J. R. Statist. Soc. B (1999)
61, *Part* 3, *pp.* 485–527

Bayesian nonparametric inference for random distributions and related functions

Stephen G. Walker,

Imperial College of Science, Technology and Medicine, London, UK

Paul Damien,

University of Michigan, Ann Arbor, USA

Purushottam W. Laud

Medical College of Wisconsin, Milwaukee, USA

and Adrian F. M. Smith

Queen Mary and Westfield College, London, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, November 11th, 1998*, Professor P. J. Green *in the Chair*]

Summary. 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 a mystery to practitioners and many open questions remain. In this paper, we discuss and illustrate the rich modelling and analytic possibilities that are available to the statistician within the Bayesian nonparametric and/or semiparametric framework.

Keywords: Dirichlet process; Exchangeability; Lévy process; Neutral to the right process; Pólya tree; Survival model

1. Introduction

1.1. 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 is 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.

Common population distributions, such as the normal and Weibull distributions, force certain assumptions concerning the underlying population: in particular, the assumptions of unimodality and an implicit inability to model population moments higher than the first two. Few statisticians would argue that this is sufficient for the analysis of complex data sets. One possibility is to search for more flexible parametric population distributions: for example, the exponential power and Student t-families have been proposed as generalizations of the normal family in Bayesian robustness studies. However, these families do not cover departures from symmetry.

Address for correspondence: Stephen G. Walker, Department of Mathematics, Huxley Building, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London, SW7 2BZ, UK.
E-mail: s.walker@ic.ac.uk

 1369–7412/99/61485

Another possibility is to turn to finite mixture distributions (Titterington *et al.*, 1985). Analyses involving such mixtures have recently received increased attention because of the advances made in simulation-based approaches to making inference (see, for example, Richardson and Green (1997) and the references cited in their paper), particularly in the Bayesian framework. Essentially, priors are constructed on a larger class of population distributions, achieved via the introduction of a larger number of parameters. The problem of working with mixtures where the number of components is taken to be unknown (random) was previously tackled by Escobar and West (1995) using 'Bayesian nonparametrics', basing the prior on the Dirichlet process.

Classical nonparametric and semiparametric methods have a measure of popularity, e.g. the Kaplan–Meier estimator, kernel density estimation and Cox regression. No population distributional assumptions are made in any of these cases, except for the proportional hazards assumption in the case of Cox regression. We argue that a state of no knowledge at all is hardly, if ever, realistic: we would typically at least have some ideas concerning location and spread. Such information can be incorporated into a Bayesian nonparametric prior. Even if there really is no information of worth, we can still construct relatively uninformative nonparametric priors, in which case inference should mimic classical nonparametric results. An example of this is the (Bayesian) nonparametric generalization of the Kaplan–Meier estimator.

Motivated by the success of the Dirichlet process prior in the important problem of a 'random number of component mixture distributions', the present paper is concerned with looking at alternative nonparametric priors (which generalize the Dirichlet process) and seeking to use them in some important areas in statistics. Bayesian nonparametric models are constructed on 'large' spaces to provide support for more eventualities than are supported by a parametric model. Technically, (to many) the off-putting aspect of the Bayesian nonparametric framework is the mathematical apparatus that is 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 knowledge into this mathematical framework. A major emphasis of this paper will therefore be an attempt to address these issues and to 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 described by Freedman (1963) who introduced tail-free and Dirichlet random measures. Subsequently, Dubins and Freedman (1965), Fabius (1964), Freedman (1965) and Ferguson (1973, 1974) formalized and explored the notion of a Dirichlet process. Early work was largely focused on stylized 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 by 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 nonparametric or semiparametric density estimation; for a recent survey of this field, see Hjort (1996). Nor shall we deal with Bayesian nonparametric regression, using, for example, random functions generated by random coefficients for a set of bases functions (see, for example, Denison *et al.* (1998)). A recent collection of Bayesian nonparametric and semiparametric papers can be found in Dey *et al.* (1998).

1.2. Outline of the paper

The paper is organized as follows. In Section 2 we summarize the fundamental 'Bayesian

nonparametric theorem'. In Section 2.1 we review the well-known Dirichlet process prior and in Section 2.2 we motivate the use of more general priors. Detailed descriptions of these more general 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 generalize the classical Kaplan and Meier (1958) nonparametric estimator. Also in Section 3.4 we consider Bayesian semiparametric approaches for the proportional hazards model (Cox, 1972). In Section 4.4 we consider an accelerated failure time model and frailty models (Clayton and Cuzick, 1985). In Section 5.4, we consider a three-state disease process model.

All the examples presented in the paper involve the analysis of data, previously studied under assumptions that are different from those made by us. Every analysis depends on assumptions about the relevant unknown function (a probability distribution or related function for the examples considered in this paper). In a Bayesian nonparametric approach we can ensure that the first two moments of the unknown function match those derived from a parametric model; see, for example, Section 3.2. This effectively creates a region in which the function is thought to be located which is the same for both parametric and nonparametric cases. The difference is that in the parametric case the shape of the unknown function is restricted whereas in the nonparametric case it is not.

2. General framework

We assume that $Y_1, Y_2, \ldots$, defined on some space Ω, is a sequence of independent and identically distributed (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 characterized 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 'centred', in some sense, on P_Θ.

The following provides the key mathematical basis for Bayesian nonparametric constructions.

Theorem 1 (Ferguson, 1973; Doksum, 1974; Dalal, 1978). Let $(\Omega, \mathcal{B})$ be a measurable space and let a system of finite dimensional distributions for

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

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

(a) $F(B)$ is a random variable on (0, 1) for all $B \in \mathcal{B}$,
(b) $F(\Omega) = 1$ almost surely and
(c)

$$(F(\cup_i B_{1,i}), \ldots, F(\cup_i B_{m,i})) \overset{\mathrm{d}}{=} \left(\sum_i F(B_{1,i}), \ldots, \sum_i F(B_{m,i}) \right)$$

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

An important seminal version of a nonparametric prior is the Dirichlet process (Ferguson, 1973, 1974), arising when the finite dimensional distributions are Dirichlet distributions. It turns out that this process has several deficiencies, but since all our nonparametric priors are generalizations of the Dirichlet process we begin by providing a brief review.

2.1. The Dirichlet process

The Dirichlet process 'generates' discrete random probability measures. The parameters of the Dirichlet process prior can be chosen so that the expected probability measure is arbitrary, say F_0. The other parameter is a scalar quantity $c > 0$, commonly interpreted as controlling the variability of the random probability measures F about F_0. We write $F \sim \mathcal{D}(cF_0)$ and the finite dimensional distribution for a measurable partition $(B_1, \ldots, B_k)$ is

$$\text{Dirichlet}\{c\,F_0(B_1), \ldots, c\,F_0(B_k)\}.$$

An immediate difficult question is whether the 'simple' Dirichlet distribution is useful bearing in mind that it assigns negative correlation between $F(B_j)$ and $F(B_l)$ for all $j \neq l$, which is counter-intuitive.

A further unsatisfactory aspect of the Dirichlet process is the role played by c. There is no clear interpretation for this parameter, owing to its dual aspect, 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, for any event A,

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

For *maximum* variability we would want $c \to 0$. However, Sethuraman and Tiwari (1982) pointed 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 $\text{var}(F)$ arbitrarily, and that the shape is determined by F_0.

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. 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 function of the observations. The naïve 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 $c = 0$, we note that the Bayes estimate for F, with respect to quadratic loss, is given by F_n which is the classical nonparametric frequentist estimator. Therefore, $c = 0$ fits in with one of the notions of a non-informative 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 (1992) is that of 'information'. Under this notion, $c = 0$ can definitely not be thought of as providing a 'non-informative' prior. As mentioned earlier, as $c \to 0$, F converges to a single atomic measure, which is strong information about the discreteness of F.

Although an experimenter may not be able to formulate a parametric model for F, he or she may have information concerning the mean and variance of F, μ and σ^2 respectively

(obviously assuming that they exist). If priors can be allocated for these parameters then coherent specifications for the Dirichlet prior involve $c = E(\sigma^2)/\mathrm{var}(\mu)$, $E(Y_0) = E(\mu)$ and $\mathrm{var}(Y_0) = E(\sigma^2) + \mathrm{var}(\mu)$, where $Y_0 \sim F_0$. Antoniak (1974) considered 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.

2.1.1. Mixture of Dirichlet process model

As mentioned earlier, one problem with the Dirichlet process is that it assigns probability 1 to the space of *discrete* probability measures. A class of priors that chooses a continuous F with probability 1 is the mixture of Dirichlet process (MDP) model, which we now briefly discuss.

MDP models are essentially Bayesian hierarchical models, one of the simplest versions taking the form considered by Lo (1984):

$$Y_i|\theta_i \sim f(\cdot|\theta_i), \qquad i = 1, \ldots, n,$$

$$\theta_1, \ldots, \theta_n|F \underset{\mathrm{IID}}{\sim} F \quad \text{and} \quad F \sim \mathcal{D}(cF_0).$$

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 centred on a parametric distribution. For applications of MDP models, see, for example, Escobar (1994), Escobar and West (1995), West *et al.* (1994), Mueller *et al.* (1996), Bush and MacEachern (1996) and MacEachern and Mueller (1998), in which priors are also assigned to c and the parameters of F_0.

MDP models have largely dominated the Bayesian nonparametric literature recently as a consequence of the realization that full posterior computation is feasible by using simulation methods (Escobar, 1994), although these can be very computer intensive and involve non-trivial sampling algorithms (particularly when $f(\cdot|\theta)$ and $F_0(\theta)$ form a non-conjugate pair). The MDP model provides a *continuous* nonparametric prior for the distribution of the Y_is. Constructively, if $F \sim \mathcal{D}(cF_0)$,

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

(Sethuraman and Tiwari, 1982; Sethuraman, 1994), where δ_θ is the measure with mass 1 at θ, leading to

$$Y_i \underset{\mathrm{IID}}{\sim} \sum_{j=1}^{\infty} V_j f(\cdot|\theta_j),$$

where $V_j = W_j(1 - W_{j-1}) \ldots (1 - W_1)$, $W_j \sim_{\mathrm{IID}} \mathrm{beta}(1, c)$ and $\theta_j \sim_{\mathrm{IID}} F_0$. This mixture model has been successfully exploited by Escobar and West (1995) and others. A further use of the constructive definition of the Dirichlet process is given by Doss (1995).

2.2. 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 focus on generalizations of the Dirichlet prior which overcome these difficulties. There are several ways of constructing a nonparametric prior to meet the requirements of theorem 1.

2.2.1. Stochastic processes

The stochastic process approach is particularly appropriate for generating random cumulative density functions on $(0, \infty)$ with application in survival data models. An important and rich class of priors is provided by *neutral to the right* (NTTR) processes (Doksum, 1974), where the distribution function is represented in the form $F(t) = 1 - \exp\{-Z(t)\}$, where Z is an independent increments (Lévy) process on $(0, \infty)$, with $Z(0) = 0$ and $\lim_{t\to\infty}\{Z(t)\} = \infty$. We shall illustrate this approach in Section 3.4 with the analysis of the well-known Kaplan and Meier (1958) data set.

2.2.2. Partitioning

In partitioning 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 $B_0 \cap B_1 = \varnothing$ 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, $F(B_{\epsilon 0}) + F(B_{\epsilon 1}) = F(B_\epsilon) \geqslant 0$ and $F(\Omega) = 1$. This is the idea behind Pólya 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, because it is easy to fix the location (median) of a random Pólya tree distribution. An application considered later in Section 4.4 includes an accelerated failure time model.

2.2.3. Exchangeability

Rather than constructing F directly, as in the stochastic process and partitioning approaches, 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, i.e. in providing the distribution of Y_{n+1} given $Y_1, \ldots, Y_n$. We illustrate this approach in Section 5.2 with an application involving 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 NTTR 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, an NTTR process called the beta-Stacy process 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).

A non-decreasing almost surely, right continuous almost surely, process $Z(t)$, with independent increments, is called an NTTR Lévy process if it satisfies

(a) $Z(0) = 0$ almost surely and
(b) $\lim_{t\to\infty}\{Z(t)\} = \infty$ almost surely.

$Z(t)$ has at most countably many fixed points of discontinuity. If $t_1, t_2, \ldots$ correspond to the fixed points of discontinuity having independent jumps $W_1, W_2 \ldots$ then the difference

$$Z_c(t) = Z(t) - \sum_j W_j I_{[t_j,\infty)}(t),$$

where $I(\cdot)$ is the indicator function, is a non-decreasing, independent increments process *without* fixed points of discontinuity. Hence, every NTTR 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 an NTTR process. In short, a random distribution function $F(t)$ on the real line is NTTR if it can be expressed as

$$F(t) = 1 - \exp\{-Z(t)\},$$

where $Z(t)$ is an NTTR Lévy process. We shall concentrate on the beta-Stacy process (Walker and Muliere, 1997) which generalizes the Dirichlet process and the simple homogeneous process. The Lévy measure for the beta-Stacy process is given by

$$\mathrm{d}N_t(z) = \frac{\mathrm{d}z}{1 - \exp(-z)} \int_0^t \exp\{-z\beta(s)\}\, \mathrm{d}\alpha(s),$$

for appropriate functions $\alpha(\cdot)$ and $\beta(\cdot)$.

3.2. Prior specifications

Ferguson and Phadia (1979) pointed out that for the NTTR processes which they considered, such as the gamma, simple homogeneous and Dirichlet processes, interpreting the prior parameters is quite difficult. Walker and Damien (1998) 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, i.e. we can find a beta-Stacy process to satisfy

$$-\log\{E[S(t)]\} = \mu(t) \quad \text{and} \quad -\log\{E[S^2(t)]\} = \lambda(t),$$

where $S(t) = 1 - F(t)$, for arbitrary μ and λ, satisfying $\mu < \lambda < 2\mu$, a consequence of the inequality $E[S]^2 < E[S^2] < E[S]$. The parameters of the beta-Stacy process are given by

$$\beta(t) = \{\xi(t) - 1\}/\{2 - \xi(t)\}$$

and $\mathrm{d}\alpha(t) = \beta(t)\, \mathrm{d}\mu(t)$, where $\xi = \mathrm{d}\lambda/\mathrm{d}\mu$. We obtain the Dirichlet process when $\beta(t) = \alpha(t, \infty)$ and we obtain the simple homogeneous process when β is a constant. The infinitesimal jumps of the beta-Stacy process follow a generalized Dirichlet distribution (Connor and Mosimann, 1969). With the constraint $\beta(t) = \alpha(t, \infty)$ the generalized Dirichlet becomes a Dirichlet distribution. These special cases might be seen as a desire to have $\mathrm{d}N_t(z)$ in closed form.

One way to provide a μ and λ is via a Bayesian parametric model. Suppose, for example, that we wish to match the nonparametric model, up to and including second moments, with the parametric exponential–gamma model, i.e. $S(t) = \exp(-at)$ with $a \sim \mathrm{ga}(p, q)$. Then we would have $\mu(t) = p \log(1 + t/q)$ and $\lambda(t) = p \log(1 + 2t/q)$. This method of specifying the prior mean and variance of the distribution function overcomes the difficulties in interpretation that were 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, i.e. 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. Further examples are given in Walker and Damien (1998).

3.3. Posterior distributions

The following establishes a key 'conjugacy' property of NTTR processes.

Theorem 2 (Doksum, 1974; Ferguson, 1974). If F is NTTR and there is a random sample from F, some of which may be right censored, then the posterior distribution of F is NTTR.

If F is a beta-Stacy process with parameters α and β then, given an IID sample from F, and/or possible right censoring, the posterior process is also beta-Stacy. The Dirichlet process is not conjugate with respect to right-censored data, and if the prior process is Dirichlet the posterior, given the presence of censored data, is beta-Stacy. The Bayes estimate for F, with respect to a quadratic loss function, is given by

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

where $N(t) = \Sigma_i I(Y_i \leqslant t)$, $M(t) = \Sigma_i I(Y_i \geqslant t)$ and $\Pi_{[0,t]}$ represents a product integral (Gill and Johansen, 1990). This estimator was first obtained by Hjort (1990). The estimator $\hat{F}$ provides the parameter update explicitly. 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, 1998).

The remaining key question is whether prior-to-posterior calculations for these models are computationally feasible. The posterior NTTR process Z splits into two independent parts: a set of fixed points of discontinuity, which occur where the uncensored observations occur, and a Lévy process without fixed points of discontinuity, Z_c. The Lévy measure for Z_c is the same as for the prior except that β is replaced by $\beta + M$. We can write

$$Z(t) = \sum_{Y_i \text{ uncensored}} W_{Y_i} I(Y_i \leqslant t) + Z_c(t),$$

where the jumps W_{Y_i} are described in detail below. From a simulation perspective it is sufficient to generate random variates from these two components separately and independently.

3.3.1. Simulating the jump component

With respect to a beta-Stacy process, let W_y denote the jump random variable corresponding to an uncensored observation at y. The density function for W_y is given by

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

so that $W_y =_{\mathrm{d}} -\log(1 - B_y)$, where $B_y \sim \text{beta}[N\{y\}, \beta(y) + M(y) - N\{y\}]$. See Walker and Muliere (1997). If $N\{y\} = 1$ then W_y has an exponential density with mean value $\{\beta(y) + M(y) - 1\}^{-1}$. Simulating the jump component is thus straightforward.

3.3.2. Simulating the continuous component

It is well known (Ferguson, 1974; Damien *et al.*, 1995) that $Z_c(t)$ will have an *infinitely divisible* (ID) distribution. Bondesson (1982), Damien *et al.* (1995) and Walker and Damien (1998) have developed algorithms to generate random variates from a large class of ID distributions. The particular choice of the algorithm might depend on the posterior process under consideration. Thus, Laud *et al.* (1996) used the Bondesson algorithm to simulate the extended gamma process; the algorithm of Damien *et al.* (1995) is exemplified for the

Dirichlet, gamma and the simple homogeneous processes; Walker and Damien (1998) provide a full Bayesian analysis for the beta-Stacy process using a hybrid of algorithms, based on an idea in Bondesson (1982). Wolpert and Ickstadt (1998) provide an algorithm for sampling the entire Lévy process; see also Ferguson and Klass (1972). For the illustrative analyses that involve NTTR processes, we shall use the Walker and Damien method.

3.4. Example

We reanalyse the data set of Kaplan and Meier (1958), 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 and 9.2 months, and censored observations at 1.0, 2.7, 7.0 and 12.1 months. For illustration, we address the problem of estimating the probability of failure before 1 month, i.e. $F(0, 1)$. Whereas Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979) could only obtain Bayesian point estimates, we can sample from the full posterior distribution. Also, we can sample from the posterior distribution of $F(0, t)$ for any t and can therefore 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

$$\mathrm{d}\alpha(s) = 0.1\ \exp(-0.1s)\,\mathrm{d}s,$$

to provide correspondence with the prior used by them. The prior is actually a Dirichlet process but, with censored observations in the data set, the posterior is a beta-Stacy process. We assume that there are no jumps in the prior process. Now

$$F(0, 1) = 1 - \exp\{-Z(0, 1)\} = 1 - \exp\{-Z_{\mathrm{c}}(0, 0.8) - W_{0.8} - Z_{\mathrm{c}}(0.8, 1)\}.$$

Therefore, we need to sample $W_{0.8}$, an exponential random variable with mean $\{\exp(-0.08) + 7\}^{-1}$, and $Z_{\mathrm{c}}(0, 0.8)$ and $Z_{\mathrm{c}}(0.8, 1)$, for which we use the algorithm described in Walker and Damien (1998). This algorithm involves sampling a Poisson process, based on

$$Z_{\mathrm{c}}(t) \underset{\mathrm{d}}{=} \int s\,\mathrm{d}P_t(s),$$

where P_t is the Poisson process which has intensity the posterior Lévy measure.

We collected 10000 samples from the posterior and the resulting histogram representation is given in Fig. 1. The mean value is given by 0.12 which is the (exact) point estimate value obtained by Ferguson and Phadia (1979).

It can be argued that the prior used in this illustrative analysis seems somewhat informative. Can we recapture the same shape by using the flexible, less informative prior that we proposed in Section 3.2? To investigate this, we reanalyse the data set by using the Bayesian parametric model described in Section 3.2 with $p = q = 1$, in an attempt to be 'relatively non-informative'. This corresponds to $\beta(t) = 1/2t$ and

$$\mathrm{d}\alpha(t) = \mathrm{d}t/2t(1 + t).$$

Again, we collected 10000 samples from the posterior and it turns out that the resulting histogram representation is essentially indistinguishable from the histogram in Fig. 1 (we do not include it).

It is of interest to see how our nonparametric analysis compares with a parametric analysis using the parametric model on which it is centred. The posterior distribution from

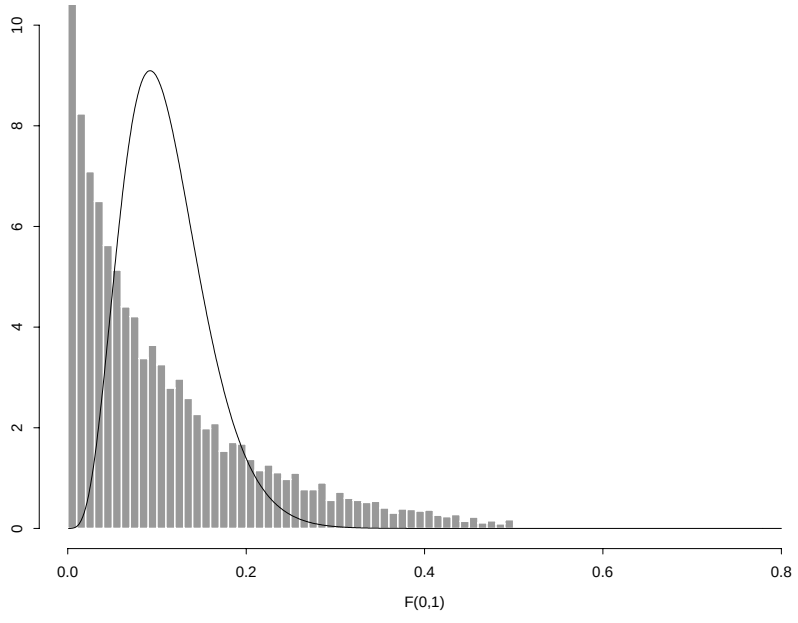

Fig. 1. Histogram representation of the posterior density of $F(0, 1)$, using a Dirichlet process prior and posterior using a parametric prior (———): Kaplan–Meier data set

the parametric model is given by $F(0, 1) = 1 - \exp(-a)$ with $a \sim \text{ga}(1 + 4, 1 + 41.3)$. We can construct this density analytically and it is shown as the curve in Fig. 1. The posterior inferences are fundamentally different, clearly showing the effects of the parametric assumption.

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 considerably more flexibility in the model and the two posteriors in Fig. 1 show very clearly the extent to which a parametric assumption can force a posterior form. With the nonparametric approach, we note that the less informative nonparametric prior leads to essentially the same result as the informative nonparametric prior. This is typical. The significant differences are between the parametric and nonparametric approaches, rather than between choices of prior within the nonparametric framework.

This example highlights a feature: we obtain similar posterior means for the two inferences (0.12 for nonparametric and 0.11 for parametric), but with appropriately greater ranges of uncertainty for the nonparametric approach (a standard deviation of 0.10 for nonparametric and 0.05 for parametric), since the assumption of a particular parametric form with probability 1 is artificially suppressing one element of uncertainty.

What else can be done with the stochastic process approach? Kalbfleisch (1978), Clayton (1991) and Laud *et al.* (1998) have provided 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). We could alternatively use the Z-process (described in Section 3.1) as a prior for the cumulative hazard function (Hjort used processes of the type $\mathrm{d}A = 1 - \exp(-\mathrm{d}Z)$). This approach was originally suggested by Laud (1977). Wild and Kalbfleisch (1981) considered the Cox regression model $\mathrm{d}Z_i(t) = \mathrm{d}Z(t)\exp(X_i\beta)$ and Walker *et al.* (1998) developed the idea to cover time-varying covariate models,

$$\mathrm{d}Z_i(t) = \mathrm{d}Z(t)\exp\{X_i(t)\beta\}.$$

Even here, the analysis is not overcomplicated because each Z_i remains a Lévy process. In particular, if Z is an extended gamma process (Dykstra and Laud, 1981) then so is each Z_i.

Dykstra and Laud (1981) considered modelling monotone hazard rates nonparametrically by using the extended gamma process. The advantage of this process is that it indexes the class of absolutely continuous functions with probability 1. Laud *et al.* (1993, 1996) developed simulation methods for the extended gamma process; Amman (1984) extended the hazard rate process to model bathtub hazard rates. Arjas and Gasbarra (1994) developed processes to model the hazard rate piecewise.

In practice, the stochastic process approach is only easy to use for relatively simple models of the kind that we have illustrated. Sampling-based 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 then 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. Pólya 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 (1997). The Pólya tree prior relies on a binary tree partitioning of the space Ω. There are two aspects to a Pólya tree: a binary tree partition of Ω and a non-negative 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} and 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 $B_{\epsilon 0} \cap B_{\epsilon 1} = \varnothing$ 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 Pólya trees, these probabilities are random and assumed to be beta variables, $(C_{\epsilon 0}, C_{\epsilon 1}) \sim \mathrm{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ with non-negative $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$. If we denote the collection of αs by $\mathcal{A} = \{\alpha_\epsilon\}$, a particular Pólya tree distribution is completely defined by Π and $\mathcal{A}$. The spot where our hypothetical particle lands is a random observation from the prior predictive.

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

(a) all the random variables in $\mathcal{C}$ are independent,

(b) for every ϵ, $C_{\epsilon 0} \sim \text{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 0})$ and
(c) for every $m = 1, 2, \ldots$ and every $\epsilon = \epsilon_1 \ldots \epsilon_m$,

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

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

A Pólya tree prior can be set to assign probability 1 to continuous distributions, unlike the Dirichlet process which has sample distribution functions which are discrete with probability 1. Additionally, the correlation structure between bins is more reasonable than it is with the Dirichlet distribution.

4.2. Prior specifications and computational issues

Problems tackled in this paper involving Pólya trees require simulating a random probability measure $F \sim \text{PT}(\Pi, \mathcal{A})$. This is done by sampling $\mathcal{C}$ using the constructive form given in Section 4.1. Since $\mathcal{C}$ is an infinite set an approximate probability measure from $\text{PT}(\Pi, \mathcal{A})$ 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) referred to as a 'partially specified Pólya tree'). From the sampled variates of $\mathcal{C}_M$ we define F_M by $F(B_{\epsilon_1 \ldots \epsilon_M})$ for each $\epsilon = \epsilon_1 \ldots \epsilon_M$. 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 Pólya 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 $B_0 = (-\infty, F_0^{-1}(\frac{1}{2}))$, $B_1 = [F_0^{-1}(\frac{1}{2}), \infty)$ and, at level m, setting, for $j = 1, \ldots, 2^m$,

$$B_{mj} = [F_0^{-1}\{(j-1)/2^m\}, F_0^{-1}(j/2^m)),$$

with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$, where $(B_{mj} : j = 1, \ldots, 2^m)$ correspond, in order, to the 2^m partitions of level m. It is then straightforward to show that $E[F(B_\epsilon)] = F_0(B_\epsilon)$ for all ϵ.

In practice, we may not wish to assign a separate α_ϵ for each ϵ. It may be convenient 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'. 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). The 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 specified choice of c.

Another idea (our preferred choice) is to define the α_ϵ to match $E_{\text{PT}}[F(B_\epsilon)]$ and $E_{\text{PT}}[F^2(B_\epsilon)]$ with those obtained from a parametric model, based on the idea discussed in Section 3.2. If the parametric model has likelihood $F_0(\cdot; \theta)$ and prior $\pi(\theta)$, then we would assign

$$\alpha_{\epsilon 0} = \frac{\mu_{\epsilon 0}(s_\epsilon - s_{\epsilon 0})}{s_{\epsilon 0}\mu_\epsilon - s_\epsilon \mu_{\epsilon 0}}$$

and

$$\alpha_{\epsilon 1} = \alpha_{\epsilon 0}(\mu_\epsilon/\mu_{\epsilon 0} - 1),$$

where $\mu_\epsilon = \int F_0(B_\epsilon;\, \theta)\, \pi(\theta)\, \mathrm{d}\theta$, $s_\epsilon = \upsilon_\epsilon/\mu_\epsilon$ and $\upsilon_\epsilon = \int F_0^2(B_\epsilon;\, \theta)\, \pi(\theta)\, \mathrm{d}\theta$, with

$$\alpha_0 = \frac{\mu_0 s_0 - \mu_0}{\mu_0 - s_0}$$

and

$$\alpha_1 = \alpha_0(1/\mu_0 - 1).$$

If analytic expressions for these α_ϵ are not available, we can evaluate them via Monte Carlo integration.

4.3. Posterior distributions

Consider a Pólya tree prior PT(Π, $\mathcal{A}$). Given an observation Y_1, the posterior Pólya tree distribution is easily obtained. Write $(F|Y_1) \sim \mathrm{PT}(\Pi, \mathcal{A}|Y_1)$ 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 $\mathcal{Y} = (Y_1, \ldots, Y_n)$, 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 $\epsilon = \epsilon_1 \ldots \epsilon_m$; then, in the absence of censoring,

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

For censored data,

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

where s_ϵ is the number of observations censored in B_ϵ. So, if we can arrange for the censoring sets to coincide with the partition sets, then we retain the conjugacy property of the Pólya tree; see, for example, Muliere and Walker (1997).

4.4. Examples

Our main example involves a linear regression model in the context of accelerated failure time data.

4.4.1. Multiple-regression example

We consider the linear model

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

where $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})$ 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, taken to have a Pólya tree prior. 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 $F(B_0) = F(B_1) = \frac{1}{2}$. If errors take values on the real line, we might typically want the median to be located at 0 and this is achieved by taking the partition point 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 the frequentist approach of Ying *et al.* (1995). If required, we could also fix the scale of F by defining $F(B_{00}), \ldots, F(B_{11})$ each equal to $\frac{1}{4}$. This would be appropriate for the alternative model

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

where F_0 could be taken to be the standard normal distribution.

We reanalyse the data set presented by Ying *et al.* (1995). This involves 121 patients suffering small cell lung cancer, each being assigned to one of two treatments: A with 62 patients; 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, coded 0 or 1, and the natural logarithm of the entry age of the patient. Ying *et al.* (1995) could only estimate the median survival time in their analysis and then test for the 'better' treatment. We are not restricted in any way about the type of inference that we can make.

In our analysis (an outline algorithm for which is provided in Appendix A.1) we took a normal prior, with mean 0 and large variance term, for β. The parameters for the Pólya tree are F_0, taken to be the normal distribution with zero mean and variance $\sigma^2 = 10^2$, and, for simplicity, $\alpha_\epsilon = cm^2$ whenever ϵ defines a set at level m, with $c = 0.1$. We took the number of levels of the Pólya tree to be fixed at 8. These specifications are chosen for illustration. A more general approach would be to treat c, M and σ^2 as unknown parameters and to 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.

For illustration, predictive survival curves are presented, the first (Fig. 2) for new patients receiving treatment A, and the second (Fig. 3) for new patients receiving 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.

4.4.2. *Frailty model example*

Walker and Mallick (1997) detail the use of Pólya trees in a frailty model (Clayton and Cuzick, 1985). We omit the 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; Fig. 4 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.

5. Exchangeable models

Let $Y_1, Y_2, \ldots$ 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

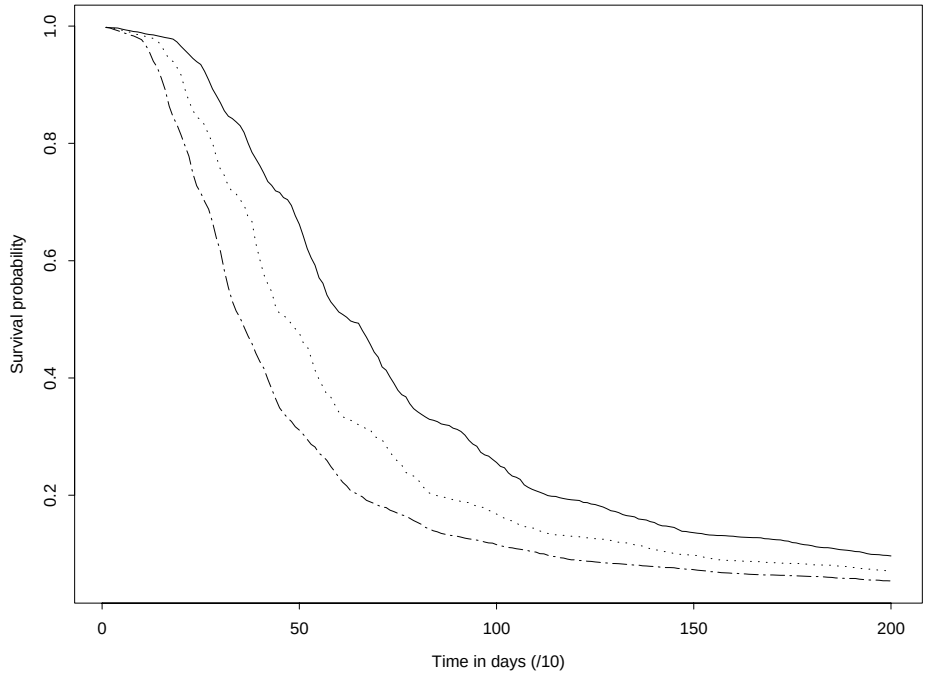

Fig. 2. Predictive survival curves for three new patients with treatment A: data set of Ying *et al.* (1995)

defined on the space of probability measures on Ω, such that the distribution of $Y_1, Y_2, \ldots$ can be obtained by first choosing $F \sim P_\Omega$ and then taking $Y_1, Y_2, \ldots | F \sim_{\text{IID}} F$, i.e.

$$P(Y_1 \in B_1, \ldots, Y_n \in B_n) = \int \left\{ \prod_{i=1}^{n} F(B_i) \right\} \mathrm{d}P_\Omega(F).$$

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

$$\begin{aligned} Y_1 &\sim F_0, \\ Y_2 | Y_1 &\sim \frac{cF_0 + \delta_{Y_1}}{c+1}, \\ &\vdots \\ Y_n | Y_1, \ldots, Y_{n-1} &\sim \frac{cF_0 + \sum_{j=1}^{n-1} \delta_{Y_j}}{c+n-1}. \end{aligned}$$

Blackwell and McQueen (1973) showed that the de Finetti measure for the sequence is the Dirichlet process. As might be expected from our earlier identification of the beta-Stacy process as a generalization of the Dirichlet process, a generalized Pólya urn scheme can be

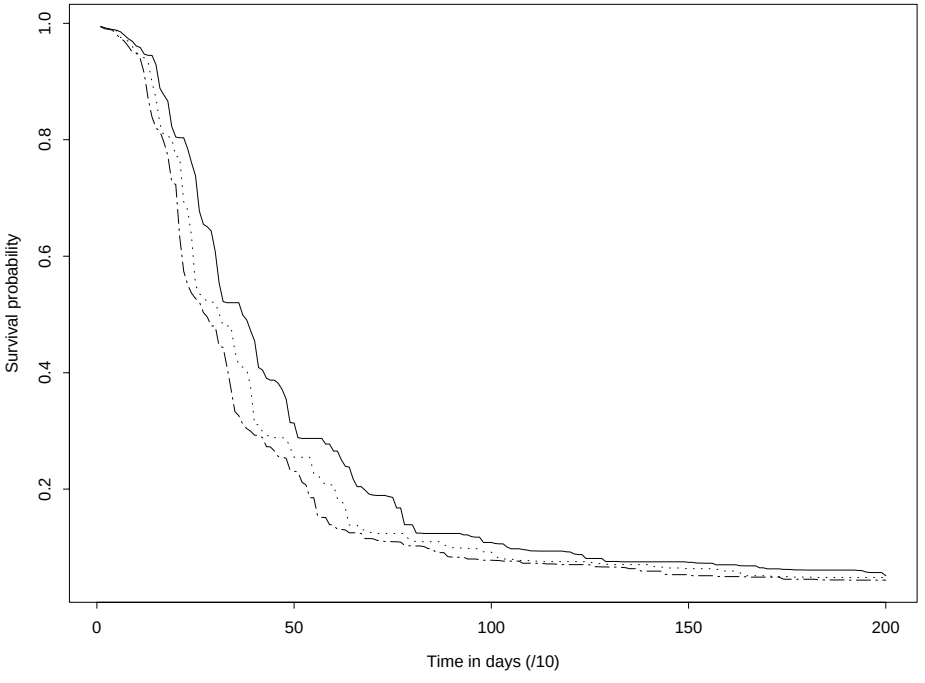

Fig. 3. Predictive survival curves for three new patients with treatment B: data set of Ying *et al.* (1995)

obtained which has the discrete beta-Stacy process as the de Finetti measure (Walker and Muliere, 1997).

There are several reasons why it is often convenient to consider the sequence $Y_1, Y_2, \ldots$ directly, marginalizing over F. First, F is an infinite dimensional parameter so the advantages in removing this is that we work 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 $Y_1, \ldots, Y_n$ is an immediate consequence. Thirdly, we are 'closer' to the data in the sense that we have the probability distribution for the data explicitly. Also the posterior parameters for P_Ω can often be determined from the sequence of predictive distributions (consider, for example, the Pólya urn sequence).

5.1. Bernoulli trips

Here we introduce a simple concept and method, the Bernoulli trip (Walker, 1998), for modelling multiple-state processes directly, using exchangeability ideas. A Bernoulli trip is a *reinforced random walk* (Coppersmith and Diaconis, 1987; Pemantle, 1988) on a 'tree' which characterizes the space for which a prior is required. An observation in this space corresponds to a unique path or branch of the tree. The path corresponding to this observation is reinforced, i.e. 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 discretize 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

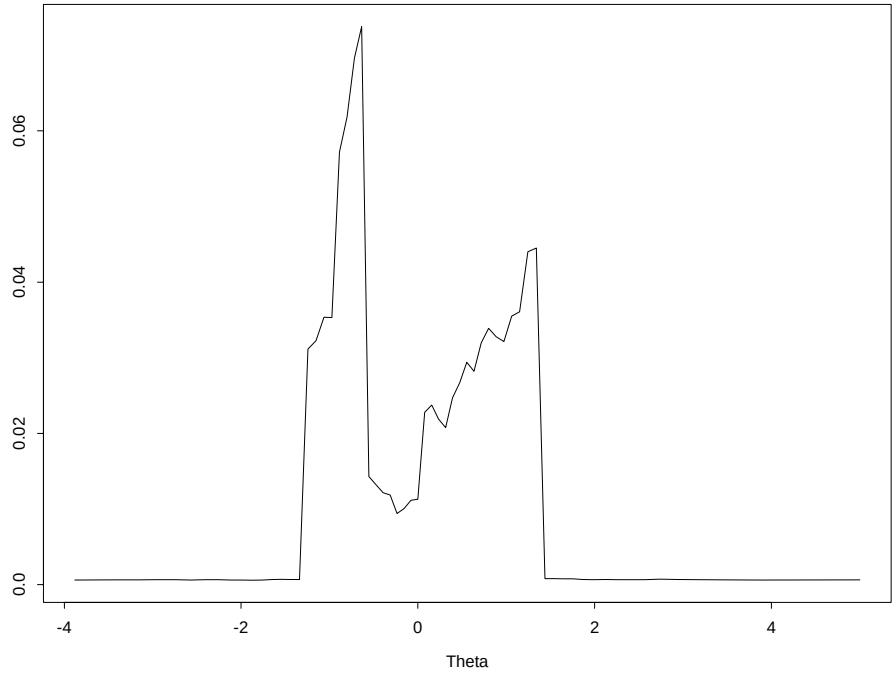

Fig. 4. Posterior expectation of the log-frailty distribution: frailty model example

and moves to one of a (finite) number of 'positions', the collection of which we shall denote by $\mathcal{M}_\epsilon$. For the first walk

$$P(\epsilon \to \epsilon' \in \mathcal{M}_\epsilon) = \alpha(\epsilon, \epsilon') \Big/ \sum_{\epsilon'' \in \mathcal{M}_\epsilon} \alpha(\epsilon, \epsilon''),$$

where each α is non-negative. There will be positions which, if reached, result in termination of the walk, 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 probability that the second walk coincides with the first walk exactly has increased (reinforcement).

If we denote the path of the first walk by Y_1 and the path of 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 $(Y_1, \ldots, Y_n)$ are exchangeable random variables for all n. Explicitly, we have

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

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

A Bayesian bootstrap procedure would be to obtain the posterior parameters and then to set the prior parameters to 0. 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, i.e. a survival model. Each walk starts at (0, 0) and on reaching say $(k, 0)$, $k = 1, 2, \ldots$, the walk can move to either $(k+1, 0)$ or $(k+1, 1)$. We assume that k indexes time points $t_1, t_2, \ldots$. If a walk 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 $k = 1, 2, \ldots$,

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

and

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

Clearly each walk is characterized by the point k at which the move to $(k, 1)$ is made; 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

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

$\alpha^*_{k0} = \alpha_{k0} + n_{k0}$ 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, i.e. $Y > k$, corresponds to a walk being censored at k. The updating mechanism for such a walk is given by $\alpha_{j0} \to \alpha_{j0} + 1$ for all $j \leqslant k$. Note that the walks remain exchangeable provided that 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, is that the space needs to be discretized. 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. Example

We reanalyse 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 blood infected with the human immunodeficiency virus (HIV). Observations take the form of health states occupied at the end of each 6-month interval. State 1 is infection free, state 2 corresponds to HIV infection

and state 3 is the onset of acquired immune deficiency syndrome (AIDS). According to current mainstream medical theory, it is not possible to have AIDS without first being HIV positive and so it is not possible to move directly from state 1 to state 3. For the illustrative results that follow, we take a Bayesian bootstrap approach, i.e. we set the prior parameters to 0. De Gruttola and Lagakos (1989) and Frydman (1992) both analysed the data non-parametrically via *self-consistent* estimators (Turnbull, 1976) but the former assumed the times in states 1 and 2 to be independent.

We define the first walk via the transition probabilities

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

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

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

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

and

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

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

$$P(T=k, S=j) = \frac{\alpha_{j1}}{\alpha_{j0}+\alpha_{j1}} \prod_{l<j} \frac{\alpha_{l0}}{\alpha_{l0}+\alpha_{l1}} \frac{\beta_{k2}}{\beta_{k1}+\beta_{k2}} \prod_{j<l<k} \frac{\beta_{l1}}{\beta_{l1}+\beta_{l2}},$$

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 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 the conditional probabilities depend only on the time spent in state 2.

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

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

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

$$\pi(M_{(\cdot)}|\text{data}) \propto \pi(\text{data}|M_{(\cdot)})\,\pi(M_{(\cdot)}),$$

where $\pi(M_{(\cdot)})$ is the prior weight assigned to model $M_{(\cdot)}$. Therefore to obtain the estimator $\hat{P}$ it only remains to evaluate $\pi(\text{data}|M_{(\cdot)})$. These are in fact straightforward to calculate based on $P(Y_1, \ldots, Y_n)$ given in Section 5.1. It remains to decide on the values of the αs and βs with which to determine $\pi(\text{data}|M_{(\cdot)})$. First, for large αs and βs the prior specifications should swamp the data and the model. This is the case and, for $\alpha = \beta = 10^6$, $\log\{\pi(\text{data}|M_{(\cdot)})\} = -578.8$ for all $M_{(\cdot)}$ (we have removed the term

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

from $\pi(\text{data}|M_{(\cdot)})$ which is common to all $M_{(\cdot)}$). To represent vague *a priori* information, we consider $\beta_{\cdot 2} = \lambda\beta_{\cdot 1} = 10^{-6}$; for $\lambda = 1, 10, 100$,

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

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

and

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

On the bases of these factors the data support $M_{(c)}$, the Markov model.

For illustration, Fig. 5 is the estimated cumulative distributions of times to HIV infections for the two groups, under the Markov assumption. These estimates are in good agreement with those of Frydman (1992), Fig. 2. Appendix A.2 considers the situation where some transition times are interval censored, relevant for the data in this example. Whereas it would appear difficult to extend the mathematical framework of Frydman to more complex models, the framework presented here is readily extended (Walker, 1998).

6. Discussion

This paper has surveyed a range of current research in the area of Bayesian nonparametrics. The work is ongoing and several 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 Pólya trees; developing exchangeable processes to cover a larger class of problems; nonparametric sensitivity analysis (Lenk, 1996).

A further question that arises is the extent to which we currently understand the potential mathematical consequences of the toolkit that we are developing. Diaconis and Freedman

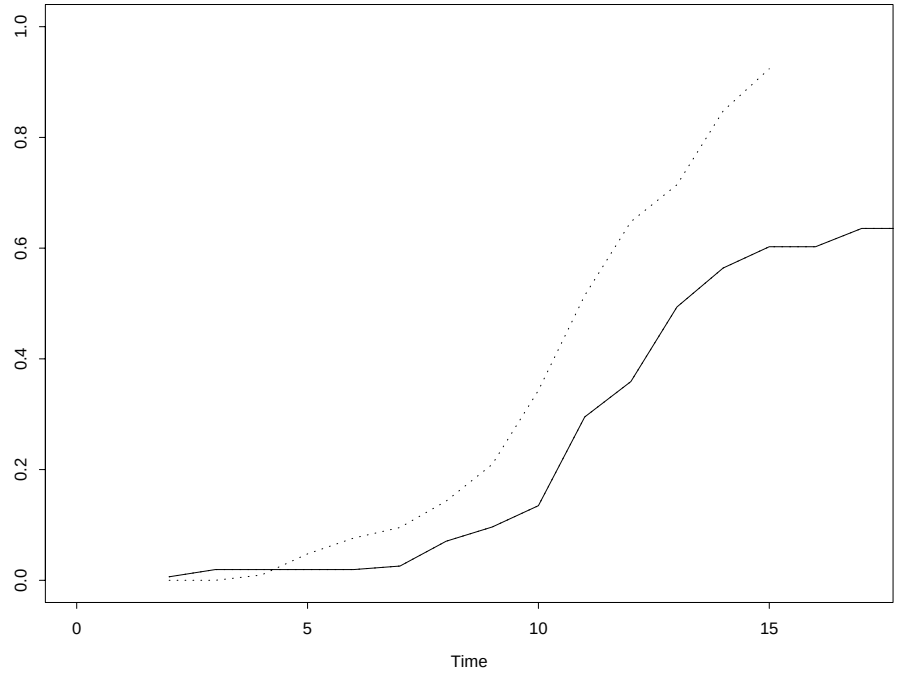

Fig. 5. Estimated cumulative distributions of times to HIV infection (———, lightly treated; ·····, heavily treated): data set of De Gruttola and Lagakos (1989)

(1986) presented a nonparametric model that uses a symmetrized Dirichlet prior for the underlying distribution and an independent prior for its median. They then demonstrated that seemingly innocuous choices for the latter led to an inconsistent Bayes estimate of the median. For the same model, they showed other reasonable priors for the median that are consistent. In the light of results such as those in Hjort (1990) and Diaconis and Freedman (1993) that give demonstrably consistent nonparametric Bayesian procedures, general theoretical advances that pin-point the pitfalls would indeed prove valuable. Recent progress has been made on these problems; see Barron *et al.* (1996), Ghosal *et al.* (1997) and Shen and Wasserman (1998).

We believe that Bayesian nonparametrics have much to offer. As far as nonparametric *versus* parametric analyses are concerned, in relatively 'well-behaved' cases, where a parametric analysis would have coped, we typically obtain similar forms of posterior inference, particularly posterior means, but with appropriately greater ranges of uncertainty (as indicated in Section 3.4). When the appropriate form of posterior should be 'badly behaved' (see, for example, Fig. 4) the nonparametric analysis will reflect this, whereas most parametric analyses would not reveal this fact. As far as Bayes *versus* non-Bayes approaches are concerned, we note

(a) the very real advantage of being able to input broad prior ideas of characteristics such as location, scale and shape,
(b) the much richer and more tractable forms of inference that are available as a consequence of the simulation-based approach to computation, where the technology of

implementation for nonparametrics is now essentially no more difficult than for the parametric case.

Acknowledgements

Research reported here was supported in part by an Engineering and Physical Sciences Research Council 'Realising our potential' award and travel grant, a National Science Foundation grant and financial support from the Business School at the University of Michigan, Ann Arbor. We are grateful to several reviewers for helpful comments on earlier versions of the paper.

Appendix A

In Appendix A.1 we outline the simulation algorithm for the example in Section 4.4 and in Appendix A.2 we detail the solution to the interval-censored observations for the example in Section 5.2.

A.1. Simulation algorithm

Here we provide an outline algorithm for the multiple-regression example in Section 4.4. The algorithm is based on a Gibbs sampler for which we need to sample from the full conditionals $p(\beta|F, \text{data})$ and $p(F|\beta, \text{data})$. In the following we let mj denote the jth partition ($j = 1, \ldots, 2^m$) in the mth level of the tree. Here $\beta = (\beta_1, \ldots, \beta_p)$ and the prior for β is a multivariate normal distribution with zero mean and covariance matrix of the form $\text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

Step 1: set the starting value for β.
Step 2: update the $\{\alpha_{mj}\}$ based on the n IID observation $Z_i = Y_i - X_i\beta$; so,

$$\alpha_{mj} \to \alpha_{mj} + \sum_{i=1}^{n} I(Z_i \in B_{mj}).$$

Step 3: if $B_\epsilon = B_{Mj}$, for $j \in (1, \ldots, 2^M)$, and $\epsilon = \epsilon_1 \ldots \epsilon_M$, then

$$F(B_{Mj}) = \left(\prod_{l=1;\epsilon_l=0}^{M} C_{\epsilon_1 \ldots \epsilon_{l-1}0} \right) \prod_{l=1;\epsilon_l=1}^{M} (1 - C_{\epsilon_1 \ldots \epsilon_{l-1}0})$$

and the $C_{\epsilon 0}$ are independent beta$(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ variables.
Step 4: the likelihood function for β, given F_M, is

$$l(\beta) = \prod_{j=1}^{2^M} F_M(B_{Mj})^{n_j},$$

where $n_j = \Sigma_i I(Z_i \in B_{Mj})$. Generate β^* from the multivariate normal distribution with mean β and covariance matrix $\text{diag}(\tau_1^2, \ldots, \tau_p^2)$. Using a random walk Metropolis–Hastings algorithm, take u from the uniform distribution on the interval (0, 1). If

$$u < \frac{l(\beta^*)}{l(\beta)} \exp\left\{ -0.5 \left(\sum_{l=1}^{p} \frac{\beta_l^{*2} - \beta_l^2}{\sigma_l^2} \right) \right\},$$

then the chain moves to β^*; otherwise it remains at β.

Repeat steps 2–4 to construct the Markov chain, resetting the $\{\alpha_{mj}\}$ to their initial values after completing step 4.

A.2. Solution for example in Section 5.2

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, i.e. S_n is known to be in the interval $[k_1, \ldots, k_L]$ ($k_L < \infty$ 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 $n_{k0} = \Sigma_{i=1}^{n-1} I(S_i = k)$ and $n_{k1} = \Sigma_{i=1}^{n-1} I(S_i > k)$. Here $\mathcal{J}_{\alpha k}$ and $\mathcal{J}_{\beta k}$ are random and defined on $\{0, 1\}$ where

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

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

and

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

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

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

where

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

and

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

for $k \in \{k_1, \ldots, k_L\}$. For more than one interval-censored observation we can proceed by sampling the missing data, conditionally 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 $S_1, \ldots, S_m$ ($m \leqslant n$) be interval censored, with $S_j \in [k_{1(j)}, \ldots, k_{L(j)}]$ ($T_j > k_{L(j)}$). The approach is to sample iteratively, for $j = 1, \ldots, m$, from

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

where $(S_{(j)}, T_{(j)})$ contains all the information in the data and from the sampled variates except on individual j. If $S_{(j)} \cap \{k_{1(j)}, \ldots, k_{L(j)}\} = \varnothing$ then S_j is taken uniformly from $\{k_{1(j)}, \ldots, k_{L(j)}\}$. At iteration t we have then sampled $\{S_j^{(t)} : j = 1, \ldots, m\}$, which, combined with the observed data, gives the estimator $\hat{P}^{(t)}$. The required estimator is then given by the average $\tau^{-1} \Sigma_{t=1}^{\tau} \hat{P}^{(t)}$, where τ is the number of iterations. 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 $[S_j | \ldots]$ replaces taking the expectation of $[S_j | \ldots]$. It is also possible to consider the situation in which T and S are both interval censored by using a modified version of the algorithm just described.

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. *Statist. Sin.*, **4**, 505–524.
Barron, A. R., Schervish, M. and Wasserman, L. (1996) The consistency of posterior distributions in nonparametric problems. *Preprint*.
Blackwell, D. (1973) The discreteness of Ferguson selections. *Ann. Statist.*, **1**, 356–358.
Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya-urn schemes. *Ann. Statist.*, **1**, 353–355.
Bondesson, L. (1982) On simulation from infinitely divisible distributions. *Adv. Appl. Probab.*, **14**, 855–869.
Bush, C. A. and MacEachern, S. N. (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–286.
Clayton, D. G. (1991) A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.
Clayton, D. and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc.* A, **148**, 82–117.
Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalisation of the Dirichlet distribution. *J. Am. Statist. Ass.*, **64**, 194–206.
Coppersmith, D. and Diaconis, P. (1987) Random walk with reinforcement. Unpublished.
Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.
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) Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. R. Statist. Soc.* B, **57**, 547–563.
———(1996) Implementation of Bayesian nonparametric inference based on beta processes. *Scand. J. Statist.*, **23**, 27–36.
De Gruttola, V. and Lagakos, S. W. (1989) Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1–11.
Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B, **60**, 333–350.
Dey, D., Mueller, P. and Sinha, D. (1998) *Practical Nonparametric and Semi-parametric Bayesian Statistics*. New York: Springer.
Diaconis, P. and Freedman, D. (1986) On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1–26.
———(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.
Doss, H. (1995) Bayesian nonparametric estimation for incomplete data via substitution sampling. *Ann. Statist.*, **22**, 1763–1786.
Dubins, L. and Freedman, D. (1965) Random distribution functions. *Bull. Am. Math. Soc.*, **69**, 548–551.
Dykstra, 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. *PhD Dissertation*. Department of Statistics, Yale University, New Haven.
———(1994) Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Ass.*, **89**, 268–277.
Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **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.
———(1974) Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**, 615–629.
Ferguson, T. S. and Klass, M. J. (1972) A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.*, **43**, 1634–1643.
Ferguson, T. S. and Phadia, E. G. (1979) Bayesian nonparametric estimation based on censored data. *Ann. Statist.*, **7**, 163–186.
de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.
Freedman, D. A. (1963) On the asymptotic behaviour of Bayes estimates in the discrete case I. *Ann. Math. Statist.*, **34**, 1386–1403.
———(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. R. Statist. Soc.* B, **54**, 853–866.
Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997) Consistency issues in Bayesian nonparametrics. *Preprint*.
Ghosh, J. K. and Mukerjee, R. (1992) Noninformative priors. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: 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. Am. 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.

———(1996) Bayesian approaches to non- and semiparametric density estimation. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.
Kalbfleisch, J. D. (1978) Non-parametric Bayesian analysis of survival time data. *J. R. Statist. Soc.* B, **40**, 214–221.
Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Statist. Ass.*, **53**, 457–481.
Laud, P. W. (1977) Bayesian nonparametric inference in reliability. *PhD Dissertation*. University of Missouri, Columbia.
Laud, P. W., Damien, P. and Smith, A. F. M. (1993) Random variate generation from D-distributions. *Statist. Comput.*, **3**, 109–112.
———(1998) Bayesian nonparametric and covariate analysis of failure time data. In *Practical Nonparametric and Semi-parametric Bayesian Statistics* (eds D. Dey, P. Mueller and D. Sinha), pp. 213–225. New York: Springer.
Laud, P. W., Smith, A. F. M. and Damien, P. (1996) Monte Carlo methods for approximating a posterior hazard rate process. *Statist. Comput.*, **6**, 77–84.
Lavine, M. (1992) Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235.
———(1994) More aspects of Pólya trees for statistical modelling. *Ann. Statist.*, **22**, 1161–1176.
Lenk, P. J. (1996) Bayesian inference of semiparametric regression and Poisson intensity functions. *Preprint*.
Lévy, P. (1936) *Théorie de l'Addition des Variables Aléatoire*. Paris: Gauthiers-Villars.
Lo, A. Y. (1984) On a class of Bayesian nonparametric estimates: I, Density estimates. *Ann. Statist.*, **12**, 351–357.
———(1993) A Bayesian bootstrap for censored data. *Ann. Statist.*, **21**, 100–123.
MacEachern, S. N. and Mueller, P. (1998) Estimating mixtures of Dirichlet process models. *J. Comput. Graph. Statist.*, **7**, 223–238.
Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992) Pólya 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. (1997) A Bayesian nonparametric approach to survival analysis using Pólya trees. *Scand. J. Statist.*, **24**, 331–340.
———(1998) Extending the family of Bayesian bootstraps and exchangeable urn schemes. *J. R. Statist. Soc.* B, **60**, 175–182.
Pemantle, R. (1988) Phase transitions in reinforced random walk and RWRE on trees. *Ann. Probab.*, **16**, 1229–1241.
Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.
Sethuraman, J. and Tiwari, R. (1982) Convergence of Dirichlet measures and the interpretation of their parameter. In *Proc. 3rd Purdue Symp. Statistical Decision Theory and Related Topics* (eds S. S. Gupta and J. O. Berger). New York: Academic Press.
Shen, X. and Wasserman, L. (1998) Rates of convergence of posterior distributions. *Technical Report 678*. Carnegie Mellon University, Pittsburgh.
Susarla, V. and Van Ryzin, J. (1976) Nonparametric Bayesian estimation of survival curves from incomplete data. *J. Am. Statist. Ass.*, **71**, 897–902.
Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.
Turnbull, B. W. (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.* B, **38**, 290–295.
Walker, S. G. (1998) A nonparametric approach to a survival study with surrogate endpoints. *Biometrics*, **54**, 662–672.
Walker, S. G. and Damien, P. (1998) A full Bayesian nonparametric analysis involving a neutral to the right process. *Scand. J. Statist.*, **25**, 669–680.
Walker, S. G., Damien, P. and Laud, P. W. (1998) Bayesian nonparametric inference for a continuous cumulative hazard function. *Preprint*.
Walker, S. G. and Mallick, B. K. (1997) Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. R. Statist. Soc.* B, **59**, 845–860.
Walker, S. G. and Muliere, P. (1997) Beta-Stacy processes and a generalisation of the Pólya-urn scheme. *Ann. Statist.*, **25**, 1762–1780.
West, M., Mueller, 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* (eds A. F. M. Smith and P. Freeman). Chichester: Wiley.
Wild, C. J. and Kalbfleisch, J. D. (1981) A note on a paper by Ferguson and Phadia. *Ann. Statist.*, **9**, 1061–1065.
Wolpert, R. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.
Ying, Z., Jung, S. H. and Wei, L. J. (1995) Survival analysis with median regression models. *J. Am. Statist. Ass.*, **90**, 178–184.

Discussion on the paper by Walker, Damien, Laud and Smith

David Draper (*University of Bath*)

This admirable paper concerns two topics of considerable importance to Bayesians and non-Bayesians alike: model selection and model robustness. In my discussion I shall begin by trying to place the subject of Bayesian nonparametrics in a slightly broader historical context than that presented by the authors; I shall then look at some of the 'small print' of Pólya trees (PTs), including some warnings for applied statisticians; and I shall conclude by making a connection between PTs and wavelet density estimation.

Model selection and robustness

Given a model, Bayes's theorem tells you how to update your uncertainty in the light of new data; but where does the model come from to begin with? Bruno de Finetti had the best answer to that question that anyone has invented so far: your model comes from considerations of similarity, or *exchangeability* (e.g. de Finetti (1930), Lindley and Novick (1981) and Draper *et al.* (1993)). An informal statement of what might be termed de Finetti's (1980) 'Fundamental theorem of Bayesian modelling' might go like this: if you are willing to treat (your uncertainty about) the real-valued observables $(y_1, \ldots, y_n)$ as exchangeable, then you may as well model them hierarchically as

$$\begin{aligned} F &\sim p(F), \\ (y_i|F) &\overset{\text{IID}}{\sim} F \end{aligned} \tag{1}$$

where F is the long run (large n) empirical cumulative density function (CDF) of the y_i. In the special case of binary outcomes, which de Finetti treated in 1930, the only possible Fs are Bernoulli distributions, differing only in their values of $\theta = P(y_i = 1)$, and model (1) becomes

$$\begin{aligned} \theta &\sim p(\theta), \\ (y_i|\theta) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta). \end{aligned}$$

This binary version of the theorem has been feasible to implement (at least approximately) for the past 250 years, since the days of the Rev. Bayes himself:

(a) you can reliably elicit a prior distribution for a quantity living on [0, 1] (e.g. one of the conjugate beta distributions or — if necessary — a mixture thereof), and
(b) you can compute things like $p(\theta|y)$ and $p(y_{n+1}|y)$ with little trouble.

But the general real-valued version of the theorem is much more difficult to implement (and even de Finetti himself did not fully know how): you must reliably elicit a prior distribution on the *function* F, and how do you compute things like $p(F|y)$ and $p(y_{n+1}|y)$? The answer to both parts of this question is Bayesian nonparametrics (e.g. PTs), with Markov chain Monte Carlo (MCMC) sampling as the computing engine. Thus, with the advent of MCMC methods and techniques like those described by the authors here, what amounts to a crucial 60-year-old foundational problem has finally been solved.

Another reason, also involving model selection, that the topic of this paper is so important is Bayesian model updating. Lindley (1972) reminded us of *Cromwell's rule*: anything to which you assign prior probability 0 must have zero posterior probability, no matter how the data come out. This is potentially embarrassing for Bayesian modelling, as follows. Suppose that you take as your prior on F (based on past experience with similar problems)

$$p(F) = \text{point mass on } N(\mu, \sigma^2) \qquad \text{(with a prior on } \mu \text{ and } \sigma), \tag{2}$$

so that, for example, your prior probability that F is multimodal is 0. Now the data arrive and are strongly bimodal. What do you do? If you retain your original prior, then $p(\text{bimodal}|\text{data}) = 0$, which may be silly; but if you go back and change your prior (naïvely) you are cheating (using the data twice), and you risk poor calibration. The problem is serious: people like Lindsey (1999) have seized on it as an apparent major nail in the Bayesian coffin.

Fortunately there appear to be at least two solutions.

(a) The first is based on Bayesian versions of cross-validation, along the following lines. Placing a prior on F (or on the structure of, for example, $E(y|x)$ for some predictor x) is a model selection problem, and — as many people have noted (see Key *et al.* (1999) and Draper (1999a)) — model selection should be approached decision theoretically: to choose a model well you must specify

the use to which it will be put, for how else will you know whether it is sufficiently good? (For instance, for some purposes getting the bimodality wrong in the example above is unimportant.) If the goal is the prediction of future observables, then dividing the data into *three* (not two) parts exchangeably (Draper, 1999b) is sufficient to permit both

(i) using the data to help to specify the prior on the model and
(ii) calculating honest (well-calibrated) uncertainty assessments, in spite of (i).

(b) The second solution to the problem posed by Cromwell's rule is Bayesian nonparametrics: if your prior on F places positive probability on all possible CDFs, then so does your posterior, and nothing in the data can unpleasantly surprise you. Methods like PTs require you to specify a prior guess for F (and something like a prior effective sample size; see below), but with sufficient data the posterior will shrug off any specification errors in the prior and adapt fully to the data.

The small print on Pólya trees

With colleagues at Bath and AEA Technologies plc (Draper *et al.*, 1998) I have recently been using Bayesian nonparametrics to solve a consulting problem in the nuclear power industry. In assessing the risk of underground storage of nuclear waste, complex computer simulation models yield independent and identically distributed predictions of radiologic dose for humans on the surface, as a function of time T since the storage facility failed. A sample of size 10000 for $T = 100$ years under one scenario had 9864 zeros and $n = 136$ positive values, 134 of which ranged smoothly from 1.059×10^{-14} to 8.522×10^{-1}, but the two largest values were 3.866 and 189.3 (!). In spite of the extreme sensitivity of the results to only one or two observations out of a large sample, public policy considerations require accurate and well-calibrated uncertainty bands for the underlying population *mean*, and ideally valid conclusions should also be obtainable with far smaller sample sizes (since the computer models are expensive to run).

A normal QQ-plot shows that the $n = 136$ non-zero dose values y_i are close to Gaussian on the log-scale. One way to write the resulting parametric Bayesian log-normal model is $\log(y_i) = \mu + \sigma e_i$, $(\mu, \sigma^2) \sim p(\mu, \sigma^2)$ and $e_i \sim^{\text{IID}} N(0, 1)$, for some prior distribution $p(\mu, \sigma^2)$. The PT idea is to replace the last part of this model, which expresses certainty about the distribution of the e_i, with a distribution on the set of possible distributions F for the e_i. The new model is

$$\left.\begin{array}{c} \log(y_i) = \mu + \sigma e_i, \\ (\mu, \sigma^2) \sim p(\mu, \sigma^2) \quad \text{(e.g. conjugate)}, \\ (e_i | F) \overset{\text{IID}}{\sim} F \quad \text{with mean 0 and standard deviation 1}, \\ F \sim \text{PT}(\Pi, \mathcal{A}_c), \end{array}\right\} \tag{3}$$

where $\Pi = \{B_\epsilon\}$ centres the prior on F on the standard normal distribution, by choosing the binary tree partition sets $\{B_\epsilon\}$ based on the quantiles of $N(0, 1)$. Like the authors we use $M = 8$ levels in the binary tree and take $\{\alpha_\epsilon = cm^2$ whenever ϵ defines a set at level $m\}$ to specify $\mathcal{A}_c$, the collection of αs in the PT's prior beta distributions for the probabilities of moving left or right in the binary partitions. As the authors note, $c > 0$ is a kind of tuning constant which is related to the prior sample size hyperparameter in Dirichlet process priors: with small c the posterior distribution for the e_i will be based almost completely on $\hat{F}_n$, the empirical CDF for the e_i, whereas with large c the posterior will be based almost completely on the prior centring distribution, $N(0, 1)$.

In Fig. 6 I present 50 simulated samples from the PT prior specified above with $c = 0.1$. It is evident that the individual samples from this prior dance around like water on a hot griddle, bearing almost no relationship to the prior centring distribution, and even aggregating across the 50 realizations produces a density trace that deviates markedly from $N(0, 1)$. This, then, is practical warning 1: attempts to be 'non-informative' with the PT prior on F should be viewed with caution. Warning 2 is that you will need to revise your intuitions about familiar objects recast in the nonparametric setting: Fig. 7 plots the joint conditional log-likelihood for (μ, σ^2) in model (3) given a particular estimate of F. In the parametric Bayesian log-normal model of which the PT model is an expansion, the joint log-likelihood is smooth and locally quadratic about a single maximum, but in the nonparametric version the analogous plot looks like a crumpled piece of paper (and is almost fractal in character: enlarging any portion of the plot produces another almost identical graph), because of the granularity imposed by the choice of $M = 8$ layers in the PT.

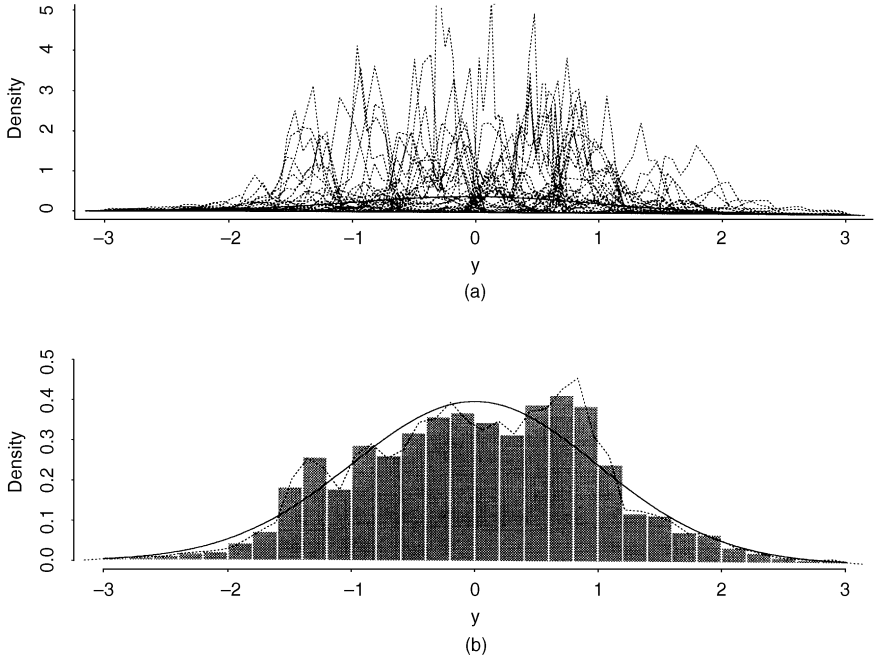

Fig. 6. (a) 50 samples from a PT(Π, $\mathcal{A}_c$) prior for F centred at $N(0, 1)$ (———) with $c = 0.1$ and (b) histogram and density trace of the aggregate of the 50 samples (———, $N(0, 1)$)

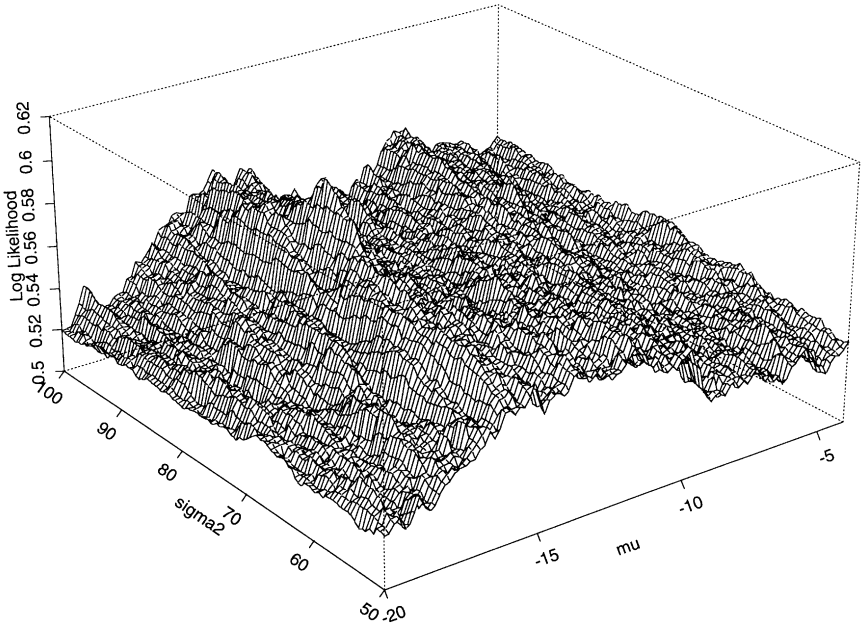

Fig. 7. Typical joint conditional log-likelihood for μ and σ^2 given F in model (3)

Warnings 3 and 4 stem from the following observation: with PTs you are essentially estimating F via random histograms with many bins (e.g. $M = 8$ yields $2^8 = 256$ histogram bars), which is like treating F as a 256-dimensional nuisance parameter, and, even though these extra parameters will exhibit a high degree of local correlation along the real number line, dozens (or more) of additional parameters are being estimated with $M = 8$. Moreover, location and scale cannot be fully unconfounded with shape in model (3), so some of the slow MCMC convergence in estimating F bleeds into (μ, σ^2), and the effective sample size for estimating location and scale is (much) smaller. Thus, warning 3: be prepared for long MCMC monitoring runs to obtain posterior summaries with decent accuracy (in spite of the examples given by the authors, in which 10000 iterations were used)—I found with the radiologic dose data that 50000–100000 iterations were needed to achieve, for example, 2.5 significant figures in reporting the posterior means of μ and σ^2 with high Monte Carlo probability. And, finally, warning 4: in specifying your priors on location and scale parameters you will probably have to recalibrate your intuition about how much information is present in the data for such parameters, i.e. people who like to be relatively uninformative in specifying such priors in parametric models should be willing to insert more prior information (from substantive sources) when the model is expanded nonparametrically to obtain the same level of accuracy for location and scale.

Wavelets

I shall conclude with a brief remark on Bayesian nonparametric density estimation: it is evident from the nature of the PT prior, with its binary splits of increasingly fine detail and its random histogram character, that there is an intimate connection between PTs and Bayesian wavelet density estimation with the Haar (step function) basis (e.g. Abramovich *et al.* (1998)), which deserves further study.

I have great pleasure in proposing the vote of thanks.

Sylvia Richardson (*Institut National de la Santé et de la Recherche Médicale, Villejuif*)

It is a pleasure to second the vote of thanks for this timely paper on nonparametric Bayes (NPB) procedures, a topic of great interest. The authors have succeeded in presenting an expert and enlightening discussion on the flexibility and the computational aspects of three central approaches to the construction of priors for random functions and consequent posterior analysis.

NPB procedures involve high dimensional or infinite dimensional spaces. As such, they require a large amount of prior information, which is paradoxical since they are precisely intended to cope with situations where we have no precise information. Adopting the point of view of a currently non-practitioner of NPB procedures, the following questions immediately come to mind.

(a) How easy is it to understand the procedures for setting the prior structure?
(b) How easy is the computation from prior to posterior?
(c) What new insights have been gained by using nonparametric *versus* more standard analyses?

The merits of the beta-Stacy process for modelling random cumulative distribution functions $F(t)$, while being able to specify $E[F(t)]$ and $E[F^2(t)]$, are convincingly explained by the authors. Thus, it will be reasonably straightforward to match these specifications with prior knowledge on the survival process. The data do not appear in the prior specifications, the set of points of discontinuity being fixed at the posterior inference stage. However, the authors, themselves, point to some of the difficulties in sampling from the posterior, though some algorithms are described for simple models. In practice, it is likely that these algorithms will not be commonly used and I shall concentrate my remarks on the other two constructions presented for which prior-to-posterior computations are fairly straightforward.

Note that both Pólya trees (PTs) and Bernoulli trips (BTs) involve *in fine* some arbitrary discretization; PTs will be partially specified to a level M, whereas BTs rely on discrete time.

PT priors require a binary tree partitioning of the space. In contrast with the Dirichlet prior (DP), it has long been recognized (Ferguson, 1974) that the points of subdivision play a part in the posterior properties of the process, which is an undesirable feature. However, in comparison with the DP, PTs are more flexible since they allow a choice of $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$ at each level, whereas for the DP $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$ is necessary (Ferguson, 1974). In particular, the parameters α of the PT can be chosen to ensure that the random probability measure F is absolutely continuous with probability 1.

To use this greater flexibility in practice, we are thus faced with the choice of the partition and that of the αs. Following Lavine (1992), the authors describe a *canonical way* of centring the PT prior on a particular probability measure F_0 by using partitions defined via the percentiles of F_0, and taking $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ for each

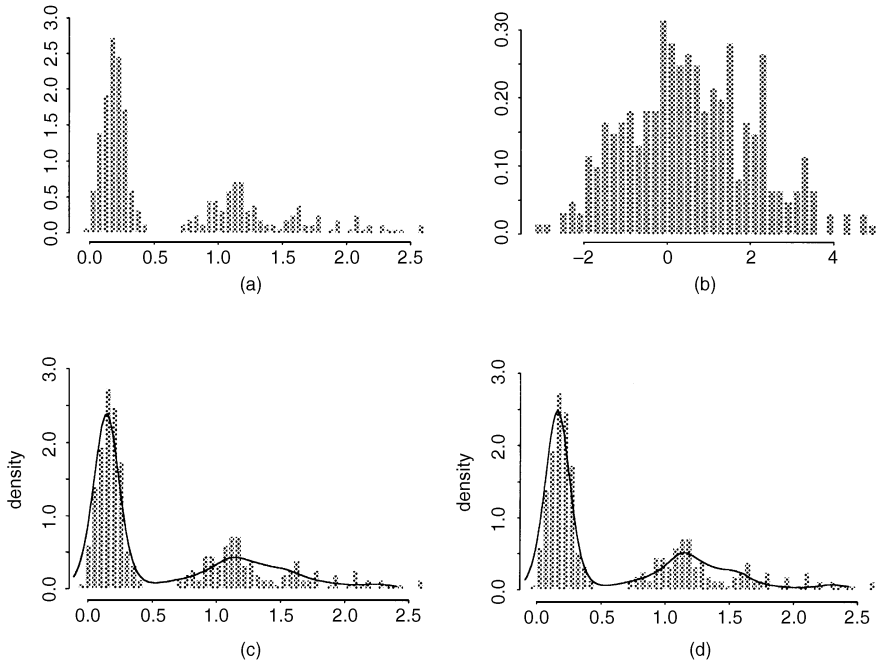

Fig. 8. (a) Latent covariate — all subjects ($n = 300$), (b) surrogate — all subjects ($n = 300$), (c) DMA estimate of the true covariate density using only 30 latent values and (d) DP estimate of the true covariate density using only 30 latent values

ϵ. This seems quite natural in regression problems, such as the example in Section 4.4. Nevertheless, values of α_ϵ must still be chosen for each level. The common choice $\alpha_\epsilon = cm^2$ gives different results for different values of c, since this parameter controls the smoothness of the predictive density, and large values of c leading to closeness to F_0. I find this disturbing and cannot see how using a prior on c, as indicated in the paper, would meaningfully reduce this arbitrariness. It would be interesting to hear the authors' experience with the use of such a prior. The alternative suggestion made in the paper of defining the αs to match $E_{\mathrm{PT}}[F(B_\epsilon)]$ and $E_{\mathrm{PT}}[F^2(B_\epsilon)]$ with those of a parametric model is not illustrated, though recommended. One wonders whether these choices which determine $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$ also ensure absolute continuity of F.

PTs have also been used in survival analyses with censored observations to define prior distributions for the time to death (Muliere and Walker, 1997). The context is that of obtaining a predictive distribution for future observations. There, instead of the canonical construction, the partition is partly determined by the censoring times observed in the data, to preserve conjugacy. Is there a justification for this choice of *data-dependent* prior structure, apart from the convenience of computation?

BTs present an interesting prior model for the time to absorption in the framework of multistate processes with an absorbing state. The dimension of the problem is reduced by time discretization and then prior parameters $\alpha(\epsilon, \epsilon')$ are set for all types of transition at each time. Simple updating rules govern the updating of parameters for prediction of time to absorption of a new observation, which is an attractive feature. The Kaplan–Meier nonparametric estimator of a survival time distribution can be obtained as a special case when all the prior parameters are set to 0.

For a simple Markov process, there are as many prior parameters as allowed transitions between the states at each time point, but, if we are rightly interested in relaxing the Markov assumption to condition on timing of previous transitions, this multiplies enormously the number of prior parameters. Moreover, how much the prior swamps the data will depend on the number of prior walks which have previously executed the same transitions. The authors interestingly use three different specifications of the transitions in their example in Section 5.2. For the semi-Markov model (a), there will clearly be fewer observed transitions of each type, and thus the prior influence will be different in comparison with the other models. Hence can we really use this approach for model comparison as done by the authors

Table 1. Performance of DMA and DP models as priors in a measurement error model†, from a simulation study with 50 replicates

True parameter values	*Parameter*	*Results from the following analyses:*	
		DMA	*DP*
$\theta = 0.5$	$\bar{\theta}$	0.51 (0.05)	0.51 (0.05)
$\beta_1 = 0.4$	$\bar{\beta}_1$	0.40 (0.40)	0.41 (0.39)
	$\text{mse}(\beta_1)$	0.07	0.07

†Regression model between outcome y and unknown latent variable x, $\text{logit}\{P(y = 1|x)\} = \beta_0 + \beta_1 x$; measurement model between surrogate u and x, $u \sim N(x, \theta^{-1})$; prior model $\pi(x)$ for the distribution of x, DP or DMA model with mixtures of Gaussian distributions. Data were simulated following these distributions with $\pi(x)$ as a mixture of three Gaussian distributions: $0.6\,N(0.19,\ 0.08^2) + 0.2\,N(1.05,\ 0.2^2) + 0.2\,N(1.63,\ 0.48^2)$. The values y and u were recorded on a sample of size 300, among which only 30 values of x were also supposed known.

and how can we calibrate it? It is a little worrying to see how the influence of λ on the posterior probabilities shown in Section 5.2 is clearly larger for model (a) which contains more prior parameters than for model (c) (assuming that the ratio λ was used in a similar fashion).

I now turn to comment on the paper's relative lack of display of posterior variability. Apart from Fig. 1, all the other graphs show predictive curves without any variability, and thus stay at a descriptive level. Nevertheless, one of the claims is that NPB procedures provide 'appropriately greater ranges of uncertainty'. It would have been interesting to have algorithmic details of the necessary steps for PT and BT models as well as seeing in the examples how a display of this variability could influence the scientific conclusions and provide new insights, particularly since two of the examples involve comparisons of treatments.

Finally, as mentioned in Section 1, another approach to making few distributional assumptions on the underlying population distribution is to turn to models involving finite mixtures of distributions. In many hierarchical models, mixture of Dirichlet process (MDP) models have commonly been used. Recently, Green and Richardson (1998) explored the connection between MDP construction and 'a variable number of components' version of the usual mixture model with Dirichlet weights and multinomial allocations, abbreviated as DMA models. They showed that the unbalancedness of the partition distribution, which exists in the DP model, persists *a posteriori*, leading to a difficulty of interpretation of the mixture components in the DP case. Overall similar predictive densities between DMA and DP models were found for a range of density shapes, as well as in the context of modelling a latent distribution in a measurement error context (Fig. 8 and Table 1), though in some cases the MDP model leads to higher variability of the latent variables for some extreme observations (Fig. 9). This supports exploring the use

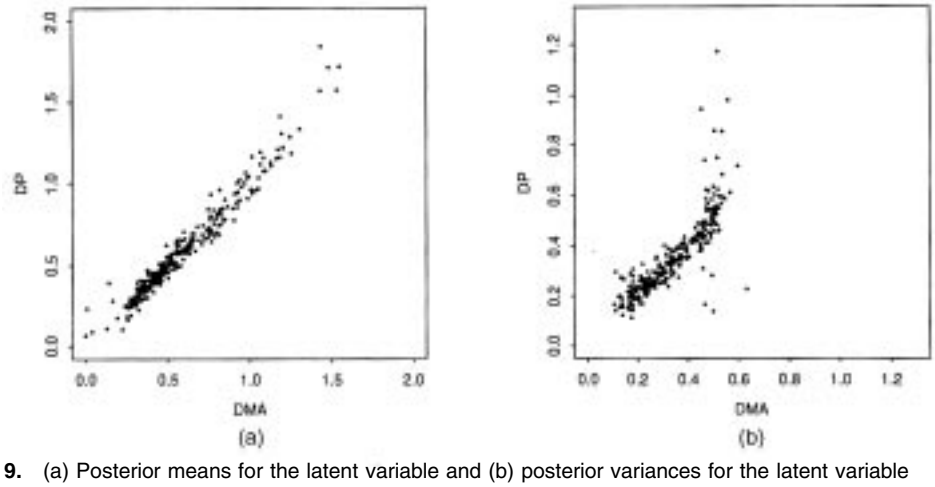

Fig. 9. (a) Posterior means for the latent variable and (b) posterior variances for the latent variable

and the relative performance of mixtures with variable numbers of components as an alternative to some NPB constructions in a variety of contexts.

In conclusion, it gives me great pleasure to second the vote of thanks. To paraphrase Milton:

'Perhaps some good will come of the tasting of the fruit of these forbidding Pólya trees'.

The vote of thanks was passed by acclamation.

J. A. Nelder (*Imperial College of Science, Technology and Medicine, London*)

I want to make a couple of relatively minor comments about two of the figures. First, in Fig. 1, the difference in the posterior density suggests to me that at least one of the priors is making a substantial contribution to the posterior. If this is so, it is important to check that the contributions from the prior and the likelihood are consistent with each other; if they are not, the model is defective. I should be interested to know whether they were consistent.

Second, Fig. 4, the frailty example: the Aitkin–Clayton trick of analysing the centring variable as a Poisson generalized linear model (GLM) can be extended via the hierarchical GLM models to include random effects. For a GLM with binary data, there are problems in interpreting the residuals, which fall into two classes and often give rise to bimodal histograms. I suspect that the same will be true for the sets of random effects that are found in the frailty model. The same bimodal effect can also be produced if an important covariate is omitted from the model. I conjecture that Fig. 4 shows the joint effect of these two causes. Certainly, the sex difference is important — its omission increases the deviance by 10 — but after allowing for it there are still signs of bimodality, which I conjecture is caused by the effect of having two subclasses corresponding to the censored and uncensored cases.

E. I. George (*University of Texas, Austin*)

This is a wonderful paper for me in that it opens up a whole new area of modelling — a new beginning. It is interesting to think about how some modelling problems might now be addressed perhaps more systematically. I wonder whether the authors can help me with a current problem on model selection on which I have been working.

I have 100 models, 90 of which are very similar. People tend to put uniform priors on things. This is really unacceptable in this case because, if 90 of the models are very similar, it would somehow put too much mass in that portion. An example of this would occur in one of my favourite model selection problems, variable selection. Imagine that I have p variables and 2^p models, and that many of the covariates are extremely highly correlated. I do not want to put a uniform prior on everything. In fact, some of the models might be so similar that I would want to divide the probability.

One way to start to think about it is perhaps that I have exchangeable subclasses, and how I might think about using one of the authors' nonparametric set-ups to put priors on those subclasses. I thought about a Dirichlet prior, but the authors have opened up a whole new realm of other possibilities, and I wonder whether the Pólya trees or Bernoulli trips might be useful in this regard.

Thomas Leonard (*University of Edinburgh*)

Are the broad classes of models so thoroughly reviewed by the authors really adequate for modelling unknown distributions? For example, Leonard (1996) demonstrated that Dirichlet processes cannot be adequately extended to hierarchical Bayes modelling; most of the models recommended by the authors contain a variety of conditional independence assumptions and do not lead to flexible posterior smoothing.

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ constitute a random sample from a distribution with multivariate density $f(\mathbf{t})$, for $\mathbf{t} \in B$, some bounded region of Euclidean space. Extending Leonard (1973, 1978), Thorburn (1986) and Hsu and Leonard (1997), let

$$f(\mathbf{t}) = \exp\{g(\mathbf{t})\} \Big/ \int_B \exp\{g(\mathbf{u})\}\, \mathrm{d}\mathbf{u} \qquad (\mathbf{t} \in B),$$

where g denotes the logistic density transform, and the integral can be evaluated by importance sampling from a mixture of multivariate t-densities. Let the prior process of g be Gaussian with mean value function $\mu(\mathbf{t})$, e.g. taking the form $\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{t})\boldsymbol{\beta}$, and covariance kernel $K(\mathbf{s}, \mathbf{t})$. I believe that this continues to provide a very general and flexible paradigm within which these problems can be beneficially investigated. Posterior estimates for f with similar smoothness to the prior covariance kernel are available, and an infinitely differentiable Gaussian covariance kernel is very convenient.

Martin B. Hansen and Steffen L. Lauritzen (*Aalborg University*)
There is another area of application of the group of ideas in the present paper. Assume that we have a random sample $Y_1, \ldots, Y_n$ from a distribution function F, which is related to an underlying distribution G via the integral equation

$$F(x) = \int k(x, y)\, \mathrm{d}G(y),$$

where $k(\cdot, \cdot)$ is an integral kernel. If the distribution of interest is F this is an instance of an infinite mixture problem (as opposed to the finite mixture case studied in Section 2.1.1). This is a familiar problem in statistics and arises in for example convolution and monotone density problems; for a general reference see for example Lindsay (1995). If G is the distribution of interest this is an instance of a statistical inverse problem, which has received much interest lately; see for example Koo and Chung (1998) and many other references therein.

We can attack this type of problem by assuming a nonparametric prior on the distribution G, and then perform the analysis under this assumption. This seems to lead to complicated posterior calculations as we are dealing with indirect observations from the prior distribution. Anyway, the problem can be resolved in some cases by Markov chain Monte Carlo techniques. An analysis along these lines is indeed carried out in Hansen and Lauritzen (1998). In that paper our main interest was to make nonparametric Bayes inference for concave distribution functions. The idea was to use a mixture of Dirichlet processes on the space of probability distributions on $(0, \infty)$ as a prior for G, and then to exploit the unique correspondence between probability distributions on $(0, \infty)$ and concave distribution functions on $(0, \infty)$ through their representation as mixtures of uniform distributions (see for example Feller (1971)).

Applications of Bayesian nonparametric inference to general inverse problems are currently being investigated.

Larry Wasserman (*Carnegie Mellon University, Pittsburgh*)
Nonparametric inference is an exciting area, rich with possibilities yet fraught with difficulties. Diaconis and Freedman (1997) and Cox (1993) showed that apparently reasonable priors may yield posteriors with essentially zero coverage. Only the most doctrinaire Bayesian would fail to be alarmed at near zero coverage. See also Shen and Wasserman (1988), Wasserman (1998) and Zhao (1998). Sometimes these problems can be fixed by using 'sieve priors' (Zhao, 1993). Here, infinite dimensional models are approximated with a sequence of finite dimensional models and the dimension is treated as a parameter. These results suggest that sieve priors might be preferred to purely infinite dimensional priors. For example, for mixture models, using a mixture with k components and treating k as a parameter (Roeder and Wasserman, 1997; Richardson and Green, 1997) may be better than a Dirichlet process mixture (DPM) which puts a prior directly on the space of countable mixtures. Note that the DPM treats the number of components appearing in the sample, not of the underlying density, as unknown.

As Robins and Ritov (1997) showed, sometimes no prior leads to good nonparametric inference. For example, if $W = (Z, X, Y)$ where Z is a high dimensional covariate, X is a binary treatment variable and $f(x|z)$ is known (e.g. X is randomized and the randomization probabilities depend on Z) then the Horvitz–Thompson estimator for the population average treatment effect

$$\theta = \int \{E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z)\}\, \mathrm{d}F(z)$$

is $\sqrt{n}$ consistent, with no assumptions on the joint law. This estimator depends on the ancillary process $f(x|z)$ so the usual Bayesian analysis will not recover anything like this estimate. Instead, the Bayesian is stuck trying to estimate nonparametrically the joint law, which is hopeless if Z is high dimensional. Perhaps in these cases we should all be frequentists.

Nonparametric Bayesian inference has great potential. But, until we know how to choose priors so that the posterior has good rates and coverage properties, we should be cautious.

Hans C. van Houwelingen (*Leiden University*)
The authors show that Bayesian nonparametric inference has made considerable progress in recent years. They start with the question 'Why nonparametrics?'. One might also ask 'Why Bayesian?'.

Bayesian methods have become popular among applied statisticians for two reasons. First, Bayesian

methodology combined with the Markov chain Monte Carlo algorithm can handle complex (hierarchical) models that are almost intractable by classical maximum likelihood. Secondly Bayesian methodology, often disguised as penalized likelihood (Eilers and Marx, 1997), can help to bridge the gap between parametric and nonparametric models. The essential element is that some parameters of the prior (the weight of the penalty; the order of the spline) are estimated from the data. Empirical or hierarchical Bayes models can help the frequentist in adapting the smoothness of his model.

In this paper, Bayesian methodology could be used to control the smoothness of the estimated distribution functions, but this needs further elaboration to show applied statisticians the potential merits of the mathematics presented here. In Section 2.1 the authors remark that the Dirichlet prior assigns counter-intuitive negative correlations. The neutral to the right process with its independent increments does not behave much better. To obtain more flexible models, the amount of correlation between the $F(B)$ of neighbouring Bs should be a parameter in the model that is not chosen *a priori* but estimated from the data in a hierarchical model. The first instance in the paper where such an extra layer could have been added is in the mixture of Dirichlet processes of Section 2.1.1. Adding a scale parameter σ to $f(\cdot|\theta_i)$ and estimating σ from the data (by adding a prior on σ) gives the data-driven smoothness control I would like to advocate.

The second instance where an extra smoothing layer could be added is in the Pólya trees of Section 4. The parameter c controls the smoothness of the distribution function. The authors remark that it could be given a prior, but they use a preset value. It would be interesting to see what the estimate of c would be and how much information about c is available from the data. Adding the extra layer in the model would add to the computational complexity. A simpler way is the kind of Bayesian mixture of models described in Section 5. A mixture of a parametric and a nonparametric model could have been considered in the examples of Sections 3 and 4. That would have enabled a comparison of the fit of the parametric and the nonparametric models and would, presumably, have made a case for Bayesian nonparametrics from a data analytic point of view. As the paper stands, it convincingly shows that it could be done, but not yet that it should.

The following contributions were received in writing after the meeting.

Mark J. Brewer (*University of Exeter*)

I confine attention to the analysis of the Kaplan and Meier (1958) data set from Section 3.4 of the paper.

The authors are concerned with estimation of the probability of failure before 1 month, $F(0, 1)$. Noting that exactly one observed failure occurs before 1 month, an oversimplistic view suggests the estimated probability to be 0.125, and of course it is no coincidence that the posterior means given for $F(0, 1)$ are close to this figure. We might also expect the bulk of the mass of the posterior to be around 0.125, but Fig. 1 shows the nonparametric version having a 'mode' near zero. Assigning such a large probability to $F(0, 1) < 0.05$, for example (around 0.37?), seems odd given the above. Do the authors have a view on this?

To study this further, consider a simplistic nonparametric approach which does not make use of a hazard function. Consider a model based on cross-validated likelihood of kernel densities as in Brewer (1998) but using simulated values for the censored observations, illustrated by the directed graph of Fig. 10(a), and where

$$f(y_j|\{x\}, \{c\}, \lambda) = f_0(y_j|\{x\}, \lambda) + f_c(y_j|\{x\}, \{c\}, \lambda),$$

$$f_0(y_j|\{x\}, \lambda) = \frac{1}{4 - I(j \leqslant 4)} \sum_{\substack{i=1 \\ i \neq j}}^{4} K_\lambda(y_j - x_i),$$

$$f_c(y_j|\{x\}, \{c\}, \lambda) = \frac{1}{4 - I(j \geqslant 5)} \sum_{\substack{i=5 \\ i \neq j}}^{8} \frac{I(y_j > c_j)\, f_0(y_j|\{x\}, \lambda)}{\int I(y_j > c_j)\, f_0(y_j|\{x\}, \lambda)\, \mathrm{d}y_j},$$

x_i represent the observed failures, c_j represent the censoring times, K_λ is a Gaussian kernel function (with reflection at the origin) and λ is the bandwidth with λ^{-2} having a standard non-informative gamma prior ($\alpha = \beta = 0.001$). Note that $y_j = x_j$ for $j = 1, \ldots, 4$.

We construct a Markov chain Monte Carlo algorithm which samples values for y_5–y_8 and λ, and uses the y-values to give a kernel density estimate from which we evaluate $F(0, 1)$ at each iteration. We make

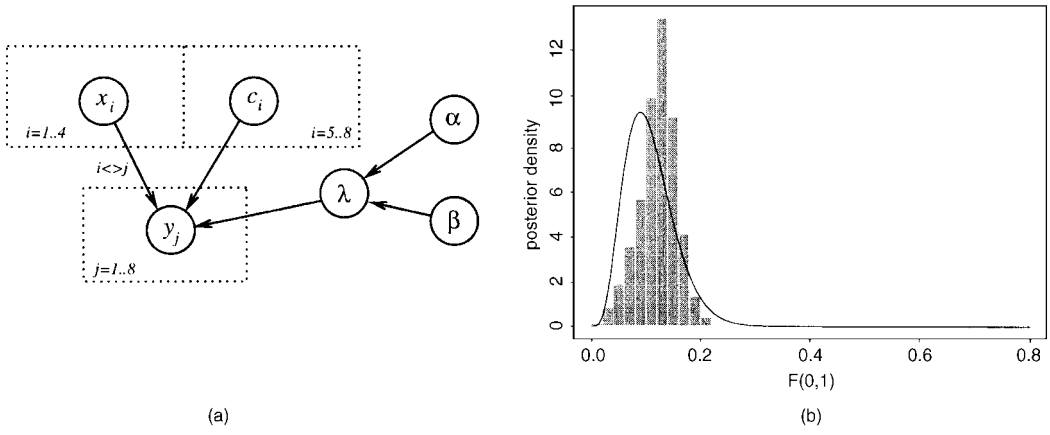

Fig. 10. (a) Independence graph for the cross-validation kernel density model and (b) parametric posterior for $F(0, 1)$ (———) and kernel nonparametric posterior (histogram)

10000 iterations, and a histogram of the sampled $F(0, 1)$ is shown in Fig. 10(b) along with the parametric posterior for comparison. Our posterior mean is 0.122, in accord with those from the paper. As can be seen, the shape of our posterior is much closer to the parametric posterior than is the authors', yet it has not been forced by any parametric assumptions. We would be interested to hear the authors' comments on this.

F. P. A. Coolen (*University of Durham*)

In relation to the presentation of the beta-Stacy process (Section 3), I would like to mention a simple construction of Kaplan and Meier's product limit estimate as presented by Efron (1967). For m observations (including non-informatively right-censored observations), at points $x_1 < \ldots < x_m$, place probability mass $1/m$ at each of these points (no ties are assumed for ease of presentation); if x_{i_1} is the smallest x_i that is censored, remove the mass at x_{i_1} and redistribute it equally among the $m - i_1$ points to the right of it, $x_{i_1+1}, x_{i_1+2}, \ldots, x_m$. If x_{i_2} is the smallest censored value among $x_{i_1+1}, x_{i_1+2}, \ldots, x_m$, redistribute its mass, which will be $\{1 + 1/(m - i_1)\}/m$, among the $m - i_2$ points to its right. Continue in this way until you reach x_m. The problem of what to do with the remaining mass at x_m if this is a right-censored value clearly reflects the general problems with defining the product limit estimate beyond the largest observation if this is a right-censored value. In my opinion, this 'redistribution of probability mass' process provides a very simple insight into the product limit estimate, in particular when presented for lifetime data by assuming that all 'individuals' receive equal probability mass (summing to 1) at the start, putting their mass where they 'die' or, if they are right censored, passing it on to individuals who are still alive (again equally shared). This redistribution of probability mass process also occurs in simple Bayesian models (see Coolen (1997)).

It seems likely that the Bayes estimate based on the beta-Stacy posterior (Section 3.3) can also be interpreted via such a construction, where part of the mass is distributed as above, and the distribution of the residual mass corresponds to the included parameters α and β. This might not only enable a better understanding of this posterior Bayes estimate but also provide more insight into the corresponding prior specifications (Section 3.2). Whether parametric or nonparametric models are used, all distribution of mass differently from according to Efron's construction occurs on the basis of assumptions or information added to the observations, included in a Bayesian analysis either via modelling assumptions or via specification of priors for assumed models.

David G. T. Denison (*Imperial College of Science, Technology and Medicine, London*) **and Bani K. Mallick** (*Texas A&M University, College Station*)

The authors have presented a timely paper bringing together many of their own ideas, and those of others, on Bayesian nonparametric estimation of random distributions. However, as they mentioned, the exposition does not include any discussion on novel Bayesian nonparametric regression methods.

The two problems are inextricably linked and we highlight one of the unanswered questions in nonparametrics and ask the authors for insight.

Consider the usual generalized autoregressive conditional heteroscedastic (GARCH) model (Bollerslev, 1986) which is common when analysing asset returns. The GARCH(1, 1) model can be written as

$$y_t = \epsilon_t \sigma_t, \tag{4}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{5}$$

where y_t and σ_t^2 are the return and volatility at time t ($t = 1, \ldots, T$) respectively and α_0, α_1 and β_1 are coefficients to be estimated. The error at time t is taken to have zero mean and unit variance. Common parametric choices for the error distribution are the standard normal or Student t-distributions even though there is some justification for skewed errors. Once this error distribution is chosen the GARCH(1, 1) model is fully parameterized.

We have extended the GARCH(1, 1) model by relaxing some of the parametric assumptions. Firstly, Denison and Mallick (1998a) modelled the volatility σ_t^2 as a nonparametric function of y_{t-1}^2 and σ_{t-1}^2 rather than the parametric form (5). However, the parametric nature of equation (4) was retained and either a normal or Student t-error distribution was used to model the noise. In later work Denison and Mallick (1998b) did the reverse. A nonparametric error distribution, namely a Pólya tree, was used to model ϵ_t in equation (4) together with the parametric form for σ_t^2 in equation (5). Fixing the variance of the error distribution in this application to be 1 was not trivial. Fixing the scale of a Pólya tree through the interquartile range may be straightforward (Section 4.4.1) but fixing its scale through its variance is problematic.

It is an open question whether modelling with nonparametric regression functions and parametric errors or vice versa is preferable and we welcome comments from the authors on this. In our experience obvious care must be taken to prevent the nonparametric functions overfitting data but it is worth the extra flexibility we gain. Wider credible intervals around the correct distribution or functional form will always be preferable to tighter intervals around a misspecified parametric model.

Using nonparametric forms to model both the errors and the regression function has not been attempted as far as we know. This model would appear to be too flexible for practical use and, we believe, the model would have difficulties in separating the truly random from the deterministic elements of the data set.

Michael D. Escobar (*University of Toronto*)

First, I would like to comment about the comparison in the discussion between parametric and nonparametric methods. Most applied statisticians would start with a formal parametric model, but after looking at the data the statistician might change the parametric model when the data are 'badly behaved'. This means that they had some uncertainty in the real mathematical form of the model. Changing the parametric model by using a mixture or including a heavier-tailed parametric form leads to results that are very similar to some of the nonparametric results. So, whereas the strict use of a parametric model might be near or not near the nonparametric model, the results produced by a statistician will often look the same regardless of the tool used. This has several implications. These nonparametric Bayesian methods potentially provide a full Bayesian justification for the statistician's *ad hoc* model fitting. Also, these methods provide a formalization so that the statistician can now quantify the uncertainty in the mathematical form of the parametric model. This quantification can then be used to guide the adjustment of the analysis in the badly behaved cases.

Second, I would like to comment on the 'several deficiencies' of the Dirichlet process. I think that the authors are being a little harsh here. Remember that the Dirichlet process is a simple two-parameter family. There is the location parameter F_0 and a precision parameter c. Yes, once these are specified, then all the other properties are specified. The advantage of only two parameters is the ease in specifying priors. If we are using the Dirichlet process to model uncertainty in the mathematical form used in a parametric model, then F_0 is that mathematical form. We would then use a distribution on c to reflect the amount of belief in the parametric model. Also, many of the limitations of the Dirichlet process seem to disappear when the Dirichlet process is not used alone. For example, when the Dirichlet process is used in a hierarchical model, then the posterior is no longer a Dirichlet process but a mixture of Dirichlet processes. Yes, I know that the authors are aware of these points. However, it might appear to the casual reader of their paper that the Dirichlet process is a mere historical curiosity instead of the simple and powerful tool that it is.

Steven N. MacEachern (*Ohio State University, Columbus*)

This fine paper exemplifies several uses of nonparametric Bayesian modelling. The benefits of such models are apparent. By fitting a larger class of models, error due to model misspecification is reduced or eliminated. The situation is analogous to a classical problem, where, in a standard regression context, a line is fitted to data that arise from a quadratic regression model. The model misspecification is swept into the error term, resulting in a greater mean-square error than with a quadratic fit. This can lead to a larger standard error for the estimated mean response at each level of the covariate. The line also provides poorer predictions. The situation is muddied in a Bayesian analysis, as the prior distribution moderates the effect of the data, and all our estimates are biased. But, for many models, with larger sample sizes, we still expect to see both a reduction in bias and a reduction in the spread of the posterior distribution for model parameters. Qin (1998) demonstrated this effect in an item response theory model, where posterior distributions for a student's ability are noticeably tighter with a nonparametric, rather than parametric, Bayesian model.

It is now well established that replacing a parametric Bayesian component with a nonparametric Bayesian component improves the fit of many models. To increase acceptance of the models, we need to use the nonparametric components in a directed fashion, developing a collection of modelling strategies that match our fundamental modelling concepts. One such fundamental concept is the distinction between fixed and random effects. For fixed effects, we focus on the effect; for random effects, we focus on the distribution from which the effects come. This distinction, captured by modelling fixed effects with parametric components and random effects with nonparametric components was, to the best of my knowledge, first successfully described and implemented in Bush and MacEachern (1996). The frailty model in Section 4.4.2 falls into this framework, with frailties as random effects, as does much other work in the field. The future of nonparametric Bayesian modelling lies in developing, refining and popularizing further fundamental modelling strategies. I am currently developing a class of correlated nonparametric models that, in addition to allowing covariates to enter the model in natural form, enable us to model collections of distributions from which random effects are drawn as being positively associated, but not identical, with each other. Applications for these models abound.

Pietro Muliere (*Università di Pavia*) **and Piercesare Secchi** (*Politecnico di Milano*)

We congratulate the authors for their interesting and stimulating review of recent developments in Bayesian nonparametrics. Our intention is to complement and reinforce the paper with two comments based on the predictive approach to Bayesian inference.

Exchangeable models

When $\{Y_n\}$ is an infinite sequence of random variables, the completely predictive approach to the construction of the law of the sequence is based on the specification of the distribution F_1 of Y_1 and of the predictive distribution F_{n+1} of Y_{n+1}, given $Y_1, \ldots, Y_n$, for all $n \geqslant 1$. Whereas the Ionescu–Tulcea extension theorem states consistency conditions which guarantee the existence of a unique law for $\{Y_n\}$ determined by the sequence $\{F_n\}$, Fortini *et al.* (1998) (see also Regazzini (1998)) give necessary and sufficient conditions on the sequence $\{F_n\}$ for the exchangeability of the law of $\{Y_n\}$. This result characterizes exchangeability in purely predictive terms; the de Finetti measure of the sequence $\{Y_n\}$ is then obtained by means of de Finetti's representation theorem. Many priors used in Bayesian nonparametrics can easily be constructed following this approach, e.g. the Dirichlet process (Regazzini, 1978; Lo, 1991), Pólya trees (Walker and Muliere, 1997a) and the beta-Stacy process (Walker and Muliere, 1997b).

Partial exchangeability

There are situations where the assumption of exchangeability for the sequence of observables is too restrictive or does not incorporate all the relevant information about the data; a weaker assumption is that of partial exchangeability introduced by de Finetti (1938) and considered also by Diaconis and Freedman (1980). When $\{Y_n\}$ is an infinite sequence of random variables with values in a discrete space, partial exchangeability and recurrence imply that the law of the sequence is that of a mixture of Markov chains (Diaconis and Freedman, 1980), i.e., conditionally on a random transition matrix Π, $\{Y_n\}$ is a Markov chain with transition matrix Π. The prior distribution for Π may often be characterized in purely predictive terms; for example, Muliere *et al.* (1998) introduced an urn scheme called a *reinforced urn process* which generates mixtures of Markov chains such that the law of Π is the product of Dirichlet processes. These processes generalize the Bernoulli trips discussed in the paper; they are also a generalization of the idea of Mauldin *et al.* (1992) for generating Pólya trees. Reinforced urn processes

have ready applications to survival analysis whenever individual specific data are modelled by a Markov chain and individuals from the population are assumed to be exchangeable.

M. A. Newton (*University of Wisconsin, Madison*) **and F. A. Quintana** (*Pontificia Universidad Catolica de Chile, Santiago*)

We echo the sentiments expressed in Section 2.1.1 on the complexity and computational intensity of available Monte Carlo methods for fitting Dirichlet-process-based models. In recent work we have investigated a simple recursive algorithm to fit such models quickly, albeit approximately (Newton and Zhang, 1999; Newton *et al.*, 1998; Tao *et al.*, 1999). This methodology may be helpful in the early stages of data analysis before Markov chain Monte Carlo implementation.

The authors discuss diffuse prior limits taken after the data are available, noting convergence of a Bayes estimator to the Kaplan–Meier estimate, for example (Section 3.3). Susarla and Van Ryzin (1976) pointed this out for the Dirichlet process prior. Here, and in other examples, the diffuse prior limit equals the nonparametric maximum likelihood estimator (NPMLE). The Dirichlet process Bayes estimator from case I interval-censored data (e.g. current status data) fails to converge to the NPMLE in the diffuse prior limit (Steve MacEachern, personal communication (1994), Newton (1994) and Newton and Zhang (1999)). Does convergence to the NPMLE occur for any of the more flexible priors?

It is interesting that a reinforced random walk on a graph — Pemantle (1988) discussed the case where the graph is an infinite tree — can be partially exchangeable and thus may be used to model dependent time series data (Diaconis, 1988; Quintana and Newton, 1998). The Bernoulli trips in Section 5.1 retain the special exchangeability structure. For example, take the graph $G = (V, E)$ with $V = \{0, 1\}$, $E = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ and let $X_1, X_2, \ldots$ denote a reinforced random walk on G, say starting at $X_1 = 0$. This sequence turns out to be a mixture of binary Markov chains, just as a binary Pólya sequence is a mixture of coin tosses. The extension to real-valued sequences seems to be an open problem.

The authors acknowledge difficulty in extending the stochastic process approach beyond relatively simple models. Given the similarity of numerical results between priors (Section 3.4), and recent work of Regazzini (1998) that any exchangeable sequence can be well approximated, in terms of the Prokhorov metric, by certain mixtures of Dirichlet processes, we wonder whether mixtures of Dirichlet processes might tend to be favoured over the constructions in Section 3. Have the authors found numerical examples where the beta-Stacy results are significantly different from those obtained by using natural mixtures of Dirichlet processes?

Sonia Petrone (*Università dell'Insubria, Milan*)

I congratulate the authors for their very interesting and useful review of Bayesian nonparametric inference and applications, particularly in survival analysis. I would like to point out a few possible directions for development.

Exchangeability is the basic framework for Bayesian inference, but more general dependence structures are often more realistic. Partial exchangeability might be appropriate for non-homogeneous data. Nonparametric priors for Bayesian inference with partially exchangeable data, which allow dependence among groups, have been proposed by Cifarelli and Regazzini (1978). An important development would be to incorporate uncertainty about the partition of data.

Most of the priors proposed for Bayesian nonparametric inference are discrete. Discreteness of the Dirichlet process originates from the beautiful form of the predictive distribution and is not, in general, a problem. However, it might have unexpected consequences outside the exchangeability framework, e.g. when data are partially exchangeable, with uncertainty about the partition (Petrone and Raftery, 1997).

Continuous nonparametric priors are reviewed in Section 2.1.1 (since 'mixtures of Dirichlet processes' is used by Antoniak (1974) with a different meaning, would 'mixtures with a Dirichlet process mixing distribution' be a clearer term?). Another proposal is to construct a continuous nonparametric prior by considering polynomial approximations of the random probability measure, as suggested by Petrone (1999a) using Bernstein polynomials for data in [0, 1]. There are connections with the idea of using mixtures of basis functions with random weights and random numbers of components (Petrone, 1999b). The 'Bernstein polynomial prior' is consistent and can be generalized to data on the positive real line or to multidimensional data.

Concerning *asymptotics*, consistency is a motivation for being nonparametric (Diaconis and Freedman, 1983). Le Cam (1986), p. 618, said that even more interesting than consistency is to study the asymptotic behaviour of the posterior. Dealing with an infinite dimensional parameter, this means to

prove an infinite dimensional version of the Bernstein–von Mises theorem (Diaconis and Freedman, 1997). For evaluating the posterior, large sample approximations could replace simulation techniques, which might be computationally expensive if the sample size is large.

Prior knowledge about the location and spread of data might be incorporated in the prior distribution of $E_F[X]$ and $V_F(X)$, whose analytic expression when F is a Dirichlet process is known (Cifarelli and Regazzini, 1990; Cifarelli and Melilli, 1999).

Finally, I believe that the predictive approach outlined in Section 5 might be a fruitful line of research also for overcoming the prior-to-posterior computation difficulties.

Gareth Roberts (*University of Lancaster*)

I would like to add my congratulations to the authors for this timely and important paper.

The concept of the neutral to the right (NTTR) process represents a natural framework for non-identifiability. However, it does endow the scale on which the data are measured a special status. In other words, the concept of an NTTR process is not preserved under monotone increasing transformations. This may or may not be appropriate. However, in the spirit of nonparametric analysis it is interesting to see whether this framework can be extended to include invariance under monotone increasing transformations without losing tractability.

A natural extension of NTTR processes replaces the Lévy process by a *time-changed* Lévy process $C(t) = Z\{\gamma(t)\}$, where Z is a Lévy process satisfying the conditions of Section 3 and γ is a monotone increasing transformation with $\lim_{t\to\infty}\{\gamma(t)\} = \infty$. Such processes are as tractable as Lévy processes.

This would provide a more flexible class of prior distributions. One simple idea would be to have a parametric family for γ. This would then allow easy implementation of Markov chain Monte Carlo algorithms by adding a γ updating step. The conjugacy property of theorem 3 is replaced by a conditional conjugacy, conditional on γ.

One natural class for γ would be the power class $\{\gamma_\beta(\cdot),\ \beta \in (0, \infty)\}$, where $\gamma_\beta(t) = t^\beta$. More complicated classes could incorporate stochastic or deterministic information about the support of the distribution.

Jean-Marie Rolin (*Université Catholique de Louvain, Louvain-la-Neuve*)

It is a pleasure to comment on this interesting paper on Bayesian nonparametric inference, as some of the aspects are covered in Rolin (1998). My comments will mainly concern the first three sections.

The introduction is convincing in motivating statisticians to use Bayesian nonparametric methodology. Concerning the general framework, theorem 1 requires some assumptions ($(\Omega, \mathcal{B})$ is a standard Borel space and the expectation of $F(B)$ is a probability measure).

Even if the Dirichlet process has deficiencies, they are not so important. If the variance has a specified form and the correlation between the probabilities of two disjoint sets is negative, the same is true for the multinomial distribution. Moreover, negative correlation is not counter-intuitive. When the probability of a set increases, the probability of another set must tend to decrease since the sum is less than or equal to 1. The prior elicitation of the variance is not easy except if it is deduced from a parametric model. The authors themselves comment that the beta-Stacy approach is indistinguishable from the Dirichlet process prior and conclude that the significant differences are not between choices of the prior.

In Section 3, the prior is a beta-Stacy process if and only if the predictable hazard process

$$Z^-(t) = \int_{[0,t]} F\{[s, \infty)\}^{-1} F(\mathrm{d}s) \tag{6}$$

is a beta process as defined by Hjort (1990). These seemingly different Bayesian models are in fact the same.

Theorem 2 with right-censored observations is due to Ferguson and Phadia (1979) but these observations do not show up in the given Bayes estimate.

The example in Section 3.4 is not very convincing. First, the simulated distribution is known because there is no censoring in the interval [0, 1] and therefore

$$F([0, 1]) \sim \mathrm{beta}\{2 - \exp(0.1), 7 + \exp(0.1)\}. \tag{7}$$

The histogram does not reveal that the density of this distribution vanishes at zero. However, kernel estimation of the density in Walker and Damien (1998) shows it. The simulation proposed by Florens and Rolin (1998) clearly shows this property and seems much more accurate and fast but is only feasible

for Dirichlet priors. Also the essential indistinguishability between histograms resulting from the beta-Stacy and the Dirichlet prior is strange because the sample is small and the (predictive) expected lifetime is infinite in the first model and 10 in the second. Consequently, a comparison between the nonparametric Dirichlet and the parametric model should at least require expected lifetimes to be the same in both, i.e. choosing *a priori* $a \sim \text{ga}(2, 10)$ (equating prior expectations of $F([0, 1])$ would give $a \sim \text{ga}(1, 9.5)$).

Finally, an important problem is to extend the simple constructive definition of Dirichlet processes of Sethuraman (1994) to beta or beta-Stacy processes. This will permit inference on more general characteristics than survival probabilities (e.g. expected residual lifetimes) and will provide a full Bayesian nonparametric analysis of models with explanatory variables (see Kalbfleisch (1978) and Florens *et al.* (1998)).

The **authors** replied later, in writing, as follows.

As Draper has reminded us, the formal need for the construction of probability distributions on (distribution, hazard, etc.) function spaces is implicit in de Finetti's representation theorem. Unfortunately, the theorem is an 'existence' theorem and — unless we impose further structural conditions — does not provide guidance on how to construct such a distribution. However, the analysis depends crucially on this construction, or prior, since the posterior is fully defined by the prior and the data.

The task then is to construct the prior. Parametric priors assign probability 1 to a subset of the function space, either for convenience or because it is genuinely felt by the modeller or analyst that this allocation of probability is warranted. However, examples given in the paper and the contributions of Draper and MacEachern make clear that it is often desirable not to force a parametric shape to the unknown. In the paper we look at a selection of types of nonparametric prior; other forms are suggested by Richardson, Leonard, Wasserman, Brewer, Muliere and Secchi, and Petrone.

Draper's illustration via a hierarchical Pólya tree model is interesting; as a general rule, we concur with his cautionary remarks on Markov chain Monte Carlo (MCMC) convergence. Regarding his warnings on the modelling aspects, for the example presented, it may be possible to avoid some of the difficulties by employing, perhaps, a different nonparametric prior, such as a mixture of Dirichlet process (MDP) prior.

As Richardson suggests, popular forms of prior tend to coincide with those which are easy to work with, in particular in the sense of being able to specify prior parameters straightforwardly and to update from prior to posterior. One way to think about the former is to have a prior for which it is possible to specify $E[f]$ and $\text{var}(f)$, where f is the unknown (random) function. We see this as a good general procedure for making use of information that is typically available to the experimenter.

For the censored data case, there should be a simple adaptation of what Richardson calls a 'canonical assignment'. Using a data-dependent partition does not automatically imply a data-dependent prior, thus rendering the Bayesian guilty of double use of data. If the distribution of the random (conditional) probability of any resulting partition is predefined via F_0, there is no double use of data: the partition points are merely for convenience and approximation.

One of the most widely used nonparametric priors is the MDP prior. Newton and Quintana highlight a result of Regazzini, namely that any exchangeable sequence can be well approximated, in the Prokhorov metric, via the MDP. Fine, but this does not provide a constructive specification of the prior.

Wasserman and Petrone mention the notion of consistency, i.e. as the sample size tends to ∞ the posterior accumulates around the true function. This may be a desirable property. However, this property must be balanced with the need to incorporate prior information fully, since posteriors are usually based on far from asymptotic samples. Constructing a prior by studying the asymptotic properties of the posterior, at the expense of prior information, does not appear to be sensible. Model identifiability from finite samples is more important than an ability to discriminate between slightly different models from very large samples.

Draper points out problems with Pólya trees and Richardson problems with the exchangeable priors. Wasserman concludes that Bayes nonparametrics do not solve everything. We are aware that the construction of probabilities on large function spaces will not be trouble free; like all other procedures in statistics, Bayesian nonparametrics must also be taken with a sizable pinch of salt. Nonparametric conditional distributions, of the kind described by Wasserman in the Robins–Ritov paradox, are particularly problematic. This said, the difficulties are worth surmounting and rapid progress in the next few years is expected.

A common way of thinking about the Bayesian approach is that it attaches a prior to the likelihood. We prefer to think in terms of a single probability on the space of distribution functions. From this perspective, Nelder's question of whether the prior and likelihood are consistent in Section 3.4 seems inappropriate. If the experimenter has assigned probability 1 to a parametric subset then it is this assignment which might be called into question in the light of data. However, what does a Bayesian do? Should he or she keep assigning probability 1 to parametric subsets until he or she is happy? Draper says that this is a serious problem. The answer is to not assign probability 1 to restricted subsets but to a class that is sufficiently large (hopefully) to contain the 'truth': hence Bayes nonparametrics. Also, in Section 3.4 we are comparing the posteriors from a nonparametric prior and a parametric prior in which the parametric prior is used to centre the nonparametric prior. It is with this in mind that we compare the densities in Fig. 1. As Brewer discovers, not all nonparametric priors lead to the same posterior.

From the public relations perspective, van Houwelingen suggests that Bayes nonparametrics might be sold in other ways to appeal to non-Bayesians.

Hansen and Lauritzen's work is similar in spirit to that of Brunner and Lo (1989), who used the Dirichlet process and the uniform distribution for modelling a unimodal density.

Coolen describes a result of Efron on the derivation of the Kaplan–Meier estimator. The redistribution of mass seems to lack theoretical back-up and it is possibly more insightful to think of the estimator as a predictive distribution for the next observation of an exchangeable sequence, the data forming the start of the sequence.

Brewer refers to his approach as 'simplistic'. Perhaps it is. But, like Coolen's, his approach is clearly *ad hoc*. Roberts's suggestion is very interesting indeed and warrants further investigation. George's questions are intriguing, but clearly require some deep thoughts.

Denison and Mallick use nonparametric priors for the GARCH(1, 1) model. A question is whether the GARCH(1, 1) model should be generalized. Its existence and form are for a specific reason and generalizations remove these reasons. One might as well start again from scratch to introduce a flexible model for economic time series.

Escobar and Rolin believe the Dirichlet process to be a powerful tool; we agree. Escobar says that one can use a distribution on the scale parameter of the Dirichlet process, to reflect the amount of belief in the parametric centring model. Allocating such a prior must be troublesome since it is difficult to quantify belief in a parametric model. Again, we re-emphasize our idea of working with priors for which $E[f]$ and $\mathrm{var}(f)$ can be specified arbitrarily, and reasons were outlined in the paper.

Rolin's, and Hjort's (1990), definition of the hazard process is

$$Z_{\mathrm{R}}(t) = \int_0^t \frac{\mathrm{d}F(s)}{F[s, \infty)}.$$

Ours is

$$Z(t) = -\log\{1 - F(t)\}.$$

Either is possible, but if F is neutral to the right the jumps of Z_{R} are bounded by 1, which might be a handicap for modelling a continuous hazard. Z is a log-beta process, rather than a beta process, if F is a beta-Stacy process.

We thank Rolin, and MacEachern, for pointing out a problem with Fig. 1 (there should be a mode in the histogram at about 0.01).

References in the discussion

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B, **60**, 725–749.

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

Bollerslev, T. (1986) Generalised autoregressive conditional heteroscedasticity. *J. Econometr.*, **51**, 307–327.

Brewer, M. J. (1998) A modelling approach for bandwidth selection in kernel density estimation. In *Compstat 1998* (eds R. Payne and P. Green), pp. 203–208. Heidelberg: Physica.

Brunner, L. J. and Lo, A. Y. (1989) Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, **17**, 1550–1566.

Bush, C. A. and MacEachern, S. N. (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–286.
Cifarelli, D. M. and Melilli, E. (1999) Some new results for Dirichlet priors. *Technical Report*. Istituto di Metodi Quantitativi, Università Bocconi, Milan.
Cifarelli, D. M. and Regazzini, E. (1978) Problemi statistici nonparametrici in condizioni di scambiabilità parziale: impiego di medie associative. *Quad. Ist. Mat. Finan. Univ. Torino* III, **12**, 1–36.
———(1990) Distribution functions of means of a Dirichlet process. *Ann. Statist.*, **18**, 429–442.
Coolen, F. P. A. (1997) An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations. *Reliab. Engng Syst. Safty*, **56**, 61–68.
Cox, D. D. (1993) An analysis of Bayesian inference for non-parametric regression. *Ann. Statist.*, **21**, 903–924.
Denison, D. G. T. and Mallick, B. K. (1998a) A nonparametric Bayesian approach to modelling nonlinear time series. *Technical Report*. Imperial College of Science, Technology and Medicine, London.
———(1998b) Analysing financial data using Pólya trees. *Technical Report*. Imperial College of Science, Technology and Medicine, London.
Diaconis, P. (1988) Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 111–125. Oxford: Oxford University Press.
Diaconis, P. and Freedman, D. (1980) de Finetti's theorem for Markov chains. *Ann. Probab.*, **8**, 115–130.
———(1983) On inconsistent Bayes estimates in the discrete case. *Ann. Statist.*, **11**, 1109–1118.
———(1997) On the Bernstein-von Mises theorem with infinite dimensional parameter. *Technical Report 492*. Department of Statistics, University of California at Berkeley, Berkeley.
Draper, D. (1999a) Discussion on Decision models in screening for breast cancer, by G. Parmigiani. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.
———(1999b) 3CV: Bayesian well-calibrated model selection via cross-validation. *Technical Report*. Statistics Group, University of Bath, Bath.
Draper, D., Cheal, R. and Sinclair, J. (1998) Fixing the broken bootstrap: Bayesian non-parametric inference with highly skewed and heavy-tailed data. *Technical Report*. Statistics Group, University of Bath, Bath.
Draper, D., Hodges, J., Mallows, C. and Pregibon, D. (1993) Exchangeability and data analysis (with discussion). *J. R. Statist. Soc.* A, **156**, 9–37.
Efron, B. (1967) The two sample problem with censored data. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (eds L. M. Le Cam and J. Neyman), vol. 4, pp. 831–853. New York: Prentice Hall.
Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with splines and penalties (with discussion). *Statist. Sci.*, **11**, 89–112.
Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, 2nd edn, vol. II. New York: Wiley.
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.
de Finetti, B. (1930) Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Linc.*, **4**, 86–133.
———(1938) Sur la condition d' "equivalence partielle". *Act. Sci. Ind.*, **739**, 5–18.
———(1980) Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (eds H. E. Kyburg, Jr, and H. E. Smokler), pp. 93–158. New York: Wiley.
Florens, J. P., Mouchart, M. and Rolin, J. M. (1998) Semi- and non-parametric Bayesian analysis of duration models. *Int. Statist. Rev.*, to be published.
Florens, J. P. and Rolin, J. M. (1998) Simulation of posterior distributions in non-parametric censored analysis. Submitted to *Scand. J. Statist.*
Fortini, S., Ladelli, L. and Regazzini, E. (1998) Exchangeability, predictive distributions and parametric models. *Technical Report 98.5*. Consiglio Nazionale delle Ricerche, Milan.
Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997) Consistency issues in Bayesian nonparametrics. Unpublished.
Green, P. J. and Richardson, S. (1998) Modelling heterogeneity with and without the Dirichlet process. *Research Report S-98-02*. Department of Mathematics, University of Bristol, Bristol.
Hansen, M. B. and Lauritzen, S. L. (1998) Nonparametric Bayes inference for concave distribution functions. *Technical Report R-98-2015*. Department of Mathematics, Aalborg University.
Hjort, N. L. (1990) Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, **18**, 1259–1294.
Hsu, J. S. J. and Leonard, T. (1997) Bayesian semi-parametric procedures for logistic regression. *Biometrika*, **84**, 85–93.
Kalbfleisch, J. D. (1978) Non-parametric Bayesian analysis of survival time data. *J. R. Statist. Soc.* B, **40**, 214–221.
Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Statist. Ass.*, **53**, 457–481.
Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999) Bayesian model choice: what and why (with discussion)? In *Bayesian Statistics 6* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.

Koo, J.-Y. and Chung, H.-Y. (1998) Log-density estimation in linear inverse problems. *Ann. Statist.*, **26**, 335–362.
Lavine, M. (1992) Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235.
Le Cam, L. (1986) *Asymptotics Methods in Statistical Decision Theory*. New York: Springer.
Leonard, T. (1973) A Bayesian method for histograms. *Biometrika*, **60**, 297–308.
——— (1978) Density estimation, stochastic processes and prior information (with discussion). *J. R. Statist. Soc.* B, **40**, 113–146.
——— (1996) On exchangeable sample distributions for uncontrolled data. *Statist. Probab. Lett.*, **26**, 1–6.
Lindley, D. V. (1972) *Bayesian Statistics: a Review*. Philadelphia: Society for Industrial and Applied Mathematics.
Lindley, D. V. and Novick, M. (1981) The role of exchangeability in inference. *Ann. Statist.*, **9**, 45–58.
Lindsay, B. G. (1995) Mixture models: theory, geometry and applications. *Regl Conf. Ser. Probab. Statist.*, **5**.
Lindsey, J. K. (1999) Some statistical heresies (with discussion). *Statistician*, **48**, 1–40.
Lo, A. Y. (1991) A characterization of the Dirichlet process. *Statist. Probab. Lett.*, **12**, 185–187.
Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992) Pólya trees and random distributions. *Ann. Statist.*, **20**, 1203–1221.
Muliere, P., Secchi, P. and Walker, S. (1998) Urn schemes and reinforced random walks for Bayesian nonparametrics. *Quaderni di Dipartimento 71(2-98)*. Dipartimento Economia Politica e Metodi Quantitativi, Università di Pavia, Pavia.
Muliere, P. and Walker, S. G. (1997) A Bayesian nonparametric approach to survival analysis using Pólya trees. *Scand. J. Statist.*, **24**, 331–340.
Newton, M. A. (1994) A diffuse prior limit in semiparametric binary regression. *Proc. Bayesian Statistical Science Sect. Am. Statist. Ass.*, pp. 181–186.
Newton, M. A., Quintana, F. A. and Zhang, Y. (1998) Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semi-parametric Bayesian Statistics* (eds D. Dey, P. Muller and D. Sinha), pp. 45–61. New York: Springer.
Newton, M. A. and Zhang, Y. (1999) A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, **86**, in the press.
Pemantle, R. (1988) Phase transitions in reinforced random walk and RWRE on trees. *Ann. Probab.*, **16**, 1229–1241.
Petrone, S. (1999a) Random Bernstein polynomials. *Scand. J. Statist.*, **26**, 1–21.
——— (1999b) Bayesian density estimation using Bernstein polynomials. *Can. J. Statist.*, to be published.
Petrone, S. and Raftery, A. E. (1997) A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statist. Probab. Lett.*, **36**, 69–83.
Qin, A. L. (1998) Nonparametric Bayesian models for item response data. *PhD Dissertation*. Ohio State University, Columbus.
Quintana, F. A. and Newton, M. A. (1998) Assessing the order of dependence for partially exchangeable binary data. *J. Am. Statist. Ass.*, **93**, 194–202.
Regazzini, E. (1978) Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *G. Ist. Ital. Att.*, **41**, 77–89.
——— (1998) Old and recent results on the relationship between predictive inference and statistical modelling either in nonparametric or parametric form. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.
Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792.
Robins, J. and Ritov, Y. (1997) Toward a curse of dimensionality appropriate asymptotic theory for semiparametric models. *Statist. Med.*, **16**, 285–319.
Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *J. Am. Statist. Ass.*, **92**, 894–902.
Rolin, J.-M. (1998) Bayesian survival analysis. In *Encyclopedia of Biostatistics*, vol. 1, pp. 271–286. New York: Wiley.
Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.
Shen, X. and Wasserman, L. (1998) Rates of convergence of posterior distributions. *Technical Report 678*. Carnegie Mellon University, Pittsburgh.
Susarla, V. and Van Ryzin, J. (1976) Nonparametric Bayesian estimation of survival curves from incomplete data. *J. Am. Statist. Ass.*, **71**, 897–902.
Tao, H., Palta, M., Yandell, B. S. and Newton, M. A. (1999) An estimation method for the semiparametric mixed effects model. *Biometrics*, **55**, 191–202.
Thorburn, D. (1986) A Bayesian approach to density estimation. *Biometrika*, **73**, 65–75.
Walker, S. G. and Damien, P. (1998) A full Bayesian nonparametric analysis involving a neutral to the right process. *Scand. J. Statist.*, **25**, 669–680.
Walker, S. and Muliere, P. (1997a) A characterization of Pólya tree distributions. *Statist. Probab. Lett.*, **31**, 163–168.
——— (1997b) Beta-Stacy processes and a generalization of the Pólya urn scheme. *Ann. Statist.*, **25**, 1762–1780.
Wasserman, L. (1998) Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds D. Dey, P. Müller and D. Sinha). New York: Springer.
Zhao, L. (1993) Frequentist and Bayesian aspects of some nonparametric estimation. *PhD Thesis*. Cornell University, Ithaca.