A DDP Model for Survival Regression

Maria De Iorio,[1,*] **Wesley O. Johnson,**[2] **Peter Müller**[3] **and Gary L. Rosner**[†3]

[1]Department of Epidemiology and Public Health,

Imperial College

London W2 1PG, U.K.

[2]Department of Statistics,

University of California

Irvine , CA 92697, U.S.A.

[3]Department of Biostatistics & Applied Mathematics,

The University of Texas, M. D. Anderson Cancer Center

Houston, TX 77030, U.S.A.

Summary. We develop a Dependent Dirichlet Process model for survival analysis data. The model extends the ANOVA DDP that was presented in De Iorio et al. (2004) to handle continuous covariates and censored data. A major feature of the proposed approach is that there is no necessity for resulting survival curve estimates to satisfy the ubiquotous proportional hazards assumption. An illustration based on a cancer clinical trial is given where survival probabilities for times early in the study are estimated to be lower for those on a high dose treatment regimen than for those on the low dose

[*] *email:* m.deiorio@ic.ac.uk

[†] Order of authors is alphabetical.

treatment, while the reverse is true later for later times, possibly due to the toxic effect of the high dose for those who are not as healthy at the beginning of the study.

KEY WORDS: Censoring, Dependent Dirichlet process, Markov chain Monte Carlo

1. Introduction

Bayesian nonparametric and semiparametric models in survival analysis have become popular recently due to the advances in computing technology and the development of efficient computational algorithms. The Dirichlet process (Ferguson, 1973; Ferguson, 1974) is probably the most frequently used tool in Bayesian nonparametric inference. The Dirichlet process (DP) is a probability model for random probability distributions. The DP is indexed with two parameters, the mass parameter M and the base measure F_0. One of the critical properties of the DP is the almost sure discreteness of the random measure $F \sim DP(M, F_0)$. Sethuraman (1994) provides a constructive definition of the DP. Let $\delta(x)$ denote a point mass at x. We write $F = \sum_{h=0}^{\infty} p_h \delta(\theta_h)$, for the discrete random probability measure F with probability masses p_h at locations θ_h. Sethuraman (1994) shows that the locations θ_h are independent, identically distributed (i.i.d.) samples from the random measure F_0, while the weights are generated by a rescaled Beta distribution, $v_h \sim Beta(1, M)$ and $p_h = (1 - \sum_{i<h} p_i)v_h$. In many data applications the discreteness of the random measure is inappropriate. Dirichlet process mixture models (DPM) avoid the discreteness in the sampling distribution by adding an additional

convolution with a continuous kernel (Antoniak, 1974). The typical DPM model assumes

$$y_i \overset{iid}{\sim} H \qquad \text{with} \qquad H(y) = \int f(y \mid \mu) dF(\mu), \qquad F \sim DP(M, F_0) \quad (1)$$

i.e., a mixture with a DP prior on the random mixing measure. One of the main attractions of DPM models is computational simplicity. In fact, posterior simulation for DPM models is well understood (Escobar and West, 1998; MacEachern and Müller, 1998; Neal, 2000).

Some of the earliest work on the Dirichlet process in the context of survival analysis models dates back to Ferguson and Phadia (1979) and Susarla and Ryzin (1976) who obtained the Bayesian estimate of the random survival function and also derived the posterior distribution of the cumulative distribution function with right censored data. Kuo and Smith (1992) proposed a Gibbs sampler in the case of a Dirichlet process prior with left, right and interval censored data. Doss (1994), Doss and Huffer (1998) and Doss and Narasimhan (1998) discussed the implementation of a mixture of Dirichlet process priors for $F(t) = 1 - S(t)$ in the presence of right censored data using the Gibbs sampler, where $S(t)$ is the survivor function. We refer to Ibrahim et al. (2001) for a thorough review of Bayesian methods in survival analysis.

An important problem with DP based models in survival analysis is the inclusion of covariates, since the DP does not have a direct representation through either the hazard or cumulative hazard function (Hjort, 1990). An alternative possibility is to specify a DP prior for the distribution of the random error in the accelerated lifetime model. For such a model to be

identifiable, we need to centre the residual distribution at zero. This is not easily accomplished with a DP prior (Ibrahim et al., 2001). However, the problem can be resolved by modelling the error distribution with a Polya tree prior (Lavine, 1992). Alternatively, Hanson and Johnson (2002) proposed a median regression model in which the error distribution is modelled as a mixture of absolutely continuous Polya trees constrained to have median zero. Gelfand and Kottas (2003) developed a semiparametric median residual life regression model. The model is induced by a semiparametric accelerated failure time regression modelling for log survival time, based on a DP mixture for the error distribution.

The main contribution of this paper is to propose a model for survival regression based on a DP prior that allows for the introduction of covariates in a computationally tractable and naturally interpretable manner. De Iorio, Müller, Rosner and MacEachern (2004) considered dependent nonparametric models for a set of related random probability distributions or functions. They proposed a model that describes dependence across random distributions in an ANOVA-type fashion. Suppose that the set of random distributions $\{F_x,\ x \in X\}$ are indexed by a p-dimensional vector $x = (x_1, \ldots, x_p)$ of categorical covariates. For example, in a clinical trial $F_{(x_1,x_2)}$ could be the random distribution of response times for patients treated at levels x_1 and x_2 of two drugs. The probability model for the collection of random distributions $(F_x, x \in X)$ is such that marginally, for each x, the random measure F_x follows a $\mathrm{DP}(M, F_{0x})$ with total mass parameter M and a base measure F_{0x}. See later for a definition of F_{0x}. The dependence for F_x across x is intro-

duced using the dependent Dirichlet process (DDP) as defined by MacEachern (1999) to allow regression on a covariate x. The random measures F_x are almost surely discrete with the point masses generated marginally from the base measures F_{0x}. MacEachern (1999) introduces dependence across random measures generated marginally by a DP by imposing dependence in the distribution of these point masses, maintaining the base measure as the marginal distribution of the point masses. The model proposed in this paper extends the work of MacEachern (1999) and De Iorio et al. (2004) to the survival regression framework. Alternative constructions of families of dependent random measures with marginal DP distributions are proposed in Griffin and Steel (2006) and Dunson and Pillai (2006).

In section 2 we will briefly review the DDP model and the ANOVA DDP model and in section 3 we extend the ANOVA DDP to include continuous covariates. In section 4 we introduce a survival regression model based on a Dependent Dirichlet process prior. In sections 5 and 6 we illustrate the proposed model on two real data examples: a cancer clinical trial and data from the Colombia National fertility survey conducted in 1976. In the appendix we give details of the Gibbs sampling algorithm needed to simulate from the model.

2. ANOVA DDP

MacEachern (1999) generalises the DP to the DDP, defining a probability model for a collection of random distributions, the realisations of which are dependent. Let $\{F_x, x \in X\}$ be a set of random distributions indexed by x,

where X is any covariate space. The collection of random distributions is then specified as follows

$$F_x = \sum_{h=1}^{\infty} p_h \delta(\theta_{xh}), \qquad \text{for each } x \in X \tag{2}$$

where $\sum_{h=1}^{\infty} p_h = 1$. The dependence across the measures F_x is introduced by assuming that the locations θ_{xh} are dependent across different levels of x, but still independent across h. Let $\theta_h = (\theta_{xh},\ x \in X)$. A probability model for θ_h defines a stochastic process indexed by x, for fixed h. The sample path of θ_{xh} provides the locations at each value of the covariate and therefore the degree of dependence among the random distributions, $\{F_x, x \in X\}$, is governed by the level of the covariate x. Let F_{0x} denote the marginal distribution of θ_{Xh} at $X = x$. The marginal distribution, F_x, follows a DP with mass M and base measure F_{0x}, $F_x \sim DP(M, F_{0x})$, for each $x \in X$.

De Iorio et al. (2004) consider an extension of the dependent Dirichlet process to multiple categorical covariates and illustrate its use as a component in modelling complex hierarchical Bayesian models. Assume that $\mathcal{F} = \{F_x, x \in X\}$ is an array of random distributions, indexed by a categorical covariate x. The dependence across the random distribution is introduced by imposing an ANOVA-type probability model for the locations θ_{xh}. To ease the explanation, assume for the moment that $x = (v, w)$ is a bivariate covariate with $v \in \{1, 2, \ldots, V\}$ and $w \in \{1, 2, \ldots, W\}$. The covariates (v, w) could be, for example, the levels of two treatments in a clinical trial and F_x could be the random distribution of the recorded measurements for each patient. The ANOVA DDP model allows us to introduce an ANOVA-type

dependence structure on $\mathcal{F}$. For example, we might want F_x and $F_{x'}$ for $x = (v_1, w_1)$ and $x' = (v_2, w_1)$ to share a common main effect due to the common factor w_1. The desired construction of dependence is achieved by imposing the following probability model on the locations θ_{xh}

$$\theta_{xh} = m_h + A_{vh} + B_{wh} \tag{3}$$

with $m_h \overset{iid}{\sim} p^o_m(m_h)$, $A_{vh} \overset{iid}{\sim} p^o_{A_v}(A_{vh})$ and $B_{wh} \overset{iid}{\sim} p^o_{B_w}(B_{wh})$ and independence across h, v and w. The joint probability model on $\mathcal{F} = \{F_x, x \in X\}$ is referred to as ANOVA DDP(M, p^o), where M is the mass parameter and p^o is the base measure on the ANOVA effects in eq. (3). Marginally for each $x = (v, w)$, F_x follows a DP process with mass parameter M and base measure F_{0x} given by the convolution of $p^o_m, p^o_{A_v}$ and $p^o_{B_w}$ and the dependence among the random distributions is defined by the covariance structure of the point masses θ_{xh} across x. One great advantage of the ANOVA DDP model is its interpretability in terms of standard ANOVA concepts. In fact, m_h can be interpreted as an "overall mean", while A_v and B_w are the "main effects" for covariate levels v and w. Moreover, the ANOVA DDP model is easily generalised to p-dimensional categorical covariates $x = (x_1, \ldots, x_p)$:

$$\theta_{xh} = m_h + \sum_{i=1}^{p} A_{ih}(x_i)$$

where $A_{ih}(x_i)$ is the main effect due to covariate x_i. The inclusion of interaction effects is straightforward as well as the imposition of constraints on some of the offsets. A crucial feature of the model is that model specification and computation are dimension independent. As in standard ANOVA models, we can introduce identifiability constraints for interpretability.

3. Extension to continuous covariates

In the described construction the ANOVA DDP model requires categorical covariates. In many data analyses this limits the applicability of the model when we wish to include continuous covariates. Discretizing the continuous covariates would involve a loss of information. For simplicity of explanation, consider the case with bivariate covariates $x = (v, z)$, where v is categorical and z is continuous. The dependence across the random distributions can be achieved by imposing a linear model on the point masses:

$$\theta_{xh} = m_h + A_{vh} + \beta_h z \tag{4}$$

with $m_h \overset{iid}{\sim} p^o_m(m_h), A_{vh} \overset{iid}{\sim} p^o_{A_v}(A_{vh})$ and $\beta_h \overset{iid}{\sim} p^o_\beta(\beta_h)$ and independence across h. As in a standard linear model β_h can be interpreted as a slope coefficient. The model is parametrised by the mass parameter M and the base measure p^o on the ANOVA effects and the slope coefficient in (4). Marginally, for each $x = (v, z)$, the random distribution F_x follows a DP. As in the ANOVA DDP case, model (4) defines dependence across x by defining the covariance structure of the point masses θ_{xh} across x. We refer to the joint probability model on $\mathcal{F}$ as $(F_x, x \in X) \sim$ LINEAR DDP(M, p^o).

The model is easily genarilised to more than one continuous covariate. Moreover, it is not restricted to univariate distributions F_x as the point masses θ_{xh} can also be multidimensional.

4. DDP Survival Regression

Let T be a continuous non-negative random variable defined on $[0, \infty)$ denoting the event times of individuals in some population of interest. Let

$f(t)$ and $F(t)$ denote the probability density function and the distribution function of T, respectively. Then the probability of an individual surviving until time t is given by the survivor function

$$S(t) = 1 - F(t) = P(T > t)$$

Let $t_1, \ldots, t_n$ be n independent and identically distributed survival times, where t_i is the survival time, or *time to event*, of the i-th individual. A particular feature of this type of data is *censoring*, i.e. some lifetimes (often a non-trivial part of the dataset) are known to have occurred only within certain intervals, while the remaining lifetimes are known exactly. In this paper we will deal only with right-censoring, which implies that the event is observed only if it occurs prior to some prespecified times. But the model can be easily generalised to account for different types of censoring.

Let ν_i be the censoring indicator for individual i, where $\nu_i = 0$ if t_i is right-censored and $\nu_i = 1$ if t_i is an observed event time. Let x_i be the p-dimensional vector of categorical and continuous covariates for individual i and let $f_x(t)$ and $S_x(t)$ denote the density and the survivor function of an individual with covariates x, respectively. Let $\boldsymbol{\phi}$ denote all model parameters and let $D = \{t_i, \nu_i, x_i\}_{i=1}^n$ denote the data. The likelihood function is

$$L(\boldsymbol{\phi} \mid D) = \prod_{i=1}^{n} f_{x_i}(t_i \mid \boldsymbol{\phi})^{\nu_i} S_{x_i}(t_i \mid \boldsymbol{\phi})^{1-\nu_i}$$

Let $\mathrm{N}(\cdot \mid \mu, \sigma^2)$ denote a Normal density function with moments μ and σ^2. For each x_i we define a mixture of Normals sampling model $f_{x_i}(t_i \mid \boldsymbol{\phi})$.

$$p(t_i \mid x_i = x, F_x) = \int \mathrm{N}(t_i \mid \mu, \sigma^2) \mathrm{d}F_x(\mu) \tag{5}$$

Using the DDP framework we then introduce dependence across covariate values x_i. We assume

$$\{F_x, x \in X\} \sim \text{LINEAR DDP}(M, F_0) \tag{6}$$

Convolving the discrete measure generated from a DP is an instance of the general DPM defined in (1).

Posterior inference in model (5)-(6) is most easily performed in the framework of DPM models. For simplicity of explanation, consider a bivariate covariate $x = (v, z)$, where v denotes a categorical covariate with levels $v = 1, \ldots, V$ and z is a continuous covariate. Let $\alpha_h = [m_h, A_{2h}, \ldots, A_{Vh}, \beta_h]$ denote the row vector of parameters corresponding to the h-th point mass in the random measures. We set A_{1h} equal to zero to ensure identifiability. Let d_i denote a design vector that represents the appropriate ANOVA effect and the value of the continuous covariate corresponding to x_i, so that $\theta_{xh} = \alpha_h d_i$ for $x = x_i$. The generalization to more than two covariates is straightforward. The model specified in eq. (4)-(6) can then be written as a DP mixture of linear models

$$p(t_i \mid x_i = x, F) = H_x(t_i), \;\; H_x(t) = \int \text{N}(t \mid \alpha d_i, \sigma^2) \text{d}F(\alpha), \quad \text{and } F \sim DP(M, F_0) \tag{7}$$

with base measure $F_0 = (p^o_m, p^o_{A_v}, p^o_\beta)$. To verify that (7) is equivalent to (5) and (6) note that $\alpha \sim F$ and $\theta = \alpha d_i$ imply $\theta \sim F_{0x}$ for $x = x_i$. This is true by definition of the base measure $F_{0x}(\theta)$ in (4). Introducing latent variables

α_i, we can rewrite model (7) as a hierarchical model

$$t_i \mid x_i, \alpha_i \sim \mathrm{N}(t_i \mid \alpha_i d_i, \sigma^2) \tag{8}$$
$$p(\alpha_i \mid F) = F \text{ and } F \sim \mathrm{DP}(M, F_0)$$

In words, the observations t_i are sampled from a mixture of linear models, with a DP prior on the unknown mixing measure. This representation implies that any Markov chain Monte Carlo (MCMC) scheme for DP mixture models can be used for posterior simulation. For example, Neal (2000), MacEachern and Müller (1998) and Jain and Neal (2004) describe specific algorithms to implement posterior MCMC simulations in DPM models, while in De Iorio et al. (2004) relevant modifications needed for the ANOVA DDP model are discussed. In the appendix we show how the existing algorithms can be modified to take in account the presence of censored observations in the data. The conjugate nature of the base measure F_0 and the kernel greatly simplifies posterior simulation.

Other choices of kernels are possible without significantly complicating inference. A natural alternative for non-negative event times is a mixture of log normal distributions for the event times and therefore specifying the model as

$$\begin{aligned} p(t_i \mid x_i = x, F_x) &= \frac{1}{t_i} \int \mathrm{N}(\log(t_i) \mid \mu, \sigma^2) \mathrm{d}F_x(\mu) \\ \{F_x, x \in X\} &\sim \text{LINEAR DDP}(M, F_0) \end{aligned} \tag{9}$$

As before, we can rewrite model (9) as a DP mixture. This modelling strategy allows for efficient computations. Other alternatives, such as assuming

an exponential or a Weibull kernel for the lifetimes, are possible. But in regression settings the lack of conjugacy would make computations more costly and require the use of less efficient MCMC schemes.

We complete model (7) and similarly model (9), by specifying a prior probability model on the remaining parameters in eq. (8):

$$\sigma^2 \sim \text{Inverse-Gamma}(\frac{\gamma}{2}, \frac{\delta}{2})$$

$$M \sim \text{Gamma}(a, b)$$

$$F_0 = \text{N}(c, C),\ c \sim \text{N}(\eta, \tau^2 I) \text{ and } C^{-1} \sim \text{Wishart}(\gamma_0, \gamma_0^{-1}\Phi_0)$$

Here, I denotes the identity matrix of appropriate dimension, while γ_0 and Φ_0 are the parameters of the Wishart distribution. Note that the mass parameter M induces a distribution on the number of clusters in which the observations fall. See Escobar (1994) for a discussion of the Gamma prior on the total mass parameter M.

5. Cancer clinical trial

We illustrate the proposed approach with inference for a cancer clinical trial. The trial is described in Rosner (2005). The data are summarised in table 1. The data record the event-free survival time in months for 761 women, i.e. t_i denotes the time until death, relapse, or treatment-related cancer. Fifty three percent of the 761 observations are censored. Researchers are interested in determining whether high doses of the treatment are more effective for treating the cancer compared to lower doses. High doses of the treatment are known to be associated with a high risk of treatment related mortality.

The clinicians hope that this initial risk is offset by a substantial reduction in mortality and disease recurrence or relapse, consequentely justifying more aggressive therapy. In the analysis we consider two categorical covariates and one continuous covariate: treatment dose (0 = low, 1 = high), estrogen receptor (ER) status (0 = positive, 1 = negative) and the size of the tumour in centimetres (cm).

[Table 1 about here.]

The primary reason for carrying out the clinical trial was to compare low versus high dose. For preliminary analysis, we carried out a Cox proportional hazards regression analysis, using tumour size and ER status as covariates and stratifying by treatment dose. Since treatment enters in the model as a stratification factor, no assumption is made about how this variable affects survival, i.e. each stratum is permitted to have a different baseline hazard function, while the coefficients for the remaining covariates are assumed to be constant across strata. In figure 1 the estimated survivor functions for the two groups are shown. The two curves intersect, indicating that the proportional hazards assumption would be inappropriate for the treatment variable. A further limitation of the proportional hazards approach is its inability to examine the effects of the treatment. In contrast, the proposed model-based Bayesian inference provides a full probabilistic description of uncertainties in addition to the point estimates of the survivor function. In particular, the model includes inference about any functional of interest of the survivor function.

[Figure 1 about here.]

We have performed a Bayesian analysis of these data using the model specified in (5) and (6). In this case, the latent variable vector α in eq. (7) is a four-dimensional vector, $\alpha = [ME, A(HD), A(ER), \beta_{TS}]$, where ME corresponds to an overall mean effect, while $A(HD)$ and $A(ER)$ are the offsets for high treatment dose and positive ER status respectively and β_{TS} denotes the coefficient for the tumour size. The design matrix could be extended to include interactions, nested effects, etc., as desired. We implemented the algorithm described in the appendix to obtain posterior simulations for these data. In the analysis, we fix the mass parameter of the Dirichlet process, M, equal to 1. We assume $F_0 = \mathrm{N}(m, C)$, $m \sim \mathrm{N}(\eta, 100I)$ and $C^{-1} \sim \mathrm{Wishart}(\gamma_0, \gamma_0^{-1} I \frac{1}{10})$, with $\eta = (0, 0, 0, 0)'$ and $\gamma_0 = 6$. The prior distribution for σ^2 is an Inverse-Gamma with mean 25 and variance 10.

Figure 2 shows the posterior estimates, $E(S \mid Data)$, of the survivor functions for different combinations of the categorical covariates and fixing tumour size at 3.8cm . The survivor functions corresponding to the two treatment groups cross (both for positive and negative ER status), showing a higher level of risk associated with high treatment dose in the first 20 months. After that, the plot shows an increase in the survival probability compared to low dose. Figure 3 illustrates posterior uncertainty, which is quite similar across treatment groups. In figure 4 we show the estimated survivor curves obtained using the LINEAR DDP model, the Cox model (stratifying by treatment dose) as well as the Kaplan-Meier curve. The curves are shown

for low treatment dose, positive ER status and tumour size equal to its mean value in the data. In the case of the Kaplan-Meier method a separate curve needs to be fitted for each covaiate level, not allowing any borrowing of information. We fitted the Kaplan-Meier curve not taking into account tumour size. The three methods lead to similar estimates of the survival probabilities, but one of the major advantages of the Bayesian framework is that it allows us to quantify the treatment effect and assess the precision of such estimates. For example, figures 5(a) and 5(b) show the difference in survival rates between the two treatment groups for positive ER status at 10 months and 40 months, respectively. Figure 6(a) shows the difference in survival probabilities over the months of the study between the two treatment groups for negative ER status.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

For comparison, we have also performed an analysis of the data using the semiparametric accelerated failure time (AFT) median regression model with a mixture of Polya trees on the error distribution. The model is described in Hanson and Johnson (2002). For the Polya tree prior we have chosen as centering distribution a Weibull distribution and the weight c was set equal

to 1. Figure 6(b) presents estimated survivor functions corresponding to the two treatment groups. The two curves are practically indistinguishable. The model fails to capture different shapes for the two groups. The DDP regression model presents greater flexibility allowing for curves with different shapes for the two groups (compare with Figure 2).

[Figure 6 about here.]

In table 2 we report the posterior inferences for 10 month survival probabilities for the two treatment groups and different tumour sizes.

[Table 2 about here.]

6. Childhood mortality in Colombia

We analyse data collected to study infant and childhood mortality in Columbia (Somoza, 1980). The data were collected in The Colombia National Fertility Survey which was carried out in 1976 by the *Corporación Centro Regional de Población* (CCRP) and the *Departmento Administrativo Nacional de Estadística* (DANE) as part of the World Fertility Survey and with the financial support of the U.S. Agency for the International Development (International Statistical Institute, 1978; Hobcraft, 1990). A questionnaire was administered to a sample of women between the ages of 15 and 49 eliciting their maternity history, educational level, age, union status and information on the sex, date of birth and survival status (at the date of interview) of all their children and, if applicable, age at death.

We consider data on a sub-sample of 1437 children (corresponding to the oldest child for each mother). The response of interest is the survival time

(in years) of a child at the time of the maternal interview. The covariates of interest are: gender (0=male; 1=female), birth cohort (1=1941-59; 2=1960-67; 3=1968-76) and a binary variable indicating whether a child was born in a rural area (1=yes; 0=no). Around 87% of the observations in the dataset are censored. The original research was conducted to investigate how patterns of childhood mortality have changed over time. Also of interest are urban/rural and gender differences.

We have fitted the model specified in (9), i.e. assuming a mixture of log normal distributions for the survival times. The baseline survivor function corresponds to a male child, born in an urban area and belonging to the third birth cohort. In this application the vector α of ANOVA effects is five-dimensional and $\alpha = (ME, S, R, BC1, BC2)'$, where ME is the overall effect and S, R, $BC1$ and $BC2$ denote the offsets corresponding to a female child, a child born in rural area, a child belonging to the first birth cohort and to the second birth cohort respectively. In performing the analysis of these data, we assume that $\sigma^2 \sim$ Inverse-Gamma$(2.5, 5)$, the base measure of the DP is Normal, $F_0 = \mathrm{N}(a, C)$, $a \sim \mathrm{N}(\eta, 100I)$ with $\eta = (-0.43, 0.32, -1.29, 0.77, 0.45)'$, C^{-1} has a Wishart distribution with 7 degrees of freedom and prior mean $E(C^{-1}) = 0.1I$ and the precision parameter M of the DP is set equal to 1.

[Figure 7 about here.]

Figure 7 shows the posterior estimates of the survivor curve for some typical children born in the third birth cohort. Male children born in urban areas have higher survival probability over time than male children born in

rural areas, but slightly lower survival probability if compared with females born in the urban area. In figure 8 we have plotted the posterior estimates of the survivor function corresponding to different age groups. While in the urban area no evident difference appears between birth cohorts, in rural areas children belonging to the first birth cohort (i.e. born between 1941 and 1959) have a lower life expectancy.

[Figure 8 about here.]

7. Discussion

We have presented a flexible nonparametric model that can be used to introduce categorical and continuous covariates in survival models based on DP priors. The inclusion of continuous covariates by imposing a linear model on the random point masses of the DP provides a natural generalization of the ANOVA DDP model (De Iorio et al., 2004). Advantages of the DDP survival regression model include ease of interpretability and computational tractability.

A limitation of the model is the need to use a conjugate base measure and mixing kernel in eq. (7) to be able to utilize efficient MCMC schemes. Models with non-conjugate base measure and mixing kernel could still be used, but at a greater computational cost (see, for example, Neal (2000) or MacEachern and Müller (1998)). Moreover, more complex DDP models could be used as priors in survival regression settings, although they would necessarily result in less efficient computational strategies.

References

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

Bush, C. and MacEachern, S. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association* **99**, 205–215.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics* **22**, 1763–1786.

Doss, H. and Huffer, F. (1998). Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet priors. Technical report, Department of Statistics, Ohio State University.

Doss, H. and Narasimhan, B. (1998). Dynamic display of changing posterior in Bayesian survival analysis. In Dey, D., Müller, P. and Sinh, D., editors, *Practical Nonparametric and Semiparamtric Bayesian Statistics*, pages 63–88. New York, NY, USA.

Dunson, D. B. and Pillai, N. S. (2006). Bayesian density regression. Technical report, Duke University ISDS, USA.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical

models. In Dey, D., Müller, P. and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 1–22. New York, NY, USA.

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

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

Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics* **7**, 163–186.

Gelfand, A. E. and Kottas, A. (2003). Bayesian Semiparametric Regression for Median Residual Life. *Scandinavian Journal of Statistics* **30**, 651–665.

Griffin, J. E. and Steel, M. F. J. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association* **101**, 179–194.

Hanson, T. and Johnson, W. O. (2002). Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association* **97**, 1020–1033.

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

Hobcraft, J. N. (1990). Illustrative analysis : evaluating fertility levels and trends in colombia. *World Fertility Survey Scientific Reports* .

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York, NY, USA.

International Statistical Institute (1978). The colombia fertility survey 1976: A summary of findings. *World Fertility Survey* .

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics* **13**, 158–182.

Kuo, L. and Smith, A. F. M. (1992). Bayesian computations in survival models via the Gibbs sampler. In Klein, J. P. and Goel, P. K., editors, *Survival Analysis: State of the Art*, pages 11–24. Boston: Kluwer Academic.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* **20**, 1222–1235.

MacEachern, S. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–239.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on bayesian Statistical Science*, Alexandria, VA. American Statistical Association.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**.

Rosner, G. L. (2005). Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics* **61**, 239–245.

Sethuraman, J. (1994). A constructive defnition of the Dirichlet process prior. *Statistica Sinica* **2**, 639–650.

Somoza, J. L. (1980). Illustrative analysis: infant and child mortality in colombia. *World Fertility Survey Scientific Reports* .

Susarla, V. and Ryzin, J. V. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American*

Statistical Association **71**, 897–902.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Discussion paper, Duke University ISDS, USA.

Appendix

Posterior simulations

We briefly describe how to implement posterior simulations for the DDP survival regression model defined in eq. (7), and we show how to modify existing MCMC algorithm to account for censoring. See, for example, MacEachern and Müller (1998) or Bush and MacEachern (1996) for a review of efficient Gibbs sampling schemes to estimate DPM models. The almost surely discreteness of the random distribution F implies that there is a positive probability for ties among the α_i. Let $\{\alpha_1^\star, \ldots, \alpha_k^\star\}$ be the set of $k \leq n$ distinct elements in $\{\alpha_1, \ldots, \alpha_n\}$ and let $s_i = j$ iff $\alpha_i = \alpha_j^\star$ denote configuration indicators. Also let n_j be the number of s_i equal to j. Note that the same type of algorithm can be applied if we model the event times with a lognormal distribution by substituting t_i with $\log(t_i)$ in the normal pdf below, and multiplying the normal pdf by $1/t_i$ (model (9)).

Resampling σ^2 given all the other parameters and the data.

$$\sigma^2 \mid \text{all the rest} \sim \text{Inverse-Gamma}\left(\frac{\gamma + n}{2}, \frac{\delta + \sum_{i=1}^n (t_i - \alpha_i d_i)^2}{2}\right)$$

Resampling s_i given all the other parameters and the data. To sample s_i, we marginalise over α_i.

Let $p(\alpha_j^\star \mid t_l, l \neq i, s_l = j, \text{ all other parameters})$ be the conditional posterior pdf for $\alpha_j^\star$.

- **Event times.** If t_i is not censored,

$$\Pr(s_i = j \mid s_{-i}, t_i, \text{ all the rest }) \propto$$

$$\begin{cases} n_j^- \int \mathrm{N}(t_i \mid \alpha_j^\star d_i, \sigma^2) p(\alpha_j^\star \mid t_l, s_l = j, l \neq i, \text{ all the rest }) \mathrm{d}\alpha_j^\star & j = 1, \ldots, k^- \\ M \int \mathrm{N}(t_i \mid \alpha d_i, \sigma^2) \mathrm{d}F_0(\alpha; c, C) & j = k^- + 1 \end{cases}$$

  Here s_{-i} denotes the vector $(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, n_j^- denotes the size of the j-th cluster with α_i removed from consideration ($n_{s_i}^- = n_{s_i} - 1$, while $n_j^- = n_j$ for $j \neq s_i$), and k^- denotes the number of clusters when α_i is removed. If $n_{s_i}^- = 0$, we relabel the remaining clusters $j = 1, \ldots, k^- = k - 1$.

- **Censored observations.** If t_i is censored, let $p_j(t)$ be the conditional predictive density for cluster j, obtained excluding the i-th observation:

$$p_j(t) = \int \mathrm{N}(t \mid \alpha_j^\star d_i, \sigma^2) p(\alpha_j^\star \mid t_l, s_l = j, l \neq i, \text{ all the rest }) \mathrm{d}\alpha_j^\star$$

  and

$$p_o(t) = \int \mathrm{N}(t \mid \alpha d_i, \sigma^2) \mathrm{d}F_0(\alpha; c, C)$$

  Then

$$\Pr(s_i = j \mid s_{-i}, t = c_i, \text{ all the rest }) \propto \begin{cases} n_j^- \int_{c_i}^\infty p_j(t) \mathrm{d}t & j = 1, \ldots, k^- \\ M \int_{c_i}^\infty p_o(t) \mathrm{d}t & j = k^- + 1 \end{cases}$$

where c_i is the actual censoring time of the i-th observation. Once s_i has been resampled impute t_i from a left truncated normal, that is sample t_i from p_{s_i} left truncated in c_i.

After resampling s_i, set $k = k^-$ if $s_i \leq k^-$ and $k = k^- + 1$ if $s_i = k^- + 1$.

Resampling $\alpha_j^\star$. The full conditional distribution for $\alpha_j^\star$ is obtained by considering the simple Bayesian model

$$\alpha_j^\star \sim F_0(c, C)$$

and

$$t_i \sim \mathrm{N}(\alpha_j^\star d_i, \sigma^2)$$

for all i such that $s_i = j$.

Resampling M. See West (1992).

Resampling c. This is straightforward update. The full conditional for c is the same as obtained in the simple Bayesian model

$$\begin{aligned} c &\sim \mathrm{N}(\eta, \tau^2 I) \\ \alpha_j^\star &\sim \mathrm{N}(c, C), \quad j = 1, \ldots, k \end{aligned}$$

Resampling C. The full conditional for $P = C^{-1}$ is a Wishart with $\gamma_1 = \gamma_0 + k$ degrees of freedom and scale matrix S^{-1} where

$$S = \gamma_0 \Phi_0^{-1} + \sum_{j=1}^{k} (\alpha_j^\star - c)(\alpha_j^\star - c)'$$

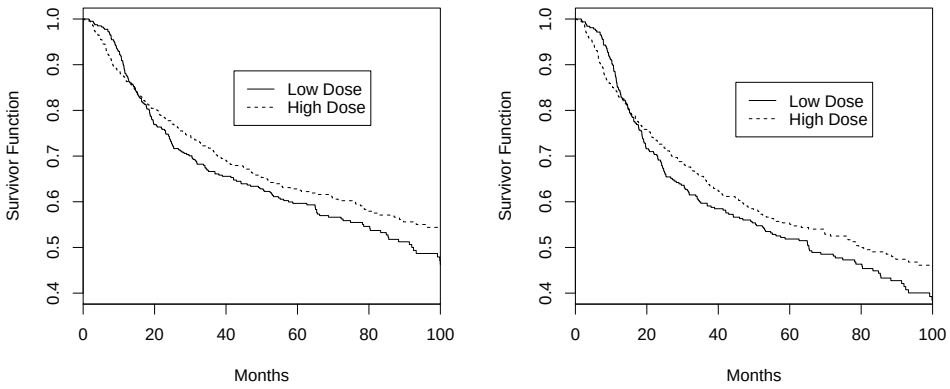


Figure 1. Survivor Curves from a Cox proportional hazards model with two strata (low and high dose). The two curves are estimated at the mean value of tumour size and for positive ER status on the left and negative ER status on the right.

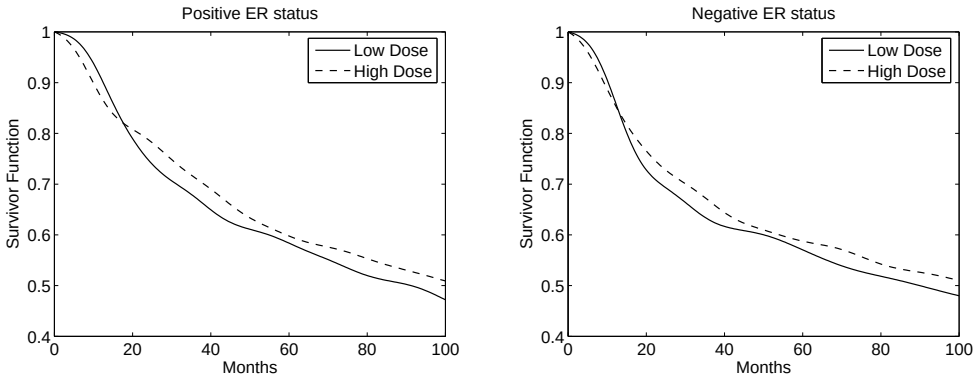


Figure 2. Posterior estimated survivor functions. The solid lines correspond to a patient in the low dose treatment group while the dashed lines correspond to the high dose group. The two curves are estimated for tumour size equal to 3.8 cm and for positive ER status (left panel) and negative ER status (right panel).

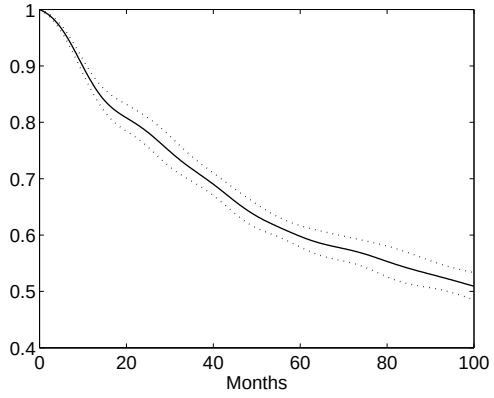


Figure 3. Posterior estimated survivor function (solid line) and 95% credible interval (dotted lines). The curve corresponds to a patient with positive ER status, who receives high treatment dose and has a tumour of size 3.8cm.

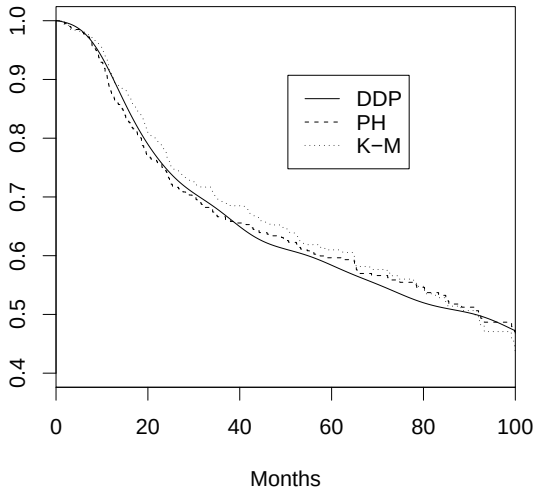
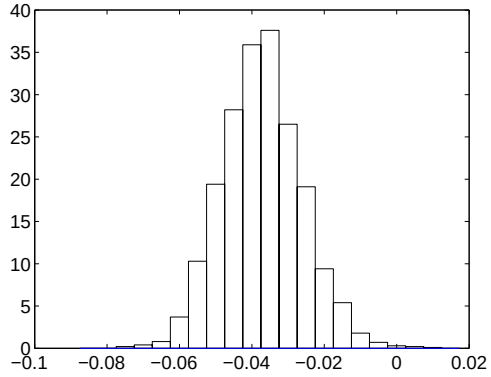
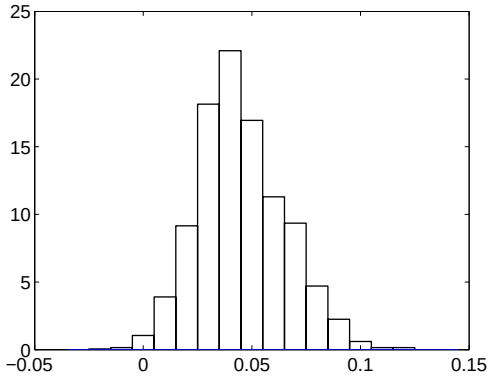


Figure 4. Estimated survivor curves from the LINEAR DDP model (solid line), the Cox proportional hazards (dashed line) model and Kaplan-Meier curve (dotted line). The curves correspond to a patient with positive ER status that receives low treatment dose. In the case of the Cox and the LINEAR DDP model tumour size was fixed equal to 3.8 cm.

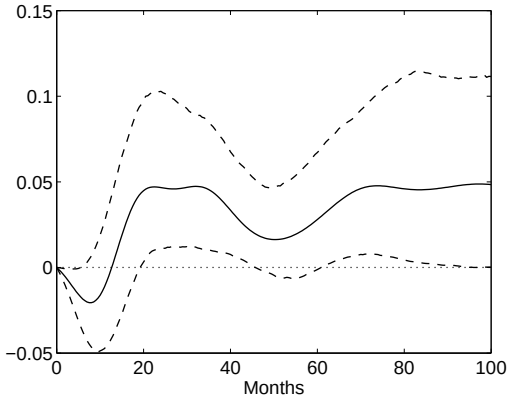


(a)

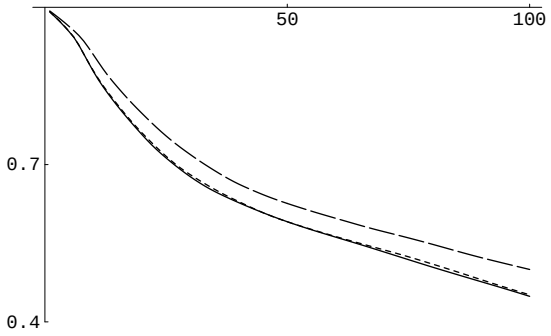


(b)

Figure 5. (a) Posterior distribution of the difference in survival probabilities at 10 months between a patient that receives high treatment dose and one that receives the low dose. The histogram is shown for positive ER status and tumour size equal to 3.8 cm. (b) Posterior distribution of the difference in survival probabilities at 40 months between a patient that receives high treatment dose and one that receives the low dose. The histogram is shown for positive ER status and tumour size equal to 3.8 cm.
Note the almost non-overlapping nature of the supports of the two marginal distributions, and the change in sign of the treatment effect.



(a)



(b)

Figure 6. (a) Posterior estimate (solid line) and 95% credible interval (dashed lines) of the difference in survival probabilities between the high dose treatment group and the low dose treatment group. The difference is shown for negative ER status and fixing tumour size at 3.8 cm. (b) Posterior survivor functions using the AFT median regression model. The curves are estimated for tumour size equal to 3.8 cm. The solid line refers to low treatment dose and negative ER status. The dashed line corresponds to high treatment dose and negative ER status, while the long dashed line shows the survival for a patient in the low dose group but with positive ER status. Compare the almost vanishing difference between the solid and the dashed line with the differences reported in panel (a).

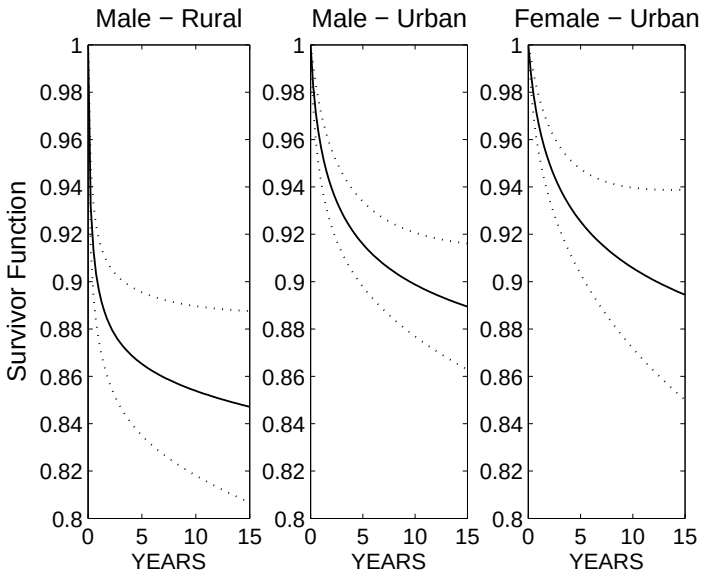


Figure 7. Posterior survivor functions (solid lines) with respective 95% credible interval (dotted lines). The figures represent the survival probability for a child belonging to the third birth cohort. The first panel shows the survivor function for a male child born in a rural area, the second one shows the survivor function of a male child born in an urban area and the third panel shows the survivor function of a female child born in an urban area.

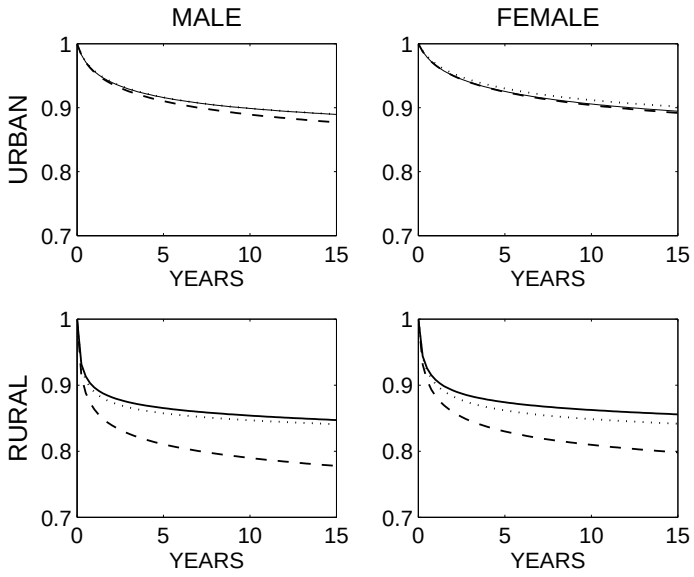


Figure 8. Posterior survivor functions. The solid line corresponds to children in the third birth cohort, the dotted line represents a child in the second birth cohort and the dashed line refers to children in the first birth cohort.

Table 1

Summary of Cancer Data

Survival time	(months)	Status	(freq.)	Dose	(freq.)	Tumour size	(cm)	ER status	(freq.)
Median	21.88	Censored	400	High	385	Mean	3.8	Positive	528
IQR	33.54	Event	361	low	376	STD	2.4	Negative	233

Table 2

10 months survival probabilities

Tumour size	1 cm	3.8 cm	8cm
LD	0.94 (0.011)	0.93 (0.008)	0.91 (0.014)
HD	0.90 (0.009)	0.89 (0.006)	0.87 (0.010)

Posterior mean (standard deviation) of 10 months survival probability for different size of tumour and treatment dose. Results are shown for positive ER status. HD stands for high treatment dose and LD for low dose.