BAYESIAN NONPARAMETRIC DURATION MODEL WITH CENSORSHIP

J. Hakizamungu, J.M. Rolin

1. INTRODUCTION

Nonparametric Bayesian methods have known a strong development after Ferguson's (1973, 1974) presentation of the Dirichlet process. Soon after number of authors brought new contributions on the subject including among others Antoniak (1974), Doksum (1974), Susarla and Van-Rysin (1976), Ferguson and Phadia (1979), Rolin (1983, 1992a, 1992b), Florens and Rolin (1988), Rolin (1997), Florens, Mouchart and Rolin (1999).

This paper is concerned with nonparametric i.i.d. durations models with censored observations and we establish by a simple and unified approach the general structure of a Bayesian nonparametric estimator for a survival function S. We then give some corollaries with Dirichlet prior random measures. These results are essentially supported by prior and posterior independence properties.

In general, the estimation of a survival function S has to cope with the technical difficulty of the partial observability of the latent duration variable X. This situation is modelized by introducing a censoring variable Y. The observable variables are then a duration variable $T=\min(X,Y)$ and an indicator variable $D=1_{\{X\leq Y\}}$ which indicates if X is observed or censored. The estimation process will be done using T and D.

This paper is organized as follows: we start by the specification of the model and some assumptions, we present afterwards nonparametric Bayesian estimators of the survival function. In an appendix, we recall some properties of random Dirichlet measures and present the Bayesian nonparametric estimator using purely discrete prior distributions.

2. SPECIFICATION OF THE MODEL

Let X and Y be two independent random variables being respectively a latent duration and a censoring time with respective survival functions S and G. The joint survival function is therefore:

$$\Phi(x, y) = P(X > x, Y > y \mid S, G) = S(x) \times G(y) \tag{1}$$

Assumption 1. A priori S and G are independent, i.e., $S \coprod G$.

Therefore, by standard properties of conditional independence, we obtain the following result:

Lemma 1. *Assumption 1 and equation(1)* are equivalent to $(X, S) \coprod (Y, G)$

If $\{(X_i, Y_i) : 1 \leq i \leq n\}$ is a sample of size n of (X, Y), i.e.,

$$\coprod_{1 \leq i \leq n} (X_i, Y_i) \mid S, G \quad \text{and} \quad (X_i, Y_i) \mid S, G \approx (X, Y) \mid S, G \quad \forall 1 \leq i \leq n,$$

the observable data are defined by

$$T_i = \min(X_i, Y_i) \text{ and } D_i = 1_{\{X_i \leq Y_i\}} \quad \forall 1 \leq i \leq n \tag{2}$$

By elementary properties of conditional independence (see, e.g., Mouchart and Rolin (1984) or Florens, Mouchart and Rolin (1990)), we deduce the following result:

Lemma 2.

i) The observed random variables are independent given $Y_1^n = \{Y_i : 1 \leq i \leq n\}$, i.e.,

$$\coprod_{1 \leq i \leq n} (T_i, D_i) \mid S, G, Y_1^n \tag{3}$$

ii) The distribution of (T_i, D_i) conditionally on (S, G, Y_1^n) only depends on S and Y_i, i.e.,

$$(T_i, D_i) \coprod (S, G, Y_1^n) \mid S, Y \tag{4}$$

and that implies

$$(T_1^n, D_1^n) \coprod G \mid S, Y_1^n \tag{5}$$

Since the distribution of Y_1^n is independent of S, we have a Bayesian cut (see Florens, Mouchart and Rolin (1990)) and the inference may be totally separated into the inference on G through the marginal model generating Y_1^n and the inference on S through the conditional model generating (T_1^n, D_1^n) given Y_1^n.

Equivalently, Y_1^n may be considered as "known" or "fixed" for the estimation of S.

As will be shown later on, we have in fact a stronger result : the inference on S does not require the knowledge of the censoring times if (T_1^n, D_1^n) are observed. This means that the knowledge of "unactive" censoring times (Y_i greater than X_i) is unnecessary.

Lemma 3. The following conditional independence relation is verified :

$$S \coprod Y_1^n \mid T_1^n, D_1^n \tag{6}$$

In conclusion, whatever is the distribution of G, the posterior distribution of S will be the same in the joint model as the posterior distribution of S in the model conditional on Y_1^n , i.e., in the model considering that the censoring times are fixed and known.

Let us now specify the prior distribution of S or equivalenty the prior distribution of the associated hazard function H defined by

$$H(t) = -\ln S(t) \tag{7}$$

Assumption 2. The measure H^* associated to the hazard function H is purely random, i.e., for all $\{B_j : 1 \le j \le k\}$, measurable finite partition of $R^+ = [0, \infty)$, we have the following independence property:

$$\coprod_{1 \le j \le k} H^*(B_j) \tag{8}$$

This assumption amounts to say that S is a neutral to the right process and Doksum (1974) has shown that the independence properties are preserved a posteriori (see also Rolin (1983)). The prior distribution is therefore only specified through independence properties. Note also that the Dirichlet process is a neutral to the right process.

3. NONPARAMETRIC BAYESIAN ESTIMATOR

We want to analyse the posterior distribution of S given (T_1^n, D_1^n). For $J \subset \{1, ..., n\}$, let $A_J = \cap_{i \in J} \{D_i = 1\} \cap \cap_{i \in J^c} \{D_i = 0\}$. On this set, $T_1^n = (X_J, Y_{J^c})$ where $X_J = \{X_i : i \in J\}$ and $Y_{J^c} = \{Y_i : i \in J^c\}$ and we have the following result:

Theorem 1. Whatever is the prior distribution of S, for any measurable function K,

$$\mathrm{E}[K(S)\,|\,T_1^n, D_1^n, Y_1^n] = \frac{\mathrm{E}\left[K(S)\prod_{i\in J^c} S(Y_i)\,|\,X_J, Y_{J^c}\right]}{\mathrm{E}\left[\prod_{i\in J^c} S(Y_i)\,|\,X_J, Y_{J^c}\right]} \quad \text{on } A_J \tag{9}$$

This proves Lemma 3, since the above formula shows that the conditional expectation only depends on the observed censoring times.

The above formula may be rewritten in terms of the order statistics of the observable lifetimes. Let $\{Z_j,\ 1 \le j \le M\}$ be the order statistics (the distinct observable lifetimes in increasing order) of T_1^n. We define the number of individuals at risk at time Z_j by

$$N_j = \sum_{1\le i\le n} 1_{\{T_i \ge Z_j\}} \tag{10}$$

and respectively, the number of individuals censored and the number of individuals uncensored at time Z_j by

$$F_j = \sum_{1\le i\le n} 1_{\{T_i = Z_j, D_i = 0\}} \text{ and } E_j = \sum_{1\le i\le n} 1_{\{T_i = Z_j, D_i = 1\}} \tag{11}$$

With these notations, Theorem 1 may be rewritten as

Corollary 1. For any measurable function K,

$$\mathrm{E}[K(S)\,|\,T_1^n, D_1^n] = \frac{\mathrm{E}\left[K(S)\prod_{1\le j\le M} S(Z_j)^{F_j}\,|\,X_J, Y_{J^c}\right]}{\mathrm{E}\left[\prod_{1\le j\le M} S(Z_j)^{F_j}\,|\,X_J, Y_{J^c}\right]} \quad \text{on } A_J \tag{12}$$

Proof. Let K and K_i, $1 \le i \le n$, be arbitrary measurable functions, by definitions and independence properties, we have

$$\mathrm{E}\left[K(S)\prod_{1\le i\le n}K_i(T_i,D_i)1_{A_J}\,|\,Y_1^n\right]$$

$$=\ \mathrm{E}\left[K(S)\prod_{i\in J}K_i(X_i,1)1_{\{X_i\le Y_i\}}\prod_{i\in J^c}K_i(Y_i,0)1_{\{X_i>Y_i\}}\,|\,Y_1^n\right]$$

$$=\ \mathrm{E}\left[K(S)\prod_{i\in J}K_i(X_i,1)1_{\{X_i\le Y_i\}}\prod_{i\in J^c}K_i(Y_i,0)S(Y_i)\,|\,Y_1^n\right]$$

because $\coprod_{1\le i\le n} X_i|S,Y_1^n$. Now, since $S\coprod Y_1^n\,|\,X_J$, if we define

$$L(X_J,Y_{J^c})=\frac{\mathrm{E}\left[K(S)\prod_{i\in J^c}S(Y_i)|X_J,Y_{J^c}\right]}{\mathrm{E}\left[\prod_{i\in J^c}S(Y_i)|X_J,Y_{J^c}\right]}$$

we have

$$\mathrm{E}\left[K(S)\prod_{i\in J^c}S(Y_i)|X_J,Y_1^n\right]=\ L(X_J,Y_{J^c})\ \ \mathrm{E}\left[\prod_{i\in J^c}S(Y_i)|X_J,Y_1^n\right]$$

Therefore,

$$\mathrm{E}\left[K(S)\prod_{1\le i\le n}K_i(T_i,D_i)1_{A_J}\,|\,Y_1^n\right]=\ \mathrm{E}\left[L(X_J,Y_{J^c})\prod_{1\le i\le n}K_i(T_i,D_i)1_{A_J}\,|\,Y_1^n\right]$$

Now, for $Z_j\le t<Z_{j+1}$, according to Docksum's result $H(t)=-\ln S(t)$ may be decomposed into the following sum of independent terms conditionally on X_J,Y_{J^c} , i.e.,

$$H(t)=\sum_{1\le l\le j}H^*((Z_{l-1},Z_l])+H^*((Z_{l-1},t]) \tag{13}$$

This is equivalent to say that $S(t)$ is a product of independent terms conditionally on X_J,Y_{J^c}, i.e.,

$$S(t) = \prod_{1 \le l \le j} \frac{S(Z_l)}{S(Z_{l-1})} \times \frac{S(t)}{S(Z_j)} \tag{14}$$

Now, if S^* is the measure associated to the survival function S and if we suppose that S^* is a random measure which follows the Dirichlet law, i.e., $S^* \approx Di(a^*)$, where a^* is a finite measure, Ferguson (1973) has shown that for uncensored observations, S^* is a posteriori a Dirichlet measure. Therefore,

$$S^* | X_J, Y_{J^c} \approx Di(a^* + N_u)$$

where

$$N_u = \sum_{i \in J} \varepsilon_{X_i} = \sum_{1 \le i \le n} D_i \varepsilon_{T_i} \tag{15}$$

is the counting measure of uncensored observations (ε_x is the Dirac probability measure at point x).

From properties of the Dirichlet measure (see Appendix 1), we deduce that,

$$\frac{S(Z_l)}{S(Z_{l-1})} | X_J, Y_{J^c} \approx Beta[(a^* + N_u)((Z_l, \infty)], a^*((Z_{l-1}, Z_l]) + E_l]$$

and that

$$\frac{S(t)}{S(Z_j)} | X_J, Y_{J^c} \approx Beta[(a^* + N_u)((t, \infty]), a^*((Z_j, t])]$$

Combining these properties with Theorem 1, we obtain the following result:

Theorem 2. If H^* is purely random (Assumption 2), then for $Z_j \le t < Z_{j+1}$, we may write $S(t)$ as a product of independent terms conditionally on (T_1^n, D_1^n), namely

$$S(t) = \prod_{1 \le l \le j} \frac{S(Z_l)}{S(Z_{l-1})} \times \frac{S(t)}{S(Z_j)}$$

Moreover if $S^* \approx Di(a^*)$ and $a(t) = a^*([0,t]), \forall t > 0$, then, for $1 \le l \le M$,

$$\frac{S(Z_l)}{S(Z_{l-1})} | T_1^n, D_1^n \approx Beta[a(\infty) - a(Z_l) + N_l - E_l, a(Z_l) - a(Z_{l-1}) + E_l] \tag{16}$$

and

$$\frac{S(t)}{S(Z_j)} \mid T_1^n, D_1^n \approx Beta[a(\infty) - a(t) + N_{j+1}, a(t) - a(Z_j)] \tag{17}$$

Proof. Note that

$$\prod_{1 \le j \le M} S(Z_j)^{F_j} = \prod_{1 \le l \le M} \left\{ \frac{S(Z_l)}{S(Z_{l-1})} \right\}^{N_c([Z_l, \infty])}$$

where

$$N_c([Z_l, \infty]) = \sum_{l \le j \le M} F_j$$

N_c denotes therefore the counting measure of censored observations, i.e.,

$$N_c = \sum_{1 \le i \le n} (1 - D_i) \varepsilon_{T_i}$$

Applying Theorem 1 provides the result if we notice that

$$N_u((Z_l, \infty]) + N_c([Z_l, \infty]) = N_l - E_l = N_{l+1} + F_l$$

and

$$N_u((t, \infty]) + N_c\ ([Z_{j+1}, \infty]) = (N_u + N_c)\ ([Z_{j+1}, \infty]) = N_{j+1}$$

The same type of computations provides the following corollary:

Corollary 2. Conditionally on (T_1^n, D_1^n),

$$\frac{S(Z_j)}{S(Z_{j-1})} = \frac{S(Z_j -)}{S(Z_{j-1})} \times \frac{S(Z_j)}{S(Z_j -)}$$

where

$$\frac{S(Z_j -)}{S(Z_{j-1})} \coprod \frac{S(Z_j)}{S(Z_j -)} \mid T_1^n, D_1^n$$

$$\frac{S(Z_j -)}{S(Z_{j-1})} \mid T_1^n, D_1^n \approx Beta[a(\infty) - a(Z_j -) + N_j, a(Z_j -) - a(Z_{j-1})] \tag{18}$$

$$\frac{S(Z_j)}{S(Z_j-)} \mid T_1^n, D_1^n \approx Beta[a(\infty)-a(Z_j)+N_j-E_j, a(Z_j)-a(Z_j-)+E_j] \tag{19}$$

This last formula shows that, if a^* is a diffuse measure, then, a posteriori, there is a jump only at the death observations ($E_j > 0$).

Taking the posterior expectation provides the Susarla-Van Ryzin estimator

Corollary 3. If $S^* \approx Di(a^*)$, then, for $Z_j \leq t < Z_{j+1}$,

$$\begin{aligned}\hat{S}_{SV}(t) &= E[S(t) \mid T_1^n, D_1^n] \\ &= \frac{a(\infty)-a(t)+N_{j+1}}{a(\infty)-a(Z_j)+N_{j+1}} \times \prod_{1\leq l\leq j} \frac{a(\infty)-a(Z_l)+N_l-E_l}{a(\infty)-a(Z_{l-1})+N_l} \\ &= \frac{a(\infty)-a(t)+N_{j+1}}{a(\infty)+n} \times \prod_{1\leq l\leq j} \{1+\frac{F_l}{a(\infty)-a(Z_l)+N_{l+1}}\}\end{aligned} \tag{20}$$

where $a(t) = a^*([0,t]), \forall t > 0$.

Considering the non informative case, i.e., $a(\infty) = 0$, we obtain the Kaplan-Meier estimator

Corollary 4. In the noninformative case, the posterior expectation is given by:

$$\hat{S}_{KP}(t) = \prod_{\{j: Z_j \leq t\}} \left\{1 - \frac{E_j}{N_j}\right\} \tag{21}$$

APPENDIX

1. *Random Dirichlet measures*

We first recall the definition of a Dirichlet measure (Ferguson (1973)).

Definition: $\tilde{m}^*$ is a random Dirichlet measure defined on (**A**, ***M***) with parameter a^*, a finite measure on (**A**, ***M***), and we write $\tilde{m}^* \approx Di(a^*)$, if for all partitions $\{B_j, 1 \leq j \leq l\}$, $B_j \in$ ***M***, the random vector $\tilde{m}^*(B_1),..., \tilde{m}^*(B_l)$ follows the Dirichlet law with parameter $a^*(B_1),..., a^*(B_l)$ characterized by

$$P\left[\bigcap_{1\le j\le l}\{\tilde{m}^*(B_j)\in dv_j\}\right]=\frac{\Gamma\left[\sum_{1\le j\le l}a^*(B_j)\right]}{\prod_{1\le j\le l}\Gamma[a^*(B_j)]}\times\prod_{1\le j\le l}v_j^{a^*(B_j)-1}dv_j\times 1_{\left\{\sum_{1\le j\le l}v_j=1\right\}} \tag{22}$$

In particular, $\forall B\in \boldsymbol{M}$

$$\tilde{m}^*(B)\approx Beta[a^*(B),a^*(B^c)] \tag{23}$$

The Dirichlet measure may also be characterized as follows (see, e.g., Rolin (1992a). Let us define the following $\boldsymbol{\sigma}$-algebra

$$\boldsymbol{F}_B=\sigma\{\tilde{m}^*(C):C\subset B,C\in M\}$$

and the conditional probability

$$\tilde{m}^*(C\,|\,B)=\frac{\tilde{m}^*(C\cap B)}{\tilde{m}^*(B)}$$

Proposition. Let $\tilde{m}^*$ be a random probability on (**A**, $\boldsymbol{M}$). Then $\tilde{m}^*\approx Di(a^*)$ if and only if, $\forall C,B\in \boldsymbol{M}$ with $0<a^*(B)<1$,

(i) $\tilde{m}^*(C\,|\,B)\coprod\ \boldsymbol{F}_{B^c}$

(ii) $\tilde{m}^*(C\,|\,B)\approx Beta[a^*(C\cap B),a^*(C^c\cap B)]$ (24)

2. *Estimator using a purely discrete prior distribution*

Let $\{t_j:1\le j\le m\}$ with $t_1<t_2<...<t_m$ be the support of S^*. The hazard rate, for $1\le j\le m$, is defined by

$$\Lambda_j=1-\frac{S(t_j)}{S(t_j-)}=1-\frac{S(t_j)}{S(t_{j-1})} \tag{25}$$

These hazard rates characterize entirely S^* since, for $1\le j\le m$,

$$S(t_j)=\prod_{1\le l\le j}(1-\Lambda_l)$$

and

$$S^*(\{t_j\}) = S(t_j) - S(t_j -) = \Lambda_j \prod_{1 \le l < j} (1 - \Lambda_l)$$

Now, if , for $1 \le j \le m$, N_j , E_j et F_j denote respectively the number of individuals at risk, uncensored and censored at time t_j, the likelihood of the sample is given by

$$L_n = \prod_{1 \le j \le m} S^*(\{t_j\})^{E_j} \times \prod_{1 \le j < m} S(t_j)^{F_j}$$

or, in terms of hazard rates, by

$$L_n = \prod_{1 \le j < m} \Lambda_j^{E_j} (1 - \Lambda_j)^{N_j - E_j} \tag{26}$$

For the prior distribution, note that H^* is a purely random measure if and only if the Λ_j 's are independent random variables. In view of the likelihood, this property will hold a posteriori. Now, if S^* is a Dirichlet measure with purely discrete parameter a^*, we have

$$\Lambda_j \approx Beta[a^*(\{t_j\}), a^*((t_j, \infty])] = Beta[c_j l_j, c_j(1 - l_j)] \tag{27}$$

where $l_j = E(\Lambda_j)$ and $c_j = a^*([t_j, \infty]) = a(\infty) - a(t_j -)$.

By taking c_j, $1 \le j < m$, to be arbitrary positive constants, we see that we may use more general priors than Dirichlet measures to obtain tractable posterior distributions. This gives the following result

Theorem 3. If a priori, $\coprod_{1 \le j < m} \Lambda_j$ and for $1 \le j < m$,

$$\Lambda_j \approx Beta\ [c_j l_j, c_j(1 - l_j)] \tag{28}$$

where $l_j = E(\Lambda_j)$ and c_j, $1 \le j < m$ are arbitrary positive constants, then, a posteriori, $\coprod_{1 \le j < m} \Lambda_j \mid T_1^n, D_1^n$ and for $1 \le j < m$,

$$\Lambda_j \approx Beta\ [c_j l_j + E_j, c_j(1 - l_j) + N_j - E_j] \tag{29}$$

Note that c_j indicates the degree of credibility that the statistician allows to the prior distribution because the variance of the hazard rates is given by

$$V(\Lambda_j) = \frac{l_j(1-l_j)}{c_j+1}$$

As a corollary, we obtain

Corollary 5. The Bayesian nonparametric estimator of the survival function for $1 \leq j < m$ is given by:

$$\hat{S}_{BN}(t_j) = E[S(t_j) \mid T_1^n, D_1^n] = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{c_l l_l + E_l}{c_l + N_l} \right\} \tag{30}$$

We note that if $c_j \to 0$ we obtain the Kaplan-Meier estimator. This is the same as to use a non-informative prior law as $f(\lambda_j) \propto \lambda_j^{-1}(1-\lambda_j)^{-1}$

Department of Mathematics, HAKIZAMUNGU JOSEPH
University of Benin, Togo

Institut de Statistique, ROLIN JEAN-MARIE
Université catholique de Louvain, Belgium

REFERENCES

C.E. ANTONIAK, (1974), *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,* "Annals of Statistics", 2, 1152-1174.

K. DOKSUM, (1974), *Tailfree and neutral random probabilities and their posterior distributions,* "Annals of Probability", 2, 183-201.

T. FERGUSON, (1973), *A Bayesian analysis of some nonparametric problems,* "Annals of statistics", 1, 209-230.

T. FERGUSON, (1974), *Prior distributions on spaces of probability measures,* "Annals of statistics", 2, 615-629.

T. FERGUSON, E.G PHADIA, (1979), *Bayesian nonparametric estimation based on censored data,* "Annals of statistics", 7, 163-186.

J.P. FLORENS, M. MOUCHART, J.M. ROLIN, (1990), *Elements of Bayesian Statistics*, Marcel Dekker, New York.

J.P. FLORENS, M. MOUCHART, J.M. ROLIN, (1999), *Semi-and non-parametric Bayesian analysis of duration models with Dirichlet priors: a survey*, "International Statistical Review", 67, 2, 187-210.

J.P. FLORENS, J.M. ROLIN, (1998), *Simulation of posterior distributions in nonparametric censored analysis,* Institut de Statistique discussion paper n°9818, , Université Catholique de Louvain, Louvain-La-Neuve, Belgique.

J. HAKIZAMUNGU, (1992), *Modèles de durées: Approche Bayésienne semiparamétrique,* Thèse de doctorat, Université Catholique de Louvain, Belgique.

M. MOUCHART, J.M. ROLIN, (1984), *A note on conditional independence with statistical applications,* "Statistica", 557-584.

J.M ROLIN, (1983), *Nonparametric Bayesian statistics: A stochastic process approach,* in *Specifying sta-*

tistical models from parametric to nonparametric using Bayesian approaches, Lecture notes in statistics, Springer-Berlin, 108-133.

J.M ROLIN, (1992a), *Some useful properties of the Dirichlet processes,* Core discussion paper n°9207. Université Catholique de Louvain, Louvain-La-Neuve, Belgique.

J.M ROLIN, (1992b), *On the distribution of jumps of the Dirichlet processes,* Core discussion paper n°9259. Université Catholique de Louvain, Louvain-La-Neuve, Belgique.

J.M ROLIN, (1998), *Bayesian survival analysis*, "Encyclopedia of Biostatistics", 1, 271-286, John Wiley, New-York.

V. SURSALA, J. VAN RYZIN, (1976), *Nonparametric Bayesian estimation of survival curves from incomplete observations,* "JASA", 71, 897-902.

RIASSUNTO

Modello non parametrico Bayesiano di durata con dati censurati

Il seguente articolo tratta di modelli i.i.d. non parametrici di durata con dati censurati e considera una struttura generale di uno stimatore bayesiano non parametrico per una funzione di sopravvivenza attraverso un approccio semplice e unificato. Per le distribuzioni a prior di Dirichlet, si descrive in modo completo la struttura della distribuzione a posteriori della funzione di sopravvivenza. Questi risultati sono essenzialmente sostenuti dalle proprietà di indipendenza a priori e a posteriori.

SUMMARY

Baesian nonparametric duration model with censorship

This paper is concerned with nonparametric i.i.d. durations models with censored observations and we establish by a simple and unified approach the general structure of a bayesian nonparametric estimator for a survival function S. For Dirichlet prior distributions, we describe completely the structure of the posterior distribution of the survival function. These results are essentially supported by prior and posterior independence properties.