ELSEVIER

Journal of Statistical Planning and Inference 111 (2003) 101–115

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Mixtures of products of Dirichlet processes for variable selection in survival analysis

Paolo Giudici[a], *, Maura Mezzetti[b], Pietro Muliere[b]

[a]*Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia, Via San Felice 5, I-27100 Pavia, Italy*
[b]*Istituto di Metodi Quantitativi, L. Bocconi University, Milan, Italy*

Abstract

A very important problem in survival analysis is the accurate selection of the relevant prognostic explanatory variables. We propose a novel approach, based on mixtures of products of Dirichlet process priors, that provides a formal inferential tool to compare the explanatory power of each covariate, in terms of the marginal likelihood attached to the induced partitions of the observations. Our proposed model is Bayesian nonparametric, and, thus, keeps the amount of model specification to a minimum, increasing robustness of the final inferences.

Keywords: Bayesian nonparametrics; Dirichlet processes; Survival analysis; Variable selection

1. Background and motivation

Consider a collection of n survival times, possibly censored, let T_i be a random variable representing the failure time of subject i, and c_i a censoring time. For each subject we observe $X_i = \min(T_i, c_i)$, and an indicator variable $\delta_i = I_{\{T_i \leqslant X_i\}}$, indicating weather the ith subject has a censored event or an event of interest. Let F_i denote the cumulative distribution function of the ith individual. For each individual we have a set of observed covariates, describing possible prognostic effects on survival times. Our motivation is to understand whether such covariates do affect the observed survival times, so as to construct a predictive model for future occurrences.

We propose an approach that provides a formal inferential tool to compare the explanatory power of each covariate, and, therefore, to select a good model for predictive

* Corresponding author. Tel.: +39-382-386224; fax: +39-382-304226.
E-mail address: giudici@unipv.it (P. Giudici).

0378-3758/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(02)00291-4

purposes. Our proposed model is Bayesian nonparametric, and, thus, keeps the amount of model specification to a minimum.

We first consider the case in which information from the covariates is at the nominal level and, therefore, the covariates are potential prognostic factors. For each given factor, we partition the individuals in as many groups as the number of observed levels of the factor. We then assume individual survival times to be homogeneous within each group and heterogeneous across groups.

Our aim is to compare the partition structures resulting from consideration of the different explanatory factors. The metric we choose for the comparison is the calculation of the marginal likelihood of each partition.

Once the groupings are accomplished, there are two important assumption that the researcher has to make, conditionally on the assumed partition. The first one concerns the dependence structure between the individual observations y_i *within* the same group. The second one regards the dependence structure *between* observations for individuals belonging to different groups.

We believe that a natural and simple modeling assumption is that, within each group, observations are considered *exchangeable*. This implies that the observations as a whole follow the *partial exchangeable* scheme proposed by De Finetti (1938).

Let g be the partition, and k the number of groups. We consider a hierarchical nonparametric approach. More specifically, we assign the distribution of the random vector $(F_1,\dots,F_k)$, assuming that $F_1,\dots,F_k$ are conditionally independent given a vector of parameters $\theta=(\theta_1,\dots,\theta_k)$ with

$$(F_i|\theta)\sim\mathscr{D}(\alpha(\theta_i)),$$

where $\mathscr{D}(\alpha(\theta_i))$ is a Dirichlet process with parameter $\alpha(\theta_i)$ (see Ferguson, 1973).

Furthermore, the parametric vector θ is taken to be a random vector with distribution function H, so that

$$(F_1,\dots,F_k)\sim\int_{\mathscr{R}^k}\prod_{i}^{k}\mathscr{D}(\alpha(\theta_i))H(\mathrm{d}\theta). \tag{1}$$

The resulting process is precisely a *mixture of products of Dirichlet processes* (MPDP), as introduced in the literature by Cifarelli and Regazzini (1978).

Some applications of MPDP processes are: Cifarelli (1979), Cifarelli et al. (1981), Consonni (1981), Muliere and Scarsini (1983), Muliere and Petrone (1993), Mira and Petrone (1996), Carota and Parmigiani (2000).

The methodology can be generalized to take into account continuous covariates. Following Cifarelli et al. (1981), the prior structure allows inclusion of continuous covariates. Comparison between continuous covariates can be addressed by looking at the marginal likelihood of each exchangeability structure determined by the design matrix corresponding to the covariate. More specifically, having observed, for each individual, besides x_i, a p-dimensional vector covariates $\underline{z_i}$, possibly continuous, we shall assume that

$$(F_i|\beta)\sim\mathscr{D}(\alpha(\underline{z_i}'\beta)),$$

where β is a p-dimensional vector of real-valued random parameters ($p \leqslant n$), with distribution function G, so that

$$(F_1,\ldots,F_n) \sim \int_{\Re^p} \prod_i^n \mathscr{D}(\alpha(\underline{z_i}'\beta))G(\mathrm{d}\beta).$$

In this paper we shall mainly refer to the model in (1).

Note that the partition g can be induced by the combination of the levels of more than one explanatory variable, allowing the evaluation of different prognostic effects at the same time as well as the consideration of confounding effects.

We also remark that the approach underlying our method is also related to that pervading survival trees models (see for instance, Mallick et al., 1999), which also leads to the identification of an optimal partition structure, starting from a different, recursive, modelization of the survival function.

The performance of our methodology will be illustrated by means of two well-known data sets. The first one is Veteran's cancer data set from Prentice (1973), the second one is the Mice data set from Grieve (1987). The former has been modeled by means of simple exponential failure time models and, thus, is well suited to illustrate our methodology, mostly by means of exact computations. The latter has often been considered in the Markov Chain Monte Carlo (MCMC) literature, to illustrate applications to Bayesian parametric survival analysis (see, e.g. Dellaportas and Smith, 1993). We shall demonstrate that our methodology can be extended to this context, and compare the results obtained with our approach with the parametric Bayesian analysis, using MCMC methods.

The paper is organized as follows: in Section 2 we present and discuss our proposed methodology; in Section 3 we consider the exponential and Weibull regression models and apply them to the two considered data sets; finally, Section 4 contains some further remarks and discussion.

2. The method

In order to investigate the possible dependencies among the observations, a well-known strategy in survival analysis is to build up a *causal* model which relates the survival times to a proper collection of covariates, say $\underline{Z} = (Z_1,\ldots,Z_p)$, whose realizations $\underline{z}_i$, for $i = 1,\ldots,n$, are known. The most frequently employed of such models is Cox's proportional hazard model.

However, in situations in which a very large number of potential explanatory variables are available, it is important to pre-screen a subset of actual explanatory variables. Otherwise, if a selection procedure is run on the whole data set, severe instability problems of the results may occur. See for instance Altman and Andersen (1989) and Sauerbrei and Schumacher (1992).

Our approach provides a formal inferential tool to tackle the above problem. In particular, we shall consider *mixtures of products of Dirichlet processes*. This allows to evaluate, in a simple and exploratory fashion, the relative importance of each potential prognostic factor.

Consider a generic explanatory covariate, say Z_l. Such a covariate partitions the data in k groups, according to its observed levels. We assume that, in each partition, observations can be deemed exchangeable, according to the partial exchangeability scheme proposed by de Finetti (1938). In other terms, each Z_l corresponds to a sequence $\{X_{i,j}^l,\ i=1,\dots,k;\ j=1,2,\dots,n_k\}$ of partially exchangeable random variables. We remark that the partial exchangeability scheme specifies the type of dependence between observations within each population group X_i^l, but does not specify anything on the nature of the stochastic dependence between observations belonging to different groups.

To ease the notation, suppose one covariate is implicitly fixed, and drop the corresponding index. Let $\underline{X}_i=(X_{i1},\dots,X_{in_i})$, for $i=1,2,\dots,k$ be the n_i observations of the ith population group which are assumed to be distributed according to an unknown cumulative distribution function (cdf) $F_i(x)$.

Assume that, conditionally on the k cdfs $F_1,\dots,F_k$, independence between observations in different groups holds, namely

$$P(\underline{X}_1 \leqslant \underline{x}_1,\dots,\underline{X}_k \leqslant \underline{x}_k|F_1,\dots,F_k)=\prod_{j=1}^{n_1}F_1(x_{1,j})\cdots\prod_{j=1}^{n_k}F_k(x_{k,j}).$$

We remark that, as the cdf $F_1,\dots,F_k$ are random quantities, the group observations $\underline{X}_i$ are not independent, even conditionally to the knowledge of the allocation into the groups. We now have to specify a distributional mechanism for the unknown cumulative distribution functions.

First of all, we assume that each cdf is distributed according to a Dirichlet process, with base measure $\alpha(u_i,\cdot)$, $i=1,\dots,k$, with u_i an unknown parameter. In order to facilitate comparisons with parametric models commonly employed in survival analysis we shall assume that

$$\alpha(u_i,\cdot)=\alpha(u_i,\mathfrak{R})\Phi(u_i,\cdot),$$

where Φ is the cdf of a distribution of a known form up to an unknown parameter u_i. Furthermore, for parsimony, and without any loss of generality, in the following we shall take $\alpha(u_i,\mathfrak{R})=M$, $i=1,\dots,k$. Note that $\Phi(\mu_i,\cdot)$ can be interpreted as a prior guess on F_i and M as a "measure of faith" in such guess.

The main assumption we make is to consider $F_1,\dots,F_k$ as random quantities drawn from a *mixture of products of Dirichlet processes*. This means to assume that, conditionally on $\underline{u}=(u_1,\dots,u_k)$, $(F_1,\dots,F_k)$ is a product of Dirichlet processes, with base measures $(\alpha(u_1,\cdot),\dots,\alpha(\mu_k,\cdot))$. In other words, observations are taken to be independent between different groups, conditionally on a random vector $\underline{u}$, which will determine the degree of dependency.

Finally, concerning the parameters $\underline{u}$, we assume that

$$u_1,\dots,u_k\sim\pi(\cdot),$$

where π is a suitable multidimensional prior distribution. Note that π need not assume the u_i to be independent.

Our task is to calculate the marginal distribution of the observations $\underline{X}=(X_1,\dots,X_k)$, namely the marginal likelihood of the considered partition. Suppose first that, for each individual, we simply observe $(x_{ij}, \delta_{ij} = 1)$, namely, no censoring is present for the time being.

Consider first the contribution of the observations belonging to group i. It turns out that

$$\begin{aligned} P(\underline{X_i} \leqslant \underline{x_i} | u_i) &= \int P(\underline{X_i} \leqslant \underline{x_i} | u_i, F_i)\,\mathrm{d}\mathscr{P}(F_i|u_i) \\ &= E\left\{ \prod_{j=1}^{n_i} F_i(x_{i,j}) | u_i \right\} \\ &= \prod_{j=1}^{n_i} \frac{M\Phi(x_{i(j)}, u_i) + j - 1}{M + j - 1}, \end{aligned} \tag{2}$$

where $x_{i(1)}, \ldots, x_{i(n_i)}$ indicates the sequence of the realized observations of the ith group, $x_{i,1}, \ldots, x_{i,n_i}$, in nondecreasing order.

From (2) it follows that

$$P(\underline{X_i} \leqslant \underline{x_i}) = \int_{\mathfrak{R}} \left\{ \prod_{j=1}^{n_i} \frac{M\Phi(x_{i(j)}, u_i) + j - 1}{M + j - 1} \right\} \mathrm{d}\Pi(u_i),$$

and, when the whole vector of observations is considered:

$$P(\underline{X}_1 \leqslant \underline{x}_1, \ldots, \underline{X}_k \leqslant \underline{x}_k) = \int_{\mathfrak{R}^k} \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{M\Phi(x_{i,(j)}, u_i) + j - 1}{M + j - 1}\,\mathrm{d}\Pi(u_1, \ldots .u_k).$$

In order to compare alternative partitions, as induced by the available covariates, we need to derive the above expression to obtain the marginal likelihood. This can be done following the procedure illustrated in Antoniak (1973); see also Petrone and Raftery (1997).

Suppose that, among the n_i observations in the ith partition group, the number of distinct observations is equal to r_i, organized in nondecreasing order, as follows: $x^*_{i(1)}, x^*_{i(2)}, \ldots, x^*_{i(r_i)}$, with each of them repeated, respectively: $n_{i1}, n_{i2}, \ldots, n_{ir_i}$ times, with $\sum_{j=1}^{r_i} n_{ij} = n_i$.

Fact. *The (conditional) likelihood of the observations in the ith group is then equal to*

$$f_i(\underline{x_i}|u_i) = \frac{M^{r_i}}{M^{n_i}} \prod_{j=1}^{r_i} (n_{i(j)} - 1)! f_0(x^*_{i(j)}).$$

Note that the previous expression can be factorized in two components. The first component is the term $(M^{r_i}/M^{n_i}) \prod_{j=1}^{r_i} (n_{i(j)} - 1)!$ which is the probability that the data follow a specific pattern of distinct and repeated values in each group, with a given ordering. The second component $\prod_{j=1}^{r_i} f_0(x_{i(j)}^*)$ is the joint density (likelihood) corresponding to the distinct values.

Now, according to our assumptions, the likelihood of g, conditional on $u=(u_1,\dots,u_k)$, can be expressed as the product of k independent terms as the previous one:

$$P(\underline{X_1}=\underline{x_1},\dots,\underline{X_k}=\underline{x_k}|u)=\prod_{i=1}^{k} f_i(\underline{x_i}|u_i).$$

In order to obtain the marginal likelihood of the observations in each group, we need to integrate the previous expression with respect to the prior distribution on $\underline{u}$, so that the marginal likelihood of g, denoted with $L(g)$, is

$$L(g)=\int_{\mathfrak{R}^k} P(\underline{X_1}=\underline{x_1},\dots,\underline{X_k}=\underline{x_k}|u)\Pi(\mathrm{d}u),$$

where π is an appropriate prior distribution for u.

Consider now the more realistic situation in which data are censored, namely let $X_{ij}=(x_{ij},\delta_{ij})$. Furthermore, let $U=\{(i,j):\delta_{ij}=1\}$ be the uncensored subjects and $C=\bar{U}$ the censored subjects.

We shall assume that, when tied observations contain *both* censored and uncensored cases, $\delta_{ij}=\max(\delta_{ij}:x_{ij}=x_{ij}^*)=1$; in other words, the repeated observation is assumed to be uncensored. When censoring is considered, the marginal likelihood of a partition g turns out to be equal to

$$L(g)=\int_{u_1}\cdots\int_{u_k}\prod_{i=1}^{k}\frac{M^{n_i^*}}{M^{[n_i]}}\prod_{j=1}^{n_i^*}[I_{\mathrm{U}}f_{u_i}(x_{ij}^*)+I_{\mathrm{C}}(1-F_{u_i}(x_{ij}^*))]\pi(u_1,\dots,u_k)\,\mathrm{d}u, \quad (3)$$

where I_{C} and I_{U} are indicator functions, respectively, for censored and uncensored subjects and π is a suitable prior distribution for $u=u_1,\dots,u_k$.

Remark 1. The above integral is, apart from simple models and prior distributions, generally intractable. We need to approximate $L(g)$ with

$$\frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{k}\frac{M^{n_i^*}}{M^{[n_i]}}\prod_{j=1}^{n_i^*}[I_{\mathrm{U}}f_{u_i^r}(x_{ij}^*)+I_{\mathrm{C}}(1-F_{u_i^r}(x_{ij}^*))]\prod_{j=1}^{n_i^*}(n_{i(j)}-1)!,$$

with R the number of draws of the k-dimensional random vector $\underline{u}$. Such vector is distributed according to Π, which may depend on a parameter vector τ. We need to choose a suitable grid of prior values for M (overall weight of the prior) and τ. In Section 3 we shall give examples, based, respectively, on static Monte Carlo and on MCMC methods.

Remark 2. In our exploratory approach, we shall compare a number of partition structures equal to the number of available covariates, and evaluate their relative importance by means of the score of each partition model. Therefore, we remark that the number of considered partitions is *not* random, but fixed in advance. We choose to report, as a model score, the marginal likelihood of each partition. More specifically, for each partition g we calculate, $D = -\log p(\underline{X}|g)$, namely the negative marginal loglikelihood of each model.

Remark 3. Note that the degree of dependence between observations in different groups is governed by the prior $\pi(u)$. The previous specification leaves the u_i independently distributed, which amount to independence between observations in different groups. Alternatively, a hierarchical prior can be taken, for instance, taking m_0 to be random, so to model flexibly the dependence.

Remark 4. As discussed in the previous Section, a more general form of dependence can be induced by means of a linear model. Let $F_i|u_i \sim \mathscr{D}(\alpha(u_i))$, where $\alpha(u_i) = M * \Phi(u_i, \cdot)$. We now assume that the u_i's can be expressed as linear functions of $p \leqslant k$ unknown, but common between groups, parameters, such as

$$u_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

with β_i a collection of random coefficients and the x_i's functions of the covariates. For our exploratory purposes, we will be typically interested in considering a simple linear model with $p = 1$.

A simple prior on the random coefficients β can be deduced through the prior on the u, specified as previously discussed. The main substantial advantage in employing a linear model formulation lies in the possibility to take a fairer account of the explanatory power of quantitative explanatory variables through a more parsimonious linear formulation of the prognostic effect. Indeed, the previous model is a special case of this when $p = k$ and each x_{ij} corresponds to an indicator variable for the jth group of the considered partition.

3. Application to failure time models

In this section we first assume that $\Phi(u_i, \cdot)$ are exponential lifetime distributions, so that

$$\alpha(u_i, \cdot) = M * (1 - \mathrm{e}^{-u_i x_i}).$$

This simple modelization will allow us to obtain analytical results, and, therefore, better illustrate our proposed methodology.

From (3) it can be shown that the marginal likelihood of the observed times, conditionally on a known partition g, is

$$\int_{u_1} \cdots \int_{u_k} \prod_{i=1}^{k} \frac{M^{n_i^*}}{M^{[n_i]}} \prod_{j=1}^{n_i^*} (n_{i(j)} - 1)![I_{\mathrm{U}}(u_i * \mathrm{e}^{-u_i x_{i(j)}}) + I_{\mathrm{C}}(\mathrm{e}^{-u_i x_{i(j)}})]\pi(u)\,\mathrm{d}u,$$

where $M^{[n_i]} = M(M+1)\cdots(M+n_i-1)$ and n_i^* is the number of distinct observations in each group.

The above integral can be solved analytically if the prior distribution $\pi(u)$ is taken of a simple form. For instance, take the u_i to be i.i.d gamma$(r_0 m_0, r_0)$. Then the marginal likelihood of g is equal to

$$L(g) = \prod_{i=1}^{k} \frac{M^{n_i^*}}{M^{[n_i]}} \frac{r^{rm}}{\Gamma(rm)} \frac{\Gamma(rm + d_i^*)}{(V_i^* + r)^{rm+d_i^*}} \prod_{j=1}^{n_i^*} (n_{i(j)} - 1)!,$$

where $d_i = \sum_{j=1}^{n_i^*} \delta_{ij}^*$ is the number of distinct observed events (deaths) in group i and $V_i = \sum_{j=1}^{n_i} x_{ij}^*$ is the total time at risk in each group, considering only distinct events.

Consider the application of our proposed methodology to the Veteran's data set discussed in Prentice (1973). In order to compare our results with the classical ones, we take a product of independent gamma priors for $\underline{u}$, each with prior expectation (m_0) equal to 1 and prior variance (m_0/r_0) also equal to 1, corresponding to one prior event. We remark that, when an independent prior for $\underline{u}$ is taken, the most important differential factor between the parametric and the nonparametric approach is the ratio $M^{n_i^*}/M^{[n_i]}$, which appears, for each group, only in the nonparametric case. As a consequence, as $M \to \infty$, the nonparametric results get closer and closer to those obtained with the parametric approach. Furthermore, the independence partition will be scored equivalently by both approaches.

In the data set six explanatory covariables are considered as potential prognostic effects for the survival times: performance status (perf), months from diagnosis (diag), age, prior therapy (ther), cell type (cell) and treatment (trt). Prentice (1973) and Raftery et al. (1996), under, respectively, a non-Bayesian and a Bayesian parametric approach, found that the two relevant explanatory variables are cell type and performance status. These results are also confirmed by the Bayesian parametric analysis of Giudici (1996).

We now apply our approach by computing the score D of the six partitions induced by the levels of the available covariates, and choosing those is relevant by means of a comparison between the obtained partition model scores. We also include the partitions corresponding to complete exchangeability and independence of the survival times, as useful comparison benchmarks.

Table 1a gives the model scores D of the partitions, for $M = 10, 100$ and 1000. We also report results from a parametric model. M is a measure of "confidence" on the parametric model. We remark that there is no general guideline on the choice of M, it does depend subjectively on the problem at hand, as well as on the amount of prior information available. We retain extremely important, when prior information is weak or absent, to carry out a sensitivity analysis of the results with respect to the choice of M. This allows to evaluate the effect of alternative choices of M explicitly. An alternative approach would be to take a prior on M as well (see, e.g., Escobar and West, 1995).

From Table 1a, note that the marginal likelihoods are sensible to M, which suggest that taking a parametric model is a strong assumption. We remark that in Table 1a we have adopted an ANOVA-like model, where the partitions corresponding to variables, age and month from the diagnosis are the most complicated ones. However, low values of M lead to a strong weight on the empirical cdf, and, therefore, more complex models are less penalised. On the other hand, as M increases, parsimonious partitions

Table 1
Model scores, $-\log p(\underline{X}|g)$, associated to the entertained partitions, for different values of M and an exponential failure model

Partition	$k(g)$	$M = 10$	$M = 100$	$M = 1000$	*Param*
(a) *Independent priors*					
g_{exc}	1	944.90	838.1	862.84	757.60
g_{trt}	2	921.06	827.9	847.07	763.15
g_{ther}	2	930.30	831.34	846.41	762.31
g_{cell}	4	880.35	803.15	811.52	756.59
g_{perf}	12	851.06	795.41	796.67	772.81
g_{diag}	28	895.52	860.01	860.98	841.52
g_{age}	40	899.21	877.19	874.54	874.54
g_{ind}	137	1091.77	1091.77	1091.77	1091.77
(b) *Dependent priors*					
g_{exc}	1	838.01	944.50	863.20	
g_{trt}	2	826.99	921.11	846.95	
g_{ther}	2	829.73	929.27	847.73	
g_{cell}	4	804.57	879.81	809.86	
g_{perf}	12	793.1	842.07	797.91	
g_{diag}	28	860.54	885.76	862.77	
g_{age}	40	866.73	873.85	856.64	
g_{ind}	137	912.92	1082.35	1019.64	
(c) *Linearized priors*					
g_{exc}	1	945.04	837.63	862.61	
g_{trt}	2	920.60	826.99	846.53	
g_{ther}	2	930.84	833.23	845.63	
g_{cell}	4	878.18	801.55	812.87	
g_{perf}	12	850.17	805.13	795.91	
g_{diag}	28	892.29	836.8	854.46	
g_{age}	40	886.73	864.12	867.90	
g_{ind}	137	1006.42	1054.51	999.24	

are more supported, especially when they correspond to within-groups homogeneity of the survival times, as is the case for cell type, therapy and treatment.

Concerning the search for prognostic variables, the parametric model seems to capture the importance of cell type, but not that of performance status, which instead emerge in the nonparametric models. Recall that both Prentice (1973) and Raftery et al. (1996), assuming a more parsimonious linear modeling formulation, select *performance status* and *cell* as the most relevant prognostic variables.

We also remark that, apart from very low values of M, the two partitions which are mostly supported are always *performance status* and *cell*. This suggests robustness of the model score with respect to M. On the other hand, we believe that a small value of M means doing no modeling at all, as one would do better in considering only the empirical cdf.

We now introduce a dependent prior for $\underline{u}$, and, consequently, make observations in different groups become marginally dependent. A simple way to induce such dependence is by means of a hierarchical prior. For instance, let $u_i \sim \gamma(r_0 m_0, r_0)$ independently; fix r_0 at 1, as previously done, but let the prior expected hazard m_0 become a random quantity, such as $\gamma(1,1)$. The marginal likelihood is then evaluated by means of static Monte Carlo. At each iteration step each u_i is drawn from a $\gamma(1,1)$ distribution, and the marginal likelihood is evaluated in the corresponding realizations. Finally, we calculate the mean marginal likelihood over the considered iterations. Table 1b reports the corresponding results, with the nonparametric model, $M = 1, 1000$ and a simulation size of $I = 100$ iterations.

Comparing Tables 1a and b, we essentially obtain, as most relevant, the same two partitions as before. However, we believe that the higher complexity of the hierarchical model requires the researcher to assign coherently a greater value for M. As a consequence, the value of M required to reach a stable selection is higher in Table 1b.

Finally, consider the stronger type of dependence induced by a linear model. In order to specify the prior model, we take a simple linear regression model, namely, for a covariate Z_j, $x_{ij} = \beta z_{ij}$, so that the prior hazard of the ith group becomes $u_i = \beta \sum_{j=1}^{n_i} z_{ij}$. In analogy with the previous prior model, we then take μ_i as a $\Gamma(1,1)$ random variable. This induces a prior on β which is a gamma random variable, with parameters $(1, 1/\sum_{j=1}^{n_i} z_{ij}))$ so that in order to obtain the marginal likelihood we need to perform a static Monte Carlo simulation, as previously.

Table 1c reports the corresponding results, with the nonparametric model, $M = 10$, 1000 and a simulation size of $I = 100$ iterations.

From Table 1c, note that the results are essentially intermediate between Table 1a and b, as the complexity of the hierarchical model is also in between. In any case, as M is sufficiently large, we obtain, as relevant covariates, always performance status and cell type.

In particular, note that, differently from what happens in Table 1a, in the regression-like linear model in Table 1c continuous variables have only one parameter and, therefore, age and months from diagnosis are less penalized.

We shall now apply our methodology to a more complex situation, where the group-specific hazard function is described by a Weibull regression model. Furthermore, different from what is done in the exponential linear model considered in the end of the previous section, the prior distribution is assigned *directly* on the regression coefficients. Clearly, such a prior is easier to specify, but makes calculations more complex.

Assume that $\Phi(u_i, \cdot)$ are Weibull lifetime distributions, so that

$$\alpha(u_i, \cdot) = M * (1 - \mathrm{e}^{-u_i x_i^r})$$

with $r > 0$. Obviously, for $r = 1$ we obtain the exponential failure time model. Furthermore, let

$$\log(u_i) = \underline{z}_i' \underline{\beta} = \beta_1 z_{i1} + \cdots + \beta_p z_{ip} \tag{4}$$

with each β_i a random coefficients and each z_{ij} the observed realization of a known covariate for each individual.

Indeed, for the data at hand, we shall consider a simplified version of (4), with $p=4$, and each covariate corresponding to the indicator function for one of four treatment groups.

As a prior distribution on $\underline{\beta}$ we shall take β_i i.i.d. $N(0, \tau)$, with $\tau > 0$.

It can be shown that the marginal likelihood of g is

$$\int_{u_1} \cdots \int_{u_k} \prod_{i=1}^{k} \frac{M^{n_i^*}}{M^{[n_i]}} \prod_{j=1}^{n_i^*} (n_{i(j)} - 1)!(r \exp\{\underline{z}_i' \underline{\beta}\})^{d_i^*}$$

$$\times \prod_{j=1}^{n_i^*} (x_{i(j)})^{r-1} (\mathrm{e}^{-\exp\{\underline{z}_i' \underline{\beta}\} \sum_{j=1}^{n_i^*} x_{i(j)}^r}) \pi(u) \,\mathrm{d}u, \tag{5}$$

where n_i^* are the number of distinct observations in each group, $M^{[n_i]} = M(M + 1) \cdots (M + n_i - 1)$; n_i^* is the number of distinct observations and $d_i = \sum_{j=1}^{n_i^*} \delta_{ij}^*$ is the number of distinct observed events (deaths) in each group i. The above integral cannot be solved analytically and, therefore, we will employ MCMC methods to approximate it.

To illustrate our methodology, we shall now consider a data set which is used in the (parametric) Bayesian literature to illustrate how MCMC methods can be employed to analyze complex parametric survival models, such as Weibull regression ones. We shall apply our methodology to such data set and compare our results with the parametric ones.

The data set is described in Dellaportas and Smith (1993), who analyze mice data from Grieve (1987) on photocarcinogenicity in four groups, each containing 20 mice, for all of which survival time in weeks and censoring are recorded. An objective of interest is to evaluate whether there is an effect of the covariate treatment against the no-effect situation, corresponding to complete exchangeability.

In order to approximate the marginal likelihood of interest, we will consider, in all cases, and for the sake of stable results, MCMC simulations length of $n = 10{,}000$ plus $n = 1000$ of burn-in.

We first compare the performance of an exponential and a Weibull regression model. The former is obtained taking $r = 1$. Take $\tau = 0.0001$, and compare the results for the parametric model with the nonparametric model with $M = 10$, 100, and 1000.

Let $D(g)$ indicate the marginal likelihood of a partition. For each partition we have calculated, as a summary performance measure, the difference with the marginal likelihood of the exchangeability partition: $D_{\mathrm{exc}} = D(\mathrm{exc}) - D(g)$.

Fig. 1 shows the behavior of such D_{exc}, for the considered models. The top left figure reports the results obtained with a Bayesian parametric model, the other three results with the Bayesian nonparametric model.

From Fig. 1 we note that there is clear evidence against a treatment effect, apart from strong prior opinion in such model (e.g. in the nonparametric case, with $M = 1000$). Note that the parametric model is more unstable, as its posterior variance is more than twice the nonparametric ones. The latter are equal, as we have run all simulations with the same initial seed, and the effect of M on D_{exc} is constant.

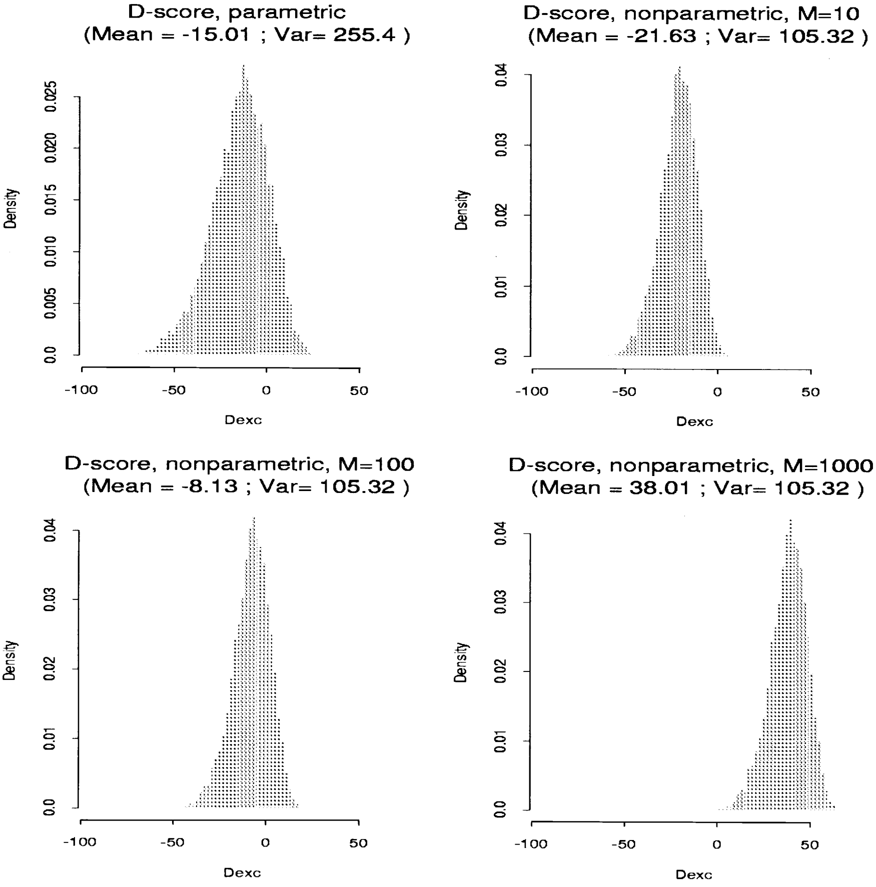


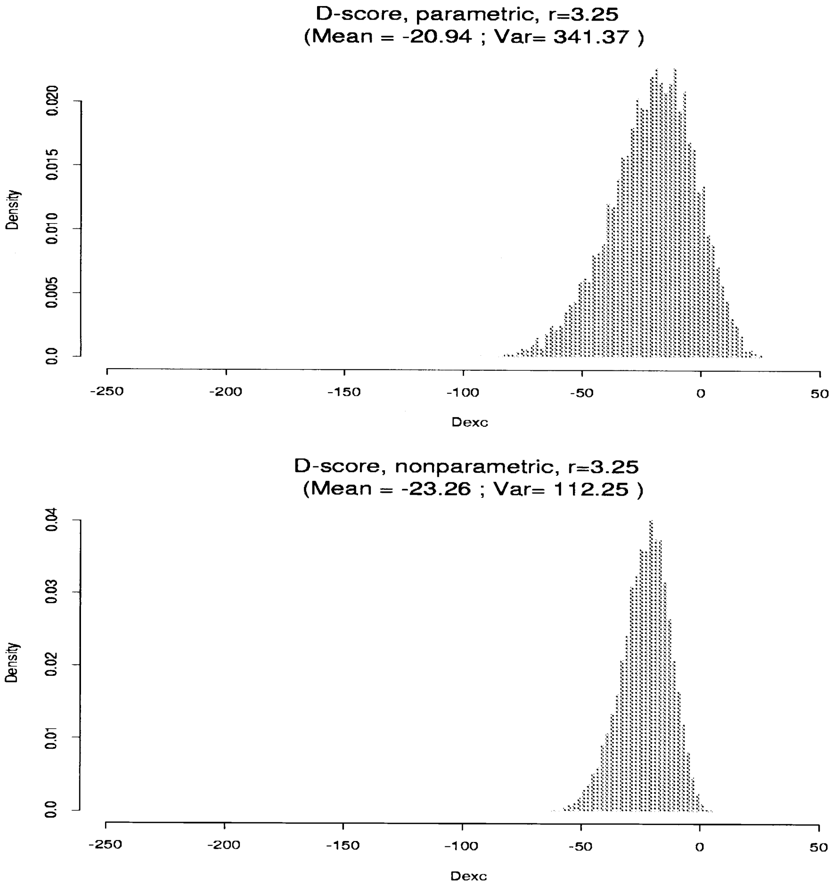
Fig. 1. D-scores against exchangeability, for the exponential regression model on the mice data set.

In order to evaluate also the robustness of the results with respect to the Weibull parameter r, consider running the same simulation, with $r = 3.25$, which is the (approximate) posterior mean of the Weibull parameter from the analysis of Dellaportas and Smith (1993).

Fig. 2 shows the approximate score D_{exc}, corresponding to the covariate treatment, against the exchangeability partition, for the considered Weibull regression model, with $r = 3.25$. The top figure reports the results obtained with a Bayesian parametric model, the bottom one with the Bayesian nonparametric model, for $M = 100$.

Comparing Figs. 1 and 2 it appears that our results are not sensible to the choice of the Weibull parameter r. This is confirmed by taking a grid of different values of r. In fact, presence of a treatment effect is now even less supported. Note again the higher posterior variance for the parametric model.

Furthermore, note that the location of the two distributions is rather similar. This signals that the parametric model we are using is well supported by the data. Note that,



D-score, parametric, r=3.25
(Mean = -20.94 ; Var= 341.37)

D-score, nonparametric, r=3.25
(Mean = -23.26 ; Var= 112.25)

Fig. 2. D-scores against exchangeability, for the Weibull regression model on the mice data set.

although this conclusion is somewhat obvious, as the choice of r was suggested by the parametric Bayesian analysis, it further shows that the nonparametric model performs quite well.

4. Concluding remarks

Our main contribution is the introduction of a nonparametric Bayesian methodology to compare explanatory variables in survival analysis on the basis of their predictive power. To achieve this aim, we have considered mixtures of products of Dirichlet processes, and provided formulae to compute the marginal likelihood of each partition, according to different dependence assumptions.

Our model can be extended in several ways, but particularly we foresee two important extensions. A first extension would be to deal with general proportional hazard models, with the aim of selecting not only the relevant covariates but also their best linear combination for use in the model.

A second extension would be to consider how to efficiently employ the selected covariates. For instance, in a model averaging perspective, research work should be devoted on the important topic of choosing the weights of the mixture density, with each component being described by a partition induced by a relevant covariate. See, in this respect, Walker et al. (2000).

We finally remark that the main objective of our analysis is to choose a covariate so as to improve the predictive ability of the model. We would like to stress that, when a hierarchical mixture of products of Dirichlet process priors is considered, the predictive cumulative distribution function of each group does depend not only on the past data from that group, but borrows strength from all observations. This can be stated more formally as follows.

Indicate with Y_i the random variable that represents the realization of a future observation in the ith group, and assume all observations are distinct. The predictive distribution of Y_i is

$$Pr(Y_i \leqslant y_i | \underline{x}_1, \ldots, \underline{x}_k)$$
$$= \frac{M}{M+n_i} \int_{\Re^k} \Phi(y_i) \pi(\underline{\mu} | \underline{x}_1, \ldots, \underline{x}_k) \, d\mu + \frac{n_i}{M+n_i} F_i(y_i),$$

where $\pi(\underline{\mu} | \underline{x}_1, \ldots, \underline{x}_k)$ is the posterior distribution of the hyperparameter μ and F_i is the empirical cdf of the ith group.

From the above expression note that the predictive cdf depends, through the posterior distribution of μ, on *all* past data. Another interesting remark is that such predictive presents discrete jumps only in correspondence of the distinct observations of the ith population.

Acknowledgements

We thank the special issue guest editor and the referees for useful comments and discussion. The research has been partially supported by C.N.R. (Italian Research Council) and M.U.R.S.T., Italy.

References

Altman, D.G., Andersen, P.K., 1989. Bootstrap investigation of the stability of a Cox regression model. Statist. Med. 8, 771–783.

Antoniak, C.E., 1973. Mixtures of Dirichlet processes, with applications to Bayesian nonparametric problems. Ann. of Statist. 2, 1152–1174.

Carota, C., Parmigiani, G., 2000. Semiparametric regression for count data. ISDS-Technical report, 97-17.

Cifarelli, D.M., 1979. Impostazione bayesiana di un problema di analisi della varianza con approccio non parametrico. Quaderni Istituto di Matematica Finanziaria, Università di Torino.

Cifarelli, D.M., Regazzini, E., 1978. Problemi statistici non parametrici in condizioni di scambiabilita' parziale e impiego di medie associative. Quaderni Istituto Matematica Finanziaria, Torino.

Cifarelli, D.M., et al. 1981. Il modello lineare nell'approccio bayesiano non parametrico. Quaderni dell'Istituto Matematico "G. Castelnuovo", Roma.

Consonni, G., 1981. Impostazione Bayesiana di un problema di analisi discriminatoria nell'ambito di un modello non parametrico. Rivista di Matematica per le Scienze Economiche e Sociali 4, 89–102.

De Finetti, B., 1938. Sur la condition d'equivalence partielle. VI Colloque geneve, Act. Sci. Ind., Vol. 739, Hermann, Paris.

Dellaportas, P., Smith, A.F.M., 1993. Bayesian inference for generalized linear and proportional hazard models via Gibbs sampling. Appl. Statist. 3, 443–459.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–588.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. of Statist. 1, 209–230.

Giudici, P., 1996. A hierarchical model to identify prognostic factors in survival analysis. Statistica Applicata, Italian J. Appl. Statist. 8, 319–326.

Grieve, A.P., 1987. Application of Bayesian software: two examples The Statistician 36, 283–288.

Mallick, B.K., Denison, D.G.T., Smith, A.F.M., 1999. Bayesian survival analysis using a MARS model. Biometrics 55, 1071–1077.

Mira, A., Petrone, S., 1996. Bayesian hierarchical nonparametric inference for changepoint problems. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, Vol. 5. Oxford University Press, Oxford.

Muliere, P., Petrone, S., 1993. A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models J. Italian Statist. Soc. 2 (3), 349–364.

Muliere, P., Scarsini, M., 1983. Impostazione Bayesiana di un problema di analisi della varianza a due criteri. Giornale degli Economisti e Annali di Economia, 519–526.

Petrone, S., Raftery, A.E., 1997. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. Statist. Probab. Lett. 36, 69–83.

Prentice, R.L., 1973. Exponential survivals with censoring and explanatory variables. Biometrika 60, 279–288.

Raftery, A.E., Madigan, D., Volinsky, C.T., 1996. Accounting for model uncertainty in survival analysis improves predictive performance. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, Vol. 5. Oxford University Press, Oxford, pp. 323–350.

Sauerbrei, W., Schumacher, M., 1992. A bootstrap resampling procedure for model building: application to the Cox regression model Statist. Med. 11, 2093–2109.

Walker, S., Gutierrez-Pena, E., Muliere, P., 2000. A decision theoretic approach to model averaging. In: J. Roy. Statist. Soc. Ser. D, to appear.