RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

MARCH 1997

BAYESIAN NONPARAMETRIC INFERENCE FOR RANDOM DISTRIBUTIONS AND RELATED FUNCTIONS

WORKING PAPER #9712-05

BY
STEPHEN WALKER
IMPERIAL COLLEGE, LONDON, UK
PAUL DAMIEN
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL
PURUSHOTTAM W. LAUD
MEDICAL COLLEGE OF WISCONSIN
ADRIAN F. M. SMITH
IMPERIAL COLLEGE, LONDON, UK

BAYESIAN NONPARAMETRIC INFERENCE FOR RANDOM DISTRIBUTIONS AND RELATED FUNCTIONS

Stephen G. Walker
Department of Mathematics, Imperial College, London, UK.

Paul Damien
School of Business, University of Michigan, Ann Arbor, MI, USA.

Purushottam W. Laud
Medical College of Wisconsin, Milwaukee, WI, USA

Adrian F.M. Smith
Department of Mathematics, Imperial College, London, UK.

Abstract

In recent years, Bayesian nonparametric inference, both theoretical and computational has witnessed considerable advances. However, these advances have not received a full critical and comparative analysis of their scope, impact and limitations in statistical modelling; many aspects of the theory and methods remain somewhat a mystery to practitioners and many open questions remain. In this paper, we discuss and illustrate the rich modelling and analytic possibilities available to the statistician within the Bayesian nonparametric and/or semiparametric framework.

Key words: Lévy processes. Bernoulli Trips, Polya Trees, Polya Urns, Bayesian Bootstraps, Hierarchical Models, Markov Chain Monte Carlo, Latent Variables, and Gibbs Sampler.

1 Introduction

1.1 From Bayesian parametrics to nonparametrics

Why nonparametrics? Obviously the answer depends on the particular problem and procedures under consideration, but many, if not most, statisticians appear to feel that it would be desirable in many contexts to make fewer assumptions about the underlying populations from which the data are obtained than are required for a parametric analysis.

Bayesian nonparametric models are constructed on 'large' spaces to provide support for more eventualities than are supported by a parametric model. A Bayesian approach to acknowledging uncertainty about a suggested parametric model is to provide a nonparametric model "centered" on the parametric model in some way. For example, to acknowledge uncertainty about the assumption of a normal parametric model, the corresponding nonparametric approach would construct a model centered on the normal assumption but including models in a neighbourhood of this normal family, with the size of the neighbourhood controlled to reflect beliefs in the normality assumption.

Technically, the off-putting aspect (to many) of the Bayesian nonparametric framework is the mathematical apparatus required for specifying distributions on function spaces and for carrying through prior to posterior calculations. A further pragmatic concern is how to incorporate real qualitative prior knowlegde into this mathematical framework. A major emphasis of this paper will be an attempt to clarify these issues and provide detailed illustrative analyses: these will demonstrate both the modelling flexibility of this framework and the ease through tailored simulation methodology with which prior to posterior analysis can be implemented.

The earliest priors for nonparametric problems seem to have been introduced by Freedman (1963) who introduced tail free and Dirichlet random measures. Subsequently, Dubins and Freedman (1965), Fabius (1964), Freedman (1965), Ferguson (1973,1974) formalized and explored the notion of a Dirichlet process. Early work was largely focussed on stylised summary estimates and tests so that comparisons with the corresponding frequentist procedures could be made. Since Ferguson (1973) the nonparametric

Bayesian literature has grown rapidly. The current focus of attention is on full Bayesian analyses of nonparametric models using simulation techniques (apparently first used in this context by Escobar, 1988). In this paper, we shall focus on nonparametric inference for random distributions and related functions. We shall not deal with Bayesian non/semi parametric density estimation; for a recent survey of this field, see Hjort (1996).

1.2 Outline of the paper

The paper is organised as follows. In Section 2.1, the well-known Dirichlet process prior for an unknown distribution function is reviewed and the limitations of this prior are noted. Linear functional estimation is considered in Section 2.2. We assume there is information concerning the functional (specifically, the mean and variance) but that information about the unknown distribution is unavailable and is modelled nonparametrically via the Dirichlet process. We show how to construct the prior for the distribution to incorporate the information on the functional. The extension to the mixture of Dirichlet process (MDP) class of models (essentially a Bayesian nonparametric approach to hierarchical models; see West et al., 1994, and references therein) is discussed in Section 2.3. Other types of prior distributions will be motivated in Section 2.4: detailed descriptions of these latter types of priors will be the focus in Sections 3 (stochastic process priors), 4 (partition model priors) and 5 (exchangeable model priors).

In particular, in the context of reliability and failure time data, interest often centres on the hazard rate and/or survival curve of the process under investigation. In Section 3.4 we consider Bayesian nonparametric survival data models, providing estimators which generalise the classical Kaplan and Meier (1958) nonparametric estimator. Also in Section 4.4 we consider Bayesian semiparametric approaches for the proportional hazards model (Cox, 1972), the accelerated failure time model, and frailty models (Clayton and Cuzick, 1985). In Section 5.4, we consider a three state disease process model.

2 General framework

We assume that $Y_1, Y_2, \cdots$, defined on some space Ω, is a sequence of iid observations from some unknown probability distribution F, assumed to be

random and assigned a prior distribution, P_Ω. In a parametric framework, F is assumed to be characterised by a *finite* dimensional unknown parameter, Θ. The prior is then assigned to Θ, and we write P_Ω as P_Θ. If we eschew the finite dimensional assumptions we enter the realms of Bayesian nonparametrics. However, if we think of the nonparametric model, P_Ω, as arising from a wish to weaken a posited parametric assumption, P_Θ, we can construct a P_Ω "centered", in some sense, on P_Θ. An important seminal version of a nonparametric prior is the Dirichlet process (Ferguson, 1973) which we now review.

2.1 The Dirichlet process prior

One approach is to define the Dirichlet process via the Dirichlet distribution:

Definition 2.1: Let Ω be a space and $\mathcal{B}$ a σ-field of subsets, and let α be a finite non-null measure on $(\Omega, \mathcal{B})$. Then a stochastic process F indexed by elements A of $\mathcal{B}$, is said to be a Dirichlet process on $(\Omega, \mathcal{B})$ with parameter α, if for any measurable partition $(A_1, \cdots, A_k)$ of Ω, the random vector $(F(A_1), \cdots, F(A_k))$ has a Dirichlet distribution with parameter $(\alpha(A_1), \cdots, \alpha(A_k))$.

We write $F \sim \mathcal{D}(\alpha)$ and $\alpha(.) = cF_0(.)$ where $c > 0$ and F_0, a probability distribution on Ω, "centres" the process in the sense that $E(F) = F_0$. The scale parameter c is commonly interpreted as controlling the variability of F about F_0. Antoniak (1974) considers a larger class of priors based on the Dirichlet process in which priors are assigned to c and the parameters of the parametric distribution F_0.

Perhaps the most unsatisfactory aspect of the Dirichlet process is the part played by c. There is actually no clear interpretation for this parameter, due to its dual role, controlling both the smoothness (or discreteness) of the random distributions and the size of the neighbourhood (or variability) of F about F_0. To illustrate this, we note that if $F \sim \mathcal{D}(cF_0)$ then

$$\mathrm{var}F(A) = \frac{F_0(A)[1 - F_0(A)]}{c+1}.$$

For *maximum* variability we would want $c \to 0$. However, Sethuraman and Tiwari (1982) point out that as $c \to 0$, F converges in distribution to a single

atomic random measure. Also, note from the expression for the variance of $F(A)$ that it is not possible to specify var$F$ arbitrarily, and that the shape is determined by F_0.

A constructive definition of the Dirichlet process is presented in Sethuraman and Tiwari (1982). If $F \sim \mathcal{D}(cF_0)$ then we can generate an F via

$$F = \sum_{j=1}^{\infty} V_j \delta_{\theta_j},$$

where

$$V_1 = W_1, \;\; V_j = W_j(1 - W_{j-1})...(1 - W_1), \; j = 2, 3, ...,$$

the $W_1, W_2, ...$ are iid beta$(1, c)$, and $\theta_1, \theta_2, ...$ are iid from F_0. This construction will be useful later. Note that F is discrete and this, combined with the problem of interpreting (and hence specifying) c, make the Dirichlet process somewhat unsatisfactory.

Bayesian inference via the Dirichlet process is attractively straightforward. Given the data (in the form of an iid sample of exact observations), the posterior is once again a Dirichlet process, so that the latter is a conjugate prior on the space of distribution functions. The prior to posterior parameter updates are $c \to c+n$ and $F_0 \to (cF_0 + nF_n)/(c+n)$, where F_n is the empirical distribution of the observations. The naive interpretation of c as a prior sample size presumably derives from the forms of these posterior parameters. But does a $c = 0$ correspond to "no information"? If $c = 0$, note that the Bayes estimate for F, with respect to quadratic loss, is given by F_n which is what a nonparametric frequentist would typically use as an estimate. Therefore $c = 0$ fits in with one of the notions of a noninformative prior discussed by Ghosh and Mukerjee (1992). Note, also, that a Dirichlet posterior under a $c = 0$ specification has the parameter nF_n which is the basis for Rubin's Bayesian bootstrap (Rubin, 1981).

An alternative notion considered by Ghosh and Mukerjee is that of "information". Under this notion, $c = 0$ can definitely not be thought of as providing a "noninformative" prior. As mentioned earlier, as $c \to 0$, F converges to a single atomic measure, which is strong information about the discreteness of F.

2.2 Functional estimation

We have seen that it is not possible to express the mean and variance of F arbitrarily using the Dirichlet process, nor is it possible to interpret the parameters satisfactorily if we seek directly to make inference about "the unknown distribution function". However, these difficulties do not arise if interest focuses on inference for a linear functional of F. Let g be a measurable function and, for illustration, consider estimation of $\phi(F) = \int g dF$, with respect to a quadratic loss function. For example, with $g(y) = y^i$, ϕ is the ith moment of F.

The Bayes estimate for ϕ is given (Ferguson, 1973) by

$$\hat{\phi}_n = \frac{c\int g dF_0 + n\int g dF_n}{c+n}.$$

Now $\phi_0 = E\phi(F) = \int g dF_0$ and so $\hat{\phi}_n$ can be written more concisely as

$$\hat{\phi}_n = \frac{c\phi_0 + n\phi_n}{c+n},$$

where $\phi_n = \int g dF_n = n^{-1}\sum_i g(Y_i)$. Assigning a value to ϕ_0 is typically not problematic since it is the prior guess for ϕ.

In the case where prior information is available, the choice of c can be made by considering $\phi(F)$ and $\lambda(F)$ where

$$\lambda(F) = \int g^2 dF - \phi^2(F).$$

Note that $\lambda(F) \geq 0$ for all F. Also note that $E\lambda(F) + \text{var}\phi(F) = \lambda(F_0)$, so that if F_0 has been assigned then a prior choice for λ can be obtained via the choice of prior variance for ϕ.

We can use the constructive definition of the Dirichlet process to show that $E\lambda(F)/\text{var}\phi(F) = c$, or, since $E\lambda(F) + \text{var}\phi(F) = \lambda(F_0)$, to show that $\text{var}\phi(F) = \lambda(F_0)/(c+1)$. Thus we obtain the following prior specifications for ϕ in terms of the parameters of the Dirichlet prior:

$$E\phi = \phi(F_0), \quad \text{var}\phi = \frac{\lambda(F_0)}{c+1}.$$

Being noninformative about ϕ, by allowing $\text{var}\phi \to \infty$, is not obtained by allowing $c \to 0$ but rather by allowing $\lambda(F_0) = \text{var}\, g(Y_0) \to \infty$. The posterior

first two moments are given by $E(\phi|\text{data}) = \hat{\phi}_n$ and

$$\text{var}(\phi|\text{data}) = \frac{1}{c+n+1}\left\{\frac{c}{c+n}\left[(c+1)\sigma^2 + \phi_0^2\right] + \frac{1}{c+n}\sum_i g^2(Y_i) - \hat{\phi}_n^2\right\},$$

where $\sigma^2 = \text{var}\phi$.

If we wish to be "noninformative" by taking $\sigma^2 \to \infty$ (with $c > 0$) then $\text{var}(\phi|\text{data}) \to \infty$, which is not what we want. To obtain a finite posterior variance under this specification (and to be coherent) we require $c = 0$ and $c\sigma^2 = 0$. Under these conditions we see that

$$\text{var}(\phi|\text{data}) = s_n^2/(n+1),$$

where $s_n^2 = n^{-1}\sum_i [g(Y_i) - \phi_n]^2$. Note that $\lambda(F) = \int [g - \phi(F)]^2 dF$ and so s_n^2 can be thought of as a sample estimate for λ. Therefore, the posterior distribution for ϕ, under this limiting form of prior, has mean ϕ_n and variance $s_n^2/(n+1)$.

As an example, let $g(y) = y$ so that $\phi(F) = \int y dF$ represents the mean of F. Under the above prior specifications we can obtain an approximate posterior distribution for ϕ given by the normal distribution with mean $\bar{Y}$ and variance $n^{-1}\sum_i [Y_i - \bar{Y}]^2/(n+1)$.

2.3 Mixture of Dirichlet process model

The Dirichlet process is one of the most "user-friendly" priors in Bayesian nonparametric inference and has received substantial coverage in the literature. However, as mentioned earlier, one problem with the Dirichlet process is that it assigns probability one to the space of *discrete* probability measures. A class of priors that chooses a continuous F with probability one is the Mixture of Dirichlet Process (MDP) model, which we now briefly discuss.

Recently, mixture models have found widespread application: see, for example, Escobar (1994); Escobar and West (1995); West et al. (1994); Mueller et al. (1996); MacEarchen and Mueller (1994); Bush and MacEarchen (1996). These are essentially Bayesian hierarchical models, one of the simplest versions taking the form:

$$Y_i|\theta_i \sim f(.|\theta_i), \;\; i = 1, \cdots, n$$

$$\theta_1, \cdots, \theta_n|F \sim_{\text{iid}} F$$

and

$$F \sim \mathcal{D}(cF_0).$$

Instead of the θ_is being assumed to be iid from some parametric distribution (as with standard Bayesian hierarchical models) greater flexibility is allowed via the introduction of the Dirichlet prior centered on a parametric distribution. In the above referenced papers, priors are also assigned to c and the parameters of F_0.

MDP models have dominated the Bayesian nonparametric literature recently as a consequence of the realisation that full posterior computation is feasible using simulation methods (Escobar, 1994), although the latter can be very computer intensive and involve non-trivial sampling algorithms (particularly when $f(.|\theta)$ and $F_0(\theta)$ form a nonconjugate pair).

The MDP model provides a *continuous* nonparametric prior for the distribution of the Y_is. Given F, we have

$$Y_i|F \sim \sum_{j=1}^{\infty} V_j f(.|\theta_j),$$

based on the constructive definition given in Section 2.1. This mixture model has been successfully exploited by Escobar and West (1995). Note, however, that centering $\sum_{j=1}^{\infty} V_j f(.|\theta_j)$ will not be easy.

2.4 Beyond the Dirichlet process

As was noted in Section 2.1, there are limitations with the Dirichlet process when it comes to prior specifications and their interpretation. In the rest of the paper, we will focus on generalisations of the Dirichlet prior which overcome these difficulties.

THEOREM 1 (Doksum, 1974; Dalal, 1978). Let $(\Omega, \mathcal{B})$ be a measurable space and let a system of finite dimensional distributions

$$\Big(F(B_{1,1}), ..., F(B_{m,k})\Big)$$

be given for each finite class $(B_{1,1}, ..., B_{m,k})$ of pairwise disjoint sets from $\mathcal{B}$. If

1. $F(B)$ is a random variable on $(0,1)$ for all $A \in \mathcal{B}$;

2. $F(\Omega) = 1$ a.s.
3.

$$\Big(F(\cup_i B_{1,i}), ..., F(\cup_i B_{m,i})\Big) =_d \Big(\sum_i F(B_{1,i}), ..., \sum_i F(B_{m,i})\Big);$$

(here $=_d$ denotes equality in distribution) then there exists a probability measure P_Ω on the space of probability measures on $(\Omega, \mathcal{B})$ yielding these finite dimensional distributions.

There are a number of ways of constructing a nonparametric prior to meet the requirements of Theorem 1:

1) *Stochastic Processes.* This approach is particularly appropriate for generating random cdfs on $(0, \infty)$ with application in survival data models. An important and rich class of priors is the *neutral to the right* process (Doksum, 1974). Briefly we have $F(t) = 1 - \exp(-Z(t))$ where Z is an independent increments or Lévy process on $(0, \infty)$, with $Z(0) = 0$ and $\lim_{t\to\infty} Z(t) = \infty$. We shall illustrate this approach with the analysis of the well known Kaplan and Meier (1958) data set.

2) *Partitioning.* Here we construct a binary tree partition of Ω denoted by $\Pi = \{(B_\epsilon)\}$, where ϵ is a binary sequence which 'places' B_ϵ in the tree. At level 1 in the partitioning process, we have sets B_0 and B_1 such that $B_0 \cap B_1 = \emptyset$ and $B_0 \cup B_1 = \Omega$. Then, at level 2, B_0 'splits' into B_{00} and B_{01} and so on. A probability distribution is assigned to $\{F(B_\epsilon)\}$ such that, for all ϵ, $F(B_{\epsilon 0}) + F(B_{\epsilon 1}) = F(B_\epsilon) \geq 0$ and $F(\Omega) = 1$. This is the idea behind Polya trees (Ferguson, 1974; Lavine, 1992,1994; Mauldin et al., 1992). Such priors seem particularly appropriate for error models, either at the first or second stage in a hierarchical model. Applications considered later in this paper include generalised linear mixed models, accelerated failure time and frailty models

3) *Exchangeability.* Rather than constructing F directly, as in 1) and 2) above, here we rely on the *Representation Theorem* (de Finetti, 1937) for a sequence of *exchangeable* random variables defined on Ω. Such an approach seems particularly appropriate when the problem is one of prediction; that is, in providing the distribution of Y_{n+1} given $Y_1, \cdots, Y_n$. Applications considered in this paper include modelling a multiple state disease process.

Each of these approaches will now be considered separately in detail (although they are by no means mutually exclusive: for example, the Dirichlet

process has a representation under all three approaches).

3 Stochastic processes

3.1 Neutral to the right process

We begin by discussing *neutral to the right* (henceforward, NTR) processes. Many well-known processes, such as the gamma and simple homogeneous processes (Ferguson and Phadia, 1979), and the Dirichlet process (Ferguson, 1973,1974) belong to this class. More recently, a NTR process called the beta-Stacy was developed by Walker and Muliere (1997). Detailed background to the following discussion can be found in Lévy (1936), Ferguson (1973, 1974), Doksum (1974), and Ferguson and Phadia (1979).

Definition 3.1: A non-decreasing almost surely (a.s.), right continuous a.s., process, $Z(t)$, with independent increments, is called a NTR Lévy process if it satisfies:

1) $Z(0) = 0$ a.s; and
2) $\lim_{t\to\infty} Z(t) = \infty$ a.s.

Fact 3.1 $Z(t)$ has at most countably many fixed points of discontinuity.
Fact 3.2 Let $t_1, t_2, \cdots$ correspond to these fixed points of discontinuity having independent jumps $W_1, W_2 \cdots$. The difference $Z_c(t) = Z(t) - \sum_j W_j I_{[t_j,\infty)}(t)$, where $I(.)$ is the indicator function, is a non-decreasing, independent increments process *without* fixed points of discontinuity and is a *pure* Lévy process. Hence, every NTR process can be written as the sum of a *jump* component and a *continuous* component. This will be useful when we later address the problem of generating random variates from a NTR process.
Fact 3.3 The Lévy formula for the log of the Laplace transform of $Z_c(t)$ has a representation given by,

$$\log E \exp\left(-\phi Z_c(t)\right) = -\phi b(t) + \int_0^\infty (e^{-\phi v} - 1) dN_t(v),$$

where N_t is a continuous Lévy measure. The "location" function $b(.)$ is not required for the following discussion: see Ferguson (1974) for details.

Definition 3.2: A random distribution function $F(t)$ on the real line is NTR if it can be expressed as $F(t) = 1 - \exp(-Z(t))$, where $Z(t)$ is a NTR Lévy process.

We will concentrate on the beta-Stacy process (Walker and Muliere, 1997). Let α be a continuous measure and β a positive function. Then F is a beta-Stacy process with parameters α and β if the Lévy measure is given by

$$dN_t(v) = \frac{dv}{(1-e^{-v})} \int_0^t \exp\Big(-v\beta(s)\Big) d\alpha(s).$$

It can be shown that F is a.s. a random probability measure under the condition $\int d\alpha(s)/\beta(s) = +\infty$. The beta-Stacy process generalises the Dirichlet process, which is obtained when α is a finite measure and $\beta(s) = \alpha(s, \infty)$. The simple homogeneous process (Ferguson and Phadia, 1979) results when β is constant. In Section 3.2, we discuss the choice of α and β to specify EF and var$F$.

In the rest of this section, the NTR process priors are discussed with reference to the beta-Stacy process, since many specific forms of interest are special cases of this prior. It is also closely related to the beta process prior, for modelling cumulative hazard functions, introduced by Hjort (1990).

3.2 Prior specifications

Ferguson and Phadia (1979) point out that for the NTR processes which they considered, such as the gamma, simple homogeneous, and Dirichlet processes, interpreting the prior parameters is quite difficult. Walker and Damien (1996) provide a way of specifying the mean and variance of the distribution function based on the beta-Stacy process. This method has the merit that the practitioner can model the prior mean and variance via a Bayesian parametric model. Let

$$\mu(t) = -\log\{ES(t)\} = \int_0^\infty (1-e^{-v}) dN_t(v)$$

and

$$\lambda(t) = -\log\Big\{E[S^2(t)]\Big\} = \int_0^\infty (1-e^{-2v}) dN_t(v),$$

where $S(t) = 1 - F(t)$. Recalling the Lévy measure for the beta-Stacy process (we assume for simplicity that there are no fixed points of discontinuity in the prior process), Walker and Damien show that there exist $\alpha(.)$ and $\beta(.)$ which provide an explicit solution satisifying the above two conditions for arbitrary μ and λ satisfying $\mu < \lambda < 2\mu$, which corresponds to $[ES]^2 < E[S^2] < ES$. Explicitly, we obtain $d\alpha(t) = d\mu(t)\beta(t)$ and $d\lambda(t)/d\mu(t) = 2 - (1 + \beta(t))^{-1}$.

Suppose, for example, we wish to centre, up to and including second moments, the nonparametric model on the parametric Bayesian model given by $S(t) = \exp(-at)$ with $a \sim \text{gamma}(p, q)$. Then we would have $\mu(t) = p\log(1 + t/q)$ and $\lambda(t) = p\log(1 + 2t/q)$ giving

$$\beta(t) = q/(2t) \text{ and } d\alpha(t) = pqdt/[2t(q + t)].$$

This method of specifying the prior mean and variance of the distribution function overcomes the difficulties in interpretation identified by Ferguson and Phadia (1979). In the absence of alternative strong prior information, this provides a flexible form of prior specification. We can specify a p and q to reflect beliefs concerning the "likely" position of S; that is, a region of high probability in which S is thought most likely to be. The unrestricted nature of the prior will then allow S to "find" its correct shape within this specified region, given sufficient data.

3.3 Posterior distributions

THEOREM 2 (Ferguson, 1974). If F is NTR and $Y_1, \cdots, Y_n$ is a sample from F, including the possibility of right censored samples, then the posterior distribution of F is NTR.

The prior distribution for $Z(t)$ (the Lévy process) is characterised by

$$M = \{t_1, t_2, \cdots\}, \ \{f_{t_1}, f_{t_2}, \cdots\},$$

the set of fixed points of discontinuity with corresponding densities for the jump components, and $N_t(.)$, the Lévy measure for the continuous component of $Z(t)$. We now give the characterisation of the posterior distribution for a single observation Y. (The case for n observations can be obtained by

repeated application.) In the following, we assume the Lévy measure to be of the type

$$dN_t(v) = dv \int_{(0,t]} K(v,s)ds,$$

which includes the beta-Stacy process. The next theorem provides the complete posterior characterisation for this class of NTR processes.

THEOREM 3 (Ferguson, 1974; Ferguson and Phadia, 1979). Let F be NTR and let Y be a random sample from F.

i) Given $Y > y$ the posterior parameters (which we denote by an asterisk) are $M^* = M$,

$$f^*_{t_j}(v) = \begin{cases} \kappa.e^{-v} f_{t_j}(v) & \text{if } t_j \leq y \\ f_{t_j}(v) & \text{if } t_j > y \end{cases}$$

and $K^*(v,s) = \exp\{-vI(y \geq s)\}K(v,s)$, where $I(.)$ is the indicator function, and κ denotes the normalising constant.

ii) Given $Y = y \in M$ the posterior parameters are $M^* = M$,

$$f^*_{t_j}(v) = \begin{cases} \kappa.e^{-v} f_{t_j}(v) & \text{if } t_j < y \\ \kappa.(1-e^{-v}) f_{t_j}(v) & \text{if } t_j = y \\ f_{t_j}(v) & \text{if } t_j > y \end{cases}$$

and, again, $K^*(v,s) = \exp\{-vI(y \geq s)\}K(v,s)$.

iii) Given $Y = y \notin M$ the posterior parameters are $M^* = M \cup \{y\}$, with $f_y(v) = \kappa.(1-e^{-v})K(v,y)$,

$$f^*_{t_j}(v) = \begin{cases} \kappa.e^{-v} f_{t_j}(v) & \text{if } t_j < y \\ f_{t_j}(v) & \text{if } t_j > y \end{cases}$$

and, again, $K^*(v,s) = \exp\{-vI(y \geq s)\}K(v,s)$.

Consequently, if F is a beta-Stacy process with parameters α and β then, given an iid sample from F, with possible right censoring, the Bayes estimate of $F(t)$, under a quadratic loss function, is given by

$$\hat{F}(t) = 1 - \prod_{[0,t]} \left\{ 1 - \frac{d\alpha(s) + dN(s)}{\beta(s) + M(s)} \right\},$$

where $N(t) = \sum_i I(Y_i \leq t)$, $M(t) = \sum_i I(Y_i \geq t)$ and $\prod_{[0,t]}$ represents a product integral (Gill and Johansen, 1990). The Kaplan-Meier estimate is obtained as $\alpha, \beta \to 0$, which is also the basis for both the censored data Bayesian bootstrap (Lo, 1993) and the finite population censored data Bayesian bootstrap (Muliere and Walker, 1997b).

Fact 3.4. The Dirichlet process is not conjugate with respect to right censored data. The beta-Stacy process is conjugate with respect to right censored data. If the prior is a Dirichlet process, then the posterior, given censored data, is a beta-Stacy process.

Fact 3.5. After a suitable transformation and selection of K, other processes such as the extended gamma (Dykstra and Laud, 1981) and the beta process (Hjort, 1990) can be obtained (for details, see Walker and Muliere, 1997).

The remaining key question is whether prior to posterior calculations for these models are computationally feasible. Below we describe a general algorithm which enables us to simulate the posterior beta-Stacy process and thus to perform fully Bayesian nonparametric calculations.

Simulating a NTR Process

Recall that any NTR process $Z(t)$ can be written as the sum of a jump random variable, say W, and a continuous component $Z_c(t)$. From a simulation perspective, given the posterior process, it is sufficient to generate random variates from these two components separately and independently.

Simulating the jump component

With respect to a beta-Stacy process, without jumps a priori, let W denote the posterior jump random variable with density $f_y^*(w)$. Then given $M^* = \{Y_i : \delta_i = 1\}$. where $\delta_i = 1$ indicates that Y_i is an uncensored observation,

$$f_y^*(w) \propto \Big(1 - \exp(-w)\Big)^{N\{y\}-1} \exp\Big(-w[\beta(y) + M(y) - N\{y\}]\Big),$$

where $y \in M^*$, $N\{y\} = \sum_{Y_i = y} \delta_i$ and $M(y) = \sum_i I(Y_i \geq y)$. Also, $W =_d -\log(1 - B)$ where

$$B \sim \text{beta}\Big(N\{y\}, \beta(y) + M(y) - N\{y\}\Big).$$

If $N\{y\} = 1$ then W has an exponential density with mean value $1/(\beta(y) + M(y) - 1)$.

Simulating the continuous component of a NTR process.

It is well known (Ferguson, 1974; Damien et al., 1995) that the continuous component will have a distribution that is *infinitely divisible* (id). Bondesson (1982), Damien et al. (1995) and Walker and Damien (1996) have developed algorithms to generate random variates from any id distribution. Here we note that simulating the continuous component is straightforward regardless of which algorithm one decides to implement. However, the particular choice of the algorithm might depend on the posterior process under consideration. Thus, Laud et al. (1996a) use the Bondesson algorithm to simulate the extended gamma process; Damien et al.'s algorithm is exemplified for the Dirichlet, gamma and the simple homogeneous processes; and Walker and Damien (1996) provide a full Bayesian analysis for a large class of NTR processes using a hybrid of algorithms. For the illustrative analyses that involve NTR processes, we will rely on the Walker and Damien method.

3.4 Example

We reanalyse the Kaplan-Meier (1958) data set, partly for its historical significance, but mainly because it has been studied extensively in recent Bayesian literature and thus provides a basis for comparing different methods and models. The data consist of exact observed failures at 0.8, 3.1, 5.4, 9.2 months, and censored observations at 1.0, 2.7, 7.0, 12.1 months. We address the problem of estimating the probability of failure before 1 month; that is $F(0,1)$. Whereas Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979) were only able to obtain Bayesian point estimates, we are able to sample from the full posterior distribution. Note, also, that we are able to sample from the posterior distribution of $F(0,t)$ for any t and are therefore able to construct a full picture of the posterior failure time distribution.

We follow Ferguson and Phadia (1979) and, within the beta-Stacy framework, take $\beta(s) = \exp(-0.1s)$ and $d\alpha(s) = 0.1\exp(-0.1s)ds$. The prior is therefore a Dirichlet process but, with censored observations in the data set, the posterior is not Dirichlet but a beta-Stacy process. We take $M = \emptyset$ a priori: i.e., there are no jumps in the prior process. The continuous component

of the posterior process, $Z_c^*(t)$, has Lévy measure given by

$$K^*(z,s)ds = \Big(1-\exp(-z)\Big)^{-1}\exp\Big(-z[\beta(s)+M(s)]\Big)d\alpha(s).$$

We consider sampling $F(0,1)$ from the posterior distribution. This involves sampling $Z_c^*(0,0.8)$ and $Z_c^*[0.8,1)$, which is achieved using the algorithm described in Walker and Damien (1996), and sampling $W_{0.8}$ from the density $f_{0.8}^*(.)$, which is the exponential density with mean $[\exp(-0.08)+7]^{-1}$. A required sample from the posterior distribution of $F(0,1)$ is then given by $1-\exp\{-Z_c^*(0,0.8)-Z_c^*[0.8,1)-W_{0.8}\}$.

We collected 1000 samples from the posterior and the resulting histogram representation with kernel density estimate is given in Figure 1. The mean value is given by 0.12 which is the (exact) point estimate value obtained by Ferguson and Phadia.

It can be argued that this prior seems somewhat informative. Can we recapture the shape of Figure 1 using the flexible, less informative prior we proposed in Section 3.2? To investigate this we reanalyse the data set using the Bayesian parametric model described in Section 3.2 with $p=q=1$, in an attempt to be "relatively noninformative". We obtain $\beta(t)=1/(2t)$ and $d\alpha(t)=dt/[2t(1+t)]$. Again, we collected 1000 samples from the posterior and the resulting histogram representation with kernel density estimate is given in Figure 2. Note that we have recovered the shape of Figure 1 extremely well.

It is also of interest to see how our nonparametric analysis compares with a parametric analysis using the parametric model on which it is centered. The posterior distribution from the parametric model is given by $F^*(0,1)=1-\exp(-a)$ with $a\sim$ gamma$(1+4,1+41.3)$. 1000 samples from this posterior were collected and the resulting histogram representation with kernel density estimate is given in Figure 3. The distributions are fundamentally different.

So what does all this add up to? With the parametric model, the first two moments define the shape of the posterior distribution. In the nonparametric model, the first two moments do not define the shape — there is still more flexibility in the model. For a nonparametric/nonparametric comparison we note that our less informative nonparametric prior leads to essentially the same result as the informative nonparametric prior, which is very encouraging — we do not need to select a fully specified distribution on which to centre the prior.

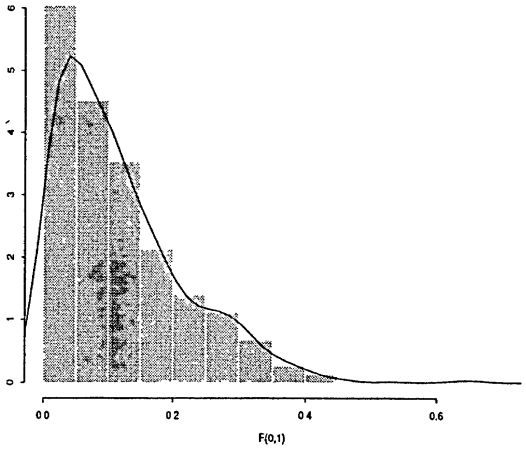

Figure 1: Histogram representation with kernel density estimate of posterior density of F(0,1) using Dirichlet process prior.

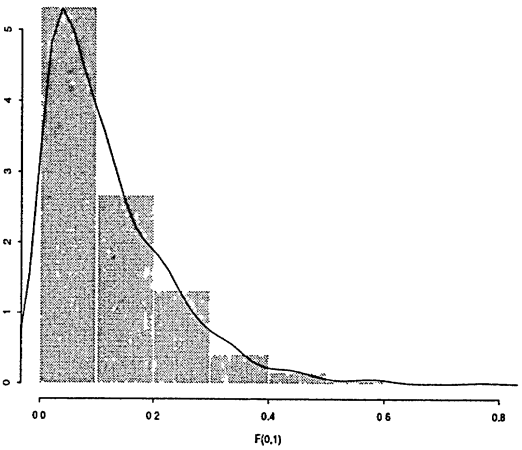

Figure 2: Histogram representation with kernel density estimate of posterior density of F(0,1) using beta-Stacy process prior.

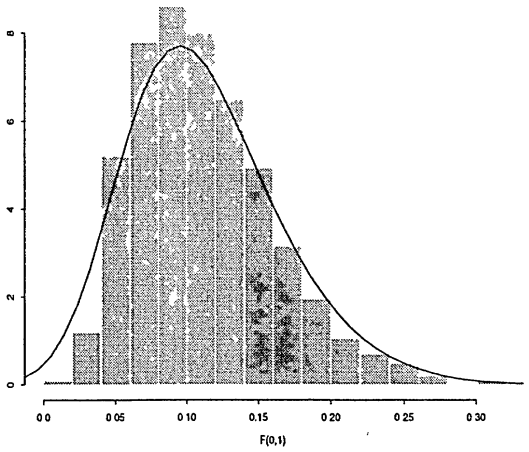

Figure 3: Histogram representation with kernel density estimate of posterior density of F(0,1) using parametric prior.

What else can be done with the stochastic process approach? Kalbfleisch (1978), Clayton (1991), Laud et al. (1996b) provide examples of the use of stochastic processes in the context of Cox regression. It is also possible to use such processes to model functions other than a distribution function. Hjort (1990) developed the beta process to model a cumulative hazard function. Simulation algorithms for carrying out prior to posterior analysis for the beta process appear in Damien et al. (1996).

Dykstra and Laud (1981) consider modelling monotone hazard rates non-parametrically by developing a class of processes called the *extended gamma* process. The advantage of this process is that it indexes the class of absolutely continuous functions with probability one. Also, as in the case of cumulative hazard processes, context motivated information can be used to specify the parameters of the hazard rate model, thus offering greater flexibility. Laud et al. (1993,1996a) develop simulation methods for the extended gamma process; Amman (1984) extended the hazard rate process to model bath-tub hazard rates. Arjas and Gasbarra (1994) develop processes to model the hazard rate piecewise.

However, in practice the stochastic process approach is only user-friendly for relatively simple models of the kind we have illustrated. Inference for

more complex models usually requires us to make some partitioning of the sample space, subsequently working with a discrete version of the process. But this suggests that we should construct the prior on a partitioned space in the first place and motivates the approach considered in the next section.

4 Partitioning Ω

4.1 Polya tree priors

Detailed background to the material of this section can be found in Ferguson (1974), Lavine (1992,1994), Mauldin et al. (1992), and Muliere and Walker (1997a). The Polya tree prior relies on a binary tree partitioning of the space Ω. There are two aspects to a Polya tree: a binary tree partition of Ω and a nonnegative parameter associated with each set in the binary partition. The binary tree partition is given by $\Pi = \{B_\epsilon\}$ where ϵ is a binary sequence which 'places' the set B_ϵ in the tree. We denote the sets at level 1 by (B_0, B_1), a measurable partition of Ω; we denote by (B_{00}, B_{01}) the 'offspring' of B_0, so that $B_{00}, B_{01}, B_{10}, B_{11}$ denote the sets at level 2, and so on. The number of partitions at the mth level is 2^m. In general, B_ϵ splits into $B_{\epsilon 0}$ and $B_{\epsilon 1}$ where $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$ and $B_{\epsilon 0} \cup B_{\epsilon 1} = B_\epsilon$.

A helpful image is that of a particle cascading through these partitions. It starts in Ω and moves into B_0 with probability C_0, or into B_1 with probability $1 - C_0$. In general, on entering B_ϵ the particle could either move into $B_{\epsilon 0}$ or into $B_{\epsilon 1}$. Let it move into the former with probability $C_{\epsilon 0}$ and into the latter with probability $C_{\epsilon 1} = 1 - C_{\epsilon 0}$. For Polya trees, these probabilities are random, beta variables, $(C_{\epsilon 0}, C_{\epsilon 1}) \sim \text{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ with non-negative $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$. If we denote the collection of αs by $\mathcal{A} = \{\alpha_\epsilon\}$, a particular Polya tree distribution is completely defined by Π and $\mathcal{A}$.

Definition 4.1 (Lavine, 1992) A random probability measure F on Ω is said to have a Polya tree distribution, or a Polya tree prior, with parameters $(\Pi, \mathcal{A})$, written $F \sim PT(\Pi, \mathcal{A})$, if there exists non-negative numbers $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \cdots)$ and random variables $\mathcal{C} = (C_0, C_{00}, C_{10}, \cdots)$ such that the following hold:

i) all the random variables in $\mathcal{C}$ are independent;

ii) for every $\epsilon, C_{\epsilon 0} \sim \text{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 0})$; and

iii) for every $m = 1, 2, \cdots$ and every $\epsilon = \epsilon_1 \cdots \epsilon_m$ define

$$F(B_{\epsilon_1 \cdots \epsilon_m}) = \Big(\prod_{j=1;\epsilon_j=0}^{m} C_{\epsilon_1 \cdots \epsilon_{j-1}0} \Big) \Big(\prod_{j=1;\epsilon_j=1}^{m} (1 - C_{\epsilon_1 \cdots \epsilon_{j-1}0}) \Big),$$

where the first terms (i.e., for $j = 1$) are interpreted as C_0 and $1 - C_0$.

Fact 4.1. The Polya tree class indexes priors which assign probability 1 to the set of continuous distributions, unlike the Dirichlet process which has sample distribution functions which are discrete with probability 1. For example, the choice $\alpha_\epsilon = m^2$, whenever ϵ defines a set at level m, leads to an F which is absolutely continuous (Ferguson, 1974).

Fact 4.2. It is easy to show that the discrete versions of the beta process (Hjort, 1990), the beta-Stacy process, and hence the Dirichlet process can all be characterised as Polya trees; see, for example, Muliere and Walker (1997a).

4.2 Prior specifications and computational issues

Problems tackled in this paper involving Polya trees require simulating a random probability measure $F \sim PT(\Pi, \mathcal{A})$. This is done by sampling $\mathcal{C}$ using the constructive form given in Definition 4.1. Since $\mathcal{C}$ is an infinite set an approximate probability measure from $PT(\Pi, \mathcal{A})$ is sampled by terminating the process at a finite level M. Let this finite set be denoted by $\mathcal{C}_M$ and denote by F_M the resulting random measure constructed to level M (which Lavine, 1992, refers to as a 'partially specified Polya tree'). From the sampled variates of $\mathcal{C}_M$ we define F_M by $F(B_{\epsilon_1 \ldots \epsilon_M})$ for each $\epsilon = \epsilon_1 \ldots \epsilon_M$ according to (iii) under Definition 4.1. So, for example, if $M = 8$, we have a random distribution which assigns random mass to $r = 2^8$ sets.

It is possible to centre the Polya tree prior, on a particular probability measure F_0 on Ω by taking the partitions to coincide with percentiles of F_0 and then to take $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ for each ϵ. This involves setting $B_0 = (-\infty, F_0^{-1}(1/2))$, $B_1 = [F_0^{-1}(1/2), \infty)$ and, at level m, setting, for $j = 1, \ldots, 2^m$, $B_j = [F_0^{-1}((j-1)/2^m), F_0^{-1}(j/2^m))$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = +\infty$, where $(B_j : j = 1, \ldots, 2^m)$ correspond, in order, to the 2^m partitions of level m. It is then straightforward to show that $EF(B_\epsilon) = F_0(B_\epsilon)$ for all ϵ.

In practice, we may not wish to assign a separate α_ϵ for each ϵ. It may be convenient, therefore, to take $\alpha_\epsilon = c_m$ whenever ϵ defines a set at level m. For the top levels (m small) it is not necessary for $F(B_{\epsilon 0})$ and $F(B_{\epsilon 1})$ to be 'close'; on the contrary, a large amount of variability is desirable. However, as we move down the levels (m large) we will increasingly wish $F(B_{\epsilon 0})$ and $F(B_{\epsilon 1})$ to be close, if we believe in the underlying continuity of F. This can be achieved by allowing c_m to be small for small m and allowing c_m to increase as m increases; choosing, for example, $c_m = cm^2$ for some $c > 0$. According to Ferguson (1974), $c_m = m^2$ implies that F is absolutely continuous with probability 1 and therefore according to Lavine (1992) this "would often be a sensible canonical choice". In what follows, we shall choose cm^2 for the αs. Note that the Dirichlet process arises when $c_m = c/2^m$, which means that $c_m \to 0$ as $m \to \infty$ (the wrong direction as far as the continuity of F is concerned) and F is discrete with probability 1 (Blackwell, 1973). This model can be extended by assigning a prior to c, but in the applications which follow we shall confine ourselves to providing illustrative analyses corresponding to a range of specified choices of c.

An alternative idea is to understand and assign the c_ms in terms of the variance of the probabilities associated with the sets on level m. These variances are all equal to

$$\frac{1}{2^m}\left\{\prod_{k=1}^{m}\frac{c_k+1}{2c_k+1}-\frac{1}{2^m}\right\}$$

which gives a procedure for assigning the c_ms based on uncertainty in the centering of $F(B_\epsilon)$. For example, if we want $\mathrm{var}F(B_\epsilon) = v_m$, whenever ϵ defines a set at level m, then we need

$$\frac{c_m+1}{2c_m+1} = \frac{4^m v_m + 1}{2(4^{m-1}v_{m-1}+1)}.$$

Note that this imposes a constraint on the v_ms given by $v_{m-1}/4 < v_m < v_{m-1}/2 + 1/4^m$.

More generally, we could define the α_ϵ to match $E_{PT}F(B_\epsilon)$ and $E_{PT}[F^2(B_\epsilon)]$ with those obtained from a parametric model. This procedure will be detailed elsewhere.

4.3 Posterior distributions

Consider a Polya tree prior $PT(\Pi, \mathcal{A})$. Following Lavine (1992), given an observation Y_1, the posterior Polya tree distribution is easily obtained. Write $(F|Y_1) \sim PT(\Pi, \mathcal{A}|Y_1)$ with $(\mathcal{A}|Y_1)$ given by

$$\alpha_\epsilon|Y_1 = \begin{cases} \alpha_\epsilon + 1 & \text{if } Y_1 \in B_\epsilon \\ \alpha_\epsilon & \text{otherwise.} \end{cases}$$

If Y_1 is observed exactly, then an α needs to be updated at each level, whereas in the case of censored data (in one of the sets B_ϵ), only a finite number require to be updated. For n observations, let $\mathcal{Y} = (Y_1, \cdots, Y_n)$, with $(\mathcal{A}|\mathcal{Y})$ given by $(\alpha_\epsilon|\mathcal{Y}) = \alpha_\epsilon + n_\epsilon$, where n_ϵ is the number of observations in B_ϵ. Let $q_\epsilon = P(Y_{n+1} \in B_\epsilon|\mathcal{Y})$, for some ϵ, denote the posterior predictive distribution, and let $\epsilon = \epsilon_1 \cdots \epsilon_m$; then, in the absence of censoring,

$$q_\epsilon = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \frac{\alpha_{\epsilon_1\epsilon_2} + n_{\epsilon_1\epsilon_2}}{\alpha_{\epsilon_1 0} + \alpha_{\epsilon_1 1} + n_{\epsilon_1}} \cdots \frac{\alpha_{\epsilon_1 \cdots \epsilon_m} + n_{\epsilon_1 \cdots \epsilon_m}}{\alpha_{\epsilon_1 \cdots \epsilon_{m-1} 0} + \alpha_{\epsilon_1 \cdots \epsilon_{m-1} 1} + n_{\epsilon_1 \cdots \epsilon_{m-1}}}.$$

For censored data,

$$q_\epsilon = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\epsilon_1 \cdots \epsilon_m} + n_{\epsilon_1 \cdots \epsilon_m}}{\alpha_{\epsilon_1 \cdots \epsilon_{m-1} 0} + \alpha_{\epsilon_1 \cdots \epsilon_{m-1} 1} + n_{\epsilon_1 \cdots \epsilon_{m-1}} - s_{\epsilon_1 \cdots \epsilon_{m-1}}},$$

where s_ϵ is the number of observations censored in B_ϵ.

4.4 Examples

We first re-analyse the Kaplan-Meier data (given in Section 3.4) using a Polya tree prior, providing a comparison with the NTR approach. The second example involves a linear regression model in the context of accelerated failure time data. The third example involves a generalised hierarchical model involving binomial data in a two-way factorial layout.

The Kaplan Meier data.

The NTR approach to the analysis of survival data is mathematically involved. Polya trees can simplify this complexity a great deal. However, what should the partitions be? Recall that with the NTR analysis in Section 3.4 the observations effectively partitioned the time axis with each partition

having to be treated separately (and independently), all the different parts subsequently being "put together".

It seems obvious that the partitions of the Polya tree should coincide with at least some of the observations. In fact the only way to make progress is to have the censoring sets, e.g. $[1.0, \infty)$, coinciding with partition sets. Consider, therefore, the following partitions:

$$B_0 = [0, 1.0), B_1 = [1.0, \infty),$$

$$B_{10} = [1.0, 2.7), B_{11} = [2.7, \infty),$$

$$B_{110} = [2.7, 7.0), B_{111} = [7.0, \infty),$$

and

$$B_{1110} = [7.0, 12.1), B_{1111} = [12.1, \infty).$$

The inclusion of these partitions is compulsory; the choice of the remainder is somewhat arbitrary. A particular partition could therefore be selected in the light of what specific questions needed to be answered; for example, for a future observation, Y_{n+1}, what is $P(Y_{n+1} > 6.0|\text{data})$? This can be answered by partitioning B_{110} appropriately. Thus an estimate of the survival curve $S(t) = 1 - F(t)$ can be found for any $t \in (0, \infty)$ and its computation for a range of ts will be sufficient to provide a clear picture of S.

Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979) assume a Dirichlet process model with an exponential base distribution, with parameter 0.12, that is, $G(B) = \int_B 0.12\exp(-0.12z)dz$ and scale parameter equal to 8 (recall Section 2.1). For comparison with their results, we set $\alpha_\epsilon = c_m G(B_\epsilon)$ whenever ϵ defines a set on level m, and took $c_m = cm^2$ for a number of $c > 0$.

The estimates for various $q_\epsilon = P(Y_{n+1} \in B_\epsilon|\text{data})$ over a selection of values for c, the Kaplan-Meier (KM) estimates, and the Susarla and Van Ryzin (SV) estimates are reported in Table 1. The estimates obtained from the Polya tree are exact and do not depend on a finite level of partitions. This is because in the absence of observations from a particular sub-branch of the tree there is no gain to be made by partitioning sets. For example, suppose we wished to estimate $P(Y_{n+1} > 6.0|\text{data})$. We would partition the interval $[5.4, 7.0)$ at 6.0 and because there are no observations in $(5.4, 7.0)$, no matter how we partitioned this interval (although at some level there would

Interval	Polya tree prior	Polya tree posterior $c = 10$	$c = 1$	$c = 0.01$	KM $c \to 0$	SV
$[0, 0.8)$	0.09	0.08	0.05	0	0	0.05
$[0.8, 1.0)$	0.02	0.03	0.06	0.11	0.12	0.07
$[1.0, 2.7)$	0.16	0.15	0.08	0	0	0.09
$[2.7, 3.1)$	0.03	0.03	0.03	0	0	0.02
$[3.1, 5.4)$	0.17	0.17	0.17	0.19	0.18	0.17
$[5.4, 7.0)$	0.09	0.09	0.13	0.18	0.18	0.13
$[7.0, 9.2)$	0.10	0.10	0.08	0	0	0.07
$[9.2, 12.1)$	0.10	0.11	0.16	0.26	0.26	0.15
$[12.1, \infty)$	0.24	0.24	0.24	0.26	0.26	0.25

Table 1: Posterior predictive probabilities for death time (the second column gives the prior predictives which are independent of c).

need to be a partition at 6.0) and to what levels we took these partitions, it would make no difference to the estimate of $P(Y_{n+1} > 6.0|\text{data})$.

Note from Table 1 that for $c = 10$ the prior dominates the information from the data and for $c = 0.01$ the data dominates the prior. This can be understood from the expressions for q_ϵ given in Section 4.3. A suitable choice for c would be somewhere in between these two extremes. For illustration we have chosen $c = 1$ and remark that the precise choice of c is somewhat problematic. For this reason we recommend the approach briefly outlined in the last paragraph of Section 4.2, which avoids the problem of defining a c altogether. Alternative approaches for dealing with c are described in the immediately following sections.

We can consider the uncertainty associated with estimates of q_ϵ by sampling $F(B_\epsilon)$ from the posterior. So, if $\epsilon = \epsilon_1 ... \epsilon_m$, we would sample C_{ϵ_1} from $p_{C_{\epsilon_1}|\text{data}}(.)$ up to $C_{\epsilon_1 ... \epsilon_m}$ from $p_{C_{\epsilon_1 ... \epsilon_m}|\text{data}}(.)$ and set $F(B_\epsilon) = C_{\epsilon_1} ... C_{\epsilon_1 ... \epsilon_m}$. It is also possible, again with the appropriate partitioning, to obtain samples from $F(t)|\text{data}$ and hence for $S(t)|\text{data}$. In Figure 4 we show the marginal posterior distribution of $S(6.0)|\text{data}$. This was obtained via the sampling strategy just described using a sample of size 5,000, and with $c = 1$.

Using Polya trees for analysing survival data is more straightforward than using NTR processes. Posterior distributions are more tractable and there

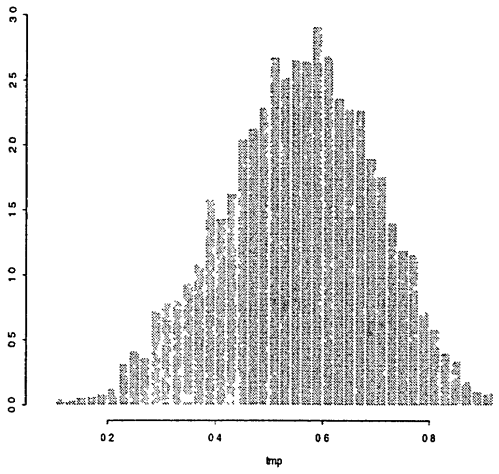

Figure 4: Posterior distribution of $S(6.0)|\text{data}$

is little need for computer intensive simulations. An additional advantage is that we can select Polya trees which are continuous, whereas NTR processes are discrete.

Using Polya trees to model error distributions

We now consider the use of Polya trees for modelling error distributions.

Multiple regression example. We start by considering the linear model

$$Y_i = X_i\beta + \Theta_i,\ i = 1, \cdots, n,$$

where $X_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,p})$ is a vector of known covariates, β is a vector of p unknown regression coefficients and Θ_i are error terms, assumed to be iid from some unknown distribution F. The parameter β is assigned a multivariate normal prior with mean μ and covariance matrix Σ. *A priori*, F and β will be taken to be independent. Since F is completely arbitrary the intercept term of β will be confounded with the location of F. This is easily overcome by fixing the median of F by defining $F(B_0) = F(B_1) = 1/2$. Typically, we will want the median to be located at 0 and this is achieved by taking the partition at level 1 to be at 0. In such cases it may be convenient to take F_0 as the normal distribution with zero mean and variance σ^2. This defines a *median* regression model instead of the more popular *mean* regression model

and parallels Ying et al.'s (1995) frequentist approach. If required, we could also fix the scale of F by defining, for example, $F(B_{00}), ..., F(B_{11})$ each equal to 1/4. This would be appropriate for the alternative model

$$Y_i = X_i\beta + \sigma\Theta_i,\ i = 1, \cdots, n,$$

where F_0 could be taken to be the standard normal distribution. Analysis is based on a MCMC algorithm outlined in the Appendix and further details can be found in Walker and Mallick (1997a).

We reanalyse the data set presented by Ying et al. (1995). This involves 121 patients suffering small cell lung cancer and each undertaking one of two treatments; A with 62 patients, or B with 59 patients. The survival times are given in days, with 98 patients providing exact survival times and the remainder right censored survival times. The covariates are the treatment type, 0 or 1, and the natural logarithm of the entry age of the patient. Ying et al. were only able to estimate the median survival time in their analysis and then test for the "better" treatment. We are not restricted in any way as to the type of inference we can make.

In our analysis we took a normal prior, with mean zero and large variance term, for β. The parameters for the Polya tree are F_0 as the normal distribution with zero mean and variance $\sigma^2 = 10^2$ and $\alpha_\epsilon = cm^2$, whenever ϵ defines a set at level m, with $c = 0.1$. We found these parameters provided a noninformative approach (see Sections 4.2 and 4.4). We took the number of levels of the Polya tree to be fixed at 8. These parameters were obtained after some preliminary analyses. Increasing σ^2 had no effect on the results and yet we were not confident about reducing σ^2 below 100. Reducing c to 0.01 gave more weight to the noninformativeness of the prior at the bottom levels rather than the continuity, and increasing c to 1 removed the noninformativeness of the prior at the top levels. Finally, we chose $M = 8$ to give us satisfactory partitions, essentially neither too big nor too small. We therefore had 254 partitions covering (approximately) the interval $(-20, +20)$, giving 0.16 as the average partition length.

An alternative approach would be to treat c, M and σ^2 as unknown parameters and assign prior distributions (though perhaps this is not necessary for M). This is a relatively straightforward idea to implement using the $Y_i = X_i\beta + \sigma\Theta_i$ model, but we shall present details elsewhere.

For illustration, predictive survival curves are presented; the first (Figure 5) for a new patient with treatment A, and the second (Figure 6) for a new

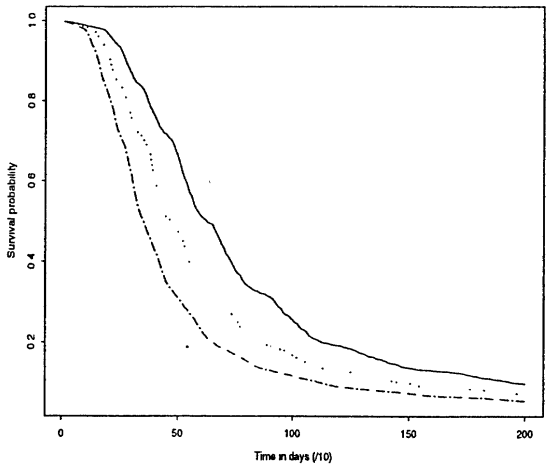

Figure 5: Predictive survival curves for three new patients with treatment A

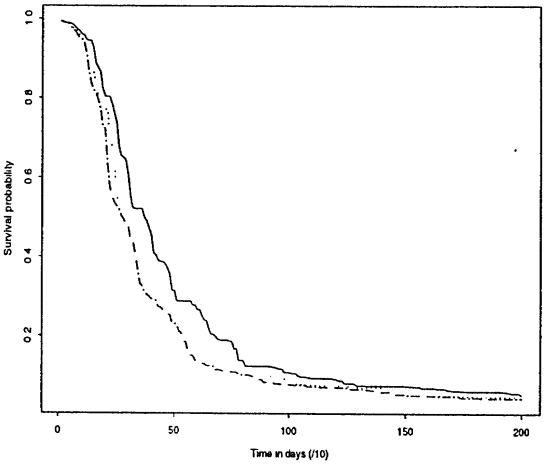

Figure 6: Predictive survival curves for three new patients with treatment B

patient with treatment B. The three curves selected for illustration are those for patients whose covariate values coincide with the quartiles of the observed values of the log entry age covariate.

Hierarchical model example. We now consider nonparametric modelling in the second stage of a hierarchical model. For illustration, we remodel and reanalyse the problem considered by Crowder (1978, Table 3), which involves binomial data in a 2×2 factorial layout. The beta-binomial model of Crowder models variation of expected proportions within cell means, by assuming that

$$Y_{ij}|Z_{ij} \sim \text{binomial}(Z_{ij}, n_{ij}), \;\; i = 1, ..., 4, \;\; j = 1, ..., n_i,$$

with

$$Z_{ij} \sim \text{beta}(\gamma_i, \delta_i),$$

where $\pi_i = \gamma_i/(\gamma_i + \delta_i)$ is the mean of Z_{ij}. The analysis proceeds by finding maximum likelihood estimates for $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$.

We remodel this by considering a random effects logit model with second stage given by

$$\text{logit}Z_{ij} = X_i\beta + \Theta_{ij},$$

where

$$\Theta_{11}, ..., \Theta_{4n_4}|F \sim_{\text{iid}} F.$$

Here, the independence assumption for the Θ_{ij} corresponds to the homogeneity of the variance term of the distribution of the Z_{ij} in the beta-binomial model given by $\gamma_i + \delta_i = \text{constant}$. The $\text{logit}Z_{ij}$ are a linear combination of the intercept term and the factorial effects, so that $X_1 = (1, 0, 0)$, $X_2 = (1, 0, 1)$, $X_3 = (1, 1, 0)$ and $X_4 = (1, 1, 1)$. A MCMC analysis is detailed in the Appendix and further information can be found in Walker and Mallick (1997b). The prior specifications for the Polya tree were taken to be the same as those for the previous example.

Posterior distributions are summarised using the samples obtained from the MCMC output. The ergodic mean estimates of π_i are evaluated as

$$\hat{\pi}_1 = 0.41, \;\; \hat{\pi}_2 = 0.64, \;\; \hat{\pi}_3 = 0.37, \;\; \hat{\pi}_4 = 0.51,$$

which compare with the values

$$\hat{\pi}_1 = 0.41, \;\; \hat{\pi}_2 = 0.65, \;\; \hat{\pi}_3 = 0.33, \;\; \hat{\pi}_4 = 0.57,$$

obtained by Crowder. The posterior mean of F is shown in Figure 7. Essentially the distribution in Figure 7 characterises the within cell mean variability, that is, the variability of the Z_{ij} about π_i. As can be seen, this is a fairly tight distribution about zero, as one would probably expect. The posterior distributions for β_1, β_2 and β_3 are shown in Figure 8. Since the density in Figure 7 does resemble a normal to some extent we reanalyse the data replacing the Polya tree by a normal distribution, with mean zero and variance σ^2, assigning σ^{-2} a noninformative gamma prior. The resulting posterior distributions for β appear in Figure 9. Note then that these posteriors have the same locations as those from the Polya tree analysis but with the posterior spread from the parametric model reduced significantly.

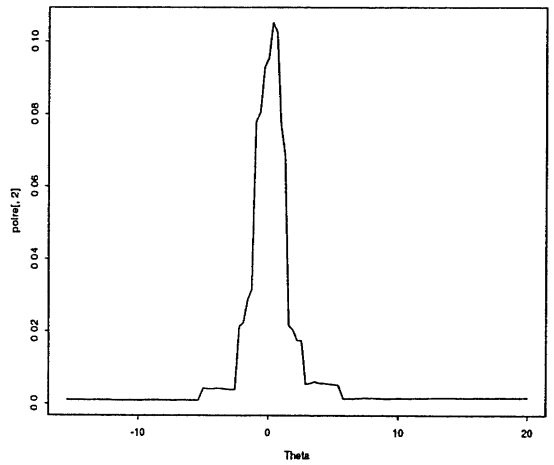

Figure 7: Posterior expectation of F in the beta-binomial model

Walker and Mallick (1997a) also detail the use of Polya trees in a frailty model (Clayton and Cuzick, 1985). Here, we omit details and simply draw attention to the posterior estimate of the log-frailty distribution obtained in that paper. In the analysis the frailties are (incorrectly) assumed to be exchangeable and not dependent on a male/female covariate; Figure 10 evidences the great flexibility of the nonparametric framework in recovering a bimodal form for the distribution of the log-frailties arising from the mixed male/female population.

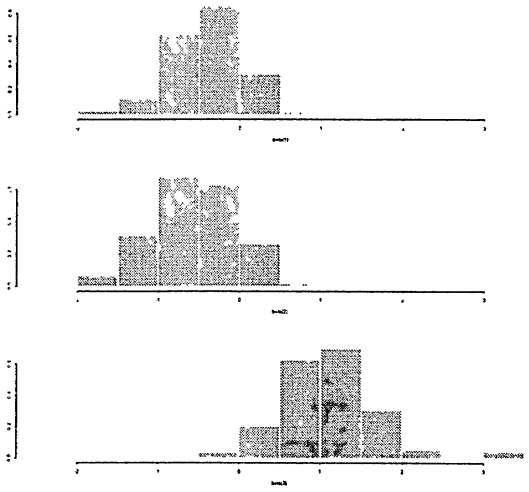

Figure 8: Histogram representations of the posterior distribution of β obtained from nonparametric model

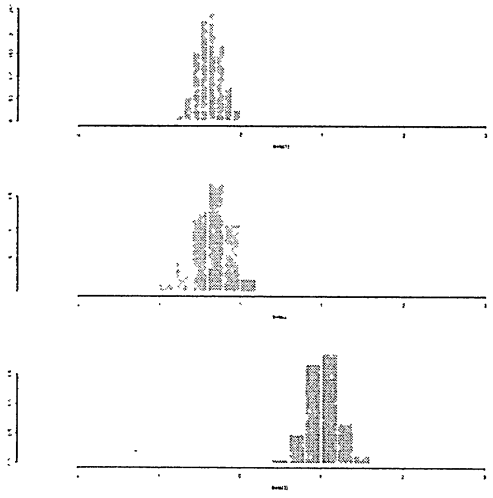

Figure 9: Histogram representations of the posterior distribution of β obtained from parametric model

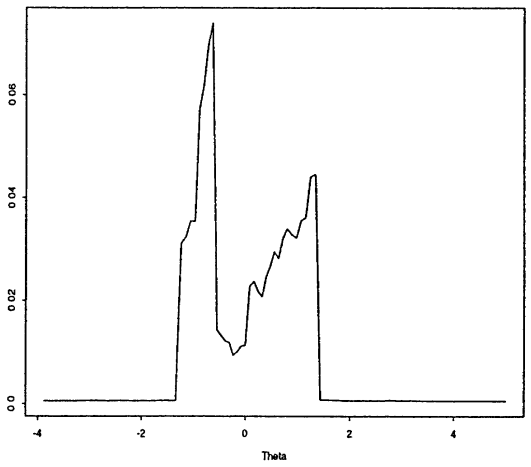

Figure 10: Posterior expectation of log frailty distribution

5 Exchangeable models

Let $Y_1, Y_2, \cdots$ be an *exchangeable* sequence of random variables defined on Ω. By de Finetti's representation theorem (de Finetti, 1937), there exists a probability measure P_Ω defined on the space of probability measures on Ω, such that the distribution of $Y_1, Y_2, \cdots$ can be obtained by first choosing $F \sim P_\Omega$ and then taking $Y_1, Y_2, \cdots | F \sim_{\text{iid}} F$. That is,

$$P(Y_1 \in B_1, \cdots, Y_n \in B_n) = \int \Big(\prod_{i=1}^{n} F(B_i) \Big) dP_\Omega(F).$$

Here P_Ω is referred to as the de Finetti or prior measure and, given the joint distribution of $Y_1, Y_2, \cdots$, this P_Ω is unique (Hewitt and Savage, 1955). An example is the general *Polya-urn* scheme (Blackwell and McQueen, 1973). Let $\alpha(.) = cF_0(.)$, where $c > 0$ and F_0 is a probability measure on Ω. The scheme for generating the exchangeable sequence $(Y_1, \cdots, Y_n)$ from Ω is given by

$$Y_1 \sim F_0,$$

$$Y_2 | Y_1 \sim \frac{cF_0 + \delta_{Y_1}}{c+1},$$

$$\cdots$$

$$Y_n|Y_1,\cdots,Y_{n-1} \sim \frac{cF_0 + \sum_{j=1}^{n-1}\delta_{Y_j}}{c+n-1}.$$

Blackwell and McQueen show that

$$n^{-1}\sum_{i=1}^{n}\delta_{Y_i} \to F \text{ with probability 1,}$$

with F from a Dirichlet process with parameter α. The de Finetti measure for the Polya-urn scheme is therefore the Dirichlet process. As might be anticipated from our earlier identification of the beta-Stacy process as a generalisation of the Dirichlet process, a generalised Polya-urn scheme can be obtained, which has the discrete beta-Stacy process as the de Finetti measure (Walker and Muliere, 1997. See also Section 5.1).

There are a number of reasons why it is often convenient to consider the sequence $Y_1, Y_2, \cdots$ directly, marginalising over F. First, F is an infinite dimensional parameter so the advantages in removing this is that one ends up working in a finite dimensional framework, making much of the mathematics simpler. Secondly, interest is often in prediction and the distribution of Y_{n+1} given $Y_1, \cdots, Y_n$ is an immediate consequence. Thirdly, we are "closer" to the data in the sense that we have the probability distribution for the data explicitly. It should also be pointed out that the posterior parameters for P_Ω can often be determined from the sequence of predictive distributions (consider, for example, the Polya urn sequence).

5.1 Bernoulli trips

Here we introduce a simple concept and method for modelling multiple state processes based on an exchangeable sampling scheme (Bernoulli trip). A Bernoulli trip is a *reinforced random walk* (Coppersmith and Diaconis, 1987; Pemantle, 1988) on a "tree" which characterises the space for which a prior is required. An observation in this space corresponds to an unique path or branch of the tree. The path corresponding to this observation is reinforced; that is, the probability of a future observation following this path is increased. Thus, after n observations, a maximum of n paths have been reinforced.

To construct a Bernoulli trip we discretise the relevant space. The walk starts at ϵ_0 and moves in one of a possible finite number of directions to reach

ϵ_1, say. From here the walk moves, again in one of a possible finite number of directions. In general, a walk reaches ϵ and moves to one of a finite number of "positions", the collection of which we will denote by $\mathcal{M}_\epsilon$. For the first walk

$$P\left(\epsilon \to \epsilon' \in \mathcal{M}_\epsilon\right) = \frac{\alpha(\epsilon, \epsilon')}{\sum_{\epsilon'' \in \mathcal{M}_\epsilon} \alpha(\epsilon, \epsilon'')},$$

where each α is nonnegative. There will be positions which, if reached, result in the walk being terminated, and this eventually happens to all walks, whatever the path. After the first walk the parameters α are updated. If during the course of the first walk a move was made from ϵ to ϵ' then we simply replace $\alpha(\epsilon, \epsilon')$ by $\alpha(\epsilon, \epsilon') + 1$. The second walk "follows" these new probabilities. After the second walk the new parameters are themselves updated in the same way and the third walk "follows" these twice updated probabilities, and so on. It is clear that the probablity of the second walk coinciding with the first walk exactly has increased (reinforcement).

If we denote the path of the first walk by Y_1 and the second walk by Y_2 and so on, then we can write down without much difficulty the joint probability for the first n walks following particular paths. From this it is straightforward to show that $(Y_1, \cdots, Y_n)$ are exchangeable random variables for all n. Explicitly, we have

$$P\left(Y_1, ..., Y_n\right) = \prod_{\epsilon} \frac{\prod_{\epsilon' \in \mathcal{M}_\epsilon} \alpha(\epsilon, \epsilon')^{[n(\epsilon, \epsilon')]}}{\{\sum_{\epsilon' \in \mathcal{M}_\epsilon} \alpha(\epsilon, \epsilon')\}^{[\sum_{\epsilon' \in \mathcal{M}_\epsilon} n(\epsilon, \epsilon')]}}$$

where $n(\epsilon, \epsilon')$ is the number of walks which move from ϵ to ϵ', $a^{[x]} = a(a+1)...(a+x-1)$ and $a^{[0]} = 1$.

A Bayesian bootstrap procedure would be to obtain the posterior parameters and then set the prior parameters to zero. Thus, $\alpha^*(\epsilon, \epsilon') = n(\epsilon, \epsilon')$. In such cases the predictives only depend on the data.

To illustrate, consider a two state process with one absorbing state; that is, a survival model. Each walk starts at $(0,0)$ and on reaching say $(k,0)$, $k = 1, 2, \cdots$, the walk can move either to $(k+1, 0)$ or $(k+1, 1)$. We assume k indexes time points $t_1, t_2, \cdots$. If it reaches $(k, 1)$, for any k, then the walk is terminated (obviously this corresponds to death at t_k). The move $(k-1, 0)$ to $(k, 0)$ indicates survival from t_{k-1} to t_k. Explicitly, for $k = 1, 2, \cdots$,

$$P\left((k-1, 0) \to (k, 0)\right) = \frac{\alpha_{k0}}{\alpha_{k0} + \alpha_{k1}}$$

and

$$P((k-1,0) \rightarrow (k,1)) = \frac{\alpha_{k1}}{\alpha_{k0}+\alpha_{k1}}.$$

Clearly each walk is characterised by the point k at which the move to $(k,1)$ is made and let Y_i represent this point for the ith walk. A priori we have

$$P(Y_1 = k) = \frac{\alpha_{k1}}{\alpha_{k0}+\alpha_{k1}} \prod_{j<k} \frac{\alpha_{j0}}{\alpha_{j0}+\alpha_{j1}},$$

and a posteriori after n observations we have

$$P(Y_{n+1} = k|Y_1,\cdots,Y_n) = \frac{\alpha^*_{k1}}{\alpha^*_{k0}+\alpha^*_{k1}} \prod_{j<k} \frac{\alpha^*_{j0}}{\alpha^*_{j0}+\alpha^*_{j1}},$$

$\alpha^*_{k0} = \alpha_{k0} + n_{k0}$, and $\alpha^*_{k1} = \alpha_{k1} + n_{k1}$, where n_{k0} is the number of walks that move from $(k-1,0)$ to $(k,0)$ and n_{k1} is the number of walks that move from $(k-1,0)$ to $(k,1)$.

We can easily deal with right censored observations within the Bernoulli trip framework. A censored observation at k, that is $Y > k$, corresponds to a walk being censored at k. The updating mechanism for such a walk is given by $\alpha_{j0} \rightarrow \alpha_{j0} + 1$ for all $j < k$. Note that the walks remain exchangeable provided the censoring mechanism is independent of the failure mechanism.

The Bernoulli trip just described can be shown to be a discrete time version of the beta-Stacy process detailed in Section 3. Whereas it would be difficult to extend the stochastic process approach to model multiple state processes it is relatively easy within the Bernoulli trip framework. The only drawback, if indeed it is one, is that the space needs to be discretised. Typically, however, data arising from multiple state processes do come in a discrete form — as information obtained each day, week, or during some other unit of time.

5.2 Prior trips

For illustration, we consider a three state (disease) process in which all patients start in state 1. From here it is possible to move directly to state 3 or move to state 3 via state 2. Random right censoring can occur at any time. We define the first walk via the transition probabilities

$$P((k-1,0) \rightarrow (k,0)) = \frac{\alpha_{k0}}{\alpha_{k0}+\alpha_{k1}+\alpha_{k2}},$$

$$P((k-1,0) \to (k,1)) = \frac{\alpha_{k1}}{\alpha_{k0} + \alpha_{k1} + \alpha_{k2}},$$

and

$$P((k-1,0) \to (k,2)) = \frac{\alpha_{k2}}{\alpha_{k0} + \alpha_{k1} + \alpha_{k2}},$$

for transition from state 1. For transtition from state 2 to state 3, we define

$$P((k-1,1) \to (k,1)) = \frac{\beta_{k1}}{\beta_{k1} + \beta_{k2}},$$

$$P((k-1,1) \to (k,2)) = \frac{\beta_{k2}}{\beta_{k1} + \beta_{k2}}.$$

The walk is completed at k whenever $(k,2)$ is reached. We can obtain the prior predictive for a particular event; for example

$$P(T = k, S = j < k) =$$

$$\frac{\alpha_{j1}}{\alpha_{j0} + \alpha_{j1} + \alpha_{j2}} \prod_{l<j} \frac{\alpha_{l0}}{\alpha_{l0} + \alpha_{l1} + \alpha_{l2}} \times \frac{\beta_{k2}}{\beta_{k1} + \beta_{k2}} \prod_{j<l<k} \frac{\beta_{l1}}{\beta_{l1} + \beta_{l2}},$$

where T denotes the time to reach state 3 and S is the time to reach state 2 (if at all). If state 2 is not visited then

$$P(T = k, \text{state 2 not visited}) = \frac{\alpha_{k2}}{\alpha_{k0} + \alpha_{k1} + \alpha_{k2}} \prod_{l<k} \frac{\alpha_{l0}}{\alpha_{l0} + \alpha_{l1} + \alpha_{l2}}.$$

Note that we need to define the parameters α and β so that the first walk will end with probability 1. Note, also, that the model described here assumes that the transition probabilities from state 2 to state 3 do not depend on the time of transition from state 1 to state 2. This is the Markov model and will be referred to as model $M_{(c)}$. The semi-Markov model, in which the transition probabilities from state 2 to state 3 do depend on the time of transition from state 1 to state 2, can be represented within the Bernoulli trip framework without difficulty. We could have model $M_{(a)}$ given by

$$P(T = k | S = j < k) = \frac{\beta_{kj2}}{\beta_{kj1} + \beta_{kj2}} \prod_{j<l<k} \frac{\beta_{lj1}}{\beta_{lj1} + \beta_{lj2}},$$

to model a direct dependence on the time of transition from state 1 to state 2, or, model $M_{(b)}$ given by

$$P(T = k|S = j < k) = \frac{\beta_{k-j\,2}}{\beta_{k-j\,1} + \beta_{k-j\,2}} \prod_{j<l<k} \frac{\beta_{l-j\,1}}{\beta_{l-j\,1} + \beta_{l-j\,2}},$$

where now the conditional probabilities depend solely on the time spent in state 2.

Here we seek an interpretation for the parameters α_{k0}, α_{k1}; we assume that $\alpha_{k2} = 0$ for simplicity. Note that a priori

$$\frac{\alpha_{k0}}{\alpha_{k0} + \alpha_{k1}} = \frac{P(S = k)}{P(S = k) + P(S > k)}$$

and

$$\frac{\alpha_{k1}}{\alpha_{k0} + \alpha_{k1}} = \frac{P(S > k)}{P(S = k) + P(S > k)}.$$

It can be seen that α_{k0} is associated with $P(S = k)$ and α_{k1} is associated with $P(S > k)$. Therefore it is reasonable to take

$$\alpha_{k0} = c_k \hat{P}\big(S = k\big) \text{ and } \alpha_{k1} = c_k \hat{P}\big(S > k\big),$$

where each c_k is a positive and $\hat{P}$ are prior guesses for the relevant probabilities. Intepretations for the c_ks can be found by considering the estimates/predictives for the conditional hazards

$$P\left(S_{n+1} = k|S_{n+1} \geq k, Y_1, \cdots, Y_n\right) = \frac{\alpha_{k0} + n_{k0}}{\alpha_{k0} + n_{k0} + \alpha_{k1} + n_{k1}}.$$

It is seen that large c_k reflects strong belief in the prior estimate/predictive $\hat{P}(S = k|S \geq k)$, and a small c_k reflects a corresponding weak belief in $\hat{P}(S = k|S \geq k)$. Similar interpretations can be found for β_{k1} and β_{k2}.

5.3 Posterior trips

A complication with obtaining the posterior trips arises if some of the observations are *interval censored.* Suppose that one observation $(i = n)$ is interval

censored, that is, S_n is known to be in the interval $[k_1, ..., k_L]$ ($k_L < \infty$ and $T_n > k_L$). The (random) updated parameters are given, for $M_{(c)}$, by

$$\alpha^*_{k0} = \alpha_{k0} + n_{k0} + \mathcal{J}_{\alpha k}$$

and

$$\alpha^*_{k1} = \alpha_{k1} + n_{k1} + I(k < k_n) + \mathcal{J}_{\beta k},$$

where $n_{k0} = \sum_{i=1}^{n-1} I(S_i = k)$ and $n_{k1} = \sum_{i=1}^{n-1} I(S_i > k)$. Here $\mathcal{J}_{\alpha k}$ and $\mathcal{J}_{\beta k}$ are random and defined on $\{0, 1\}$ where

$$I\Big(\mathcal{J}_{\alpha k} = 1\Big) = I\Big(S_n = k | k_1 \leq S_n \leq k_L, S_1, ..., S_{n-1}, T_1, ..., T_{n-1}\Big),$$

$$I\Big(\mathcal{J}_{\beta k} = 1\Big) = I\Big(S_n > k | k_1 \leq S_n \leq k_L, S_1, ..., S_{n-1}, T_1, ..., T_{n-1}\Big)$$

and

$$P\Big(\mathcal{J}_{\alpha k} = 1\Big) = \frac{P\Big(S_n = k | S_1, ..., S_{n-1}, T_1, ..., T_{n-1}\Big)}{P\Big(k_1 \leq S_n \leq k_L | S_1, ..., S_{n-1}, T_1, ..., T_{n-1}\Big)}$$

which is given, up to a constant of proportionality, by

$$\tau_k \prod_{l=k_1}^{k-1} \Big(1 - \tau_l\Big) \times \prod_{l=k+1}^{k_L} \Big(1 - \xi_l\Big),$$

where

$$\tau_k = \frac{\alpha_{k0} + n_{k0}}{\alpha_{k0} + n_{k0} + \alpha_{k1} + n_{k1}}$$

and

$$\xi_k = \frac{\beta_{k2} + \sum_{i=1}^{n-1} I(T_i = k, S_i < k)}{\beta_{k1} + \beta_{k2} + \sum_{i=1}^{n-1} I(T_i \geq k, S_i < k)}$$

for $k \in \{k_1, ..., k_L\}$. For more than one interval censored observation we can proceed by sampling the missing data, conditional on all the other observations, obtain the predictive estimate, or whatever is required, and then take the average over a number of simulations. Without loss of generality, let $S_1, ..., S_m$ ($m \leq n$) be interval censored, with $S_j \in [k_{1(j)}, ..., k_{L(j)}]$ ($T_j > k_{L(j)}$). The approach is to sample iteratively, for $j = 1, ..., m$, from

$$P\Big(S_j | k_{1(j)} \leq S_j \leq k_{L(j)}, S_{(j)}, T_{(j)}\Big),$$

where $(S_{(j)}, T_{(j)})$ contains all the information in the data and from the sampled variates except on individual j. Note that if

$$S_{(j)} \cap \left\{k_{1(j)}, ..., k_{L(j)}\right\} = \emptyset$$

then S_j is taken uniformly from $\{k_{1(j)}, ..., k_{L(j)}\}$. At iteration t we have then sampled

$$\left\{S_j^{(t)} : j = 1, ..., m\right\},$$

which, combined with the observed data, gives the estimator $\hat{P}^{(t)}$. The required estimator is then given by the average

$$\tau^{-1} \sum_{t=1}^{\tau} \hat{P}^{(t)}$$

for some large enough τ to ensure convergence of the simulated Markov chain. Such a procedure can be viewed as a stochastic version of the iterative algorithm for obtaining the self consistent estimator in Frydman (1992). Essentially the sampling from $[S_j|...]$ replaces taking the expectation of $[S_j|...]$. Note that it is also possible to consider the situation in which T and S are both interval censored using a modified version of the algorithm just described.

5.4 Example

We analyse a data set presented by De Gruttola and Lagakos (1989) and reanalysed by Frydman (1992, Table 1). 262 haemophiliacs, divided into two groups, heavily and lightly treated, were followed up over a period of time after receiving HIV infected blood. Observations are discretised into 6 months intervals. State 1 is infection free, state 2 corresponds to HIV infection and state 3 is the onset of AIDS. According to current mainstream theory, it is not possible to have AIDS without first being HIV and so it is not possible to move directly from state 1 to state 3. Therefore

$$P((k-1,0) \rightarrow (k,2)) = 0,$$

and we can achieve this by defining $\alpha_{k2} = 0$ for all k. For the illustrative results that follow, we take a Bayesian bootstrap approach; that is, we set

the prior parameters to zero. De Gruttola and Lagakos (1989) and Frydman (1992) both analysed the data nonparametrically via *self consistent* estimators (Turnbull, 1976) but the former assumed the times in states 1 and 2 to be independent.

Figure 11 is the estimated cumulative distributions of times to HIV infections for the two groups. These are similar to the results obtained by Frydman. Figures 12 and 13 are, respectively, the estimated marginal cumulative distributions for the onset of AIDS under assumptions (c) and (a). These highlight differences under the two quite valid assumptions.

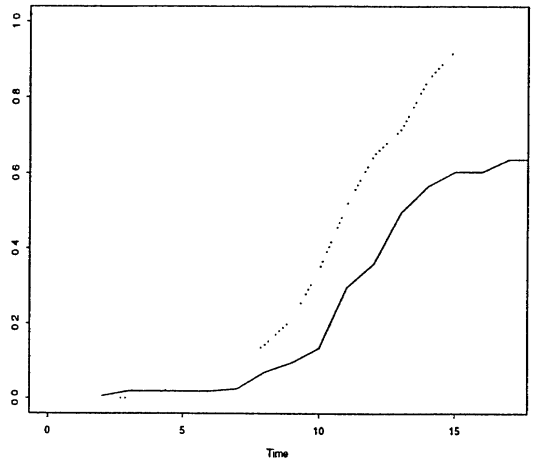

Figure 11: Estimated cumulative distributions of times to HIV infection; lightly treated—-, heavily treated- - -

If there is uncertainty in which assumption, or model, to choose then a possibility is to obtain an estimator which is comprised of a mixture of estimators under the different assumptions. Explicitly this involves taking the estimator $\hat{P}$ given by

$$\hat{P} = \hat{P}_{(a)}\pi\Big(M_{(a)}|\text{data}\Big) + \hat{P}_{(b)}\pi\Big(M_{(b)}|\text{data}\Big) + \hat{P}_{(c)}\pi\Big(M_{(c)}|\text{data}\Big),$$

where $\hat{P}_{(.)}$ is the estimator under $M_{(.)}$ and $\pi(M_{(.)}|\text{data})$ is the posterior weight assigned to $M_{(.)}$, that is,

$$\pi\Big(M_{(.)}|\text{data}\Big) \propto \pi\Big(\text{data}|M_{(.)}\Big) \times \pi\Big(M_{(.)}\Big),$$

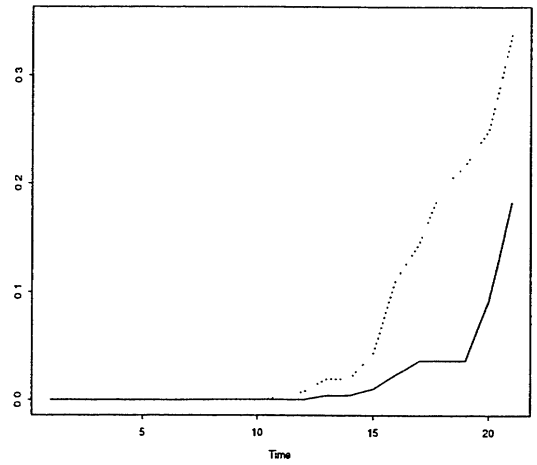

Figure 12: Estimated (marginal) cumulative distributions of times to onset of Aids, assumption (c); lightly treated—-, heavily treated- - -

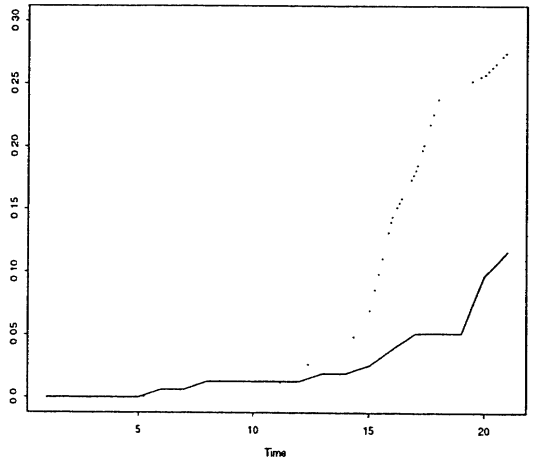

Figure 13: Estimated (marginal) cumulative distributions of times to onset of Aids, assumption (a); lightly treated—-, heavily treated- - -

where $\pi(M_{(.)})$ is the prior weight assigned to model $M_{(.)}$. Therefore to obtain the estimator $\hat{P}$ it only remains to evaluate $\pi(\text{data}|M_{(.)})$. These are in fact straightforward to calculate. It remains to decide on the values of the αs and βs with which to determine the $\pi(\text{data}|M_{(.)})$. First, for large αs and βs the prior specifications should swamp the data and the model. This is the case and for $\alpha = \beta = 10^6$

$$\log\pi\Big(\text{data}|M_{(.)}\Big) = -578.8 \text{ for all } M_{(.)}$$

(we have removed the

$$\prod_k \left[\frac{\alpha_{k0}^{[n_{k0}]}\alpha_{k1}^{[n_{k1}]}}{(\alpha_{k0}+\alpha_{k1})^{[n_{k0}+n_{k1}]}}\right]$$

from $\pi(\text{data}|M_{(.)})$ which is common to all $M_{(.)}$). To represent vague a priori information we consider $\beta_{k2} = \lambda\beta_{k1} = 10^{-6}$, for $\lambda = 1, 10, 100$,

$$\log\pi\Big(\text{data}|M_{(a)}\Big) = -390.3, -337.3, -284.4,$$

$$\log\pi\Big(\text{data}|M_{(b)}\Big) = -260.0, -234.6, -209.3,$$

and

$$\log\pi\Big(\text{data}|M_{(c)}\Big) = -249.2, -226.1, -203.1.$$

As far as Bayes factors are concerned therefore the data support $M_{(c)}$, the Markov model.

6 Discussion

Following a description of the Dirichlet process in Section 2.1, we demonstrated in Section 2.2 the use of the Dirichlet process with respect to functional estimation. We showed how it is possible to incorporate prior information on both the mean and variance on the unknown parameter. In Section 3.2 we briefly considered the MDP model.

In Section 3 stochastic process priors were described, in particular the neutral to the right process. In particular, we showed how to specify and interpret the mean and variance of an unknown survival curve, developed

the full posterior distribution and, via illustrative analysis, implemented a full Bayesian solution using simulation.

In Section 4 we described Polya tree priors for partition models. We demonstrated the use of these priors in modelling errors in both hierarchical and non-hierarchical frameworks. In particular we were able to capture both 'well behaved' and 'badly behaved' distributions.

In Section 5 priors constructed from exchangeable processes were detailed and their use demonstrated in a three state disease process model.

A natural question that arises is how to choose the appropriate approach for a given problem. Due to the difficulties involved in simulating a continuous time stochastic process, we recommend the use of such processes only when interest is in the specific function being modelled as a process, for example, the cumulative distribution/hazard function.

We have found Polya trees particularly appropriate for modelling error distributions in a large class of models, including linear models, generalised linear models and frailty models. In particular, it is straightforward to fix the location (and scale) of a random probability measure chosen from such a prior.

Exchangeable processes are more suited to predictive inference and especially useful in extending the traditional alive/death survival models to incorporate multiple states.

This paper is an exposé of the current state of the art of Bayesian nonparametrics from our perspective. The work is ongoing and a number of problems remain unresolved. In particular, more work is required in the following areas: a full Bayesian nonparametric analysis involving covariate information; multivariate priors based on stochastic processes; multivariate error models involving Polya trees; developing exchangeable processes to cover a larger class of problems; and nonparametric sensitivity analysis (Lenk, 1996).

A further question that arises is the extent to which we currently understand the potential mathematical consequences of the tool-kit we are developing. Diaconis and Freedman (1986) present a nonparametric model that uses a symmetrized Dirichlet prior for the underlying distribution and an independent prior for its median. They then demonstrate that seemingly innocuous choices for the latter lead to an inconsistent Bayes estimate of the median. For the same model, they show other reasonable priors for the median that are consistent. The source of the problem, when it occurs, appears to be the infinite dimensionality of the nuisance parameter. In light of

results such as in Hjort (1990) and Diaconis and Freedman (1993) that give demonstrably consistent nonparametric Bayesian procedures, general theoretical advances that pinpoint the pitfalls would indeed prove valuable. In the interim, we advocate the use of prudent albeit heuristic sensitivity analyses and look forward to more formal developments in this direction that would afford the practitioner a higher degree of assurance.

Acknowledgements

Research reported here was supported in part by an EPSRC ROPA and travel grant, an NSF grant, and financial support from the Business School at the University of Michigan, Ann Arbor.

References

Amman, L. (1984). Bayesian nonparametric inference for quantal response data. *Ann. Statist.* **12**, 636-645.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* 2, 1152-1174.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica* **14**, 505-524.

Blackwell, D. (1973). The discreteness of Ferguson selections. *Ann. Statist.* **1**, 356-358.

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Polya-urn schemes. *Ann. Statist.* **1**, 353-355.

Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Adv. App. Prob.* **14**, 855-869.

Bush, C.A. and MacEarchen, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275-286.

Christensen, R. and Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika* **75**, 693-704.

Clayton, D.G. and Cuzick, J. (1985). Multivariate generalisations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc., Ser. A* **148**, 82-117.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467-485.

Coppersmith, D., and Diaconis, P. (1987). Random walk with reinforcement. *Unpublished manuscript.*

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc., Ser B* **34**, 187-220.

Crowder, M.J. (1978). Beta-binomial Anova for proportions. *Appl. Statist.* **27**, 34-47.

Dalal, S.R. (1978). A note on the adequacy of mixtures of Dirichlet processes. *Sankhya* **40**, 185-191.

Damien, P., Laud, P.W. and Smith, A.F.M. (1995). Random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. Roy. Statist. Soc., Ser. B* **57**, 547-564.

Damien, P. Laud, P.W. and Smith, A.F.M. (1996). Implementation of Bayesian nonparametric inference based on beta processes. *Scand. J. Statist.* **23**, 27-36.

Damien, P. and Walker, S.G. (1996). Sampling probability densities via uniform random variables and a Gibbs sampler. *Submitted for publication.*

de Finetti, B. (1937). La prevision: Ses lois logiques, ses sources subjectives. *Ann. l'Instit. H. Poincaré* **7**, 1-68.

De Gruttola, V. and Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**, 1-11.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1-26.

Diaconis and Freedman (1993). Nonparametric binary regression : a Bayesian approach. *Ann. Statist.* **21**, 2108-2137.

Doksum, K.A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183-201.

Dubins, L. and Freedman, D. (1965). Random distribution functions. *Bull. Amer. Math. Soc.* **69**, 548-551.

Dystra, R.L. and Laud, P.W.. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356-367.

Escobar, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Assoc.* **89**, 268-277.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577-588.

Fabius, J. (1964). Asymptotic behaviour of Bayes estimates. *Ann. Math. Statist.* **35**, 846-856.
Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.
Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.
Ferguson, T.S. and Phadia, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-186.
Freedman, D.A. (1963). On the asymptotic behaviour of Bayes estimates in the discrete case I. *Ann. Math. Statist.* **34**, 1386-1403.
Freedman, D.A. (1965). On the asymptotic behaviour of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **36**, 454-456.
Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *J. Roy. Statist. Soc. B,* **54**, 853-866.
Ghosh, J.K. and Mukerjee, R. (1992). Noninformative priors. In *Bayesian Statistics 4.* J.M.Bernardo,J.O.Berger,A.P.Dawid and A.F.M.Smith (Eds.) Oxford University Press.
Gill, R.D. and Johansen, S. (1990). A survey of product integration with a view toward application in survival analysis. *Ann. Statist.* **18**, 1501-1555.
Hewitt, E. and Savage, L.J. (1955). Symmetric measures on cartesian products. *Trans. Amer. Math. Soc.* **80**, 470-501.
Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.
Hjort, N.L. (1996). Bayesian approaches to non- and semiparametric density estimation. In *Bayesian Statistics 5.* J.M.Bernardo,J.O.Berger,A.P.Dawid and A.F.M.Smith (Eds.) Oxford University Press.
Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc., Ser. B* **40**, 214-221.
Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457-481.
Laud, P.W., Damien, P. and Smith, A.F.M. (1993). Random variate generation from D-distributions. *Statist. Comput.* **3**, 109-112.
Laud, P.W., Smith, A.F.M. and Damien, P. (1996 a). Monte Carlo methods for approximating a posterior hazard rate process. *Statist. Comput.* **6**, 77-84.
Laud, P.W., Smith, A.F.M. and Damien, P. (1996 b). Bayesian nonparamet-

ric and covariate analysis of failure time data. *Submitted for publication.*
Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1203-1221.
Lavine, M. (1994). More aspects of Polya trees for statistical modelling. *Ann. Statist.* **22**, 1161-1176.
Lenk, P.J. (1996) Bayesian inference of semiparametric regression and Poisson intensity functions. *Submitted for publication.*
Lévy, P. (1936). *Theorie de l'addition des variables aleatoire.* Gauthiers-Villars, Paris.
Lo, A.Y. (1993). A Bayesian bootstrap for censored data. *Ann. Statist.* **21**, 100-123.
MacEarchen, S.N. and Mueller, P. (1994). Estimating mixtures of Dirichlet process models. *Unpublished manuscript.*
Mauldin, R.D., Sudderth, W.D, and Williams, S.C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203-1221.
Mueller,P.,Erkanli,A. and West,M. (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.
Muliere, P. and Walker, S.G. (1997a). A Bayesian nonparametric approach to survival analysis using Polya trees. *Scand. J. Statist.* To appear.
Muliere, P. and Walker, S.G. (1997b). Extending the family of Bayesian bootstraps and exchangeable urn schemes. *J. Roy. Statist. Soc., Ser. B.* To appear.
Pemantle, R. (1988). Phase transitions in reinforced random walk and RWRE on trees. *Ann. Probab.* **16**, 1229-1241.
Rubin, D.B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130-134.
Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Proceedings of the third Purdue symposium on statistical decision theory and related topics.* Gupta,S.S. and Berger,J.O. (Eds.). Academic press, New York.
Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc., Ser. B* **55**, 3-24.
Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete data. *J. Am. Statist. Assoc.* **71**, 897-902.
Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.
Turnbull,B.W. (1976). The empirical distribution function with arbitrarily

grouped, censored, and truncated data. *J. Roy. Statist. Soc., Ser. B* **38**, 290-295.
Walker, S.G. and Damien, P. (1996). A full Bayesian nonparametric analysis involving a neutral to the right process. *Submitted for publication.*
Walker, S.G. and Mallick, B.K. (1996). A Bayesian semiparametric accelerated failure time model. Revised for *J. Am. Statist. Assoc.*
Walker, S.G. and Mallick, B.K. (1997). Hierarchical generalised linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc., Ser. B.* To appear.
Walker, S.G. and Muliere, P. (1997). Beta-Stacy processes and a generalisation of the Polya-urn scheme. *Ann. Statist.* To appear.
West, M.,Muller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A tribute to D.V.Lindley.* A.F.M.Smith and P.Freeman (Eds.)
Ying, Z., Jung, S.H. and Wei, L.J. (1995). Survival analysis with median regression models. *J. Am. Statist. Assoc.* **90**, 178-184.

Appendix

Here we give a brief outline the MCMC algorithms for the models described in Section 4.4; for full details see the references cited in the main text. We will concentrate on the hierarchical generalised linear model since this is the most complex. Recall the first stage binomial model given by

$$Y_{ij}|(Z_{ij} = z_{ij}) \sim \text{binomial}(z_{ij}, n_{ij}),$$

and that, given the $\{Z_{ij}\}$, the $\{Y_{ij}\}$ are mutually independent. The second stage is given by

$$\text{logit}(Z_{ij}) = X_{ij}\beta + \Theta_{ij},$$

where

$$\Theta_{1n_1}, ..., \Theta_{nn_n}|F \sim_{\text{iid}} F.$$

Finally, the prior specifications are that

$$F \sim PT(\Pi, \mathcal{A})$$

and a normal prior with zero mean and large variance term is taken for β.

Samples from the relevant full conditional distributions, including

$$p(F|Y,Z,\beta),\ p(Z_{ij}|Y_{ij},F,\beta),\ \text{and}\ p(\beta|Y,Z,F),$$

are obtained using an MCMC algorithm and, in particular, a Metropolis/Hastings/ Gibbs method. The full conditional distribution for F is a Polya tree which has been updated, to give the posterior Polya tree, obtained from the $\sum_i n_i$ iid observations

$$\Theta_{ij} = \text{logit}(Z_{ij}) - X_{ij}\beta.$$

A random F_M is then taken as described in Section 4.2.

Sampling from the full conditional of β proceeds as follows. Recall that F is sampled to the level M giving F_M as a sequence of weights, $\{W_k : k = 1,...,2^M\}$, on the $r = 2^M$ sets at level M, say $A_1,...,A_r$. The likelihood of β, given Z and F_M, is

$$l(\beta|Z,F_M) = \prod_{ij} W_k(\beta,ij),$$

where $W_k(\beta,ij) = F_M(A_k(\beta,ij))$ and $\text{logit}(Z_{ij}) - X_{ij}\beta \in A_k(\beta,ij)$. With this likelihood established, a Metropolis/Hastings step can be used, after some preliminary work to establish a good proposal distribution, to sample the full conditional for β. An identical approach can be used for sampling the full conditional distributions of Z_{ij}.

An alternative method for sampling the full conditionals for β and Θ_{ij} (replacing Z_{ij}) uses latent variables. The joint distribution for β and Θ_{ij}, given F, is given, up to a constant of proportionality, by

$$f(\beta,\Theta_{ij}|F) \propto \frac{\exp(y_{ij}X_{ij}\beta + y_{ij}\Theta_{ij})}{\{1+\exp(X_{ij}\beta+\Theta_{ij})\}^{n_{ij}}} p(\Theta_{ij}|F)p(\beta).$$

We can write this in another way by introducing the latent variables U_{ij} and V_{ij} and defining the joint distribution

$$f(\beta,\Theta_{ij},U_{ij},V_{ij}|F) \propto p(\beta)\times$$

$$\exp\left(-u_{ij}y_{ij} - v_{ij}(n_{ij}-y_{ij})\right) I\left(u_{ij} > \log(1+e^{-Z_{ij}}), v_{ij} > \log(1+e^{Z_{ij}})\right) p(\Theta_{ij}|F),$$

where now $Z_{ij} = X_{ij}\beta + \Theta_{ij}$. It is easy to see that the marginal distribution for β and Θ_{ij} is as required. This eases the Gibbs sampler since the full conditionals for β and Θ_{ij} are now of known types, albeit restricted to particular sets (Damien and Walker, 1996).