Published in final edited form as:

Biometrics. 2009 September ; 65(3): 762–771. doi:10.1111/j.1541-0420.2008.01166.x.

Bayesian Nonparametric Non-Proportional Hazards Survival Modelling

Maria De Iorio*,

Department of Epidemiology and Public Health, Imperial College London W2 1PG, U.K

Wesley O. Johnson,

Department of Statistics, University of California, Irvine, CA 92697, U.S.A

Peter Müller, and

Department of Biostatistics & Applied Mathematics, The University of Texas, M. D. Anderson Cancer Center Houston, TX 77030, U.S.A

Gary L. Rosner

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. A major feature of the proposed approach is that there is no necessity for resulting survival curve estimates to satisfy the ubiquitous 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 treatment, while the reverse is true 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.

Keywords

Censoring; Dependent Dirichlet process; Markov chain Monte Carlo

1. Introduction

The Cox proportional hazards (PH) model is ubiquitous in Statistics. However, it often fails to fit, at least without some form of tinkering. For example, when a treatment effect is negative at the beginning of a study and positive by the end, namely when treatment interacts with time, the PH assumption fails. It is then necessary to either fit an interaction with time or to stratify on treatment. Other semiparametric models such as the accelerated failure time (AFT) model or the proportional odds (PO) model (see Klein and Moeschberger, 1997) will also fail under such circumstances since they also constrain in such a way that survival curves are not allowed to cross. In what follows, we present a Bayesian nonparametric approach to survival regression that allows curves to cross, or not, as the data dictate. Full inferences regarding the treatment effect, straightforward or not, are easily obtained using a Markov chain Monte Carlo implementation of the Bayesian model that we develop.

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

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 (DP) (Ferguson 1973, 1974) is probably the most frequently used tool in Bayesian nonparametric inference. The DP is a probability model for random probability distributions and is indexed with two parameters, the mass parameter M and the base measure F_0 . One of its critical properties is the almost sure discreteness of the random measure $F \sim DP(M, F_0)$. Sethuraman (1994) provides a constructive definition. Let δ

(x) denote a point mass at x . 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, $v_h \sim \text{Beta}(1, M)$ 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 (Lo, 1984). The typical DPM model assumes

$$y_i \stackrel{iid}{\sim} g \quad \text{with} \quad g(y) = \int f(y|\mu) dF(\mu), \quad F \sim DP(M, F_0) \quad (1)$$

i.e., a mixture of parametric probability densities 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).

Early work on the Dirichlet process in the context of survival analysis models dates back to Susarla and Ryzin (1976) and Ferguson and Phadia (1979) who obtained the Bayesian estimate of the survival function and derived its posterior distribution with right-censored data. Johnson and Christensen (1986) developed corresponding results for life table data and Kuo and Smith (1992) proposed a Gibbs sampler for right, left and interval-censored data that were modeled with a Dirichlet process prior. Doss (1994), Doss and Huffer (1998) and Doss and Narasimhan (1998) discussed the implementation of a mixture of Dirichlet processes 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 up to that point in time.

Empirical Bayesian semiparametric regression methods were first developed by Kalbfleisch (1978) for the PH model using a gamma process prior for the cumulative hazard, and by Christensen and Johnson (1988) for the AFT model using a Dirichlet process prior. Fully Bayesian semiparametric survival models based on Markov chain Monte Carlo methods were developed by Walker and Mallick (1999) using Polya Tree priors; see Lavine (1992) for development of these priors and references. More recently, Mixtures of Polya Tree (MPT) priors were also developed for use in semiparametric PH, AFT and PO models (Hanson and Johnson, 2002; Hanson et al., 2008). In each instance, MPT priors are placed on baseline survival distributions, which is straightforward since each model has a simple formula that relates survival to covariate information through this baseline and a linear function of the covariates. Kuo and Mallick (1997) and Gelfand and Kottas (2003) developed corresponding DPM approaches to the AFT model. Despite the flexibility in the baseline survivor distributions for these models, and notwithstanding the advantages of the Bayesian methods developed for them, they still involve severe constraints, in particular in the way covariates modify the baseline distribution.

The main contribution of this paper is to propose an unconstrained model for survival regression based on a DP prior that allows for the introduction of covariates in a computationally tractable and naturally interpretable manner. DeIorio et al. (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 $x = (x_1, \dots, x_p)$ 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 $DP(M, F_{0x})$ with total mass parameter M and a base measure F_{0x} . See Section 2 for a definition of F_{0x} . The dependence for F_x across x is introduced 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} . The DDP introduces dependence across the random measures F_x by imposing dependence in the distribution of these point masses, maintaining the base measure F_{0x} 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. An attractive feature of the model in De Iorio et al (2004) is that this extension is mathematically straightforward. However, from a data analysis perspective the extension is important, as it provides a straight-forward and relatively easy-to-implement model when proportional hazards or other restrictive semi-parametric settings fail to hold.

Alternative constructions of families of dependent random measures are proposed in Griffin and Steel (2006), who introduce dependence among the random distributions by making the weights in Sethuraman's representation dependent on the covariates still ensuring marginal DP distributions, and also in Dunson et al. (2007) and Dunson and Park (2007) who propose Bayesian kernel-weighted approaches for modelling random distributions conditional on covariate levels.

In Section 2 we will briefly review the general and more specific ANOVA DDP models, 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 we illustrate the proposed model on real data from a cancer clinical trial. In the Web Appendix B 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

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

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} . 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 $x \in X$. Note that the original construction of MacEachern (1999) is more general as it allows also the weights to depend on x . What we consider here is the *fixed-p* case which offers computational advantages.

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 distributions is introduced by imposing an ANOVA-type probability model for the locations θ_{xh} . For ease of explanation, assume for the moment that $x = (v, w)$ is a bivariate covariate with $v \in \{1, 2, \dots, V\}$ and $w \in \{1, 2, \dots, W\}$. 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}

$$\theta_{xh} = m_h + A_{vh} + B_{wh} \quad (3)$$

with $m_h \stackrel{iid}{\sim} p_m^o(m_h)$, $A_{vh} \stackrel{iid}{\sim} p_{A_v}^o(A_{vh})$ and $B_{wh} \stackrel{iid}{\sim} p_{B_w}^o(B_{wh})$ 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_m^o , $p_{A_v}^o$ and $p_{B_w}^o$ 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 . A crucial feature of the model is that model specification and computation are independent of the dimension of θ_{xh} . The ANOVA structure is defined for the vector-valued θ_{xh} , and inference is implemented by a posterior simulation scheme that exploits analytic availability of conditional posterior distributions of vectors of the dimension of θ_{xh} . Details appear in Web Appendix B. Finally, as in standard ANOVA models, we can introduce identifiability constraints for interpretability.

3. 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:

$$\theta_{xh} = m_h + A_{vh} + \beta_h z \quad (4)$$

with $m_h \stackrel{iid}{\sim} p_m^o(m_h)$, $A_{vh} \stackrel{iid}{\sim} p_{A_v}^o(A_{vh})$ and $\beta_h \stackrel{iid}{\sim} p_{\beta}^o(\beta_h)$ 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 $\mathcal{C}^Z$ as $\{F_x, x \in X\} \sim \text{LINEAR DDP}(M, p^o)$. The model is non-parametric in the specification of the effects of the ANOVA factors v and the sampling model. The effect of the continuous covariate is defined by the linear term in (4) and implies that in the framework of Section 4 the posterior expectation of the survival time changes linearly with continuous covariates. If desired, it could be generalized by including additional terms.

The model is easily generalised to more than one continuous covariate. Moreover, it is not restricted to univariate distributions F_x as the point masses θ_{xh} can 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. (We use $f(\cdot)$ and $F(\cdot)$ to generically indicate distributions and use the argument to specify the random variable). 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 $t_1, \dots, t_n$ 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 occurs either after an independent random departure from the study, or after the study ends. But the model can be easily generalised to account for different types of censoring.

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

$$L(\phi|D) = \prod_{i=1}^n f_{x_i}(t_i|\phi)^{\delta_i} S_{x_i}(t_i|\phi)^{1-\delta_i}.$$

Let $f(\cdot|\mu, \sigma^2)$ denote an arbitrary location scale family of parametric densities. For each x_i we define a mixture for the data:

$$f_{x_i=x}(t_i|F_x) = \int f(t_i|\mu, \sigma^2) dF_x(\mu, \sigma^2). \quad (5)$$

Here, $\phi = F_x$ parametrises the model. The integrand in (5) is often denoted as a kernel.

Three natural choices of kernel for survival data are the log normal, the log logistic or the log extreme value (Weibull) families. However, in many instances, a mixture of normals may

suffice (Lo, 1984). The mixture of normals model is sufficient to approximate arbitrary hazard functions. This follows from the approximation of a univariate density by a mixture of normals. Specifically, DP mixtures of normal kernels define a density estimate that converges almost surely in L_1 to the true pdf. See, for example, Tokdar (2006), theorem 4.2, for a precise statement and conditions. Then, assuming that the conditions of the theorem are satisfied, estimation of the hazard function inherits the same L_1 convergence properties (restricted to a finite interval). It is easy to show that the following proposition holds.

Proposition 1

Let $f(t)$ denote a p.d.f. for an event time, i.e., on R^+ . Let $f_J(t) = \sum_{j=1}^J w_j N(t|m_j, s_j)$ denote a mixture of Normals such that, for $\varepsilon \in (0, 1)$,

$$\int |f(t) - f_J(t)| dt < \varepsilon^2.$$

Let t_1 be such that the survival functions are bounded by $S(t_1) \geq \varepsilon$ and $S_J(t_1) \geq \varepsilon$. Let S, h, S_J, h_J denote the survival and hazard functions corresponding to f and f_J . Then

$$\int_0^{t_1} |h(t) - h_J(t)| dt < \varepsilon(1 + \varepsilon).$$

Proof

See Web Appendix A.

As ε tends to zero, t_1 tends to the upper limit of the support of the survival density. From the proposition it is clear that a mixture of normals can reasonably approximate any survival model, in particular, it would approximate the hazard for data that were log normal or log logistic etc. However, since survival data are positive, it remains appealing to take logs of the data first and to then implement the normal kernel.

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

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

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 example, consider a bivariate covariate $x = (v, z)$, where v denotes a categorical covariate with levels $v = 1, \dots, V$ and z is a continuous covariate. Let $\alpha_h = [m_h, A_{2h}, \dots, A_{Vh}, \beta_h]$ denote the row vector of parameters that appear in equation (4) for the h -th point mass in the random measure. 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 equation (4) can be written as $\theta_{xh} = \alpha_h d_i$ 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

$$f_{x_i=x}(t_i|F) \equiv f_x(t|\varphi) \equiv \int f(t|\alpha d_i, \sigma^2) dF(\alpha, \sigma^2), \text{ and } F \sim DP(M, F_0) \quad (7)$$

where $\varphi = F$ and with base measure $F_0 = (p_m^o, p_{A_v}^o, p_{\beta}^o, p_{\sigma^2}^o)$. To verify that (7) is equivalent to (5) and (6) note that $(\alpha, \sigma^2) \sim F$ and $\theta = \alpha d_i$ imply $(\theta, \sigma^2) \sim F_{0x}$ for $x = x_i$. This is true by definition of the base measure. Introducing latent variables α_i and σ_i^2 , we can rewrite model (7) as a hierarchical model

$$\begin{aligned} t_i | x_i, \alpha_i, \sigma_i^2 &\sim f(\cdot | \alpha_i d_i, \sigma_i^2) \\ (\alpha_i, \sigma_i^2 | F) &\sim F \text{ and } F \sim DP(M, F_0) \end{aligned} \quad (8)$$

If for example the parametric family $f(\cdot | \mu, \sigma^2)$ is normal, the observations t_i are sampled from a mixture of linear models, with a DP prior on the unknown mixing measure; if the family is log normal, we have a mixture of log linear models etc.

Due to conditional conjugacy of the base measure F_0 and the kernel, the use of normal or log normal kernels greatly simplifies the computation. In that case, any Markov chain Monte Carlo (MCMC) scheme for DP mixture models can be used. 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 Web Appendix B we show how the existing algorithms can be modified to take into account the presence of censored observations in the data.

The choice of the log normal kernel is equivalent to using a normal kernel based on log transformed data. Alternative kernels such as log logistic or Weibull are possible, but lack of conjugacy would make computations more costly and require the use of less efficient MCMC schemes.

We complete model (7) by specifying a conditionally conjugate base measure:

$$F_0 = (\alpha, \sigma^2) = \text{Ga}(\sigma^{-2} | s/2, sS/2) \text{N}(\alpha | m, B)$$

with conjugate hyperpriors

$$S \sim \text{Ga}(0.5q, 0.5q/R), \quad m \sim \text{N}(a, A) \text{ and } B^{-1} \sim \text{Wish}(c, (cC)^{-1}).$$

The Gamma and Wishart distribution are parametrized such that $E(B^{-1}) = C^{-1}$ and $E(S) = R$. Finally, we assume a Gamma prior for the total mass parameter, $M \sim \text{Ga}(a_0, b_0)$.

Here, $\text{N}(\cdot | m, s)$ denotes a Normal p.d.f. with parameters (m, s) , $\text{Ga}(\cdot | a, b)$ and $\text{Wish}(f, A)$ denote a Gamma p.d.f. with shape a and expectation a/b , and a Wishart distribution with degrees of freedom f and expectation fA respectively. 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. Simulation Study

We simulate data for two groups of hypothetical patients, for example patients in different groups who receive two different doses of a specific treatment. We denote the two doses by 'high' and 'low'. The log survival times $y = \log(t)$ for the low dose group are simulated from $p_1(y) = N(3, 0.8^2)$ while we use a mixture $p_2(y) = 0.4N(4, 1) + 0.6N(2, 0.8^2)$ for the high dose group. The survivor functions for the two groups cross around $y = 3.4$. We have fitted model (7) to the log survival times from the two groups. We define the design vectors d_i as $d_i = (1, \delta)$ with $\delta \in \{-1, 1\}$ for the high and the low dose group, respectively. This implies that $\alpha = [\text{ME}, A(\text{HD})]$ is a bivariate vector, where ME is an overall mean effect and $A(\text{HD})$ is an offset for the high treatment dose. In the analysis we set $q = s = 5$, $c = 7$ and $C = A = 0.004 I_2$, $R = 1$, $a_0 = b_0 = 1$, where I denotes the 2×2 identity matrix.

Figure 1 shows estimated survivor functions for $n = 25, 50, 100, 250, 500$ observations *per group* and $n = 500$ with censoring, comparing estimation under the DDP model versus a proportional hazards (PH) model with time dependent covariate. Since the survivor functions intersect, a proportional hazards assumption would be inappropriate for the treatment variable. When the proportional hazards assumption does not hold and interest focuses on a binary covariate (in this case treatment dose), one traditional approach consists in introducing an indicator function as a time dependent covariate (Klein and Moeschberger, 1997). In the PH regression analysis we have used a categorical variable x for treatment dose with $x = 1$ if a patient receives high dose and 0 otherwise and we have then introduced a second indicator variable z as time dependent covariate defined as follows:

$$z = \begin{cases} x & \text{if } y > y^* \\ 0 & \text{if } y \leq y^* \end{cases}$$

Therefore, the hazard rate for the PH model becomes

$$h(y|x, z) = \begin{cases} h_o(y)\exp(\beta_1 x) & \text{if } y \leq y^* \\ h_o(y)\exp\{(\beta_1 + \beta_2)x\} & \text{if } y > y^* \end{cases}$$

where $h_o(y)$ is the baseline hazard rate. Note that $\exp(\beta_1)$ is the relative risk prior to time y^* for the high dose group relative to the low dose group, while $\exp(\beta_1 + \beta_2)$ is the relative risk after time y^* . Thus $\exp(\beta_2)$ is the increase in relative risk after time y^* , that is sometimes referred to as *change point* for the relative risk (Klein and Moeschberger, 1997).

To fix y^* we proceed as follows. We first fit the model for all values of y^* on a grid and evaluate the log partial likelihood for each possible y^* . Finally y^* is fixed at the value with the largest partial likelihood. The use of a time dependent covariate introduces a jump in the hazard function and a discontinuity in the survivor function at y^* , while the DDP model naturally smoothes over that jump. In the DDP model there is no need for the indicator variable z . The proposed methodology provides data adaptive smoothing. Our method is better at recovering the shape of the true survivor functions, especially for smaller sample sizes.

A possible alternative to the PH model with time dependent covariate is a stratified analysis. First stratify the data by treatment dose. Then carry out, for example, a Cox proportional hazards regression analysis for each stratum. No assumption is needed about how the treatment variable affects survival, i.e. each stratum is permitted to have a different baseline hazard function. The limitation of this approach is the inability to formally examine the effects of

treatment dose. In contrast, the proposed model-based Bayesian inference provides a full probabilistic description of uncertainties and effects. In particular, the model includes inference about any functional of interest of the survivor function (some examples are given in the next Section).

Figure 1f shows inference for simulated data *with censoring*, including around 23% censored observations with independent censoring times. We used the same model as before, except that we fix $S = 1.0$, at the prior mean. In general, we recommend to fix S for data with a large proportion of censoring. Without fixing S posterior MCMC simulations mix badly. The chain gets trapped in a state with large clusters of very similar imputed event times, and very small σ_i^2 . Figure 2 shows the estimated probability densities for $n = 100$ under the DDP model.

6. 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 summarized 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, or censoring. 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, consequently justifying more aggressive therapy. Thus the primary reason for carrying out the clinical trial was to compare low versus high dose. We consider two categorical covariates, one continuous covariate, and one interaction term: treatment dose ($-1 = \text{low}$, $1 = \text{high}$), estrogen receptor (ER) status ($-1 = \text{negative}$, $1 = \text{positive or unknown}$), the size of the tumour (standardized to zero mean and unit variance), and a treatment/ER interaction (1 if a patient receives high treatment dose and has positive ER status and 0 otherwise).

We performed a Bayesian analysis of these data using the model specified in (7) with a normal kernel. First we carried out inference with one covariate, treatment dose, only. Figure 3(a) shows the posterior estimates, $E(S | \text{Data})$, of the survivor functions under the ANOVA DDP model. For comparison panel (b) shows the data as a Kaplan-Meier curve. The survivor functions corresponding to the two treatment groups cross, 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. In the case of the Kaplan-Meier method a separate curve needs to be fitted for each treatment level, not allowing any borrowing of information. Figure 3(c) shows the estimated survival function under a proportional hazards model with time dependent covariate as described in the previous Section, while panel (d) shows the estimates obtained fitting a proportional hazards model with time-varying coefficients (see, for example, Therneau and Grambsch (2000) chapter 6), which assumes that the effect of a covariate changes over time. In this case, the covariate itself is fixed but its effect is allowed to vary through time. Note that in Figure 3(d) the two curves are basically indistinguishable for the first few months. Simple model diagnostics, such as Cox-Snell residuals and standard residuals (Hanson et al., 2008), could be used to better assess the performance of the different models. Moreover, a measure such as the Log Pseudomarginal Likelihood (LPML) measure (Geisser and Eddy, 1979) could be employed to compare different models. LPML is based on the "leave-one-out" principle used in procedures such as cross validation and jackknife and leads to a pseudo Bayes factor that can be interpreted as traditional Bayes factors.

Next we augmented the probability model by including additional covariates, ER status, tumour size and treatment/ER-status interaction. The latent variable vector α in eq. (7) is a five-

dimensional vector, $\alpha = [ME, A(HD), A(ER), A(TS), A(DOSE \cdot ER)]$, where ME corresponds to an overall mean effect, while $A(HD)$, $A(ER)$ and $A(TS)$ are the offsets for high treatment dose, positive ER status and the linear coefficient for tumor size, respectively and $A(DOSE \cdot ER)$ is the offset for the dose/ER interaction term. The design matrix could be extended to include other interactions, nested effects, etc., as desired. We implemented the algorithm described in Web Appendix B to obtain posterior simulations for these data. In the analysis, we set $q = s = 4$, $c = 10$ and $C = A = \text{diag}(40, 40, 40, 40, 400)$, $R = 10000$, $a_0 = b_0 = 1$.

Figures 4(a) and (b) show the posterior estimates, $E(S | \text{Data})$ and $E(h | \text{Data})$, and corresponding posterior uncertainty of the survivor and hazard functions for ER positive patients with tumor size 2.0 cm (equal to the first quartile). Also in this subpopulation the survivor functions corresponding to the two treatment groups cross, showing a higher level of risk associated with high treatment dose in the first 20 months.

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, Figure 4(c) shows the difference in survival rates between the two treatment groups for positive ER status and tumour size equal 2.0 cm, across a range of times.

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 with weight set equal to 1. Figure 5a 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 (Figure 5b) presents greater flexibility allowing for curves with different shapes for the two groups.

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.

The programs are available as a function in the R package `ddpanova` at <http://odin.mdacc.tmc.edu/~pm/prog.html>. The function `ddpsurvival(.)` implements the proposed DDP survival regression model.

Supplementary Materials

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported, in part, by grant CA075981 from the U.S. National Cancer Institute. The authors thank two anonymous referees for very useful comments and suggestions that improved the presentation of the paper.

References

- Christensen R, Johnson W. Modelling accelerated failure time with a Dirichlet process. *Biometrika* 1988;75:693–704.
- De Iorio M, Müller P, Rosner GL, MacEachern SN. An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association* 2004;99:205–215.
- Doss H. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics* 1994;22:1763–1786.
- Doss, H.; Huffer, F. Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet priors. Technical report; Department of Statistics, Ohio State University. 1998.
- Doss, H.; Narasimhan, B. Dynamic display of changing posterior in Bayesian survival analysis. In: Dey, D.; Müller, P.; Sinha, D., editors. *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York, NY, USA: 1998. p. 63–88.
- Dunson DB, Park J-H. Bayesian density regression. *Biometrika*. 2007in press
- Dunson DB, Pillai NS, Park JH. Bayesian density regression. *Journal of the Royal Statistical Society, Series B (Methodological)* 2007;69:163–183.
- Escobar MD. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* 1994;89:268–277.
- Escobar, MD.; West, M. Computing nonparametric hierarchical models. In: Dey, D.; Müller, P.; Sinha, D., editors. *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York, NY, USA: 1998. p. 1–22.
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973;1:209–230.
- Ferguson TS. Prior distributions on spaces of probability measures. *Annals of Statistics* 1974;2:615–629.
- Ferguson TS, Phadia EG. Bayesian nonparametric estimation based on censored data. *Annals of Statistics* 1979;7:163–186.
- Geisser S, Eddy WF. A predictive approach to model selection. *Journal of the American Statistical Association* 1979;74:153–160.
- Gelfand AE, Kottas A. Bayesian Semiparametric Regression for Median Residual Life. *Scandinavian Journal of Statistics* 2003;30:651–665.
- Griffin JE, Steel MFJ. Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association* 2006;101:179–194.
- Hanson T, Johnson WO. Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association* 2002;97:1020–1033.
- Hanson T, Johnson WO, Laud P. A unified approach to semiparametric inference for survival models with time dependent covariates. *Canadian Journal of Statistics*. 2008in revision
- Ibrahim, JG.; Chen, M-H.; Sinha, D. *Bayesian Survival Analysis*. Springer- Verlag; New York, NY, USA: 2001.
- Jain S, Neal RM. A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics* 2004;13:158–182.
- Johnson W, Christensen R. Bayesian nonparametric survival analysis for grouped data. *Canadian Journal of Statistics* 1986;14:307–314.
- Kalbfleisch J. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* 1978;40:214–241.
- Klein, J.; Moeschberger, M. *Techniques for Censored and Truncated Data*. Springer-Verlag; New York, NY, USA: 1997. *Survival Analysis*.
- Kuo L, Mallick B. Bayesian semiparametric inference for the accelerated failure- time model. *Canadian Journal of Statistics* 1997;25:457–472.

- Kuo, L.; Smith, AFM. Bayesian computations in survival models via the Gibbs sampler. In: Klein, JP.; Goel, PK., editors. *Survival Analysis: State of the Art*. Boston: Kluwer Academic; 1992. p. 11-24.
- Lavine M. Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* 1992;20:1222–1235.
- Lo A. On a class of bayesian nonparametric estimates: I. density estimates. *Annals of Statistics* 1984;12:351–357.
- MacEachern S, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* 1998;7:223–239.
- MacEachern, SN. *ASA Proceedings of the Section on bayesian Statistical Science*. Alexandria, VA: American Statistical Association; 1999. Dependent nonparametric processes.
- Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000;9
- Rosner GL. Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics* 2005;61:239–245. [PubMed: 15737099]
- Sethuraman J. A constructive definition of the Dirichlet process prior. *Statistica Sinica* 1994;2:639–650.
- Susarla V, Ryzin JV. Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* 1976;71:897–902.
- Therneau, TM.; Grambsch, PM. *Modeling Survival Data*. Springer-Verlag; New York, NY, USA: 2000.
- Tokdar ST. Posterior consistency of dirichlet location-scale mixture of normals in density estimation and regression. *Sankhya-: The Indian Journal of Statistics* 2006;68:90–110.
- Walker S, Mallick BK. A Bayesian semiparametric accelerated failure time model. *Biometrics* 1999;55:477–483. [PubMed: 11318203]

Figure 1.

Simulation example. Estimated survival functions under the ANOVA DDP model (thick solid lines) for sample sizes $n = 25$ (panel a) through 500 (panel e) without censoring, and for $n = 500$ with 23% censoring (panel f). The initially lower curve is the “high dose group.” For comparison the panels also show inference under a proportional hazards rate model with time dependent covariate (thick dashed lines) and the simulation truth (dotted curves).

Figure 2. Simulation example, with $n = 100$. Predictive densities under the two groups (panels (a) and (b)), and estimated survival distribution with pointwise posterior intervals (panel (c)).

Figure 3.

Cancer clinical trial. (a) Estimated survivor functions for high versus low dose under the ANOVA DDP model. (b) Kaplan-Meier curves. (c) Estimated survival functions under a proportional hazards model with time dependent covariate. (d) Estimated survival function under a proportional hazards model with time-varying coefficients. The solid line refers to the low dose group, while the dashed lines refers to the high dose group.

Figure 4.

Cancer clinical trial. In the three panels the results are displayed for positive ER status and tumour size equal to 2.0cm (first quartile). (a) Estimated survivor function for high versus low dose. The grey shaded bands show point-wise central 50% credible intervals. (b) Estimated hazard function for high versus low dose. The grey shaded bands show point-wise central 50% credible intervals. (c) Posterior distribution of the difference in survival probabilities at 10, 20, 40, 60, 80 and 100 months between a patient that receives high treatment dose versus a patient that receives the low dose. Note how the difference changes sign from 20 to 40 months.

Figure 5.

Cancer clinical trial. Posterior survivor functions using the AFT median regression model (panel a) and using the ANOVA DDP model (panel b). The curves are estimated for tumour size 2.0 cm (first quartile). In both plots, 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. Note the almost vanishing difference between the solid and the dashed line. Panel (c) shows the data as a KM plot arranged by dose and ER status.

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