that contains the observed censoring times* $\{C_i\colon \Delta_i=0\}$ *smaller than* t*, with* $\bar M(t)=\sum_{i=1}^n(1-\Delta_i)\mathbb{1}\{C_i\ge t\}$*,*

$$\mathrm{E}\big[1-F(t)|\,D_n\big]=\prod_{i=1}^k\Big(1-\frac{\alpha(t_{i-1},t_i]+N(t_{i-1},t_i]}{\alpha(t_{i-1},\infty)+N(t_{i-1},\infty)+\bar M(t_i)}\Big).$$

Proof The data consists of full observations X_i, for i such that $\Delta_i=1$, and censored observations, for which it is only known that $X_i\in(C_i,\infty)$. Updating the prior with only the full observations gives a Dirichlet posterior process $\mathrm{DP}(\alpha+\sum_{i=1}^n\Delta_i\delta_{X_i})$, by Theorem 4.6. On the given partition sequence $\{\mathcal{T}_m\}$ this can be represented as a Pólya tree process $\mathrm{PT}(\mathcal{T}_m,\alpha'_\varepsilon)$ with parameters $\alpha'_\varepsilon=\alpha_\varepsilon+N(A_\varepsilon)$. Each censored observation contributes information that is equivalent to a Bernoulli experiment with possible outcomes that the survival time is in (C_i,∞) or not, with probabilities $\bar F(C_i)$ and $F(C_i)$, and that is realized in giving the first of these two outcomes. Thus this observation contributes the term $\bar F(C_i)$ to the likelihood. By construction the set (C_i,∞) is contained in the partitioning tree, whence $\bar F(C_i)$ can be written as the product of the splitting variables along the path in the partitioning tree corresponding to the set. The prior distribution of the splitting variables is given by a product of beta densities. Multiplying this product by the likelihood increases the powers in the beta densities by 1 if they refer to a set A_ε that contains (C_i,∞) (so that A_ε is in its path) and leaves the powers unchanged otherwise. In other words, the posterior is again a Pólya tree process with parameters α'_ε or $\alpha'_\varepsilon+1$ depending on whether A_ε contains the set (C_i,∞) or not. Repeating this procedure for all censored observations, we obtain the first assertion of the theorem.

To derive the formula for the posterior mean, we apply the preceding with the partition that uses the numbers $t_1,t_2,\ldots,t_k$ as the first k splitting points for the right most branch of the splitting tree: we first split at t_1 giving $A_0=(-\infty,t_1]$ and $A_1=(t_1,\infty)$, next split A_0 arbitrarily and A_1 into $A_{10}=(t_1,t_2]$ and $A_{11}=(t_2,\infty)$, next split A_{00}, A_{01}, A_{10} arbitrarily and A_{11} at t_3, etc., thus ensuring for every i that $A_{11\cdots1}=(t_i,\infty)$ for the string $11\cdots1\in\{0,1\}^i$. We continue in this manner until $t_k=t$, and let all subsequent splits be arbitrary, except that they must include the remaining censoring times in order to meet the condition in the first part of the theorem. By their definition the splitting variables V_0, V_{10}, $V_{110},\ldots$ give the conditional probabilities $F(A_0)$, $F(A_{10}|\,A_1)$, $F(A_{110}|\,A_{11}),\ldots$, whence for this special partition

$$1-F(t)=F(A_{11\cdots1})=(1-V_0)(1-V_{10})\times\cdots\times(1-V_{11\cdots10}),$$

where V_0 is attached to A_0, V_{10} to $(t_1,t_2]$, etc., and the last variable $V_{11\cdots10}$ in the product is attached to $(t_{k-1},t_k]$. Under the (Pólya tree) posterior distribution these variables are independent and possess beta distributions. By the first part of the proof the beta distribution of $V_{11\cdots10}$ with $i-1$ symbols 1 has parameters $\alpha(t_{i-1},t_i]+N(t_{i-1},t_i]$ and $\alpha(t_i,\infty)+N(t_i,\infty)+M(t_i,\infty)$. Here $M(t_i,\infty)=\bar M(t_i)$, since (t_i,∞) contains (C_j,∞) if and only if $C_j\ge t_i$. Finally we replace every term in the product representing $\bar F(t)$ by its expectation. □

The variable $M(A)$ is zero on every set A that is bounded to the right, and counts the number of observed censoring times C_i with $C_i > c$ for a partitioning set of the form $A_{11\cdots1} = (c, \infty)$. In the latter case the survival time X_i is known to be to the right of C_i. Hence $M(A)$ counts the number of censored survival times that are certain to belong to A, and the sum $N(A) + M(A)$ counts all survival times that are certain to belong to A. This is intuitively plausible: survival times are counted in a partitioning set A_ε if and only if they are certain to have fallen in this set.

The formula for the posterior mean may be applied with different partitions, to produce seemingly different representations. If the grid contains all observed values $T_j < t$ and t, then $N(t_{i-1}, t_i] = \Delta N(t_i)$, for N the counting process defined in (13.5). Since in this case also $N(t_{i-1}, \infty) + \bar{M}(t_i) = Y(t_i)$ for the at risk process Y defined in (13.6), the formula reduces to

$$\mathrm{E}[1 - F(t) | D_n] = \prod_{i=1}^{k} \Big(1 - \frac{\alpha(t_{i-1}, t_i] + \Delta N(t_i)}{\alpha(t_{i-1}, \infty) + Y(t_i)}\Big).$$

For $|\alpha| \to 0$ this reduces to the Kaplan-Meier estimator. The minimal permissible grid for the formula for the posterior mean in Theorem 13.1 consists of all observed censoring times (C_i with $\Delta_i = 0$) smaller than t augmented with t. The formula then reduces to a product "over the censored observations." This appears not to be especially useful, but is somewhat unexpected as the Kaplan-Meier estimator is a product over the *uncensored* observations.

The processes N and Y both tend to infinity at the order n, on any interval within the supports of the survival and censoring times. Therefore the difference between the posterior mean and the Kaplan-Meier estimator should be asymptotically negligible, resulting in the consistency and asymptotic normality of the posterior mean. The following result follows from a general theorem in Section 13.4.1, but is included here with a sketch of a direct proof.

Theorem 13.2 *As $n \to \infty$ almost surely $[P^\infty_{F_0,G}]$,*

(i) $\mathrm{E}[F(t) | D_n] \to F_0(t)$,
(ii) $var[F(t) | D_n] \to 0$.

In particular, the posterior distribution of $\bar{F}(t)$ is consistent.

Proof For A_n the discrete measure with $A_n\{t_i\} = \alpha(t_{i-1}, t_i]$, we can write the posterior mean of the survival distribution as

$$\mathrm{E}[1 - F(t) | D_n] = \prod_{i=1}^{k} \Big(1 - \frac{\Delta(A_n + N)(t_i)}{A_n(t_i -) + Y(t_i)}\Big).$$

By the uniform law of large numbers we have that $n^{-1}(A_n + N)$ tends almost surely to $n^{-1}\mathrm{E}[N] = \int_{(0,\cdot]} \bar{G}_- \, dF$, and that $n^{-1}(A_{n-} + Y)$ tends almost surely to $\mathrm{E}[n^{-1}Y] = \bar{F}_-\bar{G}_-$, where for a monotone function H, $H_-(t) = H(t-)$. The continuity of the maps $(U, V) \mapsto \Lambda = \int_{(0,\cdot]} V^{-1} \, dU \mapsto \prod_{(0,\cdot]}(1 - d\Lambda)$ on a domain where the function V is bounded away from zero now gives result (i).

To prove (ii) it now suffices to show that $\mathrm{E}[(1-F(t))^2 |\, D_n] \to \bar{F}_0(t)^2$, almost surely. This may be proved by calculation of second moments of product of independent beta variables. We omit the details. □

13.3 Beta Process Prior

The beta process appears to be the canonical prior for the cumulative hazard function, just as the Dirichlet process is for the distribution function. The beta process is an example of an independent increment process, which are discussed in the next section in general. In this section we derive the beta process as a limit of beta priors on the hazard function in a discrete time setup, where definition and computation are elementary.

13.3.1 Discrete Time

For a given grid width $b > 0$ consider survival and censoring variables taking values in the set of points $0, b, 2b, \ldots$ Let $f(jb) = \mathrm{P}(X = jb)$ be the probability of failure at time jb, and let $F(jb) = \mathrm{P}(X \le jb) = \sum_{l=0}^{j} f(lb)$. Define a corresponding *discrete hazard function* and cumulative hazard function by, for $j = 0, 1, \ldots,$

$$h(jb) = \mathrm{P}(X = jb |\, X \ge jb) = \frac{f(jb)}{\bar{F}(jb-)}, \qquad H(jb) = \sum_{l=0}^{j} h(lb).$$

The discrete hazard function takes its values in $[0, 1]$, and $h(jb) < 1$ except at the final atom of f, if there is one. The functions f and F can be recovered from h (or H) by

$$1 - F(jb) = \prod_{l=0}^{j} (1 - h(lb)), \qquad f(jb) = \Bigl[\prod_{l=0}^{j-1} (1 - h(lb))\Bigr] h(jb). \tag{13.10}$$

Suppose that we observe survival times $X_1, \ldots, X_n$ subject to random right censoring at the points $C_1, \ldots, C_n$, with all possible values belonging to the lattice with span b. For inference on f the censoring times may be considered fixed. A typical censored observation $(X_i \wedge C_i, \mathbb{1}\{X_i \le C_i\})$ then falls in the sample space consisting of the point $(C_i, 0)$ and the points $(jb, 1)$ for $j = 0, 1, \ldots, C_i$, and the likelihoods of these points are $1 - F(C_i)$ and $f(jb)$, respectively. The total likelihood is the product over the likelihoods for the individual observations. In terms of the counting process N and the at-risk process Y, defined in (13.5) and (13.6), this likelihood can be written as[3]

$$\prod_{j=0}^{\infty} (1 - h(jb))^{Y(jb) - \Delta N(jb)} h(jb)^{\Delta N(jb)}. \tag{13.11}$$

[3] It suffices to show this for $n = 1$. For a censored observation $Y(jb) = 1$ or 0 as $jb \le C$ or $jb > C$, respectively, while $\Delta N \equiv 0$, reducing the display to $\prod_{j:jb\le C}(1 - h(jb)) = \bar{F}(X)$, by (13.10). For an uncensored observation $Y(jb) = 1$ or 0 as $jb \le X$ or $jb > X$, while $\Delta N(X) = 1$ and $\Delta N(jb) = 0$ otherwise, reducing the display to $\prod_{j:jb<X}(1 - h(jb))h(X) = f(X)$, by (13.10).

This also has an intuitive interpretation in terms of binomial experiments: at each time jb we record the number $\Delta N(jb)$ of successes in $Y(jb)$ independent experiments with success probability $h(jb)$. Conjugacy of the binomial and beta distributions suggests to put independent beta priors on the parameters $h(jb)$: for given constants $c_{j,b} > 0$ and $0 < \lambda_{j,b} < 1$,

$$h(jb) \overset{\text{ind}}{\sim} \text{Be}\Big(c_{j,b}\lambda_{j,b}, c_{j,b}(1-\lambda_{j,b})\Big). \tag{13.12}$$

The parameters $\lambda_{j,b}$ can be interpreted as prior guesses, since $\text{E}[h(jb)] = \lambda_{j,b}$, while the parameters $c_{j,b}$ control the prior variability of the discrete hazard, since $\text{var}[h(jb)] = \lambda_{j,b}(1-\lambda_{j,b})/(1+c_{j,b})$. It is attractive to generate the hyper parameters by a single function $c\colon [0,\infty) \to \mathbb{R}^+$ and cumulative hazard function Λ on $[0,\infty)$ through

$$c_{j,b} = c(jb), \qquad \lambda_{j,b} = \Lambda(jb) - \Lambda((j-1)b). \tag{13.13}$$

The induced prior on the cumulative hazard function H is then called a *discrete beta process* with parameters c and Λ.

By the binomial-beta conjugacy (see Proposition G.8) the joint posterior distribution of the parameters $h(jb)$ is

$$h(jb)|\, D_n \overset{\text{ind}}{\sim} \text{Be}\Big(c_{j,b}\lambda_{j,b} + \Delta N(jb), c_{j,b}(1-\lambda_{j,b}) + Y(jb) - \Delta N(jb)\Big).$$

It follows that the posterior distribution of the cumulative hazard function H also follows a discrete beta process. If the parameters satisfy (13.13) and either c is constant on every interval $((j-1)b, jb]$ or Λ is discrete and supported on the grid points, then the new parameters can again be written in the form (13.13) but with c and Λ updated to $c+Y$ and $t \mapsto \int_{[0,t]}(c+Y)^{-1}\,(c\,d\Lambda + dN)$. The posterior mean and variance are given by

$$\hat{H}(jb) := \text{E}[H(jb)|\, D_n] = \sum_{l=0}^{j} \frac{c_{l,b}\lambda_{l,b} + \Delta N(lb)}{c_{l,b} + Y(lb)},$$

$$\text{var}[H(jb)|\, D_n] = \sum_{l=0}^{j} \frac{\Delta\hat{H}(lb)(1-\Delta\hat{H}(lb))}{c_{l,b} + Y(lb) + 1}.$$

By the prior independence of the $h(jb)$ and the factorization of the likelihood over these parameters, it follows that the variables $h(jb)$ are a posteriori independent, which leads to

$$1 - \hat{F}(jb) := \text{E}[1 - F(jb)|\, D_n] = \prod_{l=0}^{j}\Big(1 - \frac{c_{l,b}\lambda_{l,b} + \Delta N(lb)}{c_{l,b} + Y(lb)}\Big). \tag{13.14}$$

The "noninformative" limits as $c_{j,b} \to 0$ of the posterior means $\hat{H}(jb)$ and $\hat{F}(jb)$ are the discrete-time Nelson-Aalen estimator $\sum_{l=0}^{j} \Delta N(lb)/Y(lb)$ and the discrete time Kaplan-Meier estimator, respectively. Like in the case of a Dirichlet process, the parameters $c_{j,b}$, or function c, measure the "strength of prior belief."

13.3.2 Continuous Time

Consider now a passage to the limit as the mesh width b tends to zero. We can view the discrete beta process as a process in continuous time, that jumps at the grid points jb only. By construction it has independent, nonnegative increments; it is cadlag, provided that we define it as such on the intermediate time values. In the terminology introduced in Appendix J all jumps occur at "fixed times," and hence its continuous intensity measure ν^c vanishes. Its discrete intensity measure has the beta distributions (13.12) as its jump heights distributions $\nu^d(\{bj\}, ds)$. The following proposition shows that as $b \to 0$ the discrete beta process with parameters specified as in (13.13) tends to a *beta process* with parameters c and Λ, which is defined as follows.

Definition 13.3 (Beta process) A *beta process* with parameters (c, Λ) is an independent increment process with intensity measure $\nu = \nu^c + \nu^d$ on $(0, \infty) \times (0, 1)$ of the form

$$\nu^c(dx, ds) = c(x)s^{-1}(1-s)^{c(x)-1}\, d\Lambda^c(x)\, ds,$$
$$\nu^d(\{x\}, \cdot) = \mathrm{Be}(c(x)\Delta\Lambda(x), c(x)(1-\Delta\Lambda(x))).$$

Here $c\colon [0, \infty) \to [0, \infty)$ is a measurable function, and Λ is a cumulative hazard function, with $\Lambda^c = \Lambda - \Lambda^d$ its continuous part. The beta distribution in the second line is understood to be degenerate at zero if $\Delta\Lambda(x) = 0$.

The interpretation of the definition is as follows. A beta process is a sum

$$H(t) = \sum_{x:x\leq t} \Delta H(x) + \sum_{j:x_j\leq t} \Delta H(x_j),$$

of two independent processes that both increase by jumps ΔH only. The second component jumps only at "fixed times" $x_1, x_2, \ldots$, given by the atoms of Λ, and with "jump heights" $\Delta H(x_j) \overset{\text{ind}}{\sim} \mathrm{Be}(c(x_j)\Lambda(\{x_j\}), c(x_j)(1-\Lambda(\{x_j\})))$. The first component can be obtained by simulating a Poisson process with intensity measure ν^c on the set $(0, \infty) \times (0, 1)$ (i.e. the number of points in a set A is $\mathrm{Poi}(\nu^c(A))$) and creating a jump at x of height $s = \Delta H(x)$ for each point (x, s) in the Poisson process with $x \leq t$. In the next section this type of process is introduced in general.

The beta process may be viewed as a canonical prior for a cumulative hazard function. In Example 13.11 it will be seen that a Dirichlet process prior on the survival function $\bar{F}$ yields a beta process of a special parameterization on the cumulative hazard function.

From the general theory (see (13.18) and (13.19)), it follows that

$$\mathrm{E}[H(t)] = \Lambda(t), \qquad \mathrm{var}[H(t)] = \int_{(0,t]} \frac{1-\Delta\Lambda}{c+1}\, d\Lambda. \tag{13.15}$$

Proposition 13.4 *For c a cadlag, nonnegative function and Λ a continuous cumulative hazard function, the discrete time beta process with parameters specified as in* (13.13) *tends as $b \to 0$ in distribution in the Skorohod space $\mathfrak{D}[0, \infty)$ equipped with the Skorohod topology on compacta to a beta process with parameters c and Λ.*

Proof The discrete time beta process has intensity measure $\nu_b = \sum_j \delta_{jb} \times B_{j,b}$, for $B_{j,b}$ the beta distribution specified in (13.12) and (13.13). We first show that the measures $s\,\nu_b(dx, ds)$ converge weakly on compact time sets to the measure $s\,\nu(dx, ds)$, for ν given in Definition 13.3. For any continuous function with compact support f and $k \geq 0$, by the formula for the $(k+1)$st moment of a beta distribution, and with $a_{j,b} = c_{j,b}\lambda_{j,b}$,

$$\int\int f(x)s^k\, s\,\nu_b(dx, ds) = \sum_j f(jb)\frac{(a_{j,b}+k)\cdots(a_{j,b}+1)a_{j,b}}{(c_{j,b}+k)\cdots(c_{j,b}+1)c_{j,b}}.$$

Since $a_{j,b}/c_{j,b} = \lambda_{j,b} = \Lambda((j-1)b, jb]$ and $c_{j,b} = c(jb)$, the right side can be written as $\int f_b g_b\, d\Lambda$, for $f_b = \sum_j f(jb)\mathbb{1}\{((j-1)b, jb]\}$ and

$$g_b = \sum_j \frac{(c(jb)\Lambda((j-1)b, jb]+k)\cdots(c(jb)\Lambda((j-1)b, jb]+1)}{(c(jb)+k)\cdots(c(jb)+1)}(jb)\mathbb{1}\{((j-1)b, jb]\}.$$

By continuity $f_b \to f$ pointwise, and similarly the functions g_b converge pointwise to a limit g. By the dominated convergence theorem $\int f_b g_b\, d\Lambda$ tends to

$$\int f(x)\frac{(c\Delta\Lambda+k)\cdots(c\Delta\Lambda+1)}{(c(x)+k)\cdots(c(x)+1)}\, d\Lambda(x) = \int\int f(x)s^k\, s\,\nu(dx, ds).$$

For continuous Λ the jumps $\Delta\Lambda$ vanish, and the numerator in the left side reduces to $k!$. Then the last equality follows by evaluating the integral over s in the right side as the beta integral $B(k+1, c(x))$. By convergence of moments we conclude that the measures $A \mapsto \int\int \mathbb{1}\{s \in A\}f(x)s\,\nu_b(dx, ds)$ converge weakly to the same measures with ν replacing ν_b, for every f. Then the weak convergence $s\,\nu_b(dx, ds) \rightsquigarrow s\,\nu(dx, ds)$ on compact time sets follows.

Under the assumption that Λ is continuous the proposition follows by Proposition J.18. □

As its discrete counterpart, the beta process possesses the attractive property of conjugacy with respect to randomly right censored data. In the case that the prior parameter Λ is continuous the following theorem is a consequence of the general conjugacy of independent increment processes given in Theorem 13.15.

Theorem 13.5 (Conjugacy) *If the cumulative hazard function H follows a beta process with parameters (c, Λ), where c is continuous and bounded away from zero and Λ has at most finitely many discontinuities in any bounded interval, then the posterior distribution of H given random right censored data D_n is again a beta process, with parameters $(c + Y, \Lambda^*)$, where $\Lambda^*(t) = \int_{(0,t]}(c+Y)^{-1}(c\, d\Lambda + dN)$.*

Proof We update the parameters of the intensity measure $\nu(dx, ds) = \rho(ds|\, x)\,\Lambda(dx)$ as in Theorem 13.15 with $\rho(ds|\, x) = c(x)s^{-1}(1-s)^{c(x)-1}\, ds$. The continuous part of the intensity measure is updated to $c(x)s^{-1}(1-s)^{c(x)+Y(x)-1}\, ds\, \Lambda^c(dx)$, in which we can rewrite $c(x)\,\Lambda^c(dx)$ as $(c(x) + Y(x))\,(\Lambda^*)^c(dx)$. Thus we obtain the continuous part of the intensity measure of a beta process with parameters $(c + Y, \Lambda^*)$. In the fixed jump part, the measure $\rho(\cdot|\, x)$ must be updated to $s^{\Delta N(x)}(1-s)^{Y(x)-\Delta(x)}\rho(ds|\, x)$. If Λ had no atom at x, then this means creating a new fixed jump with a beta distribution with parameters equal to $(\Delta N(x), Y(x) - \Delta N(x) + c(x))$; in the other case the parameters $(c(x)\Delta\Lambda(x), c(x)(1 - \Delta\Lambda(x)))$ must be updated by adding $(\Delta N(x), Y(x) - \Delta N(x))$, giving the beta distribution with parameters $(\Delta N(x) + c(x)\Delta\Lambda(x), Y(x) - \Delta N(x) + c(x)(1 - \Delta\Lambda(x)))$. In both cases this corresponds to a fixed jump distribution of a beta process with parameters $(c + Y, \Lambda^*)$.

If Λ is continuous, then the assumptions of Theorem 13.15 are fulfilled. As is noted in its proof, Theorem 13.15 remains valid if Λ has finitely many jumps. □

13.3.3 Sample Path Generation

Because the (marginal) distributions of the increments of a beta process do not possess a simple, closed form, it is not straightforward to simulate sample paths from the prior or posterior distributions. In this section we list some algorithms that give approximations. Most of the algorithms extend to other independent increment processes by substituting the correct intensity measure.

As explained following Definition 13.3 the beta process can be split in a part with a continuous parameter Λ and a fixed jump part. As the parts are independent and the fixed jump part is easy to generate, we concentrate on the case that $\Lambda = \Lambda^c$ is continuous.

The first algorithm is to simulate the discrete time beta process; this tends to the beta process as the time step tends to zero by Proposition 13.4. The other algorithms are based on the representation of a beta process through a counting measure: $H(t) = \sum_{x:x\le t} \Delta H(x)$, where for $t \le \tau$ the sum has countably many terms given by the points $(x, \Delta H(x))$ of a Poisson process on $(0, \tau] \times (0, 1)$ with intensity measure given by

$$\nu(dx, ds) = c(x)s^{-1}(1 - s)^{c(x)}\, ds\, d\Lambda(x).$$

We obtain an approximation by generating a large (but regrettably necessarily finite) set of points (X_i, S_i) according to this Poisson process, or some approximation, and forming the process $H(t) = \sum_i S_i \mathbb{1}\{X_i \le t\}$. This can be achieved in various ways, leading to several algorithms, which differ in complexity, dependent on the availability of numerical routines to compute certain special functions. Algorithm (d) appears to give a particularly good trade-off between ease, efficiency and accuracy.

Algorithm a (Time discretization) For a sufficiently small b generate H as a discrete time beta process on the points $0, b, 2b, \ldots$, with parameters given by (13.13). This construction is justified by Proposition 13.4.

Algorithm b (Inverse Lévy measure I) For $L_x(s) = \Lambda(\tau)\int_s^1 c(x)u^{-1}(1 - u)^{c(x)-1}\, du$, and a sufficiently large number m, generate, for $i = 1, \ldots, m$,

$$X_i \overset{\text{iid}}{\sim} \frac{\Lambda(\cdot \wedge \tau)}{\Lambda(\tau)}, \qquad E_i \overset{\text{iid}}{\sim} \text{Exp}(1), \qquad V_j = \sum_{i=1}^{j} E_i, \qquad S_i = L_{X_i}^{-1}(V_i).$$

Then $H(t) = \sum_{i=1}^{m} S_i \mathbb{1}\{X_i \le t\}$ is approximately a beta process.

This algorithm is based on the fact that the points (X_i, S_i), for $i = 1, 2, \ldots$, are a realization from the Poisson process with intensity measure ν (see Example J.9). For $m = \infty$ the algorithm would be exact.

Algorithm c (Inverse Lévy measure II) For Q_s the probability distribution on $(0, \tau]$ satisfying $dQ_s(x) \propto c(x)(1-s)^{c(x)-1}\, d\Lambda(x)$, the nondecreasing function L given by $L(s) = \int_0^\tau \int_{(s,1]} u^{-1} c(x)(1-u)^{c(x)-1} du\, d\Lambda(x)$, and a sufficiently large number m, generate, for $i = 1, \ldots, m$,

$$E_i \overset{\text{iid}}{\sim} \text{Exp}(1), \qquad V_j = \sum_{i=1}^{j} E_i, \qquad S_i = L^{-1}(V_i), \qquad X_i \,|\, S_i \overset{\text{ind}}{\sim} Q_{S_i}.$$

Then $H(t) = \sum_{i=1}^m S_i \mathbb{1}\{X_i \le t\}$ is approximately a beta process.

As Algorithm (b), this algorithm generates the points (X_i, S_i) of the Poisson process with intensity measure ν, but this intensity measure is disintegrated in the other direction. Marginally the jump heights of the beta process follow a Poisson process on $[0, 1]$ with intensity measure $\nu((0, \tau] \times \cdot)$. As the latter process explodes at 0, it is convenient to order its points in reverse order of magnitude, downwards from 1 to 0. The function L is the cumulative intensity function of this reversed process and hence its points can be obtained by transforming the points $V_1, V_2, \ldots$ of a standard Poisson process by the inverse L^{-1}. Given the jump heights, the locations X_i in $[0, \tau]$ are independent, and can be generated from the intensity measure conditioned to the horizontal line at the specified height w, which is given by Q_w.

Algorithm d (Poisson weighting) For positive conditional probability densities $s \mapsto g(s|\,x)$ on $[0, 1]$, and $f(x, s) = c(x)s^{-1}(1-s)^{c(x)-1}$, and a sufficiently large number m, generate, for $i = 1, \ldots, m$,

$$X_i \overset{\text{iid}}{\sim} \frac{\Lambda(\cdot \wedge \tau)}{\Lambda(\tau)}, \qquad S_i \,|\, X_i \overset{\text{ind}}{\sim} g(\cdot\,|\, X_i), \qquad K_i \,|\, X_i, S_i \overset{\text{ind}}{\sim} \text{Poi}\Big(\frac{\Lambda(\tau) f(X_i, S_i)}{m\, g(S_i \,|\, X_i)}\Big).$$

Then $H(t) = \sum_{i=1}^m S_i K_i \mathbb{1}\{X_i \le t\}$ is approximately a beta process. A convenient particular special choice for g is the density of the $\text{Be}(1, c(x))$-distribution, for which the quotient $f(x, s)/g(s|\,x)$ reduces to $1/s$. However, the $\text{Be}(\epsilon, c(x))$-distribution for $\epsilon < 1$ appears to work more efficiently.

The algorithm can be understood as placing Poisson numbers K_i of points at the locations (X_i, S_i) in the jump space $(0, \tau] \times [0, 1]$. Allowing more than one point at a location differentiates it from the other algorithms and is unlike the Poisson process corresponding to the beta process. However, for large m this may be viewed as a (random) discretization, and the mean number of points is chosen so that the expected number of points is exactly right, as, for any Borel set A,

$$m\text{E}[K_i \mathbb{1}\{(X_i, S_i) \in A\}] = m \int_0^\tau \int_0^1 \mathbb{1}_A(x, s) \frac{\Lambda(\tau) f(x, s)}{m\, g(s|\,x)}\, g(s|\,x)\, ds\, \frac{d\Lambda(x)}{\Lambda(\tau)} = \nu(A).$$

A rigorous justification for the algorithm is that the process H converges in the Skorohod space to a beta process, as $m \to \infty$ (see Lee 2007 or Damien et al. 1995).

Algorithm e (ϵ-approximation) For a sufficiently small $\epsilon > 0$, generate

$$K \sim \mathrm{Poi}\Big(\frac{1}{\epsilon}\int_0^\tau c\,d\Lambda\Big), \qquad X_1,\ldots,X_K \mid K \overset{\text{iid}}{\sim} \frac{c\,d\Lambda}{\int_{[0,\tau]} c\,d\Lambda}, \qquad S_i \mid X_i \overset{\text{ind}}{\sim} \mathrm{Be}(\epsilon, c(X_i)).$$

Then $H(t) = \sum_{i=1}^m S_i \mathbb{1}\{X_i \le t\}$ is approximately a beta process.

The scheme actually generates a realization of the independent increment process with intensity measure $\nu_\epsilon(dx, ds) = \epsilon^{-1} c(x)\,\mathrm{be}(\epsilon, c(x))(s)\,ds\,d\Lambda(x)$, where $\mathrm{be}(\alpha, \beta)$ is the density of the beta distribution. Indeed, for reasonable functions c the measure ν_ϵ is finite with total mass $\nu_\epsilon([0,\tau]\times[0,1]) = \epsilon^{-1}\int_0^\tau c\,d\Lambda$ and hence the corresponding Poisson process can be simulated by first generating its total number of points and next their locations, where we may first generate the x-coordinates from the second marginal of the renormalized ν_ϵ and next the corresponding s-coordinates. The justification of the algorithm is that the intensity measures ν_ϵ converge in the appropriate sense to ν, as $\epsilon \to 0$, so that the corresponding independent increment processes converge in the Skorohod topology to the beta process with intensity measure ν, by Proposition J.18. Note that $\epsilon^{-1}\mathrm{be}(\epsilon, c)(s) \sim s^{-1}(1-s)^{c-1}$, as $\epsilon \downarrow 0$, since $\epsilon B(\epsilon, c) \to 1$.

13.3.4 Mixtures of Beta Processes

As in the case of a Dirichlet or Pólya tree process, the parameters c and Λ of a beta process may depend on a hyperparameter θ, which may be given a prior ν. This results in a *mixture of beta processes* prior for H.

The posterior distribution given right censored data D_n is again a mixture.

Theorem 13.6 *If $H \mid \theta$ follows a mixture of beta process prior with continuous parameters c_θ and Λ_θ and $\theta \sim \nu$, then the posterior distribution of H given right censored data is again a mixture of beta process: $H \mid (D_n, \theta)$ follows a beta process with parameters $(c_\theta^*, \Lambda_\theta^*)$, and $\theta \mid D_n \sim \nu^*$, with parameter updates given by $c_\theta^* = c_\theta + Y$, $d\Lambda_\theta^* = (c_\theta + Y)^{-1}(c_\theta\,d\Lambda_\theta + dN)$ and*

$$\nu^*(d\theta) \propto \exp\Big[-\sum_{i=1}^n \int_0^{T_{(i)}} \frac{c_\theta}{c_\theta + n - i}\,d\Lambda_\theta\Big] \prod_{i=1}^{K_{n,j}} \frac{c_\theta(U_j) h_\theta(U_j)}{\prod_{i=1}^j (c_\theta(U_i) + Y(U_i) - i)}\,\nu(d\theta),$$

where K_n is the number of distinct uncensored observations, $U_1, \ldots, U_{K_n}$ are the distinct uncensored observations and $K_{n,j}$ is the number of uncensored observations greater than U_j, for $j = 1, \ldots, K_n$.

The proof of the theorem is based on a calculation of the marginal distribution of the sample using the posterior conjugacy of the beta process (see Theorem 13.5), and then applying Bayes's theorem, as in case of a mixture of Dirichlet or Pólya tree process posterior; see Kim (2001) for details.

13.4 Neutral to the Right and Independent Increment Processes

The lack of conjugacy of the Dirichlet process prior for right censored data diminishes its role in survival analysis. Since the family of survival distributions is not (assumed) dominated, Bayes's theorem is not applicable, which leaves conjugacy as the most important path to computation of the posterior distribution. A large class of priors defined below will be seen to have the desirable conjugacy property.

Definition 13.7 (Neutral to the right process) A random distribution function F (or the corresponding survival function) is said to follow a *neutral to the right* (NTR) process if for every finite partition $0 = t_0 \leq t_1 < \cdots < t_k < \infty$ and $k \in \mathbb{N}$, the random variables

$$F(t_1), \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, \ldots, \frac{F(t_k) - F(t_{k-1})}{1 - F(t_{k-1})}$$

(or equivalently the variables $\bar{F}(t_j)/\bar{F}(t_{j-1})$) are mutually independent.

This definition can be alternatively and perhaps more easily posed in terms of the cumulative hazard function: a random survival function is neutral to the right if the corresponding cumulative hazard function has independent increments. We say that a stochastic process $H = (H(t): t \geq 0)$ is an *independent increment process* (or *PII*) if for every finite partition $0 = t_0 \leq t_1 < \cdots < t_k < \infty$ and $k \in \mathbb{N}$, the random variables $H(t_1) - H(t_0), H(t_2) - H(t_1), \ldots, H(t_k) - H(t_{k-1})$ are jointly independent.

Theorem 13.8 *For a random distribution F, the corresponding cumulative hazard function H, and the function $A = -\log \bar{F}$, the following assertions are equivalent:*

(i) *F follows a neutral to the right process.*
(ii) *H is an independent increment process.*
(iii) *A is an independent increment process.*

In this case, the means $F_0(t) = \mathrm{E}[F(t)]$ and $H_0(t) = \mathrm{E}[H(t)]$ are corresponding survival and cumulative hazard functions: $H_0(t) = \int_{(0,t]} dF_0/\bar{F}_{0-}$ and $\bar{F}_0(t) = \prod_{(0,t]}(1 - dH_0)$.

Proof The equivalence of (i) and (iii) is immediate from the definition of A and the fact that the logarithm turns quotients into differences. We prove the equivalence of (i) and (ii).

Fix a partition $0 = t_0 < t_1 < \cdots < t_k < \infty$, and a countable dense subset $\{s_1, s_2, \ldots\}$ of $\mathbb{R}^+$. For a given m, let $s_{1:m} < \cdots < s_{m:m}$ be the ordering of $\{s_1, \ldots, s_m\}$. Then

$$H(t_j) - H(t_{j-1}) = \lim_{m\to\infty} \sum_{t_{j-1} < s_{i:m} \leq t_j} \frac{F(s_{i:m}) - F(s_{i-1:m})}{1 - F(s_{i-1:m})}. \qquad (13.16)$$

If F follows an NTR process, then the summands on the right are mutually independent. Since for disjoint intervals $(t_{j-1}, t_j]$ the sums have no common terms, it follows that the variables $H(t_j) - H(t_{j-1})$, for $j = 1, \ldots k$, are mutually independent. Conversely, by the product-integral representation (13.3) of F in terms of H,

$$\frac{\bar F(t_j)}{\bar F(t_{j-1})}=\lim_{m\to\infty}\prod_{t_{j-1}<s_{i:m}\le t_j}(1-H(s_{i-1:m},s_{i:m}]).$$

If H has independent increments, then the factors of the product are mutually independent. Since there is no common factor for disjoint intervals $(t_{j-1},t_j]$, the variables $\bar F(t_j)/\bar F(t_{j-1})$, for $j=1,\dots,k$, are mutually independent.

To prove the final statement, observe that $\mathrm{E}(U/V)=\mathrm{E}(U)/\mathrm{E}(V)$ whenever the random variables U/V and V are independent. Applying this to the variables $U=F(s_{i:m})-F(s_{i-1:m})$ and $V=1-F(s_{i-1:m})$ in (13.16), we see that this display remains valid if H and F are replaced by their means H_0 and F_0. This leads to the representation of H_0 in terms of F_0. The converse follows as the relation between cumulative hazard and survival functions is one-to-one. □

An independent increment process with cadlag, nondecreasing sample paths is the distribution function of a random measure on $[0,\infty)$, which is a *completely random measure* (CRM) as defined in Appendix J (see Problem J.5). By Proposition J.6 it can be represented as

$$H(t)=\int_{(0,t]}\int s\,N^c(dx,ds)+\sum_{j:x_j\le t}\Delta H(x_j),$$

for a Poisson random measure N^c on $[0,\infty)\times(0,\infty)$ and arbitrary points $x_1,x_2,\dots$ in $[0,\infty)$, called the *fixed jump times* of H. The Poisson process N^c is fully characterized by its *intensity measure* $\nu_H^c(A)=\mathrm{E}[N^c(A)]$, and is independent of the “fixed jump” heights $\Delta H(x_1),\Delta H(x_2),\dots$, which are arbitrary independent nonnegative variables. The representation shows that the sample paths of a PII with nondecreasing sample paths necessarily increase by jumps only, and hence $H(t)=\sum_{x\le t}\Delta H(x)$, where for every sample path the series has at most countable many jumps. The fixed jump times are deterministic locations at which *every* sample path of H increases (by the random heights $\Delta H(x_j)$), but typically most of the jumps occur at random locations, different for different sample paths. These locations are given by the x-coordinates of the points (x,s) in the Poisson process N^c, with s the jump height $\Delta H(x)$ at location x.

For unity of notation it is attractive to encode the fixed jumps also in a counting measure, and write the representation in the form

$$H(t)=\int_{(0,t]}\int s\,N(dx,ds),\qquad N=N^c+N^d,\qquad N^d=\sum_j\delta_{x_j,\Delta H(x_j)}.$$

The sum N has mean measure the sum $\nu_H=\nu_H^c+\nu_H^d$ of the mean measures of N^c and N^d. The measure ν_H^d concentrates on the set $\cup_j\{x_j\}\times(0,\infty)$, while ν_H^c gives zero probability to this set; in particular, the measures ν_H^c and ν_H^d are identifiable from their sum. The latter measure gives the law of the fixed jump heights: for $j=1,2,\dots$ and $D\subset(0,\infty)$,[4]

$$\mathrm{P}(\Delta H(x_j)\in D)=\nu_H(\{x_j\}\times D)=\nu_H^d(\{x_j\}\times D).$$

The measures N^d and N (unless $N^d=0$) are *not* Poisson random measures, unlike N^c.

[4] The variable $\Delta H(x_j)$ in a general PII may have an atom at zero of size $1-\nu_H^d\big(\{x_j\}\times[0,\infty)\big)$, but in the present chapter this atom is always zero.

For theoretical manipulation the Laplace transform of the variable $H(t)$ is handy, as this allows explicit expression in terms of the intensity measure. For $\theta > 0$, if the fixed jump heights are strictly positive a.s.,

$$\mathrm{E}[e^{-\theta H(t)}] = e^{-\int_{(0,t]}\int(1-e^{-\theta s})\,\nu_H^c(dx,ds)} \prod_{x:x\le t} \int e^{-\theta s}\,\nu_H^d(\{x\},ds). \tag{13.17}$$

Here $\int f(s)\,\nu_H^d(\{x\},ds)$ denotes the integral of the function f with respect to the measure $A \mapsto \nu_H^d(\{x\},A)$, and the product is understood to be over the fixed jump times x, and to be 1 if there are no fixed jumps. Differentiating at $\theta = 0$ (or see (J.11) and (J.12)), we find

$$\mathrm{E}[H(t)] = \int_{(0,t]}\int s\,\nu_H(dx,ds), \tag{13.18}$$

$$\mathrm{var}[H(t)] = \int_{(0,t]}\int s^2\,\nu_H(dx,ds) - \sum_{x:x\le t}\Big(\int s\,\nu_H^d(\{x\},ds)\Big)^2. \tag{13.19}$$

This aids in the interpretation of the intensity measure. For prior modeling one will typically not include fixed atoms, but we shall see that these arise naturally in the posterior distribution.

To construct a stochastic process H with independent, nondecreasing sample paths one may start from any measure $\nu_H = \nu_H^c + \nu_H^d$ on $[0,\infty)\times(0,\infty)$, where the x-marginals of ν_H^c and ν_H^d are atomless and discrete, respectively, satisfying, for $t, x \ge 0$,

$$\int_{(0,t]}\int (s\wedge 1)\,\nu_H^c(dx,ds) < \infty, \qquad \nu_H^d(\{x\}\times[0,\infty)) \le 1. \tag{13.20}$$

A cumulative hazard function H is restricted to have jump heights ΔH smaller than 1, and corresponds to a proper probability distribution only if either $H(t)\to\infty$ as $t\to\infty$ and all jumps are strictly smaller than 1, or H has a single jump of size 1 at the end of its support. The jump heights can be controlled by restricting ν_H to $[0,\infty)\times[0,1)$; the following lemma gives a sufficient condition for the distribution function to be proper.

Lemma 13.9 *Any measure ν_H on $[0,\infty)\times[0,\infty)$ satisfying (13.20), for every $t > 0$, is the intensity measure of a stochastic process with independent, nonnegative increments. If ν_H concentrates on $[0,\infty)\times[0,1)$ and satisfies $\int_0^\infty\int_0^1 s\,\nu_H(dx,ds) = \infty$, then this process is a cumulative hazard function corresponding to a proper probability distribution F on $(0,\infty)$, almost surely. In particular, this is true for an intensity measure of the form $\nu_H(dx,ds) = \rho(ds|\,x)\,d\alpha(x)$, with $\inf_x \int_0^1 s\,\rho(ds|\,x) > 0$ and $\alpha(\mathbb{R}^+) = \infty$.*

Proof That (13.20) characterizes intensity measures follows from the general theory explained in Appendix J. In view of (13.18) the condition $\int_0^\infty\int_0^1 s\,\nu_H(dx,ds) = \infty$ implies that $\mathrm{E}[H(t)] \uparrow \infty$ as $t\to\infty$, whence $\mathrm{E}[\bar F(t)] \to 0$ by Theorem 13.8. By monotonicity $\bar F(t) \to 0$ almost surely.

For the special choice of intensity, the double integral is for every $t > 0$ bounded below by $\inf_{x:x\le t}\int s\,\rho(ds|\,x)\,\alpha(0,t]$, and hence is infinite. □

If F follows a neutral to the right process, then both H and A corresponding to F are processes with independent increments, and hence allow representations through jump measures. The following proposition connects their intensity measures.

Proposition 13.10 *The intensity measures of the independent increment processes H and A corresponding to a neutral to the right survival distribution F satisfy*

$$\nu_A(C \times D) = \nu_H(C \times \{s\colon -\log(1-s) \in D\}),$$
$$\nu_H(C \times D) = \nu_A(C \times \{s\colon 1 - e^{-s} \in D\}).$$

Proof The jumps of A and H are related by $\Delta A = -\log(\bar{F}/\bar{F}_-) = -\log(1 - \Delta H)$. It follows that the Poisson process N_A^c corresponding to A has a jump at $(x, -\log(1-s))$ if and only if the Poisson process N_H^c corresponding to H has a jump at (x, s). This implies the relationship between the mean measures ν_A^c and ν_H^c of the Poisson processes, as stated. The mean measures ν_A^d and ν_H^d of the fixed jumps transform similarly.[5] □

Example 13.11 (Dirichlet process) It follows from Theorem 4.28 that a Dirichlet process prior on a distribution F on $(0, \infty)$ is also a neutral to the right process. In fact, the theorem shows that the Dirichlet process is "completely neutral," which may be described as "neutral to the right" also if the intervals $(t_{j-1}, t_j]$ are placed in arbitrary and not their natural order.

It turns out that a survival distribution F follows a $\mathrm{DP}(MF_0)$-process with prior strength M and center measure F_0 if and only if the corresponding cumulative hazard function follows a beta process with parameters $(M\bar{F}_{0-}, H_0)$, for H_0 the cumulative hazard function that goes with F_0. Thus the Dirichlet processes form a subclass of the beta processes, with the two parameters linked.

An insightful method to verify the claim is to approximate the prior that the Dirichlet process induces on the cumulative hazard function by a discrete beta process. As $b \downarrow 0$ the discrete cumulative hazard functions H_b with jumps $\Delta H_b(jb) = F((j-1)b, jb]/F((j-1)b)$ tend pointwise to H, as seen in (13.16). If $F \sim \mathrm{DP}(MF_0)$, then the jumps $\Delta H_b(jb)$ are independent $\mathrm{Be}(M\bar{F}_0((j-1)b)\lambda_{b,j}, M\bar{F}_0((j-1)b)\,(1-\lambda_{b,j}))$ variables, for $\lambda_{j,b} := F_0((j-1)b, jb]\;/F_0((j-1)b)$ by the self-similarity of the Dirichlet process (see (4.9)). In other words, H_b is a discrete beta process with parameters $c(jb) = M\bar{F}_0((j-1)b)$ and $\lambda_{j,b}$, as given (note that $c(jb)$ is identified as the sum of the two parameters, and next $\lambda_{j,b}$ as the first parameter divided by $c(jb)$). An argument analogous to that of Proposition 13.4 will show that the processes H_b converge in distribution to a beta process with parameters $c = M\bar{F}_{0-}$ and $\Lambda = H_0$.

Example 13.12 (Beta-Stacy process) A survival distribution F is said to follow a *beta-Stacy process* with parameters (c_0, F_0) if the corresponding cumulative hazard function H follows a beta process with parameters $(c_0\bar{F}_{0-}, H_0)$ (as in Definition 13.3). In other words, the intensity measure is given by

[5] Since a Poisson process remains a Poisson process under a change of variables, the argument gives an elegant proof of Theorem 13.8 as well.

$$\nu_H^c(dx, ds) = c_0(x) s^{-1} (1-s)^{c_0(x)\bar{F}_0(x-)-1}\, dF_0^c(x)\, ds,$$
$$\nu_H^d(\{x\}, ds) = \mathrm{be}(c_0(x) F_0\{x\}, c_0(x)\bar{F}_0(x-))\, ds.$$

The parameters c_0 and F_0 are a measurable function $c_0\colon \mathbb{R}^+ \to \mathbb{R}^+$ and a cumulative distribution function F_0, and H_0 is the cumulative hazard function that goes with F_0. In view of Example 13.11 the beta-Stacy process generalizes the Dirichlet process by replacing the prior strength M by a function c_0, thus providing a "location-dependent" prior strength. On the other hand, it offers not more than a reparameterization of the beta process.

Mean and variance of the beta-Stacy process can be computed using the general equations (13.18) and (13.19), leading to (13.15) with $\Lambda = H_0$ and $c = c_0 F_0-$. In particular, the parameter H_0 is the prior mean of H, whence F_0 is the prior mean of F, in view of Theorem 13.8.

By Proposition 13.10 the process $A = -\log \bar{F}$ follows a neutral to the right process with intensity measure

$$\nu_A^c(dx, ds) = (1-e^{-s})^{-1} e^{-s c_0(x)\bar{F}_0(x-)} ds\, c_0(x)\, dF_0^c(x),$$
$$\nu_A^d(\{x\}, ds) \propto (1-e^{-s})^{c_0(x)F_0\{x\}-1} e^{-s c_0(x)\bar{F}_0(x-))}\, ds.$$

Here the discrete components $\nu_A^d(\{x\}, \cdot)$ are proper probability measures on $(0, \infty)$, and restricted to the jump points x of F_0.

Example 13.13 (Extended gamma process) A prior process is said to follow an *extended gamma process* with parameters (c_0, A_0) if its associated process $A = -\log \bar{F}$ is an independent increment process with intensity measure

$$\nu_A^c(dx, ds) = s^{-1} c_0(x) e^{-c_0(x)s}\, ds\, dA_0(x).$$

By Proposition 13.10 the corresponding cumulative hazard function H has intensity measure

$$\nu_H^c(dx, ds) = \frac{c_0(x)(1-s)^{c_0(x)-1}}{\log_-(1-s)}\, ds\, dA_0(x).$$

For c_0 a constant function we obtain the *standard gamma process*.

The increments of the standard gamma process possess gamma distributions. The extended process can be constructed from a standard process ξ as the integral $A(x) = \int_0^x c_0^{-1}\, d\xi$, where ξ must have intensity $c_0\, dA_0$, i.e. the process ξ must have independent increments with $\xi(x) \sim \mathrm{Ga}(\int_0^x c\, dA_0, 1)$, for every $x > 0$ (see Example J.15).

The support of a neutral to the right process is determined by the support of the intensity measure ν_H of the independent increment process H, and is typically very large in terms of the weak topology.

Theorem 13.14 (Support) *If H does not have fixed jumps and the support of its intensity measure ν_H is equal to $\mathbb{R}^+ \times [0, 1]$, then the weak support of the corresponding neutral to the right process F is the full space of probability measures $\mathfrak{M}(\mathbb{R}^+)$.*

Proof Since the continuous distributions are dense in $\mathfrak{M}(\mathbb{R}^+)$, it suffices to show that any F_0 with continuous cumulative distribution is in the support. We shall show that for every continuous F_0 the prior gives positive probability to every Kolmogorov-Smirnov ball $\{F\colon d_{KS}(F, F_0) < \epsilon\}$, for $\epsilon > 0$. We can restrict to compact support, because if $1 - F_0(\tau) < \epsilon/2$ and F is within uniform distance $\epsilon/2$ to F_0 on the interval $[0, \tau]$, then F is within uniform distance ϵ to F_0 on $\mathbb{R}^+$. Because the product integral $H \mapsto F = 1 - \prod(1 - dH)$ is continuous relative to the uniform norms on compact intervals,[6] it further suffices to show that for every $\tau, \delta > 0$ the set $\{H\colon \sup_{x\le\tau} |H(x) - H_0(x)| < \delta\}$ receives positive prior mass, for H_0 the cumulative hazard function of F_0.

By the uniform continuity of H_0 on $[0, \tau]$, there exists a partition $0 = t_0 < t_1 < \cdots < t_k = \tau$ such that $H_0(t_{i-1}, t_i] < \epsilon/2$, for every i. If $|H(t_{i-1}, t_i] - H_0(t_{i-1}, t_i]| < \epsilon/(2k)$, for every i, then $|H(x) - H_0(x)| < \epsilon$, for every $x \le \tau$. Now $H(t_{i-1}, t_i]$ is the sum of the heights of the points of the associated Poisson process N inside the strip $(t_{i-1}, t_i] \times [0, 1)$. For D_i equal to the interval $(H_0(t_{i-1}, t_i] - \epsilon/(2k), H_0(t_{i-1}, t_i] + \epsilon/(2k))$, we split this process as

$$N = \sum_{(x,s): s<\delta} \delta_{(x,s)} + \sum_{i=1}^{k} \sum_{(x,s)\in(t_{i-1},t_i]\times D_i, s\ge\delta} \delta_{(x,s)} + \sum_{i=1}^{k} \sum_{(x,s)\in(t_{i-1},t_i]\times D_i^c, s\ge\delta} \delta_{(x,s)}.$$

Because $\int_{0,\tau]} \int_0^\delta s\nu_H(dx, ds) \downarrow 0$, as $\delta \downarrow 0$, the contribution $\int_{(0,\tau]} \int_0^\delta s\, N(dx, ds)$ of the first term on the right to H can be made arbitrarily small by choosing δ small. Let G be the event that this contribution is smaller than ϵ. Let E_i and F_i be the events that the i terms in the two sums on the right have exactly one term and are empty, respectively, i.e. the process N has exactly one point inside every set $(t_{i-1}, t_i] \times (D_i \cap [\delta, 1))$ and zero points inside $(0, \tau] \times (D_i^c \cap [\delta, 1))$, respectively.

The events $E_1, \ldots, E_k, F_1, \ldots, F_k, G$ are independent and have positive probabilities, as they refer to disjoint parts of the Poisson process, and $N(E_i)$ and $N(F_i)$ are Poisson distributed with means $\nu_H(E_i) > 0$ and $\nu_H(F_i) > 0$, by the assumption on ν_H. On the intersection $\cap_i (E_i \cap F_i) \cap G$ the process H is within distance 2ϵ of H_0. □

The main result of this section is the conjugacy of neutral to the right process priors for random right censored data: the posterior distribution of F corresponding to an neutral to the right prior is again neutral to the right. Because the neutral to the right property is equivalent to the independent increment property of the cumulative hazard function, this fact can also be formulated in terms of PIIs and their intensity measures. The following theorem explicitly derives the intensity measure of the posterior process.

A typical prior intensity measure will be absolutely continuous. However, the posterior intensity measure will have fixed jumps at the uncensored observations. This is somewhat analogous to the Kaplan-Meier estimator, which also jumps (only) at uncensored observations.

Recall the notations N and Y for the observed death process and at risk process relating to randomly right censored data $D_n = \{(T_1, \Delta_1), \ldots, (T_n, \Delta_n)\}$, given in (13.5) and (13.6).

Theorem 13.15 (Conjugacy) *If F follows a neutral to the right process, then the posterior distribution of F given randomly right censored data D_n also follows a neutral to the right*

[6] See Theorem 7 of Gill and Johansen (1990).

process. If the corresponding cumulative hazard process H possesses intensity measure ν_H with disintegration $\nu_H(dx, ds) = \rho(ds|\, x)\, \Lambda(dx)$ such that $x \mapsto s\rho(ds|\, x)$ is weakly continuous and Λ is without atoms, then its posterior distribution possesses intensity measure $\nu_{H|D_n}$ given by

$$\nu^c_{H|D_n}(dx, ds) = (1-s)^{Y(x)}\rho(ds|\, x)\, d\Lambda(x),$$
$$\nu^d_{H|D_n}(\{x\}, ds) \propto s^{\Delta N(x)}(1-s)^{Y(x)-\Delta N(x)}\rho(ds|\, x),$$

where the set of fixed jump times is equal to the set $\{T_i\colon \Delta_i = 1\}$ of uncensored observations, and the fixed jump distributions $\nu^d_{H|D_n}(\{x\}, \cdot)$ are probability distributions on $(0, 1)$.

Proof To highlight the structural property of a neutral to the right process, we start with a simple proof of the preservation of the neutral to the right property in the posterior. (The property also follows from the explicit calculations in the second part of the proof.) It suffices to prove the result for $n = 1$, since then the conclusion can be repeated n times.

First consider the case that the observation is not censored. Denote it by T and fix a partition $0 < u_1 < \cdots < u_k < \infty$ of $\mathbb{R}^+$. Given a finer partition $0 = t_0 < t_1 < \cdots < t_{m+1} = \infty$, define $M_i = \mathbb{1}\{t_{i-1} < X \le t_i\}$ and $\theta_i := \bar F(t_i)/\bar F(t_{i-1})$. The likelihood function for observing $M = (M_1, \ldots, M_m)$ can be written as

$$\prod_{i=1}^m \mathrm{P}_\theta(M_i = m_i|\, M_{i-1} = m_{i-1}, \ldots, M_1 = m_1) = \prod_{i=1}^m (1-\theta_i)^{m_i}\theta_i^{1-\sum_{l=1}^{i-1} m_l},$$

which factorizes in separate functions of $\theta_1, \ldots, \theta_m$. Since F follows a neutral to the right process prior, the random variables $\theta_1, \ldots, \theta_m$ are independent in the prior. Since the likelihood factorizes, these variables are then also independent in the posterior given M. The same is true for the variables $\bar F(u_i)/\bar F(u_{i-1})$, which are products of disjoint sets of θ_i. By making the partitions $0 < t_1 < \cdots < t_{m+1} = \infty$ finer and finer, the latter posterior distribution tends to the posterior distribution of the variables $\bar F(u_i)/\bar F(u_{i-1})$ given T, by the martingale convergence theorem, where the independence is preserved.

If the observation is censored, then the same proof works, but we choose the partitions with t_m equal to the censoring variable C (which is independent of everything and may be considered a fixed number), so that by successively refining the partition $0 = t_0 < \cdots < t_m$ the vector M eventually contains the same information as the censored observation. The grid points u_i bigger than t_m cannot appear in the latter partitions, but the corresponding variables $\bar F(u_i)/\bar F(u_{i-1})$ also do not enter the likelihood for M. Hence the likelihood trivially factorizes and the argument can be finished as before.

For the proof of the updating formula for the intensities, we may equivalently use the intensities of the process $A = -\log \bar F$ or of the cumulative hazard function H, in view of Proposition 13.10. We shall use the former, as this allows a more accessible expression for the likelihood function.

Again it suffices to consider the case of a single observation, provided we allow the prior intensity measure to have the form of the posterior. In particular, we allow it to have finitely many fixed atoms, with strictly positive jump sizes. For a given partition $0 = t_0 < t_1 < \cdots < t_m \le t_{m+1} = \infty$, chosen to have $t_m = C$ if the observation is censored, let $M = (M_1, \ldots, M_{m+1})$ have coordinates $M_i = \mathbb{1}\{(t_{i-1}, t_i]\}(T)$. Then M is less informative than the observation (T, Δ), but in the limit as the meshwidth of the partition tends to zero (and $m \to \infty$) it generates the same σ-field. By the martingale convergence theorem the posterior distribution given M tends to the posterior distribution given (T, Δ).

The vector M possesses a multinomial distribution with parameters 1 and $F(t_{i-1}, t_i]$, for $i = 1, \ldots, m+1$. Its likelihood can be written in terms of A as $L_M(A) = \prod_{i=1}^{m+1} (e^{-A(t_{i-1})} - e^{-A(t_i)})^{M_i}$. We can identify the posterior distribution of A given M by evaluating expectations of the form, for bounded, continuous functions $f\colon \mathbb{R}^+ \to \mathbb{R}^+$,

$$\mathrm{E}(e^{-\int f\,dA} |\, M) = \frac{\mathrm{E}_A\big[e^{-\int f\,dA} L_M(A)\big]}{\mathrm{E}_A\big[L_M(A)\big]}.$$

Here the expectations E_A are relative to the prior distribution of A, for fixed M.

Exactly one coordinate of M is nonzero, and this is fixed in the conditioning event. On the event that this is the jth coordinate, we can write $e^{-\int f\,dA} L_M(A) = e^{-\int (f+e_{j-1})\,dA} - e^{-\int (f+e_j)\,dA}$, for the function $e_i = \mathbb{1}\{(0, t_i]\}$. Two applications of formula (13.17) give, with the products taken over the set of fixed atoms of A,

$$\begin{aligned}
&\mathrm{E}_A(e^{-\int f\,dA} L_M(A)) \\
&\quad = e^{-\int\int (1-e^{-(f+e_{j-1})(x)s})\,\nu_A^c(dx,ds)} \prod_x \int e^{-(f+e_{j-1})(x)s}\,\nu_A^d(\{x\}, ds) \\
&\qquad - e^{-\int\int (1-e^{-(f+e_j)(x)s})\,\nu_A^c(dx,ds)} \prod_x \int e^{-(f+e_j)(x)s}\,\nu_A^d(\{x\}, ds) \\
&\quad = e^{-\int\int (1-e^{-(f+e_{j-1})(x)s})\,\nu_A^c(dx,ds)} \prod_x \int e^{-(f+e_{j-1})(x)s}\,\nu_A^d(\{x\}, ds) \\
&\qquad \times \Bigg[1 - e^{-\int_{(t_{j-1},t_j]}\int e^{-f(x)s}(1-e^{-s})\,\nu_A^c(dx,ds)} \prod_{t_{j-1}<x\le t_j} \frac{\int e^{-f(x)s-s}\,\nu_A^d(\{x\},ds)}{\int e^{-f(x)s}\,\nu_A^d(\{x\},ds)}\Bigg].
\end{aligned}$$

The denominator $\mathrm{E}_A[L_M(A)]$ can be expressed in exactly the same way, but with f taken equal to 0. Suppose that the meshwidth of the partition tends to zero. If the observation is uncensored, so that $t_j - t_{j-1} \to 0$, then $e_{j-1} \to \mathbb{1}\{(0, T)\}$; if the observation is censored, then $t_{j-1} = C = T$ and hence $e_{j-1} = \mathbb{1}\{(0, T]\}$. If in both cases the limit function is written as e_T, then the quotient (arising from numerator and denominator of the posterior) of the leading terms, outside the square brackets, tends to

$$e^{-\int\int (1-e^{-f(x)s})e^{-e_T(x)s}\,\nu_A^c(dx,ds)} \prod_x \frac{\int e^{-f(x)s} e^{-e_T(x)s}\,\nu_A^d(\{x\},ds)}{\int e^{-e_T(x)s}\,\nu_A^d(\{x\},ds)}. \tag{13.21}$$

This shows that the intensity measure is updated by multiplying it with the density $e^{-e_T(x)s}$. It remains to analyze the expression within square brackets.

If the observation is censored, then $t_j = \infty$ and the first exponential inside the square brackets is equal to $e^{-\infty} = 0$ (this corresponds to $e^{-A(\infty)} = 0$). The proof is then complete. If the observation is uncensored, we split in two cases: T is a fixed jump time of A, or it is not.

If T is a fixed jump time of A, then for a sufficiently fine partition, it will be the only one in the interval $(t_{j-1}, t_j]$, and the product in square brackets will contain exactly one term. Since the x-marginal of ν_A^c has no atoms (by definition), the first exponential within square brackets tends to $e^{-0} = 1$. The quotient (arising from the numerator and denominator of the posterior) of the two terms in square brackets tends to

$$\frac{1 - \int e^{-f(T)s-s}\, \nu_A^d(\{T\}, ds) / \int e^{-f(T)s}\, \nu_A^d(\{T\}, ds)}{1 - \int e^{-s}\, \nu_A^d(\{T\}, ds) / \int \nu_A^d(\{T\}, ds)}.$$

This combines with the contribution of the fixed jump T to the product in the leading term (see (13.21), where $e_T(T) = 0$) to give the contribution of this atom as

$$\frac{\int e^{-f(T)s}\, \nu_A^d(\{T\},ds) - \int e^{-f(T)s} e^{-s}\, \nu_A^d(\{T\},ds)}{\int \nu_A^d(\{T\},ds) - \int e^{-s}\, \nu_A^d(\{T\},ds)} = \frac{\int e^{-f(T)s}(1-e^{-s})\, \nu_A^d(\{T\},ds)}{\int (1-e^{-s})\, \nu_A^d(\{T\},ds)}.$$

This exhibits the updated intensity measure to be proportional to $(1 - e^{-s})\nu_A^d(\{T\}, ds)$.

If T is not a fixed jump time of A, then for a sufficiently fine partition, the interval $(t_{j-1}, t_j]$ will be free of fixed atoms and the product is empty and should be interpreted as 1. The term within square brackets tends to zero, but so does the corresponding term from the denominator of the posterior distribution. By Taylor's expansion, as the meshwidth of the partition (left of the censoring time) tends to zero, the quotient of the terms satisfies

$$\frac{1 - e^{-\int_{(t_{j-1},t_j]} \int e^{-f(x)s}(1-e^{-s})\, \nu_A^c(dx,ds)}}{1 - e^{-\int_{(t_{j-1},t_j]} \int (1-e^{-s})\, \nu_A^c(dx,ds)}} \to \frac{\int e^{-f(T)s}(1 - e^{-s})\, \nu_A^c(ds|\, T)}{\int (1 - e^{-s})\, \nu_A^c(ds|\, T)},$$

where $\nu_A^c(ds|\, x) = \rho(ds|\, x)$ is the conditional distribution of the second coordinate in ν_A^c given the first coordinate. Comparison to (13.17) shows that a (strictly positive) fixed jump is added at T, with intensity measure proportional to $(1 - e^{-s})\, \nu_A^c(ds|\, T)$. □

Because the jump process ΔN is nonzero only at finitely many time points, it vanishes almost everywhere under Λ, and the updating formula for the continuous part of the intensity measure in Theorem 13.15 can also be unified to

$$\nu_{H|D_n}(dx, ds) = s^{\Delta N(x)}(1 - s)^{Y(x)-\Delta N(x)}\, [\nu_H(dx, ds) + Z(x)\, \nu_H(ds|\, x)\, dN(x)],$$

where ν_H is the (continuous part of the) intensity of the prior process, and Z gives the norming constants for the fixed jump components: $Z^{-1} = (\Delta N)^{-1} \int s^{\Delta N}(1-s)^{Y-\Delta N}\, \nu_H(ds|\,\cdot)$. Because of the different roles of continuous and fixed jump parts of the process, this rewrite may be of moderate help.

When combined with the general formulas (13.18) and (13.19), the theorem also gives formulas for the posterior mean $\mathrm{E}[H(t)|\, D_n]$ and posterior variance $\mathrm{var}[H(t)|\, D_n]$, in terms of the posterior intensity measure $\nu_{H|D_n}$.

Example 13.16 (Beta-Stacy process) Within the class of neutral to the right processes, the subclass of beta-Stacy processes also forms a conjugate class for randomly right censored data. Because a beta-Stacy process is a reparameterized beta-process, this follows from Theorem 13.5.

Example 13.17 (Extended beta process) The *extended beta process* with parameters (a, b, Λ) is the independent increment process without fixed jumps with intensity measure given by

$$\nu_H^c(dx, ds) = s^{-1}\mathrm{be}(s; a(x), b(x))\, ds\, d\Lambda(x).$$

The parameters are positive functions a and b and a continuous cumulative hazard function Λ. The posterior process given randomly right censored data D_n is again an independent increment process, with intensity measure given by

$$\nu_{H|D_n}^c(dx, ds) = s^{-1}(1-s)^{Y(x)}\mathrm{be}(s; a(x), b(x))\, d\Lambda(x)\, ds,$$
$$\nu_{H|D_n}^d(\{x\}, ds) = \mathrm{be}(s; a(x) + \Delta N(x) - 1, b(x) + Y(x) - \Delta N(x))\, ds.$$

Here the fixed atoms x appear only at the uncensored survival times. In view of (13.18) and (13.19) the posterior mean and variance of $H(t)$ are given by

$$\mathrm{E}[H(t)|\, D_n] = \int_{(0,t]} \frac{\Gamma(a+b)\Gamma(b+Y)}{\Gamma(b)\Gamma(a+b+Y)}\, d\Lambda + \int_{(0,t]} \frac{a+\Delta N-1}{a+b+Y-1}\, dN,$$
$$\mathrm{var}[H(t)|\, D_n] = \int_{(0,t]} \frac{a\Gamma(a+b)\Gamma(b+Y)\, d\Lambda}{\Gamma(b)\Gamma(a+b+Y+1)} + \int_{(0,t]} \frac{(a+\Delta N-1)(b+Y-\Delta N)\, dN}{(a+b+Y-1)^2(a+b+Y)}.$$

13.4.1 Consistency

From the variety of independent increment processes, only a small subset leads to an asymptotically consistent posterior distribution. In this section we give simple sufficient conditions in terms of the intensity measure. As is customary in survival analysis we study consistency of the cumulative hazard function H on a finite interval $[0, \tau]$ within the supports of the survival and censoring distributions. We say that the posterior distribution $\Pi_n(\cdot|\, D_n)$ is *consistent* at (F_0, G_0) in this setup if, for every $\epsilon > 0$,

$$\Pi_n\Big(H\colon \sup_{0\le t\le\tau} |H(t) - H_0(t)| > \epsilon |\, D_n\Big) \to 0, \qquad \text{a.s. } [P_{F_0,G_0}^\infty]. \tag{13.22}$$

Here (F_0, G_0) are the true survival and censoring distribution, and P_{F_0,G_0} refers to the distribution of a sample of random right censored data.

Instead of the cumulative hazard function, we may also consider the survival function, but this leads to exactly the same definition of consistency. Indeed, by the continuity of the product integral $H \mapsto \bar{F} = \prod_{(0,\cdot)}(1-dH)$ the consistency (13.22), for every $\epsilon > 0$, implies that, for every $\epsilon > 0$,

$$\Pi_n\Big(F\colon \sup_{0\le t\le\tau} |F(t) - F_0(t)| > \epsilon |\, D_n\Big) \to 0, \qquad \text{a.s. } [P_{F_0,G_0}^\infty].$$

Because the inverse map $F \mapsto H$ is also continuous, the two types of consistency are equivalent.[7]

Since the realizations of independent increment processes are discrete distributions supported on random countable sets and are not dominated (by a σ-finite measure), Schwartz's consistency theory is inapplicable. However, for the neutral to the right priors we can utilize the characterization of the posterior distribution established in Theorem 13.15. This allows computation of posterior mean and variance, and next an appeal to Lemma 6.4. Below we

[7] See Theorem 7 of Gill and Johansen (1990).

add the sample size n as an index to the notations N_n and Y_n for the counting process N and the at-risk process Y.

To illustrate that consistency is far from automatic, we first study the special case of extended beta processes, introduced in Example 13.17.

Proposition 13.18 (Consistency extended beta process) *If the prior for H follows an extended beta process prior with parameters (a, b, Λ), for continuous a, b and Λ, then the posterior distribution for H is consistent at a continuous H_0 (i.e.* (13.22) *holds for any τ such that $(\bar{F}_0\bar{G}_0)(\tau-) > 0$), if and only if a is identically* 1*, equivalently, if and only if the prior is a beta process.*

Proof The posterior mean and variance of $H(t)$ are given in Example 13.17. Since H_0 is continuous without probability one there are no ties in the survival times, and hence the jumps ΔN_n in these formulas can be replaced by 1. By the Glivenko-Cantelli theorem the processes $n^{-1}Y_n$ tend almost surely uniformly to $(F_0G_0)_-$, which is bounded away from zero on $[0, \tau]$ by assumption, with a reciprocal of bounded variation. The processes $n^{-1}N_n$ are of variation bounded by 1 and tend uniformly to the function $\int_{(0,\cdot]} G_{0-}\, dF_0$. Application of Lemma 13.19 shows that the first term in the posterior mean converges to 0 almost surely, while the second term tends almost surely $\int_{(0,\cdot]} a/(F_0G_0)_-\, G_{0-}\, dF_0 = \int_{(0,\cdot]} a\, dH_0$. By the same arguments both terms in the expression for the posterior variance tend to zero almost surely.

It follows that the posterior distribution of $H(t)$ tends to $\int_0^t a\, dH_0$, for every $t \in [0, \tau]$. Because pointwise convergence of functions of bounded variation to a continuous function of bounded variation implies uniform convergence, we can apply Lemma 6.4, to conclude that the posterior distribution also tends to $\int_0^t a\, dH_0$ as a process, in the sense of (13.22).

The limit is equal to H_0 if and only if a is identically equal to 1. □

Lemma 13.19 *If A_n are cadlag functions and B_n are cadlag functions of uniformly bounded variation and $A_n \to A$ and $B_n \to B$ uniformly on* $[0, \tau]$ *for a function of bounded variation A, then $\int_{(0,\cdot]} A_n\, dB_n \to \int_{(0,\cdot]} A\, dB$ uniformly on* $[0, \tau]$.

Proof The difference $\int A_n\, dB_n - \int A\, dB$ is the sum of $\int (A_n - A)\, dB_n$ and $\int A\, d(B_n - B)$. The first is bounded above in absolute value by $\|A_n - A\|_\infty \|B_n\|_{TV}$, and after partial integration, the second can be seen to be bounded by $2\|B_n - B\|_\infty \|A\|_{TV}$. □

Example 13.20 (Curious case of inconsistency) The extended beta process priors Π_1 and Π_2 on the cumulative hazard function H with parameters $(\frac{1}{2}, 1, H_0)$ and $(1, 1, H_0)$ both have prior mean H_0, and hence are centered perfectly. Their prior variances are given by $\mathrm{var}_1[H(t)] = \frac{1}{3}H_0(t)$, for Π_1 and $\mathrm{var}_2[H(t)] = \frac{1}{2}H_0(t)$, for Π_2. The prior Π_1 might seem preferable over Π_2 if H_0 is indeed the true cumulative hazard function, since they have the same, correct mean, but Π_1 is more concentrated. However, by Proposition 13.18 the prior Π_1 leads to an inconsistent posterior distribution, whereas Π_2 is consistent. The proof of the proposition shows that under Π_1 the posterior settles down to $\frac{1}{2}H_0$.

This example demonstrates that the prior mean and (pointwise) variance have minimal roles in determining consistency, which is rather dependent on the structure of the full prior process.

Proposition 13.18 illustrates that consistency depends on the fine properties of the intensity measure as $s \to 0$: within the class of extended beta processes the density of ν_H^c must blow up exactly at the rate s^{-1}, as is true (only) for the ordinary beta processes. In the general case of independent increment processes we obtain consistency under the same condition on the intensity.

Theorem 13.21 (Consistency) *Let the prior for H follows an independent increment process with intensity measure of the form $\nu_H(dx, ds) = s^{-1}q(x, s)\, d\Lambda(x)\, ds$ for a continuous cumulative hazard function Λ and a function q. Assume that q is continuous in x and such that $\sup_{x\in[0,\tau],s\in(0,1)}(1-s)^\kappa q(x, s) < \infty$ for some $\kappa > 0$ and $\sup_{x\in[0,\tau]} |q(x, s)-q_0(x)| \to 0$ as $s \to 0$ for a function q_0 that is bounded away from zero and infinity. Then the posterior distribution for H is consistent at every continuous H_0:* (13.22) *holds for every τ such that $(\bar F_0 \bar G_0)(\tau-) > 0$.*

Proof We show that $\mathrm{E}[H(t)|\, D_n] \to H_0(t)$ and $\mathrm{var}[H(t)|\, D_n] \to 0$, for every $t \in [0, \tau]$. Since pointwise convergence to a continuous function implies uniform convergence, by Pólya's theorem, we can then appeal to Lemma 6.4 to obtain the result.

Because F_0 is assumed continuous, all jumps ΔN_n in the process N_n have size 1, and a sum over the jumps is the same as an integral with respect to N_n.

By Theorem 13.15 and (13.18), the posterior mean $\mathrm{E}[H(t)|\, D_n]$ is given by

$$\int_0^t \int_0^1 (1-s)^{Y_n(x)} q(x, s)\, ds\, \Lambda(dx) + \int_0^t \frac{\int_0^1 s(1-s)^{Y_n(x)-1} q(x, s)\, ds}{\int_0^1 (1-s)^{Y_n(t)-1} q(x, s)\, ds}\, dN_n(x).$$

The sequence of processes $n^{-1}Y_n$ tends uniformly almost surely to the function $G_{0-}F_{0-}$, which is bounded away from zero on $[0, \tau]$. This implies that the first term is bounded above by $\int_0^t \int_0^1 (1-s)^{cn} q(x, s)\, ds\, d\Lambda(x)$ for some $c > 0$. By the assumption on q this integral is uniformly bounded in $t \le \tau$ by a multiple of $\int_0^1 (1-s)^{cb-\kappa}\, ds = O(1/n)$ and hence negligible.

The processes $n^{-1}N_n$ in the second term are of variation bounded by 1 and tend almost surely to $\int_0^\cdot G_{0-}\, dF_0$. To analyze the quotient of integrals inside the outer integral, we first note that $\int_\epsilon^1 (1-s)^{Y_n(x)-1} q(x, s)\, ds \lesssim (1-\epsilon)^{cn}$ almost surely, for some $c > 0$ and uniformly in x, for any $\epsilon > 0$. For $\epsilon_n \to 0$ slowly, this is of smaller order than any power of n^{-1}. We thus see that the quotient of integrals can be written, uniformly in x,

$$\frac{\int_0^{\epsilon_n} s(1-s)^{Y_n(x)-1} q(x, s)\, ds + o(n^{-2})}{\int_0^{\epsilon_n} (1-s)^{Y_n(x)-1} q(x, s)\, ds + o(n^{-2})} \tag{13.23}$$
$$= \frac{Y_n(x)^{-2} \int_0^{Y_n(x)\epsilon_n} u(1-u/Y_n(x))^{Y_n(x)-1}\, du\, (q_0(x) + o(1)) + o(n^{-2})}{Y_n(x)^{-1} \int_0^{Y_n(x)\epsilon_n} (1-u/Y_n(x))^{Y_n(x)-1}\, du (q_0(x) + o(1)) + o(n^{-2})}.$$

Because $n^{-1}Y_n$ converges to a positive limit, for $n\epsilon_n \to \infty$ the integrals in numerator and denominator tend to $\int_0^\infty ue^{-u}\,du$ and $\int_0^\infty e^{-u}\,du$, respectively, which are both equal to 1. We conclude that the left side is asymptotically equivalent to $Y_n(x)^{-1}$, uniformly in x. Substituting this in the preceding display and appealing to Lemma 13.19, we conclude that the posterior mean is equivalent to $\int_{(0,t]} Y_n^{-1}\,dN_n$, which converges to $H_0(t)$.

By (13.19) the posterior variance $\mathrm{var}[H(t)|\, D_n]$ can be written

$$\int_0^t \int_0^1 s(1-s)^{Y_n(x)} q(x,s)\,ds\,d\Lambda(x) \tag{13.24}$$

$$+\int_0^t \Big[\frac{\int_0^1 s^2(1-s)^{Y_n(x)-1}q(x,s)\,ds}{\int_0^1 (1-s)^{Y_n(x)-1}q(x,s)\,ds} - \Big[\frac{\int_0^1 s(1-s)^{Y_n(x)-1}q(x,s)\,ds}{\int_0^1 (1-s)^{Y_n(x)-1}q(x,s)\,ds}\Big]^2\Big]\,dN(x).$$

The first term tends to zero by the same argument as for the posterior mean. Similarly the quotient of integrals in the second term can be seen to be equivalent to $2Y_n(x)^{-2}$, uniformly in x, by arguments as before, while the quotient inside square brackets in the term was already seen to be equivalent to $Y_n(x)^{-1}$. It follows that both terms tend to zero. Hence the posterior variance tends to zero. □

Example 13.22 (Extended gamma process) The extended gamma process prior with parameters (c, A_0) is described in Example 13.13, and has q-function

$$q(x,s) = c(x)\frac{s}{\log_-(1-s)}(1-s)^{c(x)-1}. \tag{13.25}$$

If the function c is continuous and bounded away from zero and infinity on $[0,\tau]$, then this satisfies the conditions of Theorem 13.21.

To see this we note that $\log_-(1-s) \sim s$ and $(1-s)^c - 1 \sim 1$ as $s \to 0$, uniformly in c in bounded intervals $[c_0, c_1] \subset (0,\infty)$, so that $q(x,s) \sim c(x)$ uniformly in x, as $s \to 0$. Furthermore $(1-s)^\epsilon/\log_-(1-s) \to 0$ as $s \to 1$ for $\epsilon > 0$ and hence $(1-s)q(x,s)$ is uniformly bounded.

13.4.2 Bernstein–von Mises Theorem

Under slightly stronger conditions on the intensity measure the posterior distribution converges at the rate $n^{-1/2}$, and satisfies a Bernstein–von Mises theorem. Given the data D_n, the posterior distribution of $\sqrt{n}(H(t) - \mathrm{E}[H(t)|\, D_n])$ converges in distribution to the same (centered) Gaussian distribution as the sequence of centered and scaled posterior means $\sqrt{n}(\mathrm{E}[H(t)|\, D_n] - H_0(t))$, almost surely. This Gaussian distribution is the same as the limit distribution of the Nelson-Aalen estimator $\hat{H}_n$, the nonparametric maximum likelihood estimator of H, given in (13.1). In fact the difference $\sqrt{n}(\mathrm{E}[H(t)|\, D_n] - \hat{H}_n(t))$ tends to zero.

These results are true for a fixed t within the supports of F_0 and G_0, but also in a uniform sense, for the estimators as processes, viewed as elements of the Skorokhod space $\mathfrak{D}[0,\tau]$ of cadlag functions equipped with the uniform distance.

Theorem 13.23 (Bernstein–von Mises) *If the prior for H follows an independent increment process with intensity measure of the form $\nu_H(dx, ds) = s^{-1}q(x, s)\, dx\, ds$ for a function q that is continuous in x and such that $\sup_{x\in[0,\tau],s\in(0,1)}(1-s)^\kappa q(x, s) < \infty$ for some $\kappa > 0$ and and $\sup_{x\in[0,\tau]} |q(x, s) - q_0(x)| = O(s^\alpha)$ as $s \to 0$ for a function q_0 that is bounded away from zero and infinity and some $\alpha \in (1/2, 1]$, then for τ such that $(\bar F_0\bar G_0)(\tau-) > 0$ and continuous F_0:*

(i) $\sqrt{n}(H - \mathrm{E}[H|\, D_n])|\, D_n \rightsquigarrow B\circ U_0$ *a.s.* $[P^\infty_{F_0,G_0}]$ *where B is a standard Brownian motion and* $U_0 = \int_{[0,\cdot)}(\bar F_0\bar G_0)_-^{-1}\, dH_0$.

(ii) $\sqrt{n}(\mathrm{E}[H(t)|\, D_n] - \hat H_n(t)) = O_P(n^{-(\alpha-1/2)})$ *a.s.* $[P^\infty_{F_0,G_0}]$.

Consequently, if $\alpha > 1/2$, then $\sqrt{n}(H - \hat H_n)|\, D_n \rightsquigarrow B \circ U_0$ a.s.$[P^\infty_{F_0,G_0}]$.

Proof The posterior distribution is the law of an independent increment process with intensity measure consisting of a Poisson and a fixed jump part, given in Theorem 13.15. The independent increment process is the sum $H = H^p + H^f$ of two independent increment processes corresponding to these two components. We show that the continuous part and its (posterior) mean tend to zero at high speed in n, whereas the discrete part gives rise to the limit law.

By (13.18) and (13.19) (and Theorem 13.15), the posterior mean $\mathrm{E}[H^p(t)|\, D_n]$ and variance $\mathrm{var}[H^p(t)|\, D_n]$ are given by $\int_0^t\int_0^1 s^k(1-s)^{Y_n(x)}q(x, s)\, ds\, dx$, for $k = 0$ and $k = 1$, respectively. From the assumption on the function and the fact that $n^{-1}Y_n$ tends to infinity almost surely, it follows that these integrals are of the order $\int_0^1 s^k(1-s)^{cn}\, ds$, for some $c > 0$, which is $O(n^{-1-k})$. It follows that $\sqrt{n}H^p(t)$ and $\sqrt{n}\mathrm{E}[H^p(t)|\, D_n]$ tend to zero almost surely. Since both processes are nondecreasing, the pointwise convergence at $t = \tau$ actually gives uniform convergence in $\mathfrak{D}[0, \tau]$.

The fixed jump part H^f is a sum of finitely many independent random variables, one at each uncensored survival time. We prove its asymptotic normality (conditional given D_n) using Lyapunov's central limit theorem, applied with the fourth moment (see e.g. Billingsley 1979, Theorem 27.3). The conditional distribution of $\Delta H(x)$ is given by $\nu^d_{H|D_n}$ in Theorem 13.15, where we can take $\Delta N = 1$ by the assumed continuity of F_0. The fourth moments given D_n are given by

$$\mathrm{E}\big[|\Delta H(x)|^4|\, D_n\big] = \frac{\int_0^1 s^4(1-s)^{Y_n(x)-1}q(x, s)\, ds}{\int_0^1 (1-s)^{Y_n(x)-1}q(x, s)\, ds}.$$

By the same arguments as in the proof of Theorem 13.21 this expression can be seen to be equal to $Y_n(x)^{-4}\Gamma(5)$ up to lower order terms. Since $n^{-1}Y_n$ tends almost surely to a positive limit, the sum of the fourth moments over the uncensored survival times x is of order $O(n^{-3})$. Thus $\sum_x \sqrt{n}(\Delta H(x) - \mathrm{E}[\Delta H(x)|\, D_n])$ easily satisfies Lyapunov's condition.

The variance $\mathrm{var}[\Delta H(x)|\, D_n]$ is given by the second integral in (13.24). Extending the argument we see that the two quotients inside the integral are asymptotic to $2Y_n(x)^{-2}$ and

$Y_n(x)^{-1}$, respectively, so that the difference of the first minus the square of the second is equivalent to $Y_n(x)^{-2}$. This shows that

$$\sum_{x:x\le t}\int_{(0,t]}\operatorname{var}[\sqrt{n}\Delta H(x)|\, D_n] = \int_{(0,t]}\frac{n}{Y_n^2}\,dN_n + o(1) \to \int_{(0,t]}\frac{G_{0-}\,dF_0}{(F_{0-}G_{0-})^2} = U_0(t).$$

This convergence also applies to increments of the process H^f. By the Lyapounov central limit theorem the posterior process $\sqrt{n}(H^f(t) - \mathrm{E}[H^f(t)|\, D_n])$ converges marginally given D_n to a Gaussian independent increment process with variance process U_0, i.e. the process $B\circ U_0$, almost surely.

The convergence of the variances of the increments is uniform in t, and the limiting variance process U_0 is continuous. Together with the independence of the increments this gives tightness in view of Theorem V.19 of Pollard (1984). Thus assertion (i) is established.

To prove assertion (ii) we refine the approximation for the posterior mean of $H^f(t)$ in (13.23). Letting $r(x,s) = q(x,s)/q_0(x) - 1$, so that $|r(x,s)| \le Cs^\alpha$ for $s\to 0$, we can approximate $\mathrm{E}[H_n|\, D_n] - \hat{H}_n$ as, for $\epsilon_n \to 0$ sufficiently slowly,

$$\int\frac{1}{Y_n}\Big[\frac{\int_0^{Y_n\epsilon_n} u(1-u/Y_n)^{Y_n-1}(1+r(\cdot,u/Y_n))\,du + O(n^{-3})}{\int_0^{Y_n\epsilon_n}(1-u/Y_n)^{Y_n-1}(1+r(\cdot,u/Y_n))\,du + O(n^{-3})} - 1\Big]\,dN_n.$$

Because the variation of $\int Y_n^{-1}\,dN_n$ is uniformly bounded, it suffices to show that $\sqrt{n}$ times the supremum over the omitted argument x of the expression within square brackets tends to zero in probability. Since $\int_0^{y\epsilon} u^k(1-u/y)^{y-1}\,du = \Gamma(k+1) + O(1/y)$ if $y\to\infty$ such that $e^{-y\epsilon} \gtrsim y^{-1}$, after replacing r by its upper bound the numerator and denominator of the quotient expand as $\Gamma(2) + \Gamma(2+\alpha)/Y_n^\alpha + O(1/Y_n)$ and $\Gamma(1) + \Gamma(1+\alpha)/Y_n^\alpha + O(1/Y_n)$, respectively. Their quotient differs from 1 by a terms of the order $O(1/Y_n^\alpha) = O(n^{-\alpha})$, almost surely. □

Corollary 13.24 *Under the conditions of Theorem 13.23 with* $\alpha > 1/2$, *we have* $\sqrt{n}(\bar{F} - \hat{\bar{F}}_n)|\, D_n) \rightsquigarrow -\bar{F}_0\, B\circ U_0$ *in* $[P^\infty_{F_0,G_0}]$*-probability, where* $\hat{\bar{F}}_n$ *is the Kaplan-Meier estimator for the survival function.*

Proof This is an immediate consequence of the Hadamard differentiability of the product integral $H \mapsto \bar{F} = \prod(1-dH)$ and the delta-method for conditional distributions (Theorem 3.9.11 in van der Vaart and Wellner 1996). □

Remark 13.25 If $\frac{\partial}{\partial s}g(x,s)$ exists and is bounded in a (right) neighborhood of 0, then the function $q_0(x) = q(x,0)$ satisfies the condition in Theorem 13.23, with $\alpha = 1$.

Example 13.26 (Beta process) The beta process prior on the cumulative hazard function with parameter (c,Λ) possesses q-function given by $q(x,s) = c(x)(1-s)^{c(x)-1}$. If c is continuous and bounded away from 0 and ∞, then q satisfies the conditions of Theorem 13.23 with $\alpha = 1$.

It suffices to note that $(1-s)^{c-1} = 1-(c-1)s+o(s)$, as $s\to 0$, uniformly in c belonging to a bounded interval in $(0,\infty)$, whence $q(x,s) = c(x) + O(s)$, as $s\to 0$.

Example 13.27 (Dirichlet process) By Example 13.11 the cumulative hazard function H corresponding to $F \sim \mathrm{DP}(MF_0)$ is a beta process with parameters parameters $c = M\bar{F}_{0-}$ and H_0. By the preceding example the conditions of Theorem 13.23 are satisfied if H_0 is continuous.

For this special example the Bernstein–von Mises theorem for the survival function was earlier obtained in a more abstract setting in Theorem 12.2 (but without censoring).

Example 13.28 (Extended gamma process) The extended gamma process prior with parameters (c, A_0) is described in Examples 13.13 and 13.22. If the function c in its intensity function (13.25) is continuous and bounded away from zero and infinity on $[0, \tau]$, then this satisfies the conditions of Theorem 13.23. See Example 13.22.

13.5 Smooth Hazard Processes

In view of the discreteness of its sample paths, an independent increment process is not an adequate model for a smooth cumulative hazard function. Smoothing through a kernel may be natural, yielding analogs of Dirichlet mixture processes. The resulting priors for smooth hazard functions may be useful even outside the setting of survival analysis.

For a given measurable *kernel function* $k\colon [0,\infty) \times (0,\infty) \to \mathbb{R}^+$ and a finite Borel measure Φ on $(0,\infty)$, consider a hazard function h_Φ of the form

$$h_\Phi(t) = \int k(t, v)\, d\Phi(v).$$

We do not require that $k(\cdot, v)$ is a probability density or that Φ is a probability measure, but do assume that the mixture is finite for all t. We obtain a prior on hazard functions by placing a prior on the mixing measure Φ. Specifically, we consider processes with independent nonnegative increments Φ, identified with measures through their cumulative distribution function, as discussed in Section 13.4. Although we take the mixing variable v in this section to be a positive number, the setup can be extended to general Polish spaces with Φ a general completely random measure.

The cumulative hazard function for h_Φ is a mixture over the integrated kernel:

$$H_\Phi(t) = \int K(t, v)\, d\Phi(v), \qquad K(t, v) := \int_0^t k(u, v)\, du.$$

The corresponding survival function and probability densities are $1 - F_\Phi(t) = e^{-H_\Phi(t)}$ and $f_\Phi(t) = h_\Phi(t) e^{-H_\Phi(t)}$. If $K(t, v) \to \infty$ as $t \to \infty$ for every v, then $H_\Phi(t) \to \infty$ as $t \to \infty$, and F_Φ is a proper probability distribution on $(0, \infty)$.

Common choices of kernels are

(i) Dykstra-Laud (DL) kernel: $k(t, v) = \mathbb{1}\{t \geq v\}$;
(ii) Rectangular kernel: $k(t, v) = \mathbb{1}\{|v - t| \leq \tau\}$, for "bandwidth" $\tau > 0$;
(iii) Ornstein-Uhlenbeck (OU) kernel: $k(t, v) = 2\kappa e^{-\kappa(t-v)} \mathbb{1}\{t \geq v\}$;
(iv) Exponential kernel: $k(t, v) = v^{-1} e^{-t/v}$.

The Dykstra-Laud kernel generates increasing hazard functions, whereas the exponential kernel makes the hazard decreasing. The rectangular kernel is a "local smoother," whereas the Dykstra-Laud and Ornstein-Uhlenbeck kernels smooth “over the entire past,” and the exponential kernel smooths “over the entire time axis.” Among these kernels, only the exponential kernel gives infinitely smooth sample paths.

Example 13.29 (Dykstra-Laud and extended gamma) Within the present class of priors on hazard functions, the Dykstra-Laud kernel combined with the extended gamma process (see Example J.15) as the mixing measure is particularly tractable. The extended gamma process can be obtained as $\Phi(v) = \int_{(0,v]} b\, d\xi$ from a gamma process ξ and a given positive function b. Combination with the Dykstra-Laud kernel yields the prior hazard function $h_\Phi(t) = \int_{(0,t]} b\, d\xi$ and hence $\mathrm{E}[h_\Phi(t)] = \int_{(0,t]} b\, d\alpha$ and $\mathrm{var}[h_\Phi(t)] = \int_{(0,t]} b^2\, d\alpha$, for $\alpha(t) = \mathrm{E}[\xi(t)]$ the mean function of ξ. The explicit expressions for prior expectation and variance are helpful for eliciting b and α. The cumulative hazard function can be similarly expressed in the integrated kernel as $H_\Phi(t) = \int K(t, v)\, b(v)\, d\xi(v)$. For the Dykstra-Laud kernel we have $K(t, v) = (t - v)_+$, whence by Example J.7, for $\theta > 0$,

$$\log \mathrm{E}[e^{-\theta H_\Phi(t)}] = -\int_0^t \log\left(1 + \theta(t - v)_+ b(v)\right) d\alpha(v). \tag{13.26}$$

The prior mean $\mathrm{E}[\bar{F}_\Phi(t)]$ of the survival function is obtained by setting $\theta = 1$ in this formula and exponentiating.

The tractability of the Laplace transform of H_Φ allows us to derive a formula for the posterior distribution given (censored) data from F_Φ, which expresses the cumulative hazard function as a mixture of extended gamma processes, with varying α measures. This representation next leads to an expression for the posterior expectation of $h_\Phi(t)$, similar to the one in Proposition 4.7 for the Dirichlet process. The large number of terms makes the result difficult to apply. Simulation methods, which are also available for other kernels and priors, often are preferable.[8]

Consider the posterior distribution based on random right censored data D_n when the cumulative hazard function of the survival times is a priori modelled as H_Φ. Hence, D_n consists of a sample $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, where T_i is the minimum of a survival X_i time and a censoring time, Δ_i indicates censoring, and given Φ the survival times $X_1, \ldots, X_n$ are i.i.d. with cumulative hazard function H_Φ. It is convenient to treat the mixing random measure Φ rather than the hazard function as the primary parameter. The likelihood of Φ given the censored data D_n can be written as

$$\exp\Big[-\sum_{i=1}^n \int K(T_i, v)\, d\Phi(v)\Big] \prod_{i=1:\Delta_i=1}^n \int k(T_i, v)\, d\Phi(v).$$

The integrals with respect to v can be eliminated if we introduce latent variables $V_1, \ldots, V_n$ “realized from the CRM Φ” as additional observations. Since Φ is a CRM, and hence is a.s. discrete, there will typically be repetitions in $V_1, \ldots, V_n$. Let $\tilde{V}_1, \ldots, \tilde{V}_k$ stand for the distinct values, appearing with multiplicities $N_1, \ldots, N_k$. They define a partition of

[8] See Theorem 3.3 of Dykstra and Laud (1981) for details.

$\{1, 2, \ldots, n\}$, given by the sets $S_j = \{i : V_i = \tilde{V}_j\}$, for $j = 1, \ldots, k$. The following theorem describes the posterior distribution of Φ in a form that is also suitable for MCMC computations.

Theorem 13.30 (Posterior distribution) *If the prior for Φ follows a CRM with intensity measure $\nu(dv, ds) = \rho(ds|\, v)\, \alpha(dv)$, where α is a σ-finite measure on $\mathbb{R}^+$, then the posterior distribution of Φ given D_n is described as follows, where $K_n(v) = \sum_{i=1}^n K(T_i, v)$ and $\tau_m(v) = \int_0^\infty s^m e^{-sK_n(v)}\, \rho(ds|\, v)$:*

(i) *The conditional distribution of Φ given D_n and the auxiliary variables $V_1, \ldots, V_n$ is equal to the distribution of the CRM with intensity measure*

$$\nu^c_{\Phi|D_n,V}(dv, ds) = e^{-sK_n(v)} \rho(ds|\, v)\, \alpha(dv),$$
$$\nu^d_{\Phi|D_n,V}(\{\tilde{V}_j\}, ds) \propto s^{N_j} e^{-sK_n(\tilde{V}_j)} \rho(ds|\, \tilde{V}_j).$$

(ii-1) *The conditional distribution of $(\tilde{V}_1, \ldots, \tilde{V}_k)$ given D_n and configuration $(k, S_1, \ldots, S_k)$ has density proportional to*

$$(v_1, \ldots, v_k) \propto \prod_{j=1}^k \tau_{N_j}(v_j) \prod_{i \in S_j, \Delta_i = 1} k(T_i, v_j).$$

(ii-2) *The probability of the configuration $(k, S_1, \ldots, S_k)$ given D_n is proportional to*

$$\prod_{j=1}^k \int \tau_{N_j}(v) \prod_{i \in S_j, \Delta_i = 1} k(T_i, v)\, \alpha(dv).$$

Proof See Theorem 4.1 of James (2005). The proof uses arguments similar to those used in the proofs of Theorem 14.56 and Lemma 14.62 and Theorem 5.3 (for part (ii)). □

Based on corresponding results on Dirichlet mixtures one may expect the posterior distribution of a smooth hazard prior to be consistent if the true hazard is also smooth, even at a rate, that will depend on a bandwidth in the kernel. The following theorem is only a first step in this direction: it gives weak consistency for the cumulative hazard function (or, equivalently, the survival function). Consistency is of course limited to an interval within the range of the censoring variables. We assume that the censoring distribution G is supported on a compact interval within the support of the survival function and then obtain consistency for the survival function on the same interval.

Theorem 13.31 (Consistency) *Suppose that the censoring distribution G is supported on $[0, \tau]$ with density bounded away from zero. Consider a smooth hazard process h_Φ for a CRM Φ such that $\liminf_{t \downarrow 0} t^{-r} h_\Phi(t) = \infty$ almost surely $[\Pi]$ for some $r > 0$ and such that $\Pi(\Phi\colon \sup_{0<t\leq\tau} |h_\Phi(t) - h_0(t)| < \delta) > 0$, for all $\delta > 0$. Then the posterior distribution of the cumulative distribution function F is consistent with respect to the Kolmogorov-Smirnov distance on $[0, \tau]$ at any F_0 with $F_0(\tau) < 1$ and with a hazard function h_0 that is bounded away from zero on every interval $[\delta, \tau]$, for every $\delta > 0$, and is such that $\int_0^1 (\log h_0(t) + \log\_\, t)\, dF_0(t) < \infty$.*

Proof The distribution P_F of an observation (T, Δ) on $(0, \infty) \times \{0, 1\}$ has a density $p_F(t, \delta) = f(t)^{\delta} \bar{F}(t)^{1-\delta} g(t)^{1-\delta} \bar{G}(t)^{\delta}$ with respect to the product of Lebesgue measure and counting measure. Here we consider the censoring distribution G and density g as fixed and given. Weak convergence of a sequence P_{F_m} to P_{F_0} can be seen to imply the weak convergence of F_m to F_0 in the space of subprobability measures on $[0, \tau]$. Since F_0 is continuous by assumption, this next implies the uniform convergence of F_m to F_0 on $[0, \tau]$, by Pólya's theorem. Therefore it is sufficient to show the weak consistency of the posterior of P_F. By Example 6.20 this is ensured by the Kullback-Leibler property of the prior of P_F at P_{F_0}.

The censoring model is an information loss model, of the type considered in Lemma B.11, where the distribution F of X is transformed to the distribution P_F of the observation (X, Δ). Since G is supported on $[0, \tau]$ every observation is censored at τ and hence the full data model can also be taken to have survival time $X \wedge \tau$ instead of X; in other words rather than $X \sim F$ we can take $X \sim \tilde{F}$, for $\tilde{F}$ the mixture distribution of the density f on $[0, \tau)$ and a point mass of size $1 - F(\tau)$ at τ. Then it follows, for h_0, h, H_0, H the hazard functions and cumulative hazard functions corresponding to F_0, F,

$$\begin{aligned} K(P_{F_0}; P_F) \le K(\tilde{F}_0; \tilde{F}) &= \int_{[0,\tau)} \log\frac{f_0}{f}\, dF_0 + \log\frac{1 - F_0(\tau)}{1 - F(\tau)}\, \bar{F}_0(\tau) \\ &= \int_0^{\tau} \log\frac{h_0}{h}\, dF_0 - \int_0^{\infty} (H_0 - H)(t \wedge \tau)\, dF_0(t). \end{aligned}$$

It suffices to show that the right side evaluated at $h = h_{\Phi}$ is bounded above by ϵ with positive probability under the prior of Φ, for every $\epsilon > 0$.

For given positive numbers δ, η let $\mathfrak{H}(\delta, \eta)$ be the set of all hazard functions h such that $\inf_{t\in[0,\delta]} (h(t)t^{-r}) \ge 1$ and $\sup_{t\in[0,\tau]} |h(t) - h_0(t)| < \eta$. For $h \in \mathfrak{H}(\delta, \eta)$ we have $|\log(h_0/h)| \le |h - h_0|/(h \wedge h_0) \le \eta/(c_0(\delta, \tau) - \eta)$, on the interval $[\delta, \tau]$, for $c_0(\delta, \tau)$ the minimum value of h_0 on the interval. Furthermore, on the same interval $|H - H_0| \le \eta\tau$. By splitting the integral in the expression for $K(\tilde{F}_0; \tilde{F})$ over the ranges $(0, \delta)$ and $[\delta, \tau]$, we see that for $h \in \mathfrak{H}(\delta, \eta)$ (where $\delta < 1$):

$$K(\tilde{F}_0; \tilde{F}) \le \int_0^{\delta} (\log h_0)\, dF_0 + r \int_0^{\delta} \log_{-} t\, dF_0(t) + \int_{\delta}^{\tau} \frac{\eta}{c_0(\delta, T) - \eta}\, dF_0 + \eta\tau.$$

The first and second integrals on the right become smaller than an arbitrary positive constant, as $\delta \downarrow 0$. For given δ the third integral and the last term become arbitrarily small if η is small. We conclude that the Kulback-Leibler property holds at P_{F_0} if h_{Φ} belongs with positive prior probability to $\mathfrak{H}(\delta, \eta)$, for sufficiently small $\delta, \eta > 0$.

By assumption $\liminf_{t\downarrow 0} h_{\Phi}(t)t^{-r} = \infty$ almost surely under the prior. Therefore the event $\cup_{\delta>0} A_{\delta}$, for A_{δ} the event that $h_{\Phi}(t) \ge t^r$ for every $t \in (0, \delta]$, has prior probability equal to 1. This implies that the prior probabilities of the events $h_{\Phi} \in \mathfrak{H}(\delta, \eta)$ tend to the prior probability that $\sup_{t\in[0,\tau]} |h_{\Phi}(t) - h_0(t)| < \eta$. Since the latter is positive by assumption, the proof is complete. □

Both the Dykstra-Laud and Ornstein-Uhlenbeck kernels satisfy $h_{\Phi}(t) \gtrsim \Phi((0, t])$. Then the condition that $\liminf_{t\downarrow 0} t^{-r} h_{\Phi}(t) = \infty$ almost surely is implied by the similar condition on the CRM that $\liminf_{t\downarrow 0} t^{-r} \Phi((0, t]) = \infty$ almost surely. This holds for the generalized extended gamma process with intensity measure $\nu(dv, ds) = (\Gamma(1 - \sigma))^{-1} s^{-(1+\sigma)} e^{-c(v)s}\, dv\, ds$. For the other two kernels $h_\Phi(0) > 0$ almost surely and hence the required condition holds automatically.

13.6 Proportional Hazard Model

The *proportional hazard model* or *Cox model* is frequently used to investigate the dependence of survival on a covariate. It was considered previously in Section 12.3.3. The Cox model postulates that the hazard function of an individual characterized by the covariate vector $z \in \mathbb{R}^d$ is equal to $e^{\beta^\top z}$ times a *baseline hazard function*. Presently we study priors on the cumulative hazard function and do not assume existence of a hazard function. In this situation one possible definition of proportional hazards is that the cumulative hazard function of the survival time X given the covariate Z takes the form $x \mapsto e^{\beta^\top Z} H(x)$, for an unknown "baseline cumulative hazard function" H. For a function H with jumps this would be awkward, as the jumps of a cumulative hazard function are bounded by 1, which would only be achieved by limiting the jumps of H by the minimum of the multiplicative factors $e^{-\beta^\top Z}$. A definition of proportional hazards that avoids this difficulty is that the negative log-conditional survival function (in the preceding denoted by the symbol A instead of H) satisfies the proportionality requirement: for a baseline negative log-survival function A we postulate that the survival distribution of X given Z satisfies

$$-\log\left(1 - F(x|\, Z)\right) = e^{\beta^\top Z} A(x). \tag{13.27}$$

For large samples from a continuous survival distribution, the difference between these possible definitions should be minor as the posterior would place most of its weight on cumulative hazard functions with small jumps. Because it is conceptually cleaner, in the following we adopt the second definition, as given in the preceding display.

We allow for right censoring, and take the data as $D_n = \{(T_i, \Delta_i, Z_i) \colon i = 1, \ldots, n\}$, for $T_i = \min(X_i, C_i)$ and $\Delta_i = \mathbb{1}\{X_i \le C_i\}$ and a random sample of triplets (X_i, C_i, Z_i) of a survival time, censoring time and covariate.

We assume that the survival time X and censoring time C are conditionally independent given the covariate Z. We choose a prior such that the pair of parameters (θ, H), the conditional distribution of C given Z, and the marginal distribution of Z are a priori independent. Then the priors of the latter two components do not enter the posterior distribution of (θ, H) and will be left unspecified. We consider an independent increment process as a prior on the baseline cumulative hazard function H, given by an intensity measure of the form $\nu_H(dt, ds) = g(x, s)\, dx\, ds$, and an independent prior with density π on β. By Theorem 13.8 the first is equivalent to choosing an independent increment process prior on the function A.

13.6.1 Posterior Distribution

The expression for the posterior distribution in the Cox model extends Theorem 13.15, to which it reduces if no covariate is present.

Theorem 13.32 (Cox posterior) *If the prior on H is an independent increment process with intensity measure $\nu_H(dx, ds) = g(x, s)\, dx\, ds$ independent of θ for a function g that is continuous in x, then the conditional posterior distribution of H given β and the data D_n is the law of an independent increment process with intensity measure*

$$\nu^c_{H|D_n,\beta}(dx, ds) = \prod_{i:T_i \geq x} (1-s)^{e^{\beta^\top Z_i}}\, g(x, s)\, dx\, ds,$$

$$\nu^d_{H|D_n,\beta}(\{x\}, ds) \propto \prod_{i:T_i=x,\Delta_i=1} [1-(1-s)^{e^{\beta^\top Z_i}}] \prod_{i\in R^+_n(x)} (1-s)^{e^{\beta^\top Z_i}} g(x, s)\, ds,$$

where the fixed jump measures are probability measures on $(0, 1)$*, the fixed jump times range over the set* $\{T_i \colon \Delta_i = 1\}$ *of uncensored observations, and* $R^+_n(t) = \{i \colon T_i \geq t\} \setminus \{i \colon T_i = t, \Delta_i = 1\}$*. Furthermore, the marginal posterior density of β given D_n satisfies*

$$\pi(\beta|\, D_n) \propto \pi(\beta) e^{-\rho_n(\beta)} \prod_{x:\Delta N(x)>0} k(x, \beta),$$

where $k(x, \beta)$ is the norming constant for the the jump measure $\nu^d_{H|D_n,\beta}(\{x\}, ds)$ as given (the integral over $(0, 1)$ *of the right side of the proportionality equation), and*

$$\rho_n(\beta) = \sum_{i=1}^n \int_0^{T_i}\int_0^1 [1-(1-s)^{e^{\beta^\top Z_i}}] \prod_{j<i:T_j\geq x} (1-s)^{e^{\beta^\top Z_j}} g(x, s)\, dx\, ds.$$

Proof By adding the observations one-by-one we can reduce to the case of a single observation (T, Δ, Z), provided we allow a prior for H of the postulated form of the posterior distribution for H. Thus given β the cumulative hazard function is an independent increment process with intensity measure $\nu_{H|\beta}$ possibly depending on β, and it may contain a fixed jump part.

In the first step we condition on β and Z and hence the observation is equivalent to just a single observation (T, Δ) in the random censoring model. In view of our definition of proportional hazards, it is convenient to parameterize the model by $A = -\log \bar{F}$, for $\bar{F}$ the baseline survival function. By (13.27) the negative log survival function given β and Z of the survival time X is equal to $e(\beta)A$, for $e(\beta) = e^{\beta^\top Z}$ and hence given β and Z is a priori an independent increment process with intensity measure given by $\nu_{e(\beta)A|\beta}(dx, ds) = \nu_{A|\beta}(dx, ds/e(\beta))$. By Theorem 13.15, the posterior distribution of $e(\beta)A$ (given β and Z and (T, Δ)) is also an independent increment process, and its intensity process can be expressed in $\nu_{e(\beta)A|\beta}$. This can next be translated back to see that the posterior distributions of A and H are independent increment processes, where the intensity measure for H is given in the statement of the theorem.

For the derivation of the marginal posterior distribution of β we retake the proof of Theorem 13.15, which consists of discretizing the observation to a multinomial vector M and taking limits as the discretization becomes fully informative. The likelihood for M takes the form $L_M(A, \beta) = \prod_{i=1}^{m+1} (e^{-A(t_{i-1})e(\beta)} - e^{-A(t_i)e(\beta)})^{M_i}$. By Bayes's rule a version of the posterior density of β is given by

$$\pi(\beta|\, M) = \frac{\mathrm{E}_{A|\beta} L_M(A, \beta)\pi(\beta)}{\mathrm{E}_\beta \mathrm{E}_{A|\beta} L_M(A, \beta)} \propto \mathrm{E}_{A|\beta} L_M(A, \beta)\, \pi(\beta).$$

Here the expectations $\mathrm{E}_{A|\beta}$ are relative to the prior distribution of A given β, for fixed M. On the event that the jth coordinate of M is nonzero, the right side is equal to, by two applications of formula (13.17) and arguing as in the proof of Theorem 13.15,

$$e^{-\int_{(0,t_{j-1}]}\int (1-e^{-e(\beta)s})\, \nu^c_{A|\beta}(dx,ds)} \prod_{x\le t_{j-1}} \int e^{-e(\beta)s}\, \nu^d_{A|\beta}(\{x\}, ds)$$
$$\times \Bigl[1 - e^{-\int_{(t_{j-1},t_j]}\int (1-e^{-e(\beta)s})\, \nu^c_{A|\beta}(dx,ds)} \prod_{t_{j-1}<x\le t_j} \int e^{-e(\beta)s}\, \nu^d_{A}(\{x\}, ds)\Bigr].$$

If the observation is censored, then the partitions are constructed so that $t_{j-1} = C$ and $t_j = \infty$. The expression is fixed and no refinement limit need be taken. If the observation is uncensored, then the leading term, outside brackets, tends to the same expression but with t_{j-1} replaced by its limit T. For the limit of the term within square brackets, we split in two cases. If the observation is uncensored and equal to a fixed jump time of $\nu_{A|\beta}$, then the exponential within square brackets tends to $e^{-0} = 1$ and the product is eventually equal to the product over the single fixed jump time T. If the observation is uncensored and not equal to a fixed jump time of $\nu_{A|\beta}$, then the interval $(t_{j-1}, t_j]$ will eventually be free of fixed jump times and the product inside square brackets should be read as 1. Because the integral in the exponent tends to zero, the term in square brackets can be expanded as

$$\int_{(t_{j-1},t_j]}\int (1 - e^{-e(\beta)s})\, \nu^c_{A|\beta}(dx, ds)(1 + o(1)).$$

In the sequential updating scheme the intensity measure $\nu^c_{A|\beta}$ will correspond to the conditional posterior of A based on the preceding observations and hence possess density $(x, s) \mapsto \prod_{i:T_i\ge x} e^{-se_i(\beta)}$ relative to the original prior intensity ν_A, which is free of β. As a function of x it changes values only at the T_i. If T is not an existing fixed jump, then there is no other uncensored observation in $(t_{j-1}, t_j]$, eventually. If we construct the partitions to include the preceding censored times, then a censored T_i in $(t_{j-1}, t_j]$ will be necessarily equal to t_j. In that case the density is constant in its first argument and equal to $q(s, \beta) = \prod_{i:T_i\ge T} e^{-se_i(\beta)}$ throughout the interval. The preceding display becomes

$$\int_{(t_{j-1},t_j]}\int (1 - e^{-e(\beta)s})g(s, \beta)\, \nu^c_{A}(dx, ds)(1 + o(1)).$$

The quotient of this expression with its value at $\beta = 0$ is asymptotically proportional (as a function of β) to $\int (1 - e^{-e(\beta)s})g(s, \beta)\, \nu^c_A(ds|\, T)$.

This finishes the derivation of the updating formula by Scheffe's theorem, since the expression at $\beta = 0$ does not change with refining discretization. By some bookkeeping, n rounds of updating can be seen to lead to the formulas as claimed. □

Example 13.33 (Extended gamma process) Let the baseline function $A = -\log \bar{F}$ follow an extended gamma process with parameter A_0 such that $dA_0(x) = \lambda(x)\, dx$, as described by Example 13.13, and let B be the corresponding cumulative hazard function. By Theorem 13.32 the intensity measure of H in the posterior is given by

$$\nu^c_{H|D_n,\beta}(dx, ds) = \frac{c(x)}{\log_-(1-s)} \prod_{i\in R_n(x)} (1-s)^{c(x)-1+e^{\beta^\top Z_i}} \lambda(x)\, dx\, ds,$$

$$\nu^d_{H|D_n,\beta}(\{x\}, ds) \propto \frac{1}{\log_-(1-s)} \prod_{i\in D_n(x)} \left[1-(1-s)^{e^{\beta^\top Z_i}}\right] \prod_{i\in R_n^+(x)} (1-s)^{c(x)-1+e^{\beta^\top Z_i}}.$$

The posterior process for H is not extended gamma, since the jump sizes are not gamma distributed.

Example 13.34 (Beta process) For H a beta process prior with parameters (c, Λ), with Λ having density λ, Theorem 13.32 yields the posterior intensity measure

$$\nu^c_{H|D_n,\beta}(dx, ds) = c(x)s^{-1} \prod_{i\in R_n(x)} (1-s)^{c(x)-1+e^{\beta^\top Z_i}} \lambda(x)\, dx\, ds,$$

$$\nu^d_{H|D_n,\beta}(\{x\}, ds) \propto s^{-1} \prod_{i\in D_n(x)} [1-(1-s)^{e^{\beta^\top Z_i}}] \prod_{i\in R_n^+(x)} (1-s)^{c(x)-1+e^{\beta^\top Z_i}}.$$

The marginal posterior density of β is obtained from the expression for $\rho_n(\beta)$ given by

$$\sum_{i=1}^n \int_0^{T_i}\int_0^1 c(x)s^{-1}\left[1-(1-s)^{e^{\beta^\top Z_i}}\right] \prod_{j<i:T_j\ge x} (1-s)^{c(x)-1+e^{\beta^\top Z_j}} \lambda(x)\, dx\, ds.$$

Prior information about the regression coefficient β is often not available, and hence a uniform improper prior for β is natural. Under mild conditions on the covariates and the density of the intensity measure the posterior distribution is proper. The conditions on the prior intensity measure in the following proposition are satisfied by the gamma and beta processes. The remaining condition is a form of “linear independence of the covariates.” This is also used to show the uniqueness of the maximum likelihood estimator and the log-concavity of the partial likelihood function (see Andersen et al. 1993).

Proposition 13.35 *If the function $(x, s) \mapsto s(1-s)^{1-c_1} g(x, s)$ is bounded over $[0, T_{(n)}] \times [0, 1]$ and $s\, g(x, s) \ge M(1-s)^{c_2-1} a_0(x)$ for some $M, c_1, c_2 > 0$ and a continuous function a_0, and every $z \in \mathbb{R}^d$ is expressible as $\sum_{k:\Delta_k=1}\sum_{i:T_i=T_k,\Delta_i=1}\sum_{j\in R_n^+(T_k)} \lambda_{ijk}(Z_i - Z_j)$ for some choice of $\lambda_{ijk} \ge 0$, then the posterior for β under the improper uniform prior is proper.*

Proof We only describe the main idea behind the proof, and refer to Kim and Lee (2003a) for the details. Under the given conditions the posterior density of β can be bounded by a constant multiple of the minimum over $\exp\left[\beta^\top(Z_i - Z_j)\right] \wedge 1$ for all pairs (i, j) such that either $\Delta_i = 1$ and $T_j > T_i$ or $T_j = T_i$ and $\Delta_j = 0$. By the third condition of the proposition, it follows that the posterior has exponentially decaying tail in every direction. □

13.6.2 Bernstein–von Mises Theorem

For large samples the joint posterior distribution of β and H obtained in Theorem 13.32 possesses a Gaussian approximation, under a similar condition on the small jump sizes in the intensity measure of the cumulative hazard function as in Theorem 13.23. When restricted to the marginal posterior distribution of β this gives a Bernstein–von Mises theorem in the spirit of Section 12.3.3, but using an independent increment process prior on the baseline hazard. As this prior is not supported on a dominated model, the techniques developed in Chapter 12 are not applicable, but the theorem is derived using the explicit, conjugate form of the posterior distribution established in Theorem 13.32.

The semiparametric maximum likelihood estimator for (β, H) are the maximizer $\hat\beta_n$ of *Cox partial likelihood*

$$L_n(\beta) = \prod_{\substack{i=1\\ \Delta_i=1}}^{n} \frac{\exp(\beta^\top Z_i)}{\sum_{j:T_j\geq T_i}\exp(\beta^\top Z_j)}. \tag{13.28}$$

and *Breslow's estimator*

$$\hat H_n(t) = \int_{(0,t]}\Bigl(\sum_{i:T_i\geq x} e^{\hat\beta_n^\top Z_i}\Bigr)^{-1}\,dN(x).$$

These are known to be asymptotically normally distributed in the sense that $\sqrt{n}(\hat\beta_n - \beta_0, \hat H_n - H_0) \rightsquigarrow (V, W\circ U_0 - V^\top e_0)$, for $V \sim \mathrm{Nor}_d(0, \tilde I_\beta^{-1})$ and W a standard Brownian motion independent of V, where

$$\begin{aligned}
U_0(t) &= \int_{(0,t]} \frac{1}{\mathrm{E}_0[e^{\beta^\top Z}\mathbb{1}\{T\geq x\}]}\,dH_0(x),\\
e_0(t) &= \int_{(0,t]} \frac{\mathrm{E}_0[Ze^{\beta^\top Z}\mathbb{1}\{T\geq x\}]}{\mathrm{E}_0[e^{\beta^\top Z}\mathbb{1}\{T\geq x\}]}\,dH_0(x),\\
\tilde I_\beta &= \int_{(0,\tau]} \mathrm{E}_0\bigl[(ZZ^\top - e_0(x)e_0(x)^\top)e^{\beta^\top Z}\mathbb{1}\{T\geq x\}\bigr]\,dH_0(x).
\end{aligned} \tag{13.29}$$

The last quantity is the *efficient Fisher information* for estimating β.

Theorem 13.36 (Bernstein–von Mises) *Assume that the covariates lie in a bounded subset of $\mathbb{R}^d$ and are not concentrated in any lower-dimensional subset, and assume that the prior density on β is positive and continuous at β_0. If the prior for H follows an independent increment process with intensity measure of the form $\nu_H(dx, ds) = s^{-1}q(x, s)\,dx\,ds$ for a function q that is continuous in x and such that $\sup_{x\in[0,\tau],s\in(0,1)}(1-s)q(x,s) < \infty$ and $\sup_{x\in[0,\tau]}|q(x,s) - q_0(x)| = O(s^\alpha)$ as $s\to 0$ for a function q_0 that is bounded away from zero and infinity and some $\alpha > 1/2$, then for τ such that $(\bar F_0\bar G_0)(\tau-) > 0$ and $G_0(\tau) = 1$, and continuous F_0:*

$$\sqrt{n}(\beta - \hat\beta_n, H - \hat H_n)|\,D_n \rightsquigarrow (V, W\circ U_0 - V^\top e_0), \qquad a.s.\ [P^\infty_{\beta_0,H_0}],$$

where $V \sim \mathrm{Nor}_d(0, \tilde I_\beta^{-1})$ and W is a standard Brownian motion independent of V, and the convergence is in the product of the $\mathbb{R}^d$ and the Skorohod space $\mathfrak{D}[0,\tau]$ equipped with the uniform norm.

Proof The proof of the theorem is long; we refer to Kim (2006) for details. The proof can be based on the characterization of the posterior distribution given in Theorem 13.32. In particular, the theorem gives an explicit expression for the marginal posterior density of β. The asymptotic analysis of the posterior density of $\sqrt{n}(\beta - \hat{\beta}_n)$ proceeds as in the proof of the parametric Bernstein–von Mises theorem for smoothly parameterized models, with two main differences. First the log-likelihood is not a sum of independent terms, and more importantly, the difference between the derivatives of the marginal log-likelihood and the partial log-likelihood need to be shown to be uniformly $o(n)$ in order to expand the partial log-likelihood in a Taylor expansion and develop a normal approximation.

The joint asymptotics of $\sqrt{n}(\beta - \hat{\beta}_n)$ and $\sqrt{n}(H - \hat{H}_n)$ can be derived from the characterization of the conditional posterior distribution of $\sqrt{n}(H - \hat{H}_n)$ given $\sqrt{n}(\beta - \hat{\beta}_n)$, using the structure of the independent increment process prior. □

The conditions of the theorem hold in particular for the extended beta and gamma process priors on H; see Examples 13.26 and 13.28.

13.7 The Bayesian Bootstrap for Censored Data

In this section we study analogs of the Bayesian bootstrap appropriate for censored data.

13.7.1 Survival Data without Covariates

The Bayesian bootstrap (BB) was introduced in Section 4.7 for i.i.d. complete data. It can be viewed as a smooth alternative to Efron's bootstrap, as the noninformative limit of the posterior distribution based on a conjugate prior, or as the posterior distribution based on the empirical likelihood with a Dirichlet prior. In this section we extend these three approaches to the censored data setting. Since censored data contain complete data as a special case, it is imperative that the definition reduces to the ordinary Bayesian bootstrap if no observations are censored, but otherwise we allow some freedom of definition.

The Bayesian bootstrap for uncensored data replaces the weights $(n^{-1}, \ldots, n^{-1})$ in the empirical distribution $\sum_{i=1}^n n^{-1}\delta_{X_i}$ by a random vector $(W_1, \ldots, W_n)$ from the $\mathrm{Dir}(n; 1, 1, \ldots, 1)$-distribution, the uniform distribution on the n-simplex. In both cases (and also in the case of Efron's bootstrap, which has $(nW_1, \ldots, nW_n) \sim \mathrm{MN}_n(n; n^{-1}, \ldots, n^{-1})$) the expected weight of every observation is $1/n$. One extension to censored data is to replace the weights in the Kaplan-Meier estimator (13.9), the analog of the empirical distribution, by random variables with the same expectations. This leads to the definition of the Bayesian bootstrap for censored data:

$$\bar{F}^*(t) = \prod_{j:T_j^* \le t} \Big(1 - \frac{\sum_{i:X_i = T_j^*} \Delta_i \Gamma_i}{\sum_{i:X_i \ge T_j^*} \Gamma_i}\Big), \tag{13.30}$$

where $\Gamma_i \overset{\text{iid}}{\sim} \mathrm{Ex}(1)$ and $T_1^* < T_2^* < \cdots < T_k^*$ are the distinct values of the uncensored observations. This definition reduces to the ordinary Bayesian bootstrap when all observations are uncensored; see Problem 13.15. Furthermore, this definition has a stick-breaking representation; see Problem 13.16.

The second approach is to define the Bayesian bootstrap as a noninformative limit of the posterior distribution corresponding to conjugate priors. In the context of survival analysis, beta processes form a natural conjugate family of priors for the cumulative hazard function. In Definition 13.3, this family is defined to have two parameters c and Λ; the noninformative limit is to let $c \to 0$. By Theorem 13.5 in this case the posterior distribution tends to a beta process with parameters Y and $\hat{H}$, for $\hat{H}$ the Nelson-Aalen estimator, given by (13.7). A beta process with a discrete Λ is the cumulative sum of its fixed jump heights, which are beta distributed with parameter $(c\Delta\Lambda, c(1-\Delta\Lambda)$. Since $\hat{H}$ is supported (only) on the uncensored observations and $Y\Delta\hat{H} = \Delta N$, the Bayesian bootstrap should be defined by

$$H^*(t) = \sum_{j=1}^{k} W_j \mathbb{1}\{T_j^* \leq t\}, \tag{13.31}$$

for $W_j \overset{\text{ind}}{\sim} \text{Be}(\Delta N(T_j^*), Y(T_j^*) - \Delta N(T_j^*))$. As in the case of uncensored data, the limit does, not depend on the second parameter Λ of the beta process prior. The corresponding distribution on the survival function $\bar{F}$ coincides with that given by (13.30); see Problem 13.17.

The third approach to the Bayesian bootstrap is to apply Bayes's theorem to the empirical likelihood with a noninformative prior on the empirical likelihood weights. In the uncensored case, letting $F = \sum_{i=1}^n W_i \delta_{X_i}$ and updating the posterior distribution of $(W_1, \ldots, W_n)$ from the improper prior density at $(w_1, \ldots, w_n)$ proportional to $w_1^{-1} \cdots w_n^{-1}$ on $\mathbb{S}_n$ to the posterior proportional to $(\prod_{i=1}^n w_i) \times (\prod_{i=1}^n w_i^{-1}) \equiv 1$ on $\mathbb{S}_n$, we obtain the Bayesian bootstrap distribution. This can be modifed to censored data, as we illustrate for the proportional hazard model in the next section.

The second interpretation of the Bayesian bootstrap shows that the Bayesian bootstrap is a special case of the beta process posterior. Hence, in view of Theorem 13.23, it gives rise to a Bernstein–von Mises theorem; see Problem 13.18 for a precise statement.

13.7.2 Cox Proportional Hazard Model

A probability density f and its survival function $\bar{F}$ can be expressed in the corresponding hazard and cumulative hazard functions as $f = he^{-H}$ and $\bar{F} = e^{-H}$. As in the Cox model these hazard functions are multiplied by $e^{\beta^\top Z}$, the likelihood of a single right censored observation (T, Δ, Z) in this model is $e^{\beta^\top Z} h(T) e^{-e^{\beta^\top Z} H(T)}$ if the observation is complete $(\Delta = 1)$ and $e^{-e^{\beta^\top Z} H(T)}$ otherwise. This likelihood makes sense in the absolutely continuous case, but to define a Bayesian bootstrap we require an *empirical likelihood*, which is valid also for general distributions. There are two possibilities in the literature, the Poisson and the binomial likelihood.

For the Poisson empirical likelihood, the baseline hazard function h is replaced by the jump ΔH of the cumulative hazard function, leading to

$$\left(e^{\beta^\top Z_i} \Delta H(T)\right)^{\Delta} e^{-e^{\beta^\top Z} H(T)}.$$

The product of this expression evaluated at n observations (T_i, Δ_i, Z_i) gives an overall *empirical likelihood*, which we shall refer to as the "Poisson form." Maximization of this empirical likelihood over (β, H) leads to the Cox partial likelihood estimator and the Breslow estimator, and in fact the resulting profile likelihood for β is exactly the Cox partial likelihood function given in (13.28) (see e.g. Murphy and van der Vaart 2000). In particular, the maximizer for H is supported only on the uncensored observations $\{T_i\colon \Delta_i = 1\}$. For discrete H with jumps only in the latter set, the "Poisson empirical likelihood" can be written as

$$L_n^P(\beta, H) \prod_{t:\Delta N(t)>0} \Big[e^{\sum_{i:T_i=t,\Delta_i=1} e^{\beta^\top Z_i}} \, \Delta H(t)^{\Delta N(t)} e^{-\Delta H(t) \sum_{i:T_i\geq t} e^{\beta^\top Z_i}} \Big]. \qquad (13.32)$$

The parameters in this likelihood are β and the jump sizes $\Delta H(t)$. The form of the likelihood suggests a prior for the jump sizes proportional to $\prod_t \Delta H(t)^{-\alpha}$. Theorem 13.21 suggests to take $\alpha = 1$ to ensure consistency. Under this prior, by Bayes's rule, given β the jumps $\Delta H(t)$ are also independent under the posterior distribution, with

$$\Delta H(t) |\, (\beta, D_n) \overset{\text{ind}}{\sim} \mathrm{Ga}\Big(\Delta N(t), \sum_{i\in R_n(t)} e^{\beta^\top Z_i}\Big). \qquad (13.33)$$

By integrating out $\Delta H(t)$ from the likelihood under the prior, the posterior density of β is found as

$$\pi_n(\beta |\, D_n) \propto \pi(\beta) \prod_{t\in \mathcal{T}_n} \frac{\exp(\sum_{i\in D_n(t)} \beta^\top Z_i)}{(\sum_{i\in R_n(t)} \exp(\beta^\top Z_i))^{\Delta N_n(t)}}. \qquad (13.34)$$

This is precisely the prior density $\pi(\beta)$ times Cox's partial likelihood (13.28).

An alternative to the Poisson empirical likelihood derives from a different discretization of the continuous likelihood. The starting point is the "counting process form" of the continuous likelihood for one observation (T, Δ, Z):[9]

$$\prod_t d\Lambda(t)^{dN(t)} (1 - d\Lambda(t))^{Y(t)-dN(t)}.$$

Here Λ is the cumulative intensity of the counting process $N(t) = \mathbb{1}\{T \leq t, \Delta = 1\}$, which in the Cox model is given by $t \mapsto \int_{(0,t]} \mathbb{1}\{T \geq x\} e^{\beta^\top Z} \, dH(x)$. The "counting process likelihood" is exact for absolutely continuous H, but suggests a "discretized" extension in which the infinitesimal $d\Lambda$ are replaced by the jump heights $\Delta\Lambda$. The resulting formula resembles (13.11). Deviating from the preceding paragraph we take "proportional hazards" to mean that the survival function of the life time X is of the form $\bar F_X = \bar F^{\exp(\beta^\top Z)}$, for $\bar F$ the baseline survival function, which implies $1 - \Delta\Lambda = \bar F_X / \bar F_{X-} = (1 - \Delta H)^{\exp(\beta^\top Z)}$ (rather than $\Delta\Lambda = e^{\beta^\top Z} \Delta H$). As before, given n observations we consider cumulative baseline hazard distributions H supported on only the points $\{T_i\colon \Delta_i = 1\}$. Substituting $(1 - \Delta H)^{\exp(\beta^\top Z)}$ for $1 - d\Lambda$ in the counting process likelihood, we find the "binomial empirical likelihood"

$$\prod_{i=1}^n \Big(1 - (1 - \Delta H(T_i))^{e^{\beta^\top Z_i}}\Big)^{\Delta N_i(T_i)} (1 - \Delta H(T_i))^{e^{\beta^\top Z_i} (Y_i(T_i) - \Delta N_i(T_i))}.$$

[9] The formula must be interpreted as formal notation for product integration, see Andersen et al. (1993).

The binomial form of this likelihood suggests the prior distribution on the jump heights $\Delta H(t)$ given by

$$\prod_t \Delta H(t)^{-1}(1-\Delta H(t))^{-1}. \tag{13.35}$$

This results in a posterior that can also be found as the limit as $c \to 0$ of the posterior distribution based on a beta process prior on H (see Problem 13.19). The marginal posterior distribution of β does not reduce to a closed form expression, but it can be calculated numerically using a Metropolis-Hastings algorithm.

Since the prior on H is improper, the marginal posterior distribution for β may be improper even when the prior distribution on β is proper. It is known that Cox's partial likelihood is log-concave under the general assumption that the positive linear span of the differences $\{Z_i - Z_j : T_i = t, \Delta_i = 1, j \in R_n^+(t)\}$ is the full space $\mathbb{R}^d$ (Jacobsen 1989), where $R_n^+(t) = \{i : T_i \geq t\} \setminus \{i : T_i = t, \Delta_i = 1\}$. Under this condition the Poisson likelihood leads to a proper posterior if π is improper uniform. Propriety of the posterior distribution corresponding to the binomial likelihood is more complicated; see Problem 13.21 for a counterexample.

Given the Bayesian bootstrap distribution for H the corresponding distribution on the survival function can be obtained by the transformations $\bar F(t) = \exp[-\sum_{T_i \leq t, \Delta_i = 1} \Delta H(T_i)]$ and $\bar F(t) = \prod_{T_i \leq t, \Delta_i = 1}(1 - \Delta H(T_i))$, in the Poisson and binomial cases, respectively. In the Poisson case the increments $\Delta H(t)$ may be larger than 1 (and hence H is not a genuine cumulative hazard function), which renders the second correspondence between H and $\bar F$ unfeasible.

As in Section 13.6.2 the posterior distribution described by the Bayesian bootstrap process admits a Bernstein–von Mises theorem. We concentrate on the parametric part β, even though a joint Bernstein–von Mises theorem for (β, H) in the spirit of Theorem 13.36 is possible also.

Theorem 13.37 *If the true cumulative hazard function H_0 is absolutely continuous with $H_0(\tau) < \infty$ for some $\tau > 0$ with $G(\tau-) < 1 = G(\tau)$, and the support of the covariate Z is a compact subset of $\mathbb{R}^d$ with nonempty interior, then the Bayesian bootstrap posterior distribution given by the density $\pi_n(\cdot|\, D_n)$ in* (13.34) *for a proper prior density π that is continuous and positive at β_0, satisfies $\|\Pi_n(\cdot|\, D_n) - \mathrm{Nor}(\hat\beta_n, n^{-1}\tilde I_{\beta_0}^{-1})\|_{TV} \to 0$ a.s. $[P^\infty_{H_0,\beta_0}]$ as $n \to \infty$, for $\hat\beta_n$ the partial likelihood estimator (which maximizes* (13.28)*) and $\tilde I_{\beta_0}$ is given by* (13.29).

Proof The proof proceeds as for the classical Bernstein–von Mises theorem for the parametric case, by expanding the partial likelihood in a Taylor series and using Cramér-type regularity conditions to bound the terms. A noticeable difference is that the partial likelihood (13.28) is not of a product of i.i.d. terms, which necessitates a nonstandard law of large numbers.

The posterior density (13.34) is proportional to $\pi(\beta)L_n(\beta)$, for L_n the partial likelihood function (13.28). It suffices to show that $\int |g_n(u) - \pi(\beta_0)\exp\{-u^\top \tilde I_{\beta_0} u/2\}|\, du \to 0$ a.s., where $g_n(u) = \pi(\hat\beta_n + n^{-1/2}u)L_n(\hat\beta_n + n^{-1/2}u)/L_n(\hat\beta_n)$. For given $B, \delta > 0$ we can bound this $\mathbb{L}_1$-distance by the sum $I_1 + I_2 + I_3 + I_4$ given by

$$\int_{\|u\|\le B} |g_n(u) - \pi(\beta_0)e^{-u^\top \tilde{I}_{\beta_0}u/2}|\,du + \int_{B<\|u\|\le\delta\sqrt{n}} g_n(u)\,du \\ + \int_{\|u\|>\delta\sqrt{n}} g_n(u)\,du + \int_{\|u\|>B} \pi(\beta_0)e^{-u^\top \tilde{I}_{\beta_0}u/2}\,du.$$

The term I_4 can clearly be made arbitrarily small by choice of B. We shall argue that I_1 tends to zero for every B, the term I_2 can be arbitrarily small by choice of δ, and I_3 tends to zero for any δ and B.

To bound I_1 we expand $u \mapsto \log L_n(\hat{\beta}_n + n^{-1/2}u)$ in a second order Taylor expansion around 0 with third order remainder. The linear term of this expansion vanishes as $\hat{\beta}_n$ maximizes L_n. Because n^{-1} times the third order partial derivatives of $\beta \mapsto \log L_n(\beta)$ are uniformly bounded in a neighborhood of 0, and $\hat{\beta}_n$ is consistent, the remainder term is bounded by a multiple of $n^{-1/2}$. Since also π is continuous at β_0, the term I_1 can be shown to tend to zero by standard arguments used in proving parametric Bernstein–von Mises theorems.

The matrix of second derivatives of $\beta \mapsto \log L_n(\beta)$ at β_0 approaches the positive definite matrix $\tilde{I}_{\beta_0}$, by a law of large numbers (see Tsiatis 1981 or Andersen et al. 1993 for details). Together with the boundedness of n^{-1} the third-order partial derivatives, this shows that there exists $\lambda > 0$ such that $\log L_n(\hat{\beta}_n + n^{-1/2}u) - \lambda\|u\|^2/2$, for $B < \|u\| \le \delta\sqrt{n}$ and sufficiently small $\delta > 0$. In view of the boundedness of the ratio $\pi(\beta)/\pi(\beta_0)$ over a small neighborhood around β_0, this implies that I_2 can be made small.

To bound I_3, we establish an exponential bound on $L_n(\beta)/L_n(\hat{\beta}_n)$, for $\|\beta\| > \delta$. The function $l_n(\beta)$ is concave, attains its maximum at $\hat{\beta}_n$, and for every fixed β the sequence $n^{-1}l_n(\beta)$ tends almost surely to

$$l(\beta) = \beta^\top \mathrm{E}(Z\mathbb{1}\{\Delta = 1\}) - \int_0^\tau \mathrm{E}[e^{\beta_0^\top Z}\mathbb{1}\{T \ge t\}] \log(\mathrm{E}[e^{\beta^\top Z}\mathbb{1}\{T \ge t\}])\,dH_0(t).$$

By concavity, the convergence is automatically uniform over compact sets. Let $\rho = \inf\{l(\beta_0) - l(\beta) \colon \|\beta - \beta_0\| = \delta\}$, and determine a sufficiently large n such that

$$\sup_{\|\beta-\beta_0\|\le\delta} |n^{-1}l_n(\beta) - l(\beta)| + |n^{-1}l_n(\hat{\beta}_n) - l(\beta_0)| < \frac{\rho}{2}.$$

By concavity, the supremum of $l_n(\beta) - l_n(\hat{\beta}_n)$ over $\|\beta - \beta_0\| \ge \delta$ is attained on the circle $\|\beta - \beta_0\| = \delta$. By the triangle inequality, it follows $l_n(\beta) - l_n(\hat{\beta}_n) < -n\rho/2$, for $\|\beta - \beta_0\| \ge \delta$, eventually. Hence $I_3 \le n^{d/2}e^{-n\rho/2} \to 0$, as π is proper. □

13.8 Historical Notes

Bayesian estimation under censoring was addressed by Susarla and Van Ryzin (1976). They obtained the mixture of Dirichlet process representation for the posterior distribution given a Dirichlet prior, and computed the corresponding posterior mean. The Pólya tree representation of the Dirichlet process posterior was obtained by Ghosh and Ramamoorthi (1995). The gamma process was first used in Bayesian survival analysis by Kalbfleisch (1978), the generalized gamma process by Ferguson and Phadia (1979), and the extended gamma process by Dykstra and Laud (1981). The beta process was constructed as a conjugate prior distribution for the cumulative hazard function for randomly right censored data by Hjort (1990). His motivation of a beta process through a discrete time setting leads also to the time discretization algorithm (a) in Section 13.3.3. Wolpert and Ickstadt (1998) proposed the inverse Lévy measure algorithm I. The Poisson weighting algorithm is based on the general method for sampling from an infinitely divisible distribution by Damien et al. (1995, 1996); Section 13.3.3 provides the refined version by Lee (2007). The ϵ-approximation algorithm is due to Lee and Kim (2004). Neutral to the right processes were introduced by Doksum (1974). Ferguson and Phadia (1979) obtained conjugacy for the posterior distribution under random right censoring. Several aspects of neutral to the right processes including Proposition 13.10 were obtained by Dey et al. (2003); see also the problems section. The beta-Stacy process was introduced by Walker and Muliere (1997). Conjugacy of independent increment processes in the form of Theorem 13.15 is due to Hjort (1990). It was generalized by Kim (1999) to the more general framework of multiplicative counting processes, which allows for observing a general counting process N with compensator $\int_0^t Y\,dH$ for some predictable increasing process Y, a setup that can treat left-truncated data as well. The proof presented here is indicated as an alternative, more-intuitive approach in Hjort (1990) and inspired by Prünster (2012). Theorem 13.18 characterizing consistency for the extended beta process prior and the example of inconsistency are due to Kim and Lee (2001). They also obtained Theorem 13.21 for a general independent increment process prior on the cumulative hazard function. Dey et al. (2003) obtained a somewhat similar result for neutral to the right process prior on distributions. The Bernstein–von Mises theorem for the cumulative hazard function was derived by Kim and Lee (2004). Kernel smoothing on a hazard function to construct a prior was first used by Dykstra and Laud (1981). They obtained the expression for the posterior moment generating function of the cumulative hazard function using the extended gamma process and the Dykstra-Laud kernel. Consistency for such a prior distribution was obtained by Drăghici and Ramamoorthi (2003). The representation of the posterior distribution given by Theorem 13.30 was obtained by James (2005). Consistency for a general smooth hazard process prior was obtained by De Blasi et al. (2009), together with a pathwise central limit theorem for the tail of linear and quadratic functionals of the posterior smooth hazard function. An analogous result on the prior process was obtained earlier by Peccati and Prünster (2008). Consistency of the posterior distribution in the Cox model for independent increment process priors was obtained by Kim and Lee (2003a), and the Bernstein–von Mises theorem in Kim (2006), using an elegant approach based on counting process likelihoods. The Bayesian bootstrap for censored data with covariates was first defined by Lo (1993). The Bayesian bootstrap was extended to the Cox model by Kim and Lee (2003b). They also obtained the Bernstein–von Mises theorem for the Bayesian bootstrap in the Cox model.

Problems

13.1 (Ghosh and Ramamoorthi 2003) Let $\mathfrak{M}_0$ be the set of all pairs of distribution functions (F, G) such that $F(t) < 1, G(t) < 1$ for all t and F and G have no common points of discontinuity. Let $\varphi(x, y) = (x \wedge y, \mathbb{1}\{x \leq y\})$ and $\mathfrak{M}_0^* =$

$\{(F, G)\varphi^{-1}: (F, G) \in \mathfrak{M}_0\}$. Let Π^* be the prior on $(F, G)\varphi^{-1}$ induced from a prior Π on (F, G), and if $\Pi^*(\cdot|D_n)$ is (strongly) consistent at $(F_0, G_0)\varphi^{-1}$ with respect to the weak topology, then show that $\Pi(\cdot|\, D_n)$ is also (strongly) consistent at (F_0, G_0) with respect to the weak topology.

Let $S_u(t) = \mathrm{P}(X > t, X \leq C|\, X \sim F, C \sim G)$ and $S_c(t) = \mathrm{P}(C > t, X > C|\, X \sim F, C \sim G)$ be the sub-survival functions corresponding respectively to uncensored and censored observations. If the induced posterior for S_u and S_c are (strongly) consistent with respect to the Kolmogorov-Smirnov norm, then show that the posterior for $(F, G)\varphi^{-1}$ is (strongly) consistent with respect to the weak topology.

13.2 (Ghosh and Ramamoorthi 2003) Consider an interval censoring mechanism where one observes only $X_i \in (L_i, R_i]$, $i = 1, \ldots, n$, and $(L_1, R_1), \ldots, (L_n, R_n)$ are independent. Let the distribution F for X follow a Dirichlet process. Obtain the expressions for the posterior distribution and Bayes estimate of F given $(L_1, R_1), \ldots, (L_n, R_n)$.

13.3 (Walker and Muliere 1997, Ramamoorthi et al. 2002) Let X be an observation following cumulative distribution function F and F be given a prior Π. Show that Π is an NTR process prior if and only if for any $t > 0$, $\Pi(F(t)|\, X) = \Pi(F(t)|\, \mathbb{1}\{X \leq s\}, 0 < s < t)$.

13.4 (Dey et al. 2003) The connection between the means of the survival function and the cumulative hazard function given by Theorem 13.8 characterizes NTR processes. Let Π be a prior for a survival distribution F such that $\Pi(0 < \bar{F}(i)/\bar{F}(i-1) < 1) = 1$ for all i. Let $H(F)$ be the cumulative hazard function associated with F. Suppose that for all $n \in \mathbb{N}$ and $X_1, \ldots, X_n$, $\mathrm{E}\big[H(F)|\, X_1, \ldots, X_n\big] = H(\mathrm{E}[F|\, X_1, \ldots, X_n])$ and $\mathrm{E}\big[H(F)|\, X > x\big] = H(\mathrm{E}[F|\, X > x])$ for all x. Show that Π is NTR.

13.5 (Dey et al. 2003) Let Π_1 and Π_2 be two NTR processes, and let ν_1 and ν_2 be the intensity measures for the corresponding independent increment processes on the c.h.f. Assume that $\nu_1 \ll \nu_2$ and $f = d\nu_1/d\nu_2$. Then Π_1 and Π_2 are mutually singular if either one of the following conditions hold:

(a) for some $c > 0$, $\int_{|f-1|>c} |f-1|\, d\nu_2 = \infty$;
(b) for all $c > 0$, $\int_{|f-1|<c} |f-1|^2\, d\nu_2 = \infty$.

13.6 (Dey et al. 2003) Let H_1 and H_2 be continuous c.h.f.s and let c_1 and c_2 be positive measurable functions such that the measure $c_1 H_1$ and $c_2 H_2$ are not identical. Then the beta processes with parameters (c_1, H_1) and (c_2, H_2) are mutually singular.

13.7 Show that if the distribution of the discrete hazard rates $h(jb)$ are given by (13.12), with $c(j)(1 - h(j)) = c(j+1)$ for all j, then $(f(jb): j = 0, 1, \ldots)$ follows $\mathrm{DP}_{(c(0), c(1), \ldots)}$.

13.8 (Dey et al. 2003) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, a cumulative distribution function on $[0, \infty)$. Let F be given an NTR prior distribution Π. If the posterior $\Pi(\cdot|\, X_1, \ldots, X_n)$ is consistent at every continuous true distribution F_0, then show that $\Pi(\cdot|\, D_n)$ is consistent at every continuous true distribution F_0.

13.9 (Lo 1993) Let α, β be finite measures on $[0, \infty)$ and let Γ_α and Γ_β be standard gamma processes with parameters α and β respectively. Show that the process $H(t) = \int_0^t (\Gamma_\alpha(x, \infty) + \Gamma_\beta(x, \infty))^{-1} \Gamma_\alpha(dx)$ follows a beta process with parameters (c, H_0) given by $c(t) = \alpha(t, \infty) + \beta(t, \infty)$ and $H_0(t) = \int_0^t (\alpha(x, \infty) + \beta(x, \infty))^{-1} \alpha(dx)$.

Note that $c(t)$ is monotone decreasing, so not all beta processes can be represented in this way.

13.10 If a beta process with parameter (c, H_0) is homogeneous, i.e. $c(x) = c > 0$, a constant, then show that $L(J_j) = \sum_{i=1}^j V_i / H_0(\tau)$, where $[0, \tau]$ is a given interval where jumps are being studied, $V_1, V_2, \ldots$ are i.i.d. standard exponential and $L(s) = \nu_H([0, \tau \times (s, 1))$, for ν_H the intensity measure of the beta process. In particular, if $c = 1$, show that J_j is a product of j i.i.d. $\mathrm{Be}(H_0(\tau), 1)$.

13.11 (Dey et al. 2003) For independent increment process priors, if observations are all uncensored, then the following shows that consistency of the posterior mean of the distribution actually implies posterior consistency. (Problem 13.8 shows that the assumptions that all observations are uncensored can be dropped if F_0 is continuous.)

Let $X_i \overset{\text{iid}}{\sim} F$, a distribution on $[0, \infty)$. Let the c.h.f. H follow an independent increment process prior Π with intensity measure given by $\nu(dt, ds) = a(x, s)\, ds\, \Lambda(dt)$, where Λ itself is a c.h.f. If $\mathrm{E}[F(t) |\, X_1, \ldots, X_n] \to F_0(t)$ a.s., then $\mathrm{var}[F(t) |\, X_1, \ldots, X_n] \to 0$ a.s.

If the condition holds for all t, then in view of Lemma 6.4, $\Pi(\cdot |\, X_1, \ldots, X_n) \rightsquigarrow \delta_{F_0}$ a.s.

13.12 (Kim and Lee 2004) Adopt the notations of Section 13.4.2. Consider an independent increment process prior on a c.h.f. H with intensity measure given by $\nu(dt, ds) = s^{-1}(1 + s^\alpha)\, ds\, dt$. Let $J_\alpha(t) = \alpha \Gamma(\alpha + 1) \int_0^t (Q(u))^{-\alpha} d H_0(u)$. Show that a.s. as $n \to \infty$:

(a) If $0 < \alpha \le \frac{1}{2}$, then $\sup\{|n^\alpha [\mathrm{E}(H_d(t) |\, D_n) - \hat{H}_n(t)] - J_\alpha(t)| : t \in [0, \tau]\} \to 0$;

(b) If $0 < \alpha < \frac{1}{2}$, then $n^\alpha (H - \hat{H}_n) |\, D_n \rightsquigarrow \delta_{J_\alpha(\cdot)}$;

(c) If $\alpha = \frac{1}{2}$, then $n^{1/2}(H - \hat{H}_n) |\, D_n \rightsquigarrow B(U_0(\cdot) + J_\alpha(\cdot))$;

(d) If $\alpha > \frac{1}{2}$, then $n^{1/2}(H - \hat{H}_n) |\, D_n \rightsquigarrow B(U_0(\cdot))$.

Thus the value of the index α determining the rate at which $g(t, s)$ approaches $q(t)$ as $s \to 0$ is critical for the Bernstein–von Mises theorem.

13.13 (De Blasi et al. 2009) Consider a CRM Φ on $[0, \infty)$ with intensity $\nu(dv, ds) = \rho(ds |\, v)\, \alpha(dv)$ satisfying $\rho(\mathbb{R}^+ |\, v) = \infty$ a.s. $[\alpha]$, and $\mathrm{supp}(\alpha) = \mathbb{R}^+$. Show that for any Lebesgue-Stieltjes distribution function G_0 $\mathrm{P}(\sup_{x \le M} |\Phi([0, x]) - G_0(x)| < \epsilon) > 0$ and $M, \epsilon > 0$.

13.14 (Kim and Lee 2003a) Let A be an independent increment process with intensity measure of the form $g(t, s)\, ds\, dt + \sum_{i=1}^k \delta_{v_j}(dt)\, G_j(ds)$ and let $h\colon [0, 1] \to [0, 1]$ be a strictly increasing continuously differentiable bijection. Show that $B(t) = \sum_{x \le t} h(\Delta A(x)) \mathbb{1}\{\Delta A(x) > 0\}$ is an independent increment process with intensity measure $(h^{-1})'(s) g(t, h^{-1}(s))\, ds\, dt + \sum_{j=1}^k \delta_{v_j}(dt)\, (G_j \circ h^{-1})(ds)$.

13.15 Show that (13.30) reduces to Rubin's Bayesian bootstrap when all the observations are uncensored.

13.16 Show that the probability measure corresponding to (13.30) can be written as $\sum_{i=1}^{n} W_j \delta_{T_j^*}$, where the variables W_j are the stick-breaking weights based on the relative stick lengths given by $V_j \overset{\text{ind}}{\sim} \text{Be}(\sum_{i:X_i=T_j^*} \Delta_i, \#\{i\colon X_i \geq T_j^*\})$, for $j = 1, \ldots, k$, and $V_k = 1$.

13.17 Show that the distribution on $\bar{F}$ induced from (13.31) on H is identical with (13.30).

13.18 Taking the limit $c(\cdot) \to 0$ in Theorem 13.23 for a beta process, obtain a Bernstein–von Mises theorem for the Bayesian bootstrap for censored data.

13.19 (Kim and Lee 2003b) Show that the limit of the beta process posterior as $c(\cdot) \to 0$ in the Cox model with binomial likelihood coincides with the posterior based on the binomial likelihood and the improper prior (13.35).

13.20 (Kim and Lee 2003b) Show that the Bayesian bootstrap distribution based on the binomial form of the likelihood reduces to (13.31) when no covariate is present.

13.21 (Kim and Lee 2003b) Consider observations of the form (T, Δ, Z) and form the binomial likelihood. Suppose that $n = 3$ and the observations are $(1, 1, -1)$, $(2, 1, -1.9)$ and $(3, 1, -1.5)$. Show that the likelihood is bounded below, and hence the posterior is improper for the improper uniform prior. Further, show that the posterior is improper for some proper priors with sufficiently thick tails.