Semi-and Non-parametric Bayesian Analysis of Duration Models[1]

J-P. Florens [2]
M. Mouchart [3]
J-M. Rolin[3]

March 23, 1995

[1] *Acknowledgement:* The work of the second author has been made during a visit in Toulouse. Supports from GREMAQ and IDEI, Université des Sciences Sociales, Toulouse and from the Fond National de la Recherche Scientifique are gratefully aknowledged.

Key words: Bayesian Non-parametric, Bayesian semi-parametric, Dirichlet process, Beta process, duration data, right-censoring, heterogeneity

[2] GREMAQ and IDEI, Université des Sciences Sociales, Toulouse, France.

[3] CORE and Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgique.

Abstract

The object of this paper is to review the main results obtained in the Semi-and non-parametric Bayesian analysis of duration models. Standard non-parametric Bayesian models for I.I.D. observations are just reviewed in line with Ferguson's pioneering papers but presenting also recent results on the characterization of Dirichlet processes. Next we present recent results on heterogeneity in the context of mix terms of Dirichlet processes and on non-parametric treatement of censoring. The final section considers a Bayesian semi-parametric version of the proportional hazard model.

1 Introduction

Duration models are not only important for the statistical analysis of data in many fields, they also display characteristic features of particular interest for the statistician. Among these features, one may mention a prominent interest for non-parametric methods, made possible by the availability of large individual data sets. Another specific feature of duration models comes from the natural importance of alternative characterizations of the sampling distributions, namely their survivor function and their integrated or instantaneous hazard functions. This particular role may be appreciated, for instance, in the specification of conditional models, a most usual class of models in econometrics, where the proportional hazard model is most naturally written in the terms of a hazard function and displays general properties different from those of the standard regression models. Although not specific to duration models, the necessity of taking into account unobservable components of heterogeneity or censoring problems, actually present in most data sets, makes still more evident the particularities of duration models.

The object of this paper is to review the main results obtained in this field from a Bayesian point of view with a particular interest for non-parametric and semi-parametric models the prior specification of which often involves a Dirichlet process.

Non-parametric Bayesian methods have known a strong impetus after Ferguson's (1973, 1974) presentation of the Dirichlet process as a prior specification natural conjugate to the empirical (non-parametric) process. Very soon, the extension to mixtures of Dirichlet processes has been realized by Antoniak (1974) and the treatment of censored data has been handled in Suzarla and Van Ryzin (1978). Later on many contributions on these themes have come to the light, in particular Lahiri and Dong Ho (1988), Phadia and Susarla (1983) and Tsai (1986). A noticeable extension, by Hjort (1990), treats prior specifications based on the beta process. In chapter VIII of Florens, Mouchart and Rolin (1990) we give some general tools of analysis and

in Florens, Mouchart and Rolin (1992) we consider a rather general model of mixtures for handling problems of heterogeneity whereas Rolin (1983, 1992a and b) contributes to the analysis of finer properties of the Dirichlet process. This review gives also credit to the original work of two of our doctoral students, Ruggiero (1989) and Hakizamungu (1992) whose dissertations have been most useful for our understanding.

This review is organized in four steps. In next section, we review the standard non-parametric Bayesian model for I.I.D. observations, in line with the pioneering papers of Ferguson (1973, 1974). This section also includes more recent results of Rolin (1992a and b) giving more precise characteristics of the Dirichlet process from the point of view of random measures. Section 3 treats the heterogeneity problem in the context of mixtures of Dirichlet processes based on our 1992 paper and other related papers. Section 4 handled the I.I.D. case with censoring and review, in particular, recent results of Hakizamungu (1992) and an illuminating survey by Ruggiero (1992). The last section considers a semi-parametric version of the proportional hazard model presenting, among others, some original contributions and extensions of Ruggiero (1989).

The presentation is made simpler by treating separately those problems even though, in actual empirical works, heterogeneity, censoring and exogenous variables should be combined into a single model.

2 Non-parametric Duration Models without Censoring.

2.1 A non-parametric Bayesian Model.

Although interest for non-parametric methods has been present in Bayesian contributions in earlier time, Ferguson's (1973, 1974) papers have been most influential for motivating new contributions in the last decade. In this section we summarize the basic results obtained in that direction, limiting the presentation for the particular case of duration models, *viz* models for non-negative observations, although the basic model is considerably more general.

Whereas a statistical model, in sampling theory, is a family of *sampling distributions* indexed by a *parameter*, a Bayesian model is characterized by a unique probability measure on the product space parameter $\times$ observation, obtained by endowing a sampling theory model with a probability measure on the parameter space, to be called the *prior probability*, and by treating the sampling model as a conditional probability on the sample space conditionally on the parameter. From this joint probability measure, Bayesian methods aim at evaluating and analysing the conditional probabilities on the parameter given the observation to be called the *posterior distribution*, and

the marginal distribution on the observation to be called the *predictive distribution* and obtained as an average of the sampling distributions weighted by the prior probability.

In the case of duration models, the sample space is $\mathbb{R}_+$, the positive part of the real line, endowed with its Borel sets. In the case of non-parametric models, the parameter space is a functional space representing a "large" subset of the set of all probability measures on the sample space or of some of their transformations, such as the survivor function or the hazard functions (cumulative or instantaneous).

Thus, the basic non-parametric Bayesian model for analyzing duration data may be described as follows. The sampling process is I.I.D and the sample space for a sample of size one, is $(\mathbb{R}_+, \mathcal{B}_+)$. For a given sample size n, the data are therefore constituted as $t_1 \cdots t_n$ where each t_i $(1 \leq i \leq n)$ is independently generated by a common sampling probability, say Φ, an element of the set of all probability measures on $(\mathbb{R}_+, \mathcal{B}_+)$. Thus one may simply write $\Phi(A)$ instead of $P(t \in A|\Phi)$.

For the prior specification, note that the prior probability makes Φ a random probability measure on $(\mathbb{R}_+, \mathcal{B}_+)$ and should therefore be viewed as a stochastic process, the trajectories of which are probability measures on $(\mathbb{R}_+, \mathcal{B}_+)$. After Ferguson (1973, 1974) a workable choice, which is also natural conjugate to the empirical process, is the Dirichlet process. The basic intuition and the main features of this specification may be approached through the finite case.

Thus, let us assume as a first step in the presentation, that the sampling probability Φ is restricted to give positive probabilities to a fixed finite set say $(a_1 \cdots a_k)$. Thus Φ is characterized by a point, θ, of the simplex S_k of $\mathbb{R}^k : \theta = (\theta_1 \cdots \theta_k)$ such that $\sum_j \theta_j = 1$ with the interpretation that $\theta_j = \Phi(\{a_j\}), \quad 1 \leq j \leq k$. An observed duration t may be represented by a vector of binary variables $x = (x_1 \cdots x_k)$ with $x_j = \mathbb{I}_{\{t=a_j\}}$ and the sampling probability may then be written as :

$$p(x|\theta) = \prod_j \theta_j^{x_j}. \tag{2.1}$$

Making of the Dirichlet distribution a natural conjugate one to the sampling process, its density may be written as:

$$\begin{aligned} p(\theta) &= f_{Di}(\theta|n_0, p_0) \\ &= \Gamma(n_0) \prod_j \frac{\theta_j^{n_0 p_{0j} - 1}}{\Gamma(n_0 p_{0j})} \mathbb{I}_{\{\theta \in S_k\}} \end{aligned} \tag{2.2}$$

where $(n_0, p_0) \in \mathbb{R}_+ \times S_k$ are natural parameters for a Dirichlet distribution. If we have n I.I.D observations of x i.e. $x_i = (x_{i1} \cdots x_{ij} \cdots x_{ik})$ for $i =$

$1, \cdots, n$, the vector of sample proportions $p = (p_1 \cdots p_j \cdots p_k)$ with $p_j = n^{-1} \sum_{1 \leq i \leq n} x_{ij}$, is a sufficient statistics the distribution of which is a multinomial distribution divided by n. The natural conjugate property implies that the posterior distribution of θ is again Dirichlet with parameter:

$$n_* = n_0 + n \tag{2.3}$$

$$p_* = \frac{n_0 p_0 + np}{n_0 + n} \quad \left(\text{ or } p_{*j} = \frac{n_0 p_{0j} + np_j}{n_0 + n} \right). \tag{2.4}$$

Suppose now that instead of restricting Φ to have of finite fixed support $\{a_1 \cdots a_k\}$ we consider a finite fixed partition of $I\!R_+$, namely $\{B_1 \cdots B_j \cdots B_k\}$, and restrict the parameter of interest to be $\theta = (\theta_j)$ with $\theta_j = P(t \in B_j | \Phi) = \Phi(B_j)$. It is then natural to retain from an I.I.D sample of durations $t_1 \cdots t_n$ the proportions of observations in each B_j i.e. $p_j = n^{-1} \sum_{1 \leq i \leq n} I\!\!1_{\{t_i \in B_j\}}$. If the prior specification of Φ is such that it implies a Dirichlet distribution on θ, the analysis of the case of finite support may be exactly repeated without any change. The transition from the Dirichlet distribution to the Dirichlet process is operated by switching from a fixed partition to an arbitrary one and by replacing the prior parameter $p_0 = (p_{01} \cdots p_{0k}) \in S_k$ by an probability measure P_0 on the sample space $(I\!R_+, \mathcal{B}_+)$. More specifically, the sampling probability Φ on $(I\!R_+, \mathcal{B}_+)$ is said to be distributed as a Dirichlet process with parameter (n_0, P_0) denoted by $\Phi \sim \mathcal{D}i(n_0, P_0)$ if for any measurable partition of $I\!R_+, (B_1 \cdots B_k)$ the random vector $(\Phi(B_1) \cdots \Phi(B_k))$ is distributed as a Dirichlet distribution with parameters n_0 and $(P_0(B_1) \cdots P_0(B_k))$. From the properties of the Dirichlet distibution it may be verified that the system $(\Phi(B_1) \cdots \Phi(B_k))$, defined for any finite partition of $I\!R_+$ induces a projective system and therefore, by Kolmogorov theorem, uniquely characterizes the law of the process generating Φ.

Let us write P_n for the empirical process, viz :

$$P_n(B) = \frac{1}{n} \sum_{1 \leq i \leq n} I\!\!1_{\{t_i \in B\}} = \frac{1}{n} \sum_{1 \leq i \leq n} \delta_{t_i}(B) \qquad B \in \mathcal{B}_+ \tag{2.5}$$

where δ_{t_i} is the unit mass at t_i.
The statistic P_n is sufficient for an I.I.D. sampling and the posterior probability of Φ is again a Dirichlet process with parameter:

$$n_* = n_0 + n \tag{2.6}$$

$$P_* = \frac{n_0 P_0 + nP_n}{n_0 + n}. \tag{2.7}$$

The description of this Bayesian model is now completed by describing the predictive process through the following sequence.

(i) the first observation t_1 is predictively distributed as P_0

(ii) the observation t_{i+1} conditionally on $(t_1 \cdots t_i)$ is predictively distributed according to $\frac{1}{n_0+i}(n_0 P_0 + iP_i)$ where P_i is the empirical process of $(t_1 \cdots t_i)$, namely:

$$P_i(B) = \frac{1}{i} \sum_{1 \le j \le i} I\!\!I_{\{t_j \in B\}} \qquad B \in \mathcal{B}_+. \tag{2.8}$$

2.2 Some properties of the Dirichlet process.

The previous section suggests that the structure of the non-parametric Bayesian model under a Dirichlet prior specification is basically simple and provides a workable approach for the evaluation of the posterior distribution as well as of the predictive distribution. In this section, some properties of the model reinforce this idea but other properties shed light on some subtle aspects of the Dirichlet process and require specific care to be handled with.

2.2.1 Moments.

Let Φ be a random probability on $(I\!R_+, \mathcal{B}_+)$ distributed as a Dirichlet process with parameter (a, M); thus in the context of section 2.1, a may be taken as n_0 or n_* and M as P_0 or P_*. From the definition of the Dirichlet process it should be clear that, for any $B \in \mathcal{B}_+$, the random variable $\Phi(B)$ follows a beta distribution with parameters $(aM(B), aM(B^c))$; in particular one has:

$$E[\Phi(B)] = M(B) \tag{2.9}$$

$$V[\Phi(B)] = \frac{M(B)M(B^c)}{a+1}. \tag{2.10}$$

Thus, P_0 may be interpreted as a "prior guess" on Φ and n_0 as a measure of prior precision and, similarly, P_* is the posterior expectation of the process and n_* characterizes its posterior precision. Consequently, P_* may be viewed as a natural Bayesian estimator of Φ built as a convex combination of the prior guess P_0 and the empirical process P_n and convergent in as far as it has the same asymptotic behavior as P_n.

2.2.2 The trajectories of the Dirichlet process.

Understanding properly the non-parametric aspect of a Bayesian model under a Dirichlet prior specification requires a serious analysis of its trajectories, both their structure and their support. When a random probability Φ on $(I\!R_+, \mathcal{B}_+)$ is distributed as a Dirichlet process with parameters (a, M), the structure, and the support, of its trajectories crucially depend on the

location parameter M. To show this, let us decompose M into its purely atomic and diffuse parts, i.e., let

$$S = \{x \in \mathbb{R}_+ : M(\{x\}) > 0\} = \{a_j : j \in I\}$$

(where $I \subset \mathbb{N}$ is a finite or a countable set)

$$a_d = aM(S); \quad a_c = aM(S^c)$$

$$M_d(B) = M(B|S); \quad M_c(B) = M(B|S^c) \quad \forall\ B \in \mathcal{B}_+.$$

Clearly, one has

$$M = a_c M_c + a_d M_d. \tag{2.11}$$

Let us now operate the same decomposition on Φ, namely let

$$\alpha = \Phi(S)$$

$$\Phi_d(B) = \Phi(B|S); \quad \Phi_c(B) = \Phi(B|S^c),$$

so that

$$\Phi = (1-\alpha)\Phi_c + \alpha\Phi_d. \tag{2.12}$$

It has been shown in Rolin (1992a) that

(i) α, Φ_c, Φ_d are independent

(ii) α has a beta distribution with parameter (a_d, a_c)

(iii) Φ_c is distributed as a Dirichlet process with parameter (a_c, M_c)

and

(iv) Φ_d is distributed as a Dirichlet process with parameter (a_d, M_d)

Since M_d has a finite or a countable support, namely S, Φ_d has the same support with probability 1 and $\{\Phi_d(\{a_j\}) : j \in J\}$ for all J finite subset of I has a Dirichlet distribution with parameter $\{aM(\{a_j\}) : j \in J\}$. In particular, $\Phi_d(\{a_j\})$ has a beta distribution with parameter $(aM(\{a_j\}), a - aM(\{a_j\}))$. Furthermore, the support of this Dirichlet process is the set of all probabilities with support S. Now, Ferguson (1973) has shown that the trajectories of Φ_c are almost surely discrete. In other words, Φ_c may be represented as

$$\Phi_c = \sum_{1 \le j < \infty} \gamma_j \delta_{\tau_j} \tag{2.13}$$

almost surely (and not only in distribution). A first description of the distribution of $(\tau_j)_{1 \le j < \infty}$ and $(\gamma_j)_{1 \le j < \infty}$ has been provided by Ferguson (1973):

(i) $(\tau_j)_{1\leq j<\infty}$ is an infinite I.I.D. sample from M_c

(ii) the sequence $(\gamma_j)_{1\leq j<\infty}$ is normalized and decreasing i.e. such that $\sum_j \gamma_j = 1$ and $0 < \gamma_{j+1} < \gamma_j$ with probability one and may be represented as follows:

$$\gamma_j = (\sum_{1\leq l<\infty} J_l)^{-1} J_j \tag{2.14}$$

where the sequence $(J_j)_{1\leq j\leq\infty}$ is also decreasing $(J_{j+1} < J_j)$ and its distribution is as follows:

$$P[J_1 \leq t] = \exp\{-a_c \int_t^\infty u^{-1} e^{-u} du\} \qquad t \geq 0 \tag{2.15}$$

$$P[J_{j+1} \leq t | J_1 \cdots J_j] = \exp\{-a_c \int_t^{J_j} u^{-1} e^{-u} du\} \quad 0 \leq t \leq J_j \tag{2.16}$$

(iii) furthermore, the sequences $(\tau_j)_{1\leq j\leq\infty}$ and $(J_j)_{1\leq j\leq\infty}$ are independent and, eventually, so are also the sequences $(\tau_j)_{1\leq j\leq\infty}$ and $(\gamma_j)_{1\leq j\leq\infty}$.

A second description has been obtained in Rolin (1992b) where it is shown that there exists an infinite permutation of $\mathbb{N}$ such that, keeping the same representation of Φ_c as (2.13), (i) and the second part of (iii) remain valid and (ii) becomes:

(ii*) the sequence $(\gamma_j)_{1\leq j<\infty}$ may be represented as follows:

$$\gamma_j = \beta_j \prod_{1\leq \ell\leq j-1} (1 - \beta_\ell) \tag{2.17}$$

where $(\beta_j)_{1\leq j<\infty}$ is an infinite I.I.D. sample of the beta distribution with parameter $(1, a_c)$.

It should be remarked that (i) the distribution of (τ_j)-resp. of $(\gamma_j)-$ depends only on M_c-resp on a_c. (ii) even though M_c is diffuse the trajectories of Φ_c are almost surely discrete but the infinitely many jumps are randomly located and the support of Φ_c is almost surely dense in the support of M_c. In other words, for any set B such that $M_c(B) > 0$, $\Phi_c(B)$ is almost surely a strictly positive random variable. Furthermore, any probability absolutely continuous with respect to M_c is in the pointwise support of the Dirichlet process (recall that the pointwise convergence is defined by $M_n(B) \to M(B) \quad \forall\ B \in \mathcal{B}_+$).

Let us now consider what does this imply for the Bayesian model under a Dirichlet prior specification. When P_0 is a discrete probability measure, with jumps at a fixed set $S = \{a_j : j \in I\} \subset I\!R_+$ with $I \subset \mathbb{N}$, the sampling probabilities Φ will be almost surely discrete with the same support S, both a priori and a posteriori for almost any observation $t_1 \cdots t_n$ because P_* will have the same support as P_0. Furthermore, the support of the Dirichlet process is the set of all probabilities on S, both a priori and a posteriori. Also the sequence of the predictive distributions generating t_1 and $(t_{i+1}|t_1 \cdots t_i)$ will all be probabilities on S. Note that in such a case, the model is unable to handle an observation falling outside S.

When P_0 is diffuse, the prior trajectories are characterized as in (2.13). For the posterior trajectories, let us remark that $P_{*c} = P_0$ and $P_{*d} = P_n$. Therefore, if $(t_{(j)})_{1 \le j \le p}$ is the set of distinct values taken by $(t_1, \ldots, t_n)$ and if n_j is the number of t_i's that are equal to $t_{(j)}$, $1 \le j \le p$, P_n may be written as

$$P_n = \sum_{1 \le j \le p} \frac{n_j}{n} \delta_{t_{(j)}}$$

Therefore according to (2.12), Φ may be represented a posteriori as

$$\Phi = (1 - \alpha_n)\Phi_c + \alpha_n \Phi_d$$

where

$$\alpha_n = \sum_{1 \le j \le p} \Phi(\{t_{(j)}\})$$

and

$$\begin{aligned} \Phi_d &= \alpha_n^{-1} \sum_{1 \le j \le p} \Phi(\{t_{(j)}\}) \delta_{t_{(j)}} \\ &= \sum_{1 \le j \le p} \Phi_d(\{t_{(j)}\}) \delta_{t_{(j)}}. \end{aligned}$$

In this representation, conditionally on $(t_1, \ldots, t_n)$,

(i) α_n, Φ_c, Φ_d are independent.

(ii) α_n has a beta distribution with parameter (n, n_0)

(iii) Φ_c is a Dirichlet process with parameter (n_0, P_0)

(iv) Φ_d is a Dirichlet process with parameter (n, P_n)

or equivalently,

(iv*) $\{\Phi_d(\{t_{(j)}\}) : 1 \le j \le p\}$ has a Dirichlet distribution with parameter $(n_j : 1 \le j \le p)$

It should be remarked that: (i) the normalized part of Φ outside the observations, Φ_c, is not revised by the observations, i.e. is distributed as Φ a priori, (ii) the normalized part of Φ at the observations, Φ_d, has a distribution independent of the prior distribution, i.e. does not depend on (n_0, P_0). Finally, Φ is a convex combination of Φ_c and Φ_d the coefficient of which (α_n) has a distribution depending only on the prior precision and the sample size, i.e., (n_0, n).

"Non-informative" prior specification raises problems. A natural suggestion has indeed been to consider that $n_0 \to 0$. In this latter case, the behaviour of the posterior process is rather natural: the posterior process tends to a Dirichlet process with parameters (n, P_n). Note however the discontinuity at $n_0 = 0$ where the prior has the pathological representation of a random jump process: $\Phi = \delta_t$ with $t \sim P_0$.

2.2.3 Using Dirichlet process in duration models

When modelling duration data, it is rather natural to give structure to one of the following transformations of the sampling probability Φ: either the survivor function

$$\Sigma(t) = \Phi((t, \infty)) \tag{2.18}$$

or the cumulative hazard function defined in either one of the following two non equivalent ways:

$$\begin{aligned} \Lambda(t) &= -ln\Sigma(t) \\ &= \sum_{0 \le s \le t} \ln \frac{\Sigma(s-)}{\Sigma(s)} \end{aligned} \tag{2.19}$$

$$\begin{aligned} \hat{\Lambda}(t) &= \int_{[0,t]} [\Sigma(u-)]^{-1} \Phi(du) \\ &= \sum_{0 \le s \le t} \left[1 - \frac{\Sigma(s)}{\Sigma(s-)} \right] \\ &= \sum_{0 \le s \le t} [1 - \exp -(\Lambda(s) - \Lambda(s-))] \end{aligned} \tag{2.20}$$

Second expressions of (2.19) and (2.20) correspond to the situation where Φ is purely atomic.

Once Φ is distributed as a Dirichlet process with parameter (a, M), the law of the survivor process Σ is easily characterized. Indeed for any k and

any ordered k-tuple $t_1 < t_2 < \cdots t_k$, the joint distribution of the random vector $(\Sigma(t_1) \cdots \Sigma(t_k))$ is obtained from the joint distribution of $1 - \Sigma(t_1) = \Phi((0, t_1]), \Sigma(t_1) - \Sigma(t_2) = \Phi((t_1, t_2]), \cdots, \Sigma(t_{k-1}) - \Sigma(t_k) = \Phi((t_{k-1}, t_k])$, $\Sigma(t_k) = \Phi((t_k, \infty))$, which follows a Dirichlet distribution with parameters $(n_0[1 - S_0(t_1)], n_0[S_0(t_1) - S_0(t_2)], \cdots, n_0[S_0(t_{k-1}) - S_0(t_k)], n_0 S_0(t_k))$ where S_0 is the survivor function associated to P_0.

For the laws of Λ and of $\hat{\Lambda}$ note first that both Λ and $\hat{\Lambda}$ are a.s. non decreasing, that $\Lambda(0) = \hat{\Lambda}(0) = 0$ and $\Lambda(\infty) = \hat{\Lambda}(\infty) = \infty$ and therefore both can be viewed as a cumulative distribution function of a random σ-finite measure process on $I\!R_+$. Finally both Λ and $\hat{\Lambda}$ have independent non negative increments. For the case of Λ, this is a direct consequence of a basic property of the Dirichlet process . Indeed for any k and any ordered k-tuple $t_1 < t_2 \cdots < t_k$, the Dirichlet process makes $\Sigma(t_{j+1})(\Sigma(t_j))^{-1}$ independent of $(\Sigma(t_1), \Sigma(t_2), \cdots, \Sigma(t_j))$ and consequently of $(\Sigma(t_{i+1}))(\Sigma(t_i))^{-1} \quad \forall i < j)$; on taking logarithm one obtains the independence of the nonnegative increments of Λ. The third expression of $\hat{\Lambda}$ shows that $\hat{\Lambda}$ has the same properties. A stochastic process indexed by $t \geq 0$, starting at 0 for $t = 0$ with independent non negative increments is called a *Levy process* and the associated random measure is a *purely random* measure. Furthermore, as $(\Sigma(t_{j+1}))(\Sigma(t_j))^{-1}$ has a beta distribution with parameter $(aM((t_{j+1}, \infty)), aM((t_j, t_{j+1}]))$ (see Rolin 1992a), the distribution of each increment is log-beta with the same parameter.

Note that in a more general situation than the Dirichlet process, (first expressions of definitions (2.19) and (2.20)) the continuous parts of Λ and $\hat{\Lambda}$ are equal and therefore Λ and $\hat{\Lambda}$ are equal when Φ is diffuse. This is not the case when, as in the Dirichlet process, Φ is not diffuse. Definition (2.19) has the advantage to be easily inverted, indeed

$$\Sigma(t) = e^{-\Lambda(t)} \tag{2.21}$$

But inversion of definition (2.20) is more complicated, since

$$\Sigma(t) = \prod_{0 \leq s \leq t} [1 - \{\hat{\Lambda}(s) - \hat{\Lambda}(s-)\}] \tag{2.22}$$

However, the process $\hat{\Lambda}(t)$ has some nicer properties than the process $\Lambda(t)$. One of them, as seen in the second expression of definition (2.20), relies on the fact that $\hat{\Lambda}(t)$ has jumps of size smaller than one. This simplifies the Levy representation of the process and hence the modelling of more general prior than Dirichlet processes, e.g., Beta processes. (see Hjort (1990))

2.3 Properties of the predictive process.

Let us give more details on the structure of the predictive process as sketched at the end of section 2.1, namely:

$$t_1 \sim P_0 \tag{2.23}$$

$$(t_{i+1}|t_1 \cdots t_i) \sim \frac{n_0}{n_0+i} P_0 + \frac{1}{n_0+i} \sum_{1 \le j \le i} \delta_{t_j} \tag{2.24}$$

An important feature of this predictive process is to generate ties. More specifically, one has, for instance:

$$P(t_2 = t_1) = \frac{1}{n_0+1}, \tag{2.25}$$

when P_0 is diffuse. As a consequence, the predictive distribution of $t_1 \cdots t_n$, described sequentially in (2.23) (2.24) may also be characterized globally in an alternative way based on the following remark. The information contained in $(t_1 \cdots t_n)$ is equivalently described by $(C_n, (t_{(j)})_{1 \le j \le p})$ where C_n is a partition of $\{1, 2, \cdots n\}$ into p non empty elements $(1 \le p \le n)$, namely $C_n = \{I_j : 1 \le j \le p\}$, I_j is a configuration of ties, i.e. a non-empty subset of $\{1, \cdots n\}$ corresponding to the indices $i \in \{1, \cdots n\}$ for which the t_i's take a same value and $(t_{(j)})_{1 \le j \le p}$ is the vector of different values taken by $(t_1 \cdots t_n)$. Note that p is a function of C_n, namely $p = |C_n|$ where $|C_n|$ is the cardinality (the number of elements) of C_n.

Therefore the distribution of $(t_1 \cdots t_n)$ is equivalently described by the distribution of C_n, running over all partitions of $\{1, \cdots n\}$ into $1, 2, \cdots n$ elements, and the distribution of $(t_{(j)})_{1 \le j \le p}$, conditionally on C_n.

The marginal distribution of C_n is somewhat involved and is given with some details on its properties in Rolin (1992b). The distribution of $(t_{(j)})_{1 \le j \le p}$, conditionally on C_n may be described more easily. It depends essentially on p, the number of distinct values, and each $t_{(j)}$ is otherwise I.I.D. with distribution P_0 when this one is diffuse.

3 Heterogeneity and Mixture of Dirichlet Processes.

3.1 Introduction.

Problems raised by the non-observability of some exogenous variables is a recurrent theme in the literature of micro-econometrics, particularly when dealing with duration data. This topic is known under the heading of "heterogeneity". In this section we review a simple multiplicative model that offers a workable solution to this kind of difficulty. It should be noted that economic theory would typically give poor information on the precise effect of those non-observable variables although statistical analysis shows that neglecting such features may produce serious specification errors jeopardizing

the reliability of any inference. It is therefore natural that non-parametric methods have attracted attention in this topic in as far as they may provide robust specification around aspects of the specification that are typically of a nuisance character.

The discrete aspect of the trajectories of a Dirichlet process with smooth location parameter may be considered as a natural counterpart to its natural conjugate property with respect to the empirical process which is also genuinely discrete. Many non-parametric procedures, particularly those based on kernel methods, aim at smoothing the empirical process to permit the estimation of a smooth characteristic of the population distribution, as for instance the estimation of the density function or the estimation of a derivative of the regression function.

It so happens that the simple model to be presented in this section is powerful enough to deal both with the problem of heterogeneity and with the discreteness of the Dirichlet process.

3.2 A simple model.

Let us consider n observed durations $(t_1 \cdots t_i \cdots t_n)$ along with the following multiplicative decomposition:

$$t_i = \theta_i \tau_i \qquad 1 \leq i \leq n \tag{3.1}$$

Both θ_i and τ_i are unobservable and I.I.D. but the common distribution of the θ_i's is assumed to be known, for reasons related to the identification problem, and to admit a density whereas the common distribution of the τ_i's is unknown and is assumed to be distributed a priori as a Dirichlet process.

In the context of the heterogeneity problem, model (3.1) may be interpreted as follows. The duration τ_i of each individual i is assumed to be generated by a same unknown distribution but accelerated by a factor θ_i reflecting unobservable individual characteristics. Assuming known the distribution of the factor θ_i is a simple solution to the standard problem of identification in accelerated time models.

It is interesting to notice that because the t_i's are representable as the product of two quantities their sampling distribution is a.s. smooth, i.e. admits a density a.s., in spite of the Dirichlet specification. In others words the introduction of θ_i acts similarly to a smoothing kernel , a main difference being a multiplicative convolution rather than an additive one. The multiplicative model is indeed more natural than the additive one when dealing with non negative random variables. One of the results of this model is to produce a Bayesian analogue to kernel estimators.

Let us now be more specific about the basic assumptions underlying (3.1)

(A.1) $(\theta_i)_{1\leq i\leq n}$ are I.I.D.; their common distribution is denoted Q, is known, is supported by $\mathbb{R}_+$ and admits a known density q.

(A.2) $(\tau_i)_{1\leq i\leq n}$ are I.I.D.; the common distribution is denoted Γ, is unknown and distributed a priori as a Dirichlet process with parameters (n_0, G_0) where G_0 is a probability measure supported by $\mathbb{R}_+$.

(A.3) The θ_i's and the τ_i's are jointly independent in the sampling, i.e.:

$$(\theta_i)_{1\leq i\leq n} \perp\!\!\!\perp (\tau_i)_{1\leq i\leq n} | \Gamma \tag{3.2}$$

These assumptions imply that the t_i's are I.I.D. in the sampling, the common distribution denoted Φ, being described as follows:

$$\begin{aligned} \Phi([0,t]) &= \int_{\mathbb{R}_+} \Gamma([0,\frac{t}{\theta}]) Q(d\theta) \\ &= \int_{\mathbb{R}_+} Q([0,\frac{t}{\tau}]) \Gamma(d\tau) \end{aligned} \tag{3.3}$$

The multiplicative convolution, defined by (3.3), is denoted $Q.\Gamma$. Because Q admits a density, q, (3.3) may be rewritten as:

$$\Phi([0,t]) = \int_0^t \int_{\mathbb{R}_+} \frac{1}{\tau}\ q(\frac{u}{\tau})\ \Gamma(d\tau)du \tag{3.4}$$

Thus Φ is dominated by Lebesgue measure and admits φ as a density:

$$\varphi(t) = \int_{\mathbb{R}_+} \frac{1}{\tau}\ q(\frac{t}{\tau})\ \Gamma(d\tau) \tag{3.5}$$

When $\Gamma \sim \mathcal{D}i(n_0, G_0)$ the distribution of Φ, derived from (3.3), is an example (a particular case) of a mixture of Dirichlet processes. One may easily check that:

$$E[\Phi] = Q.E(\Gamma) = Q.G_0 \tag{3.6}$$

$$\frac{d}{dt}E[\Phi([0,t])] = \int_{\mathbb{R}_+} \frac{1}{\tau}\ q(\frac{t}{\tau})\ G_0(d\tau) \tag{3.7}$$

The posterior distribution of Φ is somewhat involved. Let us have a look on its expectation and evaluate it first conditionally on $(\tau_1 \cdots \tau_n)$:

$$E[\Phi | t_1 \cdots t_n, \tau_i \cdots \tau_n] = \frac{n_0}{n_0+n}(Q.G_0) + \frac{n}{n_0+n}(Q.G_n) \tag{3.8}$$

where G_n is the empirical distribution of the τ_i's:

$$G_n = n^{-1} \sum_{1\leq i\leq n} \delta_{\tau_i} \tag{3.9}$$

In order to integrate the τ_i's out, it is convenient to represent $(\tau_1 \cdots \tau_n)$ through three components namely: p, the number of the different values of

the τ_i's, a partition C_n of $\{1, \cdots n\}$ into p elements $(I_j)_{1\leq j\leq p}$ where each I_j is a set of indices corresponding to identical values of τ_i (thus C_n represents the partition of the configuration of ties in the τ_i's) and finally the p-vector of distinct values of the τ_i's: $(\tau_{(j)})_{1\leq j\leq p}$. Note that G_n may also be written as:

$$G_n = \frac{1}{n} \sum_{1\leq j\leq p} n_j \delta_{\tau_{(j)}} \tag{3.10}$$

where $n_j = |I_j|$ and consequently $\sum_{1\leq j\leq p} n_j = n$. Note also that

$$t_i = \tau_{(j)}\theta_i \qquad \forall i \in I_j \tag{3.11}$$

where this expression is based on the configuration C_n. Let us now denote G_n^j for the conditional probability of $\tau_{(j)}$ conditionally on $t_1 \cdots t_n$ and C_n. Thus:

$$G_n^j(A) = E[\mathbb{1}_A(\tau_{(j)}) \mid t_1 \cdots t_n, C_n] \tag{3.12}$$

It may be shown, and it is intuitively rather natural, that G_n^j depends only on those t_i's for which $i \in I_j$ and may be evaluated easily using standard techniques for evaluating posterior distributions; this is so, in particular, because each θ_i is I.I.D. with known distribution Q and each $\tau_{(j)}$ is a priori distributed according to G_0. Later on we give an example for evaluating G_n^j.

We may now integrate, at least formally, $\tau_1 \cdots \tau_n$ out of (3.8) as follows:

$$\begin{aligned} E[\Phi|t_1 \cdots t_n] &= \frac{n_0}{n_0+n}(Q \cdot G_0) \\ &+ \frac{1}{n_0+n} \sum_{C_n \in \mathcal{C}_n} P(C_n|t_1 \cdots t_n) \sum_{1\leq j\leq p} n_j \, (Q \cdot G_n^j) \end{aligned} \tag{3.13}$$

where $\mathcal{C}_n$ is the set of all partitions of $\{1, \cdots n\}$. Unfortunately, an exact evaluation of (3.13) is close to impossible for reasonably large sample size n. Indeed, the evaluation of $P(C_n|t_1 \cdots t_n)$ is rather involved in view of the analysis given in section 2.3 and of the fact that $|\mathcal{C}_n|$ increases dramatically with n.

Different strategies can be envisaged for facing this difficulty. Two of these are the following. A first strategy consists in selecting arbitrarily a configuration C_n of ties. This is leaving a strict Bayesian framework as far as one is conditioning on a non-available information. Although general theorems may ensure the convergence of $E(\Phi|t_1 \cdots t_n)$, the convergence of $E(\Phi|t_1 \cdots t_n, C_n)$ requires specific hypotheses, an example of which will be examined later.

A second kind of strategies relies on simulation methods (post-data sampling). Here, Gibbs sampling is a natural candidate because of two arguments. Firstly the evaluation of $E[\Phi|t_1 \cdots t_n, C_n]$ is easy because, in view of (3.3) it is equivalent to evaluate:

$$\begin{aligned} E[\Gamma|t_1 \cdots t_n, C_n] &= \frac{n_0}{n_0+n} G_0 + \frac{1}{n+n_0} \sum_{1 \le j \le p} n_j G_n^j \\ &= G_n^*. \end{aligned} \tag{3.14}$$

To evaluate G_n^* by simulation, we have to simulate C_n conditionally on $(t_1, t_2, \ldots, t_n)$ or more generally, $(\tau_1, \tau_2, \ldots \tau_n)$ conditionally on $(t_1, t_2, \ldots, t_n)$. This last distribution relies on the distribution of $(\tau_1, \tau_2, \ldots \tau_n)$ which, as seen in section 2.3, is rather complicated and almost unmanageable if n is large.

However, the conditional distributions are rather simple (formula (2.24)). Here is the second argument in favour of the Gibbs sampler since this simulation only uses the conditional distributions. More precisely, the Gibbs sampler will be based on the conditional probabilities

$$P[\tau_i \in A | \tau_j, j \neq i, t_j, 1 \le j \le n] = q_i(t_i) G_0^{t_i}(A) + \sum_{j \neq i} q_j(t_i) \delta_{\tau_j}(A) \tag{3.15}$$

where $G_0^{t_i}$ is the distribution of τ_i conditionally on t_i, $q_i(t_i)$ is proportional to $n_0 g(t_i)$ where $g(t_i)$ is the density of $(Q \cdot G_0)$ and $q_j(t_i)$ is proportional to $\frac{1}{\tau_j} q\left(\frac{t_i}{\tau_j}\right)$ for $j \neq i$ and $\sum_{1 \le j \le n} q_j(t_i) = 1$. The Gibbs sampler will then generate $(\tau_1, \tau_2, \ldots \tau_n)$ from the distribution conditional on $(t_1, t_2, \ldots, t_n)$ from which G_n^* will be easily computed.

3.3 A particular case.

Let us give a simple example of such computations.

Let Q be the inverse gamma distribution with parameters $(\theta_0, \nu_0) \in \mathbb{R}_+^2$ whose density is given by

$$q(\theta) = \frac{\theta_0^{\nu_0}}{\Gamma(\nu_0)} \theta^{-(\nu_0+1)} e^{-\theta_0/\theta} \tag{3.16}$$

and let G_0 be the gamma distribution with parameters $(\tau_0, \mu_0) \in \mathbb{R}_+^2$ whose density is given by

$$g_0(\tau) = \frac{1}{\Gamma(\mu_0)} \tau_0^{-\mu_0} \tau^{\mu_0 - 1} e^{-\tau/\tau_0} \tag{3.17}$$

Clearly the distribution of t_i conditionally on τ_i is the inverse gamma with parameters $(\theta_0 \tau_i, \nu_0)$ since its density is given by $\frac{1}{\tau_i} q\left(\frac{t}{\tau_i}\right)$.

Simple Bayesian computations show that the distribution of τ_i conditionally on t_i, denoted above as $G_0^{t_i}$, is the gamma distribution with parameters $\left[\left(\frac{1}{\tau_0}+\frac{\theta_0}{t_i}\right)^{-1}, \mu_0+\nu_0\right]$. Furthermore the distribution of t_i, denoted above as $(Q \cdot G_0)$ is the Fisher distribution with parameters $(\theta_0\tau_0, \mu_0, \nu_0)$ whose density is given by

$$g(t) = \frac{\Gamma(\mu_0+\nu_0)}{\Gamma(\mu_0)\Gamma(\nu_0)} \frac{(\theta_0\tau_0)^{-\mu_0} t^{\mu_0-1}}{\left(1+\frac{t}{\theta_0\tau_0}\right)^{\mu_0+\nu_0}} \tag{3.18}$$

Now conditionally on $\tau_{(j)}$ and C_n, the t_i's for $i \in I_j$ are distributed independently following an inverse gamma distribution with parameters $(\theta_0\tau_{(j)}, \nu_0)$. Therefore, the likelihood is proportional to

$$\prod_{i\in I_j} \tau_{(j)}^{\nu_0} e^{-\theta_0\tau_{(j)}/t_i} = \tau_{(j)}^{n_j\nu_0} e^{-\theta_0\tau_{(j)}\sum_{i\in I_j} t_i^{-1}}$$

Therefore, by Bayes' theorem, the distribution of $\tau_{(j)}$ conditionally on $\{t_i; 1 \leq i \leq n\}$ and C_n denoted before as G_n^j is the gamma distribution with parameters

$$\left[\left(\frac{1}{\tau_0}+\theta_0\sum_{i\in I_j}\frac{1}{t_i}\right)^{-1}, \mu_0+n_j\nu_0\right]$$

Finally, by the same computation as before, $Q \cdot G_n^j$ is the Fisher distribution with parameters

$$\left[\left(\frac{1}{\theta_0\tau_0}+\sum_{i\in I_j}\frac{1}{t_i}\right)^{-1}, \mu_0+n_j\nu_0, \nu_0\right]$$

This implies that the posterior expectation of Γ conditionally on C_n has a density given by a convex combination of gamma densities and that the posterior expectation of Φ conditionally on C_n has a density given by a convex combination of Fisher densities.

4 Non-Parametric iid duration models with censored observations.

One of the main features of duration data sets is the presence of censored observations. In this section, we want to show the limits of the extensions of the results presented in section 2 to non stochastically right-censored durations. Let us first recall that, in this case, the sample is generated as follows: for

any $i = 1, \cdots, n$, c_i is a fixed duration, τ_i is a latent duration identically and independently distributed from an unknown probability Φ and we observe:

$$\begin{aligned} t_i &= \min(\tau_i, c_i) \\ d_i &= \mathbb{1}_{\{\tau_i \leq c_i\}} \end{aligned} \tag{4.1}$$

The unknown functional parameter is provided with a prior probability (a Dirichlet process in our presentation) and we want to analyse the posterior probability of Φ given $(t_i, d_i)_{1 \leq i \leq n}$.

The main results of this analysis are the following:

i) the family of Dirichlet processes is not closed for such a sampling. In others words the posterior probability deduced from a Dirichlet prior and a censored sample is not a Dirichlet probability. However it is possible to construct a class of prior probability measures on Φ, larger than the Dirichlet family, and called the class of neutral processes, which is closed for the Bayesian inference.

ii) Elements of this class are not so simple as Dirichlet processes but some of their characteristics are obtained analytically. This is in particular the case of the expectation of the survivor function. An application of this computation is the posterior expected survivor function constructed through a Dirichlet prior (see Susarla and Van Ryzin (1978a)).

iii) Recently Hjort (1990) introduces a subclass of neutral process that he calls Beta processes which generalize the Dirichlet processes and that are also closed for sampling with right-censoring.

In order to point out the essential elements of this construction we will start by the treatment of the "finite" case in which the support of Φ is finite. If we assume that the values c_i are also elements of this set, we may describe the sample using different counting statistics.

Let $\{a_1, \cdots a_k\}$ be the support of Φ (with $a_1 < a_2 < \cdots a_k$) and n the sample size. For any $j = 1, \cdots, k$ we define:

- e_j the number of non censored durations equal to a_j

- h_j the number of censored or non censored durations equal to a_j, i.e.:

$$h_j = \sum_{1 \leq i \leq n} \mathbb{1}_{\{t_i = a_j\}}, \quad e_j = \sum_{1 \leq i \leq n} \mathbb{1}_{\{t_i = a_j, d_i = 1\}} \tag{4.2}$$

- n_j is the number of individuals at risk in a_j, i.e.

$$n_j = \sum_{j \leq \ell \leq k} h_\ell = \sum_{1 \leq i \leq n} \mathbb{1}_{\{t_i \geq a_j\}} \tag{4.3}$$

Note that, in particular, $n_1 = n$ and that $n = \sum_{1\le j\le k} h_j$ and $h_j - e_j$ is the number of censored data at a_j.
Also, the statistic $(h_j, e_j)_{1\le j\le k}$ is obviously sufficient.

The probability Φ on $\{a_1, \cdots, a_k\}$ may be described in different ways:

i) the sequence $(\theta_j)_{1\le j\le k}$, $\theta_j > 0$ $\sum_{1\le j\le k} \theta_j = 1$ of the probabilities of the a_j's.

or,

ii) from θ_j we construct the survivor function:

$$\Sigma(t-) = \sum_{a_j \ge t} \theta_j = \sum_j \theta_j \mathbb{1}_{\{a_j \ge t\}} \tag{4.4}$$

iii) Φ is also characterized by the hazard rates sequence $(\lambda_j)_{1\le j\le k}$ defined by:

$$\lambda_j = \frac{\theta_j}{\Sigma_j} \quad \text{where } \Sigma_j = \Sigma(a_j-). \tag{4.5}$$

In particular, the well known product formula connect θ_j and Σ_j to λ_j:

$$\theta_j = \lambda_j \prod_{\ell<j}(1-\lambda_\ell) \quad \Sigma_j = \prod_{\ell<j}(1-\lambda_\ell). \tag{4.6}$$

The likelihood of the sufficient statistic may be written in terms of the probabilities or in terms of the hazard rates:

$$\ell((h_j, e_j)_{1\le j\le k} \mid (\theta_j)_{1\le j\le k}) = \prod_{1\le j\le k} \theta_j^{e_j} \Sigma_j^{h_j - e_j} \tag{4.7}$$

or

$$\ell((h_j, e_j)_{1\le j\le k} \mid (\lambda_j)_{1\le j\le k}) = \prod_{1\le j\le k-1} \lambda_j^{e_j} (1-\lambda_j)^{n_j - e_j}. \tag{4.8}$$

Let us now consider the prior probability. In term of θ_j, a Dirichlet prior distribution has density:

$$m((\theta_j)_{1\le j\le k}) \propto \prod_{1\le j\le k} \theta_j^{n_0 p_{0j} - 1} \tag{4.9}$$

with respect to the uniform density restricted to the simplex in $\mathbb{R}^k$. Using previous results given in section 2.2.3 or after some manipulations, the prior density defined in (4.9) may be rewritten as:

$$m((\lambda_j)_{1\le j\le k}) \propto \prod_{1\le j\le k-1} \lambda_j^{n_0 p_{0j} - 1} (1-\lambda_j)^{n_0 S_{0\, j+1} - 1} \tag{4.10}$$

with respect to the uniform measure restricted to the set of hazard rates in $\mathbb{R}^k$. In (4.10) S_0 is the expected prior survivor function:

$$S_0(t-) = \sum_{1 \leq j \leq k} p_{0j} \mathbb{1}_{\{a_j \geq t\}} \text{ and } S_{0j} = S_0(a_j-) \tag{4.11}$$

The parameters $(\lambda_1, \cdots, \lambda_{k-1})$ are independently Beta distributed with parameter $(n_0 p_{0j}, n_0 S_{0\,j+1})$, $1 \leq j \leq k-1$. An interesting property is the following:

$$E(\lambda_j) = \frac{n_0 p_{0j}}{n_0 p_{0j} + n_0 S_{0\,j+1}} = \frac{p_{0j}}{S_{0j}} = \lambda_{0j}. \tag{4.12}$$

In others words the prior expected hazard rates are the hazard rates of the prior expected distribution.

The posterior probability on the θ_j's has the following density:

$$m((\theta_j)_{1 \leq j \leq k} | (h_j, e_j)_{1 \leq j \leq k}) \propto \tag{4.13}$$

$$\prod_j \theta_j^{n_0 p_{0j} + e_{j-1}} \Sigma_{j+1}^{h_j - e_j}, \tag{4.14}$$

which is clearly not a Dirichlet distribution and the lack of a closedness property of the class of the Dirichlet distributions is then pointed out.

The posterior density on the λ_j's is equal to:

$$m((\lambda_j)_{1 \leq j \leq k-1} \mid (h_j, e_j)_{1 \leq j \leq k}) \propto \tag{4.15}$$

$$\prod_{1 \leq j \leq k} \lambda_j^{n_0 p_{0j} + e_j - 1} (1 - \lambda_j)^{n_0 S_{0\,j+1} + n_j - e_j - 1} \tag{4.16}$$

It results from (4.16) that the $\lambda_j (1 \leq j \leq k-1)$ are a posteriori independent. Any hazard rate λ_j follows a posteriori a Beta distribution whose parameters are equal to $n_0 p_{0j} + e_j$ and $n_0 S_{0\,j+1} + n_j - e_j$. This sequence of $k-1$ independent Beta probabilities looks like the prior specification but, excepted in the case of non censored data, these distributions cannot be derived from a Dirichlet posterior on $(\theta_j)_{1 \leq j \leq k}$. However the product computation of the survivor function and the posterior independence of the λ_j imply that:

$$\begin{aligned} S_{nj} = E(\Sigma_j | (h_j, e_j), 1 \leq j \leq k) &= E(\textstyle\prod_{\ell<j}(1-\lambda_\ell) | (h_j, e_j), 1 \leq j \leq k) \\ &= \textstyle\prod_{\ell<j}(1 - E(\lambda_\ell | (h_j, e_j), 1 \leq j \leq k) \\ &= \textstyle\prod_{\ell<j}(1 - \frac{n_0 p_{0j} + e_j}{n_0 S_{0j} + n_j}) \end{aligned} \tag{4.17}$$

The Bayesian estimation $S_n(t)$ is deduced from the S_{nj} using the property that the survivor function is constant between the jumps in a_j and is right continuous.

It is easily seen how we can generalize the Dirichlet process in such a way that one obtains the closedness property for the Bayesian inference. Indeed, if a priori the λ_j's are independently and Beta distributed with parameters $(c_{0j}\lambda_{0j}, c_{0j}(1-\lambda_{0j}))$ where c_{0j}, $1 \leq j \leq k-1$, are arbitrary constants, then a posteriori the λ_j's are independently and Beta distributed with parameters $(c_{*j}\lambda_{*j}, c_{*j}(1-\lambda_{*j}))$ where $c_{*j} = c_{0j} + n_j$ and $\lambda_{*j} = \dfrac{c_{0j}\lambda_{0j} + e_j}{c_{0j} + n_j}$.
This is at the basis of the definition of Beta processes.

Let us consider the general case in which the functional parameter Φ is not constrained by a finite support condition. The previous example shows that the family of Dirichlet prior measures is not closed by the inference in presence of censored durations. The new class of prior measures is the class of neutral processes which extends the Dirichlet family.

By definition Φ is provided with a neutral prior process if its associated integrated hazard function $\Lambda(t) = \ln \Sigma(t)$ (where $\Sigma(t) = \Phi((t, \infty))$ is a Levy process. Such a process has been defined in section 2.2.3. Let us recall that a Levy process is increasing with independent increments. For any sequence $t_1 < t_2 < \cdots < t_k$, the random variables $r_j = \Lambda(t_j) - \Lambda(t_{j-1})$ are independently distributed with densities $m_j(r_j)$ $1 \leq j \leq k$. In the case of a Dirichlet prior distribution, we have seen that the r_j are log-beta distributed. It is sufficient to prove that for any observation t (censored or not) the posterior measure given t is still a Levy process.

First let us assume that t is noncensored. In order to compute the posterior distribution of Λ restricted to any finite family of increments, we just have to consider the marginalized likelihood given this family. This likelihood reduces to the probability of the interval $(t_\ell, t_{\ell+1}]$ to which belongs the observed t. Then the posterior probability of $(r_j)_{1 \leq j \leq k}$ is proportional to :

$$
\begin{aligned}
&[\prod_{1 \leq j \leq k} m_j(r_j)](\Sigma(t_\ell) - \Sigma(t_{\ell+1})) \\
&= [\prod_{1 \leq j \leq k} m_j(r_j)][\exp - \sum_{1 \leq j \leq \ell} r_j][1 - \exp -r_{\ell+1}] \qquad (4.18) \\
&= [\prod_{1 \leq j \leq \ell} m_j(r_j) \exp -r_j][m_{\ell+1}(r_{\ell+1})(1 - \exp -r_{\ell+1})][\prod_{\ell+2 \leq j \leq k} m_j(r_j)]
\end{aligned}
$$

The r_j's are a posteriori independent. The posterior density of the first ℓ are proportional to $m_j(r_j)\exp(-r_j)$, the posterior density of $r_{\ell+1}$ is proportional to $m_{\ell+1}(r_{\ell+1})(1 - \exp -r_{\ell+1})$ and the distribution of the last increments is identical a priori and a posteriori.

An identical computation can be done in the case of a censored obervation $t \in (t_\ell, t_{\ell+1}]$. The marginalized likelihood on the finite family of increments is equal to $\Sigma(t_\ell) = \exp - \sum_{1 \leq j \leq \ell} r_j$ and the posterior distribution of the r_j's is identical to the previous one excepted for $r_{\ell+1}$ which becomes unrevised by the observation.

Given a prior distribution on $\Lambda(t)$ one may derive the prior distribution of any sequence of increments and compute the posterior using previous argument sequentially on the sample. We briefly sketch the arguments which can be used to obtain the posterior expectation of $\Sigma(t)$. Given a Levy process $\Lambda(t)$ the generating function associated to $\Lambda(t)$ is defined by:

$$g(s,t) = -lnE[\exp -s\Lambda(t)] \quad \forall s,t \in I\!R^+ \tag{4.19}$$

Then:

$$E(\Sigma(t)) = \exp -g(1,t). \tag{4.20}$$

The problem reduces to the computation of the generating function of a Levy process. This function is easy to obtain if $\Lambda(t)$ is a Dirichlet process (see e.g. Hakizamungu (1992)) but as it was noticed before, for the posterior distribution, this is only the case in the absence of censored data. In the case of possibly censored data we don't detail the computation and we just give the result:

$$\begin{aligned} E(\Sigma(t)|(t_i,d_i)_{1\leq i\leq n}) \quad &= \frac{n_0 S_0(t) + n_\ell}{n_0 + n} \\ &\prod_{1\leq j\leq \ell} \frac{n_0 S_0(u_j) + n_j + f_j}{n_0 S_0(u_j) + n_j} \qquad t \in [u_\ell, u_{\ell+1}) \end{aligned} \tag{4.21}$$

where $u_1, u_2, \cdots, u_k$ are the distinct observed values of the sample, n_j $(j = 1, \cdots, k)$ is the number of individuals at risk after u_j i.e. $n_j = \sum_{1\leq i\leq n} I\!\!1_{\{t_i > u_j\}}$, f_j is the number of censored observations at u_j and S_0 is the prior survivor function. Let us remark that, in absence of censoring, (4.21) reduces to the usual result given in section 2.

5 Semi-parametric model with proportional hazard.

The last section of this article is devoted to a Bayesian treatment of the semi-parametric analysis of duration models conditional to observed explanatory variables. We restrict our attention to fixed explanatory variables (i.e. only

dependent on the individuals but not on time) acting through a proportional hazard model. Moreover we assume for the sake of exposition that the data are observed without censoring and we don't introduce unobserved heterogeneity component.

The sample is now defined by the sequence (t_i, z_i) $i = 1, \cdots, n$ where t_i is a duration and z_i a vector of explanatory variables. The observations are assumed to be independent and the sampling process is characterized by the distribution of t_i conditionally on z_i. This conditional probability Φ_i is defined through its survivor function Σ_i which verifies:

$$\Sigma_i(t) = \Sigma(t)^{a(\beta, z_i)} \tag{5.1}$$

where Σ is an unknown survivor function of a baseline probability Φ and $a(\beta, z_i)$ is a known positive function of an unknown vector of parameters β and of the explanatory variables z_i. A common choice for a is $a(\beta, z_i) = \exp \beta' z_i$.

The specification (5.1) is equivalent to:

$$\Lambda_i(t) = a(\beta, z_i)\Lambda(t) \tag{5.2}$$

where $\Lambda_i(t)$ is the conditional integrated hazard associated to $\sum_i($ i.e.$\Lambda_i(t) = -ln \sum_i(t))$ and $\Lambda(t)$ the integrated hazard associated to the baseline survivor function $\sum(t)$. Relation (5.2) justifies the name " proportional hazard model".

The sampling model is then indexed by a functional parameter (the baseline probability Φ or equivalently $\sum$ or Λ) and by a vector of real parameters β. The Bayesian specification is completed by the choice of a prior distribution. Given β, Φ is endowed with a Dirichlet process with parameters (n_0, Φ_0) that are possibly dependent on β. The vector β has a prior density $m(\beta)$.

This section is essentially devoted to the computation of the posterior distribution of β. This means that the functional parameter is treated as a nuisance parameter and integrated out analytically. However the posterior distribution of β must be treated numerically (in order to compute, for example, its moments or its marginal densities). There is no suitable choice of $m(\beta)$ which simplifies the posterior computations and this density will be kept unspecified.

In order to simplify the computations we assumed that the observed sample does not present ties and has n distinct durations. This assumption may be easily verified and the following derivations may be extended to samples with ties (see Ruggiero(1992)). Under the independence property one may assume without loss of generality that $t_1 < t_2 < \cdots < t_n$.

Let us start with the joint sampling survivor function:

$$\begin{aligned}\prod_{1\leq i\leq n}\Sigma_i(t_i) &= \prod_{1\leq i\leq n}\Sigma(t_i)^{a(\beta,z_i)}\\ &= \exp-\sum_{1\leq i\leq n}a(\beta,z_i)\Lambda(t_i)\\ &= \exp-\sum_{1\leq i\leq n}a(\beta,z_i)\sum_{1\leq j\leq i}\gamma_j\end{aligned}\tag{5.3}$$

where $\gamma_j = \Lambda(t_j) - \Lambda(t_{j-1})$ for $j > 1$ and $\gamma_1 = \Lambda(t_1)$; therefore:

$$\prod_{1\leq i\leq n}\Sigma_i(t_i) = \prod_{1\leq i\leq n}\exp-\gamma_i A_i(\beta)\tag{5.4}$$

where $A_i(\beta) = \sum_{i\leq j\leq n} a(\beta, z_j)$. Note that the joint sampling survivor function depends only on the functional parameter Λ through the sequence γ_i. As a consequence, integrating out the functional parameter Λ may be performed by integrating out the vector of positive random variables γ_i's out of (5.4). Furthermore, the Dirichlet prior on Φ implies that the γ_i are a priori independent with logBeta prior density

$$\begin{aligned}m(\gamma_i) &= \frac{\Gamma(n_0\Sigma_0(t_{i-1}))}{\Gamma(n_0\Sigma_0(t_i))\Gamma(n_0\Sigma_0(t_{i-1}) - n_0\Sigma_0(t_i))}\\ &\times\ (e^{-\gamma_i})^{n_0\Sigma_0(t_i)}(1-e^{-\gamma_i})^{n_0(\Sigma_0(t_{i-1})-\Sigma_0(t_i))-1}\end{aligned}\tag{5.5}$$

The marginalized joint survivor function is therefore equal to:

$$\begin{aligned}S(t_1,\cdots,t_n|\beta) &= \prod_{1\leq i\leq n}\int_0^\infty m(\gamma_i)\exp-\gamma_i A_i(\beta)d\gamma_i\\ &= \frac{\Gamma(n_0)}{\Gamma(n_0\Sigma_0(t_n))}\prod_{1\leq i\leq n}\frac{\Gamma(n_0\Sigma_0(t_i)+A_i(\beta))}{\Gamma(n_0\Sigma_0(t_{i-1})+A_i(\beta))}\\ &= \frac{\Gamma(n_0)}{\Gamma(n_0+A_1(\beta))}\prod_{1\leq i\leq n}\frac{\Gamma(n_0\Sigma_0(t_i)+A_i(\beta))}{\Gamma(n_0\Sigma_0(t_i)+A_{i+1}(\beta))}\end{aligned}\tag{5.6}$$

The marginalized sampling density conditionally on β is obtained through the computation of the derivative of the marginalized survivor function:

$$\ell(t_1, \cdots, t_n|\beta)$$

$$= (-1)^n \frac{\partial^n}{\partial t_1 \cdots \partial t_n} S(t_1, \cdots, t_n|\beta)$$

$$= (-1)^n \frac{\Gamma(n_0)}{\Gamma(n_0 + A_1)} \prod_{1 \leq i \leq n} n_0 \Sigma'_{0i} \tag{5.7}$$

$$[\Gamma'(n_0\Sigma_{0i} + A_i)\Gamma(n_0\Sigma_{0i} + A_{i+1}) - \Gamma'(n_0\Sigma_{0i} + A_{i+1})\Gamma(n_0\Sigma_{0i} + A_i)]$$

$$[\Gamma(n_0\Sigma_{0i} + A_{i+1})]^{-2}$$

where Γ' is the derivative of the gamma function and $\Sigma_{0i} = \Sigma_0(t_i)$.

The posterior density of β follows from Bayes rule:

$$m(\beta|t_i, \cdots, t_n) \propto m(\beta)\ell(t_1, \cdots, t_n|\beta). \tag{5.8}$$

We have computed the joint marginalized survivor function in the (open) subset of $(I\!R_+)^n$ in which all the durations are different and this function is differentiable on this subset. However the marginalized survivor function is not differentiable on the subsets defined by configurations of ties and the application of Bayes theorem becomes more complicated if the sample has nties; (see Ruggiero (1992) for a general description of this computation). A completed analysis of the marginalized density in the case of n distinct values (computation of score and second derivative) can be found in Hakizamungu (1992).

The marginal posterior probability of Φ given β cannot be computed using the technique of section 2 because, in the proportional hazard model, there is in general no transformation of the t_i whose distribution is the baseline probability (see Kalbfeisch (1978) and Hakizamungu (1992)).

A last natural question concerns the relation between the Bayesian marginalized likelihood (5.7) and the Cox marginalized likelihood. Note first that the word "marginalized" has different signification in the two approaches: in the Bayesian analysis the marginalization is obtained through an integration of the nuisance parameter Φ using a prior probability and in Cox analysis the marginalization is realized on the rank statistic in the sampling distribution. However some connection between the two results would have been expected in the case of "non informative" prior measure Φ (or on Γ).

A natural way is to compute the posterior distribution of β with a uniform prior measure on β, i.e., $m(\beta) = 1$ in (5.8) and to take its limit when $n_0 \to 0$. The result becomes:

$$m(\beta|t_1,\cdots,t_n) \propto \frac{1}{\Gamma(A_1)} \prod_{i=1}^{n} \frac{\Gamma'(A_i)\Gamma(A_{i+1}) - \Gamma'(A_{i+1})\Gamma(A_i)}{[\Gamma(A_{i+1}]^2} \tag{5.9}$$

Note that in (5.9) individual duration t_i disappears and the rank statistics becomes sufficient but this posterior density is finally different from the Cox density on β.

BIBLIOGRAPHIE

Antoniak, C.E. (1974), Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, *Annals of Statistics*, **2**, 1152-1174.

Bunke, O. (1981), Bayesian Estimators in Semiparametric Problems, Preprint 102, Humboldt University, Berlin.

Ferguson, T. (1973), A Bayesian Analysis of Some Nonparametric Problems, *Annals of Statistics*, **1**, 209-230.

Ferguson, T.(1974), Prior Distributions on Spaces of Probability Measures, *Annals of Statistics*, **2**, 615-629.

Ferguson, T. and E.G.. Phadia (1979), Bayesian Nonparametric Estimation Based on Censored Data, *Annals of Statistics*, **7**, 163-186.

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

Florens, J.P, Mouchart, M. and J.M. Rolin (1992), Bayesian Analysis of Mixtures: Some Results on Exact Estimability and Identification. In Bernado, J.M. and al., *Bayesian Statistics 4.* Oxford Science Publications, 127-145.

Hakizamungu, J. (1992), Modèles de Durée : Approche Bayésienne Semiparamétrique, Unpublished Ph.D. Dissertation, Faculté des Sciences, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Hjort, N. (1990), Nonparametric Bayes Estimators Based on Beta Processes in Models For Life History Data, *Annals of Statistics*, **18**, 1259-1294.

Kalbfeisch, J.D. (1978), Nonparametric Bayesian Analysis of Survival Time Data, *Journal of the Royal Statistical Society Ser.B*, 214-221.

Lahiri, P. and P. Dong Ho (1988), Nonparametric Bayes and Empirical Bayes Estimation of the Residual Survival Function at Age *t*, *Commun. Statist. Theory Meth.*, 4085-4098.

Lo, A.Y and C.S. Weng (1989), On a Class of Bayesian Nonparametric Estimates: II. Hazard Rates Estimates, *Annals of the Institute of Statistical Mathematics*, 227-245.

Phadia, E. and V. Susarla (1983), Nonparametric Bayesian Estimation of a Survival with Dependent Censoring Mechanism, *Annals of the Institute of Statistical Mathematic*, 389-400.

Rolin, J.M. (1983), Nonparametric Bayesian Statistic : A Stochastic Process Approach. In: *Specifying Statistical Models, From Parametric to Nonparametric Using Bayesian Approaches.* Lecture Notes in Statistics, Springer, Berlin, 108-133.

Rolin, J.M. (1992,a), Some Useful Properties of the Dirichlet Process. Core Discussion Paper n○ 9207, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Rolin, J.M. (1992,b), On the Distribution of Jumps of the Dirichlet Process, Core Discussion Paper, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Ruggiero, M. (1989), Analyse semi-paramétrique des modèles de durées: L'apport des méthodes bayésiennes, Thèse de doctorat , Université d'Aix-Marseille II, France.

Ruggiero, M. (1992), A priori non paramétriques sur la distribution des données de survie. Document du GREQE, Marseille.

Susarla, V. and J. Van Ryzin (1978a), Empirical Bayes Estimation of a Distribution (Survival) Function from Right-Censored Observations, *Annals of Statistics*, **6**, 740-754.

Tsai, W.Y. (1988), Estimation of Survival Curves from Dependent Censorship Models via a Generalized Self-Consistent Property with Nonparametric Bayesian Estimation Application, *Annals of Statistics*, **14**, 238-249.