RT-MAE 2005-10

On Bayesian estimation of a survival curve: Comparative study and examples.

by

Victor H. Salinas,
José S. Romeo
and
Aléxis Peña

Palavras-Chave: Dirichlet process, randomly censored data, subsurvival function, Peterson's formula, numerical methods, case study in cancer.
Classificação AMS: 62N02, 62G05, 62P10.

- Março de 2005 -

ON BAYESIAN ESTIMATION OF A SURVIVAL CURVE: COMPARATIVE STUDY AND EXAMPLES

VICTOR H. SALINAS, JOSÉ S. ROMEO, AND ALEXIS PEÑA

ABSTRACT. This paper is concerned with a nonparametric Bayesian approach applied to estimate a survival curve by means of a functional of the subsurvival functions associated to censored and non-censored event. In order to actually compute the Bayesian estimator, a numerical algorithm based on the Runge-Kutta fourth-order method is introduced. It provides good accuracy and it is simple to program. Using a simulated data set, the performance of the Bayesian estimator is compared to the Product-Limit. A descriptive analysis of the results from the simulations is presented. The conclusions are given in terms of the proportion of the censored data and sample size. Also, the numerical methodology is illustrated considering the original Kaplan-Meier data. Finally, the Bayesian analysis is applied to a real case of cervix uterine cancer, where the elicitation of the prior distribution considers the high proportion of censoring in the sample.

1. INTRODUCTION

Consider the following right censored data model or in a reliability framework, a series system with two components. Let X_1, X_2 be independently distributed random variables, with survival functions $S_j(t) = \Pr(X_j > t)$, $j = 1, 2$, the vector of observed values is (Z, δ), where

$$Z = \min(X_1, X_2)\ , \quad \text{and } \delta = j \text{ if } Z = X_j\ , \quad j = 1, 2. \tag{1.1}$$

Let

$$S_j^*(t) = \Pr(Z > t, \delta = j), \tag{1.2}$$

be the respective subsurvival function, $j = 1, 2$. Viewing the model as a series systems with two components, the system survival function is given by

$$S(t) = \Pr(Z > t) = S_1^*(t) + S_2^*(t). \tag{1.3}$$

Let (X_{i1}, X_{i2}), $i = 1, \ldots, n$, be n independent latent or imaginary observations on (X_1, X_2). The actual observations are the pairs (Z_i, δ_i), $i = 1, \ldots, n$, which form a sample on (Z, δ).

Key words and phrases. Dirichlet process, randomly censored data, subsurvival function, Peterson's formula, numerical methods, case study in cancer.

PARTIALLY SUPPORTED BY FONDECYT-CHILE GRANT 1030787 AND CNPQ-BRASIL

The estimation of $S_1(t) = \Pr(X_1 > t)$ have been considered by several authors. [1] derived the product-limit (PL) estimator for S_1. [2] showed, using a functional representation, that the later is a maximum likelihood estimator.

Peterson's formula is given by

$$S_1(t) = \varphi(S_1^*(\cdot), S_2^*(\cdot); t), \text{ for } t \leq t^* = \min(t_{S_1}, t_{S_2}), \tag{1.4}$$

where

$$\varphi(F(\cdot), G(\cdot); t) = \exp\left\{\oint_0^t \frac{dF(s)}{F(s)+G(s)}\right\} \prod_t \frac{F(s_+)+G(s_+)}{F(s_-)+G(s_-)}, \tag{1.5}$$

$t_{S_j} = \sup\{t : S_j(t) > 0\}$, $j = 1, 2$, and $\oint_0^t$ is the integral over the union of intervals of points less than t for which $F(\cdot)$ is continuous. $\prod_t$ indicates the product over the set $\{s \leq t : s \text{ is a jump point of } F\}$. For (1.4) to be well defined it is assumed that S_1 and S_2 have no common discontinuities. If S_1 and S_2 are continuous, then (1.4) reduces to equation (7.5) of [3]:

$$S_1(t) = \exp\left[\int_0^t \frac{dS_1^*(s)}{S_1^*(s)+S_2^*(s)}\right].$$

[4] used the Dirichlet process prior of [5] for S_1, i.e. $S_1 \sim D_\alpha$, and obtained a nonparametric Bayes estimator. This estimator reduces to PL estimator as $\alpha(\mathbb{R}^+)$ tends to zero. [6] complemented this result by showing that the posterior distribution is a mixture of Dirichlet processes, but the representation of the mixture is somewhat cumbersome. For a survey of works on the Bayesian estimation of survival function using Dirichlet processes, see [7]. Also, see [8].

Other works consider different approaches. In particular, [9] studies the problem of finding Bayes estimators for cumulative hazard rates and related parameters, considering a class of Beta processes as a prior distribution. Then the Bayes estimates of the survival function is calculated.

The main purpose of this paper is to estimate $S_1(t)$ in a Bayesian nonparametric context under a certain Dirichlet bivariate process prior for the vector of subsurvival functions. We apply the approach introduced by [10] which considers a series system with r components or a competing-risks model. Certainly the case when the series system has only two components corresponds to the random right censored data model as described above. In order to actually compute the Bayesian estimator of $S_1(t)$, a numerical method is formulated. We are also interested in comparing the performance of the our Bayesian estimator with others estimators already in the literature.

In Section 2, Bayes estimates of the (sub)survival functions corresponding to the randomly censored data are calculated. By substituting the Bayes estimator of $\mathbf{S}^* = (S_1^*, S_2^*)$ in Peterson's (1977) formula we obtain a nonparametric estimator of S_1. Following [11], this nonparametric estimator of S_1 is in fact a Bayes estimator under a bivariate Dirichlet process prior and quadratic loss function, i.e., the posterior mean of S_1 given the data. For the numerical computation of the Bayes estimator of S_1, in Section 3 we develop an algorithm based on the Runge-Kutta fourth-order method, see [12].

In Section 4 we apply the numerical method to a simulated data set. In this section we compare the performance of the Bayesian alternative with the product-limit estimator. A descriptive statistics analysis is reported. We present our results in terms of the proportion of censored data and the sample size. Also, using the Kaplan-Meier data set we report the survival probabilities that are useful for comparison purposes, in particular, among nonparametric Bayesian methodologies. Furthermore, a real case study coming from cervix uterine cancer which presents high level of censoring is analyzed. This last information is incorporated in the prior distribution of the subsurvival functions. Then, the Bayesian estimator of the survival function is computed and survival probabilities on different times are calculated. Finally, the work concludes with a detailed discussion on theoretical aspects, applications and future works.

2. Bayesian Analysis of Censored Data

The purpose of this section is to consider a Bayesian approach to the censored data problem, specifically, to calculate the Bayes estimators of the subsurvival and survival functions, (S_1^*, S_2^*) and S_1, under a certain Dirichlet bivariate process prior and quadratic loss function.
Let $\rho = \Pr(\delta = 1) = 1 - \Pr(\delta = 2)$ and $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$ be a random sample on (Z, δ). Recall that $\mathbf{S}^*(t) = (S_1^*(t), S_2^*(t))$ is the vector of subsurvival functions, and $S(t) = S_1^*(t) + S_2^*(t) = \Pr(Z > t)$ is the survival function of the system.

2.1. **Dirichlet Bivariate Processes.** We give the definition of a class of Dirichlet bivariate process priors as defined in [10]. For a general treatment of a Dirichlet process and properties of the Dirichlet distribution, see [5] and [13], respectively.
Let $(\mathcal{X}, \mathcal{A})$ be a measurable space such as $(R_r, \mathcal{B}_r)$ and $(\Omega, \mathcal{F}, Q)$ be a probability space. Consider a stochastic process $\{\mathbf{P}^*(A) = (P_1^*(A), P_2^*(A)) : A \in \mathcal{A}\}$ defined on $(\Omega, \mathcal{F}, Q)$, indexed with sets A in $\mathcal{A}$, and assuming values in the simplex

$$\mathcal{S}_2 = \{(x_1, x_2) : x_1, x_2 \geq 0\ ,\ \ x_1 + x_2 \leq 1\}$$

of R_2. Note that $\mathbf{P}^*$ is a random two-dimensional set function.

Definition 1. Let α_1 and α_2 be finite, nonnull, and nonnegative measures on $(\mathcal{X}, \mathcal{A})$. A random two-dimensional set function $\mathbf{P}^* = (P_1^*, P_2^*)$ with values in $\mathcal{S}_2$ is a Dirichlet

bivariate process on $(\mathcal{X}, \mathcal{A})$ with parameter (α_1, α_2), denoted by $\mathbf{P}^* \sim DB(\alpha_1, \alpha_2)$, if for every $k > 0$ and measurable partition $\{A_1, \ldots, A_k\}$ of $\mathcal{X}$, the distribution of $2(k-1)$-dimensional random vector $(\mathbf{P}^*(A_1), \ldots, \mathbf{P}^*(A_{k-1}))$ is the nonsingular Dirichlet with parameter $(\alpha_1(A_1), \alpha_2(A_1), \ldots, \alpha_1(A_{k-1}), \alpha_2(A_{k-1}); \alpha_1(A_k)+\alpha_2(A_k))$.

That is $\mathbf{P}^* \sim DB(\alpha_1, \alpha_2)$

We note that the marginal process P_1^* is such that for any partition $\{A_1, \ldots, A_k\}$ of $\mathcal{X}$, the $(k-1)$-dimensional random vector $(P_1^*(A_1), \ldots P_1^*(A_{k-1}))$ is distributed as the nonsingular Dirichlet with parameter $(\alpha_1(A_1), \ldots, \alpha_1(A_{k-1}); \alpha_1(A_k) + \alpha_2(\mathcal{X}))$. Consequently, P_1^* is not exactly a Dirichlet process as defined by [5].

2.2. **Bayes Estimation.** First, note that observing the censored data $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ is equivalent to observe, for each $t > 0$, the random counting vector $n\mathbf{S}_n^*(t) = (nS_{1n}^*(t), nS_{2n}^*(t), n(1 - S_n(t))$ from a trinomial distribution with sample size n and parameters $(S_1^*(t), S_2^*(t); 1 - S(t))$, where

$$S_{jn}^*(t) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i > t, \delta_i = j), \ j = 1, 2, \tag{2.1}$$

is the empirical subsurvival function of the j-th component, $j = 1, 2$ and $I(\cdot)$ is the indicator function. $S_n(t) = S_{1n}^*(t) + S_{2n}^*(t)$ is the empirical survival function of the system.

Thus, at this stage, the parameter of interest is the vector of subsurvival functions $\mathbf{S}^* = (S_1^*, S_2^*)$ such that the sum of its components is the survival function of the system. Then our approach considers putting a prior distribution on a functional space of the form

$$\Theta = \{(S_1^*, S_2^*) : S_j^* \text{ is such as (1.2), } j = 1, 2, \text{ and (1.3) is a survival function}\}.$$

To define the prior for the vector $\mathbf{S}^*$, consider the positive real line $\mathcal{X} = (0, \infty)$ with its respective Borel $\sigma-$algebra, $\mathcal{A} = \mathcal{B}_{(0,\infty)}$. Assume that all random elements are defined on a common probability space $(\Omega, \mathcal{F}, Q)$. The following lemma, obtained as a direct consequence of the definition of the Dirichlet bivariate process, gives the prior for $\mathbf{S}^* = (S_1^*, S_2^*)$.

Lemma 1. *Let α_1 and α_2 be finite, non-null, and nonnegative measures on $(\mathcal{X}, \mathcal{A})$. Let $\rho = \Pr(\delta = 1) \sim Beta(\alpha_1(\mathcal{X}), \alpha_2(\mathcal{X}))$, $T_j^*(t) = \Pr(Z > t|\delta = j) \sim Beta(\alpha_j(t, \infty), \alpha_j(0,$ $j = 1, 2$. Suppose that $\rho, T_1^*(t), T_2^*(t)$ are mutually independent. Then*

$$\mathbf{S}^*(t) \sim D(\alpha_1(t, \infty), \alpha_2(t, \infty); \alpha_1(0, t] + \alpha_2(0, t]), \tag{2.2}$$

where $D(a, b; c)$ is the non-singular Dirichlet distribution of parameters $(a, b; c)$.

The marginal prior for S_j^* is given by

$$S_j^*(t) \sim Beta\left(cS_{j,0}^*(t),\, c(1-S_{j,0}^*(t))\right),\ t>0, \tag{2.3}$$

where $c=\alpha_1(\mathcal{X})+\alpha_2(\mathcal{X})$ and $S_{j,0}^*(t)=\alpha_j(t,\infty)/c$ is the prior mean of S_j^*, $j=1,2$. Also, $S_0(t)=S_{1,0}^*(t)+S_{2,0}^*(t)$ is the prior mean of S.

Then using the Bayes rule, the posterior distribution of $\mathbf{S}^*(t)$ is an updated Dirichlet distribution given by

$$\begin{aligned}\mathbf{S}^*(t)|n\mathbf{S}_n^*(t) \sim\ & D(\alpha_1(t,\infty)+nS_{1n}^*(t),\alpha_2(t,\infty)+nS_{2n}^*(t);\\ & \alpha_1(0,t]+\alpha_2(0,t]+n(1-S_n(t))).\end{aligned} \tag{2.4}$$

Consider the quadratic loss function

$$L(\mathbf{S}^*,\widehat{\mathbf{S}}^*)=\int_0^\infty \|\,\mathbf{S}^*(t)-\widehat{\mathbf{S}}^*(t)\,\|^2\,dW(t), \tag{2.5}$$

where $\|\cdot\|$ is the usual R_2 norm, $\widehat{\mathbf{S}}^*=(\widehat{S}_1^*,\widehat{S}_2^*)$ is an estimator of $\mathbf{S}^*=(S_1^*,S_2^*)$ and $W(\cdot)$ is a weight function.
Let the $m(\le n)$ distinct order statistics of Z be $Z_{(1)}^*<\ldots<Z_{(m)}^*$. Set $n_j=\sum_{i=1}^n I(Z_i\ge Z_{(j)}^*)$ and $d_j=\sum_{i=1}^n I(Z_i=Z_{(j)}^*,\delta_i=1)$, $j=1,\ldots,m$.
Define

$$i(t)=\exp\left\{\frac{-1}{c+n}\int_0^t\frac{d\alpha_2(s,\infty)}{\widehat{S}(s)}\right\},$$

and

$$\pi(t)=\prod_{i:Z_{(i)}^*\le t}\frac{\alpha_1(Z_{(i)}^*,\infty)+\alpha_2(Z_{(i)}^*,\infty)+n_i-d_i}{\alpha_1(Z_{(i)}^*,\infty)+\alpha_2(Z_{(i)}^*,\infty)+n_i}.$$

The following proposition is a direct consequence of the conjugated Bayesian analysis and the main result of [11]. It gives the Bayes estimators for the subsurvival and survival functions.

Proposition 1. *Under the prior (2.2) for* $\mathbf{S}^*$ *and loss function (2.5),*

a) *the Bayes estimator of* S_j^*, *and* S *are given by*

$$\widehat{S}_j^*(t)=\frac{c}{c+n}S_{j,0}^*(t)+\frac{n}{c+n}S_{jn}^*(t),\ j=1,2, \tag{2.6}$$

$$\widehat{S}(t)=\widehat{S}_1^*(t)+\widehat{S}_2^*(t), \tag{2.7}$$

and

b) *suppose that the functions* $f_j(s) = \alpha_j(s,\infty)$, $j = 1,2$, *are continuous on* $(0,t)$, *for each* $t > 0$, *and* S_1 *and* S_2 *have no common discontinuities. The Bayes estimator of* S_1 *is given by*

$$\widehat{S}_1(t) = \varphi(\widehat{S}_1^*, \widehat{S}_2^*; t) = \widehat{S}(t)i(t)\pi(t). \tag{2.8}$$

Remark 1.

(a) [2] proved that the PL estimator of S_1 is obtained evaluating the functional (1.4) on the vector of empirical subsurvival functions $(S_{1n}^*(\cdot), S_{2n}^*(\cdot))$. This implies that if $\alpha_1(0,\infty)$ and $\alpha_2(0,\infty)$ tend to zero, the estimator $\widehat{S}_1$ reduces to the Kaplan-Meier estimator.

(b) The Bayes estimators $\widehat{\mathbf{S}}^* = (\widehat{S}_1^*, \widehat{S}_2^*)$ and $\widehat{S}$ are strongly consistent. For instance, using Glivenko-Cantelli Theorem and the fact that $p_n = \frac{c}{c+n} \downarrow 0$, it can be shown that $\widehat{\mathbf{S}}^*$ converges to $\mathbf{S}^*$ uniformly w.p. 1.

(c) The strong consistency of $\widehat{S}_1$ follows from the continuity of the functional φ in (1.4) and the strong consistency of $\widehat{\mathbf{S}}^* = (\widehat{S}_1^*, \widehat{S}_2^*)$.
Also, $||\widehat{S}_1 - \mathrm{PL}|| \to 0$ w.p. 1, where PL is the Kaplan-Meier estimator of S_1.

3. Numerical computation of the Bayes Estimator $\widehat{S}_1$

In equation (2.8) of the estimator $\widehat{S}_1(t)$, the second term in the product, $i(t)$, must be numerically computed. So, we are interested in an approximation via numerical algorithm of the integral $\phi(t) := \int_0^t \frac{d\alpha_2(s,\infty)}{\widehat{S}(s)}$.

Let $z_{(1)}^*, \ldots, z_{(m)}^*$ be the realizations of the random variables $Z_{(1)}^*, \ldots, Z_{(m)}^*$ and suppose that $\frac{d\alpha_2(s,\infty)}{\widehat{S}(s)} = g_2(s)ds$, with $g_2(s) = \frac{f_2'(s)}{\widehat{S}(s)}$, i.e., $\alpha_2 \ll \lambda$, where λ is the Lebesgue measure on the real line, then we must solve the differential equation

$$\begin{cases} \phi'(t) = g_2(t), & t \in [0, z_{(m)}^*] \\ \phi(0) = 0. \end{cases}$$

The Runge-Kutta fourth-order method [12] and its variations for solving ordinary differential equations are very popular. It provides good accuracy, it is simple to program, it requires minimum storage and it is stable. In our case, we apply the method considering a differential equation for each one of the intervals, $[0, z_{(1)}^*), [0, z_{(2)}^*), \ldots, [0, z_{(m)}^*)$.

First, we solve the differential equation $\phi'(t) = g_2(t)$, $\phi(0) = 0$; $t \in [0, z_{(1)}^*]$. Then we solve the differential equations $\phi'(t) = g_2(t)$, $\phi(z_{(j-1)}^*) = \zeta_{j-1}^*$ in the intervals $[0, z_{(j)}^*)$, where ζ_{j-1}^* is obtained from the solution of the differential equation in the previous interval $[0, z_{(j-1)}^*)$, $j = 2, \ldots, m$.

For simplicity, let $[0, b)$ be one of the intervals $[0, z^*_{(j)})$, and $\{t_0, \ldots, t_k\}$ a partition of $[0, b)$ such that $t_l - t_{l-1} = \frac{b}{k} = h,\ l = 1, \ldots, k$. Then a numerical approximation of the function $\phi(t)$ on the interval $[0, b)$ is obtained using the formula:

$$\begin{cases} \zeta_l = \zeta_{l-1} + \frac{h}{6}[g_2(t_{l-1}) + 4g_2(t_{l-1} + \frac{h}{2}) + g_2(t_{l-1} + h)];\ l = 1, \ldots, k \\ \zeta_0 = \phi(0), \quad t_0 = 0. \end{cases} \tag{3.1}$$

where ζ_l is the solution evaluated on t_l.

Thus, the expression $\widehat{S}_1(t) = \widehat{S}(t)i(t)\pi(t)$ in the equation (2.8) is numerically approximated in the interval $[0, b)$ by pairs $\{(t, \widehat{S}(t)\,\widetilde{\zeta}_l\,\pi(t))\}$, where $\widetilde{\zeta}_l = \exp\{-\zeta_l/(\alpha_1(0,\infty) + \alpha_2(0,\infty) + n)\},\ l = 1, \ldots, k$.

Through the methodology described above, only point estimates of the survival probability $\widehat{S}_1$ are obtained. With the purpose of finding a full Bayesian solution to inference questions we make use of Monte Carlo methods for evaluating $\widehat{S}_1(t)$. For each t, we draw d times (S_1^*, S_2^*) from it posterior distribution (2.4) and compute $\widehat{S}_1$ from the functional relation (2.8), i.e., calculating $S(t)\pi(t)$ and evaluating $i(t)$ as it was mentioned previously.

4. Comparative Study

We apply the numerical method described in the previous section. First, we consider a simulated data set from a specific probability distribution. Then, we compare the performance of the our Bayesian alternative to PL-estimator via the L_2-norm calculated on the sample range. The second part is devoted to the analysis of two examples of applications. We compare our estimator with other Bayesian approaches involving the well-know Kaplan-Meier data set and an application to a real case study of cervix uterine cancer in patients treated at the Instituto Nacional del Cancer of Chile.

4.1. **Simulation Study.** We implement a numerical example through a simulation study. A comparison is made between the Bayes estimator obtained in the previous section and the Kaplan-Meier estimator, in terms of the fit to a theoretical distribution. For this purpose we consider two independent random samples of size n of the failure time variable X_1 with exponential distributions of mean 3, and the censoring variable X_2, with gamma distributions of parameters $(\mu, 1)$ for different values of μ. Thus, the survival functions are $S_1(t) = \exp(-t/3)$ and $S_2(t) = 1 - I(\mu, t)$, where $I(k,t) = \Gamma(k)^{-1} \int_0^t u^{k-1}e^{-u}du$ is the incomplete gamma function.

It follows a description of the class of prior distributions used in the simulation study. According to equation (2.2), we consider

$$\alpha_j(t,\infty) = M_j - t,\ 0 < t < M_j,\ j = 1,2, \tag{4.1}$$

measures in $[0,\infty)$. Note that $\alpha_j(M_j,\infty) = 0$, $\alpha_j(0,t) = t$, $\alpha_j(0,M_j) = M_j$, $0 < t < M_j$, $j = 1,2$. Thus, the priori distribution of $\mathbf{S}^*(t)$ is a non-singular Dirichlet distribution of parameters $(M_1 - t, M_2 - t; 2t)$. Then the marginal distribution of $S_j^*(t)$ is

$$S_j^*(t) \sim Beta\,(M_j - t\,, M_i + t), \qquad t > 0, \quad j \neq i, \quad j,i = 1,2. \tag{4.2}$$

Note that, $S_{j,0}^*(t) = \frac{M_j - t}{c}$ is the marginal prior guess of $S_j^*(t)$, $j = 1,2$; $c = M_1 + M_2$.

We simulate random samples of size $n = \{50, 100, 200\}$ from X_1, the variable of interest, and from X_2. This process is repeated r=1000 times for $\mu = 1,\ 2,\ 4$, obtaining, after observing (Z,δ), 75, 55 and 30 percent of censoring, respectively. We consider $M_1 = M_2 \geq 13$ with different values for each sample size and percentage of censoring. Note that $\Pr(X_1 > 13)$ and $\Pr(X_2 > 13)$ are less than 0.015, and the prior distribution of the proportion of failures $\rho = \Pr(X_1 < X_2)$ is a symmetric Beta distribution reflecting non-information about this proportion.

Table 1 presents the different values for $M_1 (= M_2)$ that were used in the simulations. Tables 2, 3 and 4 present a descriptive statistic analysis of $L_2 - norm$ calculated on $[0, z_{(m)}^*)$, which is given by

$$\| S_1 - \widehat{S}_1 \|_2 = \left\{ \int_0^{z_{(m)}^*} (S_1(t) - \widehat{S}_1(t))^2 dt \right\}^{1/2},$$

where $S_1(t)$ is the true survival (exponential distribution) and $\widehat{S}_1(t)$ is the respective estimator of $S_1(t)$. Specifically, we calculate mean, median, minimum, maximum, the 25th percentile, the 75th percentile, and standard deviation of the $r = 1000$ values of the $L_2 - norm$ for Bayesian and frequentist estimates, for the different proportions of censoring and sample size.

TABLE 1. Selection of M_1 for each sample size and percent of censoring

percent of censoring	$n = 50$	$n = 100$	$n = 200$
75	15	30	52
55	13	20	30
30	13	16	30

TABLE 2. Descriptive statistics of $L_2 - norm$, n=50

75% of censoring	Mean	Median	Min	Max	P25	P75	S.D.
$\| S_1 - \widehat{S}_1 \|_2$	0.165	0.154	0.032	0.730	0.120	0.191	0.071
$\| S_1 - \mathrm{PL} \|_2$	0.208	0.176	0.053	1.233	0.132	0.249	0.122
55% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.151	0.142	0.025	0.531	0.104	0.182	0.066
$\| S_1 - \mathrm{PL} \|_2$	0.200	0.174	0.057	1.343	0.129	0.236	0.110
30% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.164	0.153	0.037	0.445	0.113	0.201	0.068
$\| S_1 - \mathrm{PL} \|_2$	0.177	0.160	0.049	0.582	0.122	0.214	0.078

TABLE 3. Descriptive statistics of $L_2 - norm$, n=100

75% of censoring	Mean	Median	Min	Max	P25	P75	S.D.
$\| S_1 - \widehat{S}_1 \|_2$	0.201	0.171	0.071	0.853	0.137	0.231	0.098
$\| S_1 - \mathrm{PL} \|_2$	0.185	0.158	0.043	0.874	0.119	0.218	0.103
55% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.164	0.157	0.036	0.563	0.122	0.194	0.065
$\| S_1 - \mathrm{PL} \|_2$	0.161	0.143	0.051	0.628	0.108	0.196	0.076
30% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.133	0.132	0.037	0.424	0.100	0.160	0.045
$\| S_1 - \mathrm{PL} \|_2$	0.136	0.125	0.044	0.651	0.096	0.159	0.061

TABLE 4. Descriptive statistics of $L_2 - norm$, n=200

75% of censoring	Mean	Median	Min	Max	P25	P75	S.D.
$\| S_1 - \widehat{S}_1 \|_2$	0.232	0.211	0.097	0.678	0.166	0.271	0.091
$\| S_1 - \mathrm{PL} \|_2$	0.156	0.138	0.042	0.568	0.105	0.185	0.076
55% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.191	0.187	0.078	0.511	0.161	0.214	0.047
$\| S_1 - \mathrm{PL} \|_2$	0.134	0.118	0.042	0.922	0.088	0.162	0.069
30% of censoring							
$\| S_1 - \widehat{S}_1 \|_2$	0.157	0.149	0.052	0.347	0.121	0.185	0.051
$\| S_1 - \mathrm{PL} \|_2$	0.105	0.094	0.033	0.381	0.074	0.124	0.045

From the Tables 2-4, we conclude that for the simulated data from the exponential of mean 3, the Bayes estimator fits better than the Kaplan-Meier estimator, for the case of low sample size and every level of censoring. The inverse situation happens when large sample is observed. When we consider a sample size of 100 and a level of censoring greater than 50 percent, the PL estimator has a better performance, and

when the number of failures corresponds to approximately 70 percent of the sample both estimators have about the same behavior.

4.2. Illustrative Examples.

Example 1. *For illustration, using the Kaplan-Meier data set we apply the methodology described in Section 2. This data set has been extensively examined by many authors in a Bayesian nonparametric context, given a practical method of comparison between different methodologies. For example see* [14], [15] *and* [8].
Based on a simulation of d=2000 Dirichlet random variates, for each t, Table 5 present the posterior mean, standard deviation and a 90% interval for the survival probabilities. We consider $M_1 = M_2$=24, *i.e., approximately two times the maximum time of the Kaplan-Meier data set.*

TABLE 5. Posterior mean, [standard deviation] and (90% interval) of the survival probabilities for the Kaplan-Meier data

t	$\widehat{S}_1(t)$	S.D.	90% interval
0.8	0.932	[0.039]	(0.857; 0.982)
1.0	0.896	[0.048]	(0.807; 0.963)
2.7	0.784	[0.064]	(0.672; 0.881)
3.1	0.743	[0.071]	(0.619; 0.850)
5.4	0.612	[0.073]	(0.489; 0.726)
7.0	0.514	[0.075]	(0.389; 0.642)
9.2	0.408	[0.074]	(0.289; 0.536)
12.1	0.280	[0.066]	(0.177; 0.394)

Considering a particular class of Beta process and an algorithm based in the Lévy formula, [15] *report for* $t = 1.0$, *0.8911, [0.0724] and (0.7486, 0.9760). Note that our estimates for the same time have less dispersion. In Figure 1 is plotted the posterior density of* $\widehat{S}_1(1.0)$. *We observe that the shape of the posterior density is asymmetric as is also showed in* [15] *and* [8].

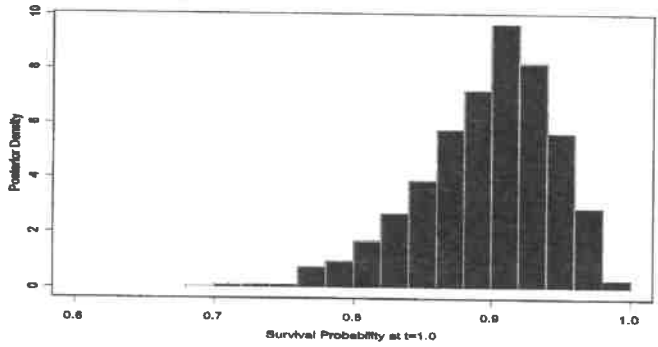

FIGURE 1. Posterior distribution of $\widehat{S}_1(1.0)$

Example 2. *The Bayesian nonparametric methodology is applied to the statistical analysis of a real-life case. The data set consists of survival times of 31 post-operated women, who were diagnosed with cervix uterine cancer of type IBI accordingly to International Federation of Gynecology and Obstetrics (FIGO). The analyzed patients correspond to a homogeneous subset of patients that presented commitment of borders when the operation was carried out and they did not present commitment of ganglions. The data were collected between 1997-2000, in the Instituto Nacional del Cancer of Chile and were measured in a range of 14 to 175 months. The average age of the patients at the time of the surgery was 52.2 (627 months approximately). The lifetime data present 61 percent of censoring. This information is incorporated in the prior distribution of the subsurvival functions $S_1^*(t)$ and $S_2^*(t)$.*

According to equation (2.2), we consider $\alpha_1(t,\infty) = M_1 \exp(-t/M_1)$, $t > 0$ and $\alpha_2(t,\infty) = M_2 - t$, $0 < t < M_2$. If the patients do not die in 7 years ($M_1 = 84$ months), the expectancy maximum life is 85 years, i.e., $M_2 = (1020 - 627) = 393$ months. From this selection of prior distribution, the prior of the proportion of censorship $\Pr(\delta = 2) = \Pr(X_1 > X_2)$ is a nonsymmetric Beta distribution with mode greater than 0.5.

In Figure 2 is displayed the Bayesian estimator for the survival function for the post-operated women with cervix uterine cancer. We considered a simulation of d=2000 Dirichlet random variates, for each t. Since the Kaplan-Meier estimator remains constant between the last time of failure and the following times of censorship, the Bayesian estimador can be calculated on all observed times, including censoring time. Note that these cancer data presents high proportion of censoring, then it is

reasonable to apply the Bayesian methodology for this case. Table 6 presents the posterior mean, standard deviation and a 90% interval for the estimated survival probability for 2, 3, 5 and 10 years.

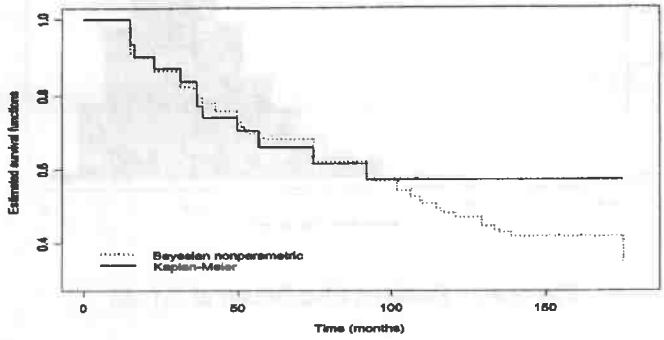

FIGURE 2. Bayesian estimated survival function for the post-operated women

TABLE 6. Posterior mean, [standard deviation] and (90% interval) of the survival probabilities for the post-operated women data

t (months)	$\widehat{S}_1(t)$	*S.D.*	*90% interval*
24	*0.866*	*[0.024]*	*(0.822; 0.905)*
36	*0.823*	*[0.027]*	*(0.776; 0.866)*
60	*0.685*	*[0.033]*	*(0.631; 0.740)*
120	*0.484*	*[0.036]*	*(0.426; 0.546)*

5. DISCUSSION

In this paper we were interested in the estimation problem of a survival curve. Several aspects were considered: a nonparametric Bayesian framework for randomly censored data under a Dirichlet bivariate process prior, numerical computation of the Bayesian estimator, a comparative study with the PL-estimator, and an application of the Bayesian approach to a case study in cancer.

First, the Bayesian setup proposed here considers a certain Dirichlet prior distribution placed on the vector of the subsurvival functions. Then the Bayesian estimator is calculated via the Peterson formula. We remark that the analysis is conjugated.

Our approach is different from others existing in the literature. For instance, assuming the censoring times to be constant [4] analyzed a model considering an usual Dirichlet process for the survival of interest. The subjectivity of the choice of the hyperparameters did not allow a comparison with our estimator. Also, their representation of the posterior is somewhat cumbersome. On the other hand the model proposed by [9] places a Beta process prior with non-negative independent increments for the cumulative hazard function. Below we will refer to some numerical aspects of this estimator in relation to our results.

In order to actually compute the nonparametric Bayesian estimator of the survival function we have introduced an algorithm based on Runge-Kutta fourth-order method, which is very popular for solving ordinary differential equations. It provides good accuracy and it is stable. In our case the application of the method has been very direct and simple to program. It is interesting to remark that the Runge-Kutta method was applied on each one of the intervals $[0, z^*_{(j)})$, $j = 1, \ldots, m$; where $z^*_{(1)} < z^*_{(2)} \ldots < z^*_{(m)}$ are the realizations of the distinct order statistics of the observable variable $Z = \min(X_1, X_2)$. Extensions from the right randomly censored model to a competing risks model can be considered in a similar way. The essential idea of the method is the same.

The comparative study has been divided in two parts. From the simulations (exponential distribution competing with gamma distribution) we conclude that the Bayesian estimator of the exponential survival fits better than the PL-estimator for low sample size and every censoring level. So our estimator is a possible alternative to be considered in that conditions.

The Bayesian methodology was illustrated considering the original data of [1]. A comparison is made with the results obtained by [15], whom implemented the approach introduced by [9]. For instance $\hat{S}_1(t = 1.0)$ is similar in both cases, but in our case the standard deviation of the estimator is smaller.

The other example has been devoted to apply the Bayesian framework to a real case of cervix uterine cancer. In this case the lifetime data present a high proportion of censoring. This fact is incorporated as a relevant information in the elicitation of the prior distribution. Some survival probabilities of medical interest were computed.

The scheme of subsurvival functions presented in this work, i.e., to consider as space of parameters $\Theta = \{(S^*_1, S^*_2) : S^*_1 + S^*_2 \text{ is a survival function}\}$, allows directly compute the Bayes estimator of the proportion of censoring $q = \Pr(\delta = 2)$. Note that the number of $\{\delta_i = 2\}$ follows a Binomial distribution of parameters $(n, \Pr(\delta = 2))$. It concludes $\hat{q} = \lim_{t \to 0^+} \widehat{S}^*_1(t)$ is the Bayes estimator of q under a Beta prior distribution and quadratic loss function.

It should be possible to apply the subsurvival scheme to the estimation problem of a cured fraction in survival analysis. This topic is theme of our current research.

REFERENCES

1. Kaplan, E.L. and Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**, 87–90, 457–481.
2. Peterson, A.V. (1977), 'Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions', *Journal of the American Statistical Association* **72**, 854–858.
3. Breslow, N.E. and Crowley, J. (1974), 'A large sample study of the life table and product limit estimates under random censorship', *The Annals of Statistics* **13**, 437–453.
4. Susarla, V. and Van Ryzin, J. (1976), 'Nonparametric Bayesian estimation of survival curves from incomplete observations', *Journal of the American Statistical Association* **71**, 897–902.
5. Ferguson, T.S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1** , 209–230.
6. Blum, J. and Susarla, V. (1977), 'On the posterior distributions of a Dirichlet process given randomly right censored observations', *Stochastic Processes and their Applications*. **5**, 207–211.
7. Ferguson, T.S., Phadia, E.G. and Tiwari, R.C. (1992), 'Bayesian nonparametric inference. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh and P.K. Pathak, eds.)', 127–150. *IMS Lecture Notes & Monograph Series*, Vol. 17.
8. Walker, S.G., Damien, P., Laud, P.L. and Smith, A.F.M. (1999), 'Bayesian nonparametric inference for random distributions and related functions', *Journal of the Royal Statistical Society B* **61**, 485–527.
9. Hjort, N.L. (1990), 'Nonparametric Bayes estimators based on Beta processes in models for life history data', *The Annals of Statistics* **18**, 1259–1294.
10. Salinas-Torres, V.H., Pereira, C.A.B. and Tiwari, R.C. (1997), 'Convergence of Dirichlet measures arising in context of Bayesian analysis of competing risks models', *Journal of Multivariate Analysis* **62**, 24–35.
11. Salinas-Torres, V.H., Pereira, C.A.B. and Tiwari, R.C. (2002), 'Bayesian nonparametric estimation in a competing risks model or a series system', *Journal of Nonparametric Statistics* **14**, 449–458.
12. Davis, P. and Ravinowitz, P. (1984), *Methods in Numerical Integrations*. 2nd Ed., Wiley, New York.
13. Wilks, S.S. (1962), *Mathematical Statistics*. Wiley, New York.
14. Ferguson, T.S. and Phadia, E.G. (1979), 'Bayesian nonparametric estimation based on censore data', *The Annals of Statistics* **7**, 163–186.
15. Damien, P., Laud, P.L. and Smith, A.F.M. (1996), 'Implementation of Bayesian non-parametric inference based on Beta processes', *Scandinavian Journal of Statistics* **23**, 27–36.

DEPARTMENT OF MATHEMATICS, U. OF SANTIAGO OF CHILE. P.O. BOX 307/2, SANTIAGO, CHILE.

E-mail address: vsalinas@usach.cl

INSTITUTE OF MATHEMATICS AND STATISTICS, U. OF SÃO PAULO. P.O. BOX 66281, 05315-970, SÃO PAULO-SP, BRASIL.

E-mail address: jromeo@ime.usp.br

DEPARTMENT OF MATHEMATICS, U. OF SANTIAGO OF CHILE. P.O. BOX 307/2, SANTIAGO, CHILE.

E-mail address: jpena@mat.puc.cl

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2005-01 - DE SOUZA BORGES,W., GUSTAVO ESTEVES, L., WECHSLER, S. Process Parameters Estimation in The Taguchi On-Line Quality Monitoring Procedure For Attributes. 2005.19p. (RT-MAE-2005-01)

2005-02 - DOS ANJOS,U.,KOLEV, N. Copulas with Given Nonoverlapping Multivariate Marginals. 2005.09p. (RT-MAE-2005-02)

2005-03 - DOS ANJOS,U.,KOLEV, N. Representation of Bivariate Copulas via Local Measure of Dependence. 2005.15p. (RT-MAE-2005-03)

2005-04 - BUENO, V. C., CARMO, I. M. A constructive example for active redundancy allocation in a k-out-of-n:F system under dependence conditions. 2005. 10p. (RT-MAE-2005-04)

2005-05 - BUENO, V. C., MENEZES, J.E. Component importance in a modulated Markov system. 2005. 11p. (RT-MAE-2005-05)

2005-06 - BASAN, J. L., BRANCO, M.D´E., BOLFARINE, H. A skew item response model. 2005. 20p. (RT-MAE-2005-06)

2005-07 - KOLEV, N., MENDES, B. V. M., ANJOS, U. Copulas: a Review and recent developments. 2005. 46p. (RT-MAE-2005-07)

2005-08 - VENEZUELA, M. K., BOTTER, D. A., SANDOVAL, M. C. Diagnostic techniques in generalized estimating equations. 2005. 16p. (RT-MAE-2005-08)

2005-09 - BOLFARINE, H., LACHOS, V.H. Skew-probit measurement error models. 2005. 10p. (RT-MAE-2005-09)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05311-970 - São Paulo, Brasil