Parametric Nonparametric Statistics: An Introduction to Mixtures of Finite Polya Trees

Ronald CHRISTENSEN, Timothy HANSON, and Alejandro JARA

We present an introduction to an exciting approach to Bayesian nonparametrics, mixtures of Polya trees (MPTs). MPTs can be viewed as a simple generalization of standard parametric statistical distributions. MPTs use a partition of the support of the original distribution's density. The more general density retains the shape of the original distribution on each partition set but adds new parameters that are conditional probabilities. This provides a highly flexible family of distributions, one that is appropriate for nonparametric fitting. MPTs allow for data-driven features to emerge that are sometimes surprising, such as multimodality and skewness, and can vastly improve model fit relative to the original parametric family. Polya tree models are broadly applicable and easily programmed given existing MCMC schemes for fitting the original parametric model.

Our examples include Paraguayan monkey hunting and toenail fungus treatment. In the first of these examples, we find that a normal theory model works quite well, but that there is little price to be paid for the extra generality of fitting a mixture of Polya trees.

KEY WORDS: Bayesian; Generalized linear mixed model; GLMM.

1. INTRODUCTION

When stranded in the wilds of Paraguay, who should you make friends with? When stuck in the wilds of Belgium, how do you get rid of your toenail fungus? We use mixtures of finite Polya trees (MPTs) to help answer these questions.

With the advent of inexpensive, high-speed computing, datasets are getting larger and our ability to perform sophisticated analyses has increased. In particular, nonparametric methods are increasingly supplanting traditional parametric approaches to data analysis. Ironically, nonparametric statistics deals with families of sampling distributions that involve huge numbers of parameters. We discuss a method for generalizing standard parametric families of distributions into broader classes involving many more parameters. These broader classes are useful for dealing with nonparametric problems in general and are particularly appropriate when the original parametric family is plausible but not certain. The classes are also highly amenable to Bayesian analysis.

In recent years, Bayesian analysis has become much more widely used. Despite its philosophical virtues (see, e.g., Christensen 2005), the reason for Bayesian analysis' increasing popularity seems to be its ability to deal with complicated models in a cogent fashion. This is seen in many venues such as multistate models (Huzurbazar 2005), hierarchical models (Gelman et al. 1995), and nonparametric problems as discussed here. It relies on the ability to use Markov chain Monte Carlo (MCMC) methods to approximate posterior distributions, see Casella and George (1992) and Robert and Casella (2004).

A non-Bayesian analysis of these generalized distributions would treat all of the parameters on an equal footing. One advantage to our Bayesian approach is that as we add more parameters, the more recently added parameters are required to have more data-based evidence to justify departures from the original parametric family. Thus, the Bayesian analysis gives good answers with numbers of parameters that equal or even exceed the number of independent observations in the data. In other words, the Bayesian approach allows great flexibility but it does not use that flexibility unless the data clearly call for it. A non-Bayesian approach would have to use fewer parameters and thus allow less flexibility or at least introduce a substantial penalty for small scale deviations away from the original parametric family.

A Bayesian analysis requires prior probabilities for the parameters of the sampling distribution. It can be difficult to choose reasonable priors in problems with many parameters, so the parameters here are broken into two convenient groups: (1) those associated with the original parametric family, and (2) those associated with the generalization of the original family. Standard methods can be used to elicit a prior for the original parameters; see Bedrick, Christensen, and Johnson (1996) for an approach in generalized linear models. Convenient reference priors exist for the parameters from the generalization. A Polya tree is a random probability distribution. Conditional on a fixed member of the original parametric family (e.g., a normal distribution with mean μ and variance σ^2), the generalized family of distributions together with the reference prior is a finite Polya tree. Integrating over a prior on the parameters of the original family (e.g., μ and σ^2) forms a mixture of Polya trees.

Ronald Christensen is Professor of Statistics, Department of Mathematics and Statistics, University of New Mexico (E-mail: *fletcher@stat.unm.edu*). Timothy Hanson is Associate Professor of Biostatistics, Division of Biostatistics, University of Minnesota. Alejandro Jara is Associate Professor, Department of Statistics, Universidad de Concepción, Chile. The authors thank Novartis, Belgium, for permission to use their dermatological data for statistical research and the reviewers for their valuable comments.

  DOI: 10.1198/000313008X366983

MPTs are very broadly applicable but both of our examples involve the use of generalized linear mixed models (GLMMs), see Breslow and Clayton (1993), and both involve generalizing the normal distributions commonly assumed with them. GLMMs provide a popular framework for the analysis of longitudinal measures and clustered data that arise in many areas such as agriculture, biology, epidemiology, economics, and geophysics. The models account for correlation among clustered observations by including random effects in the linear predictor component of the model. Although GLMM fitting is typically complex, standard random intercept and random intercept/slope models with normally distributed random effects can now be routinely fitted in such commercial software packages as SAS and Stata. Such models are quite flexible in accommodating heterogenous behavior, but they suffer from the same lack of robustness against departures from distributional assumptions as other statistical models based on Gaussian distributions.

Lack of robustness to normality may be a more severe problem with GLMMs than with linear mixed models. While linear mixed models typically assume normality, the estimates and predictions are also best linear unbiased estimates and predictions. This linear optimality requires only assumptions about second moments, rather than normality, thus giving some measure of robustness to the linear mixed model analysis. No comparably appealing property seems to hold for nonlinear GLMMs.

An obvious strategy for guarding against inappropriate normality assumptions is to incorporate more flexible distributional assumptions for the random effects into the model. Therefore, a nonparametric extension of parametric GLMMs seems appealing. We illustrate these extensions and the consequences of the incorrect use of traditional model assumptions in GLMMs with our two real-life examples. For the sake of comparison, we fitted all models under both the assumption of normally distributed random effects and our generalization.

Section 2 illustrates the development of finite Polya trees generalizing the normal family of distributions. The development for other parametric families follows a similar pattern. Section 3 uses MPTs to address the toenail and Paraguayan problems. Section 4 examines applications that have been made of mixtures of finite Polya trees along with future directions.

Polya trees are random distributions generated using trees of Polya urn models. The general definition of mixtures of finite Polya trees is quite broad and mathematical and will not be given here. The theory goes back to the seminal works of Freedman (1963) and Fabius (1964) that defined tail-free processes. Ferguson (1974) provided an early review of Bayesian nonparametrics. Blackwell and McQueen (1973) showed that general Polya trees include Dirichlet processes as a special case. Unlike Dirichlet processes, the family of Polya trees can assign probability one to the set of continuous distributions; see Dubins and Freedman (1967), Kraft (1964), and Metivier (1971). Polya trees were later investigated by Mauldin and Williams (1990), Mauldin, Sudderth, and Williams (1992), and Lavine (1992, 1994), who also introduced mixtures of Polya trees. Hanson and Johnson (2002) and Hanson (2006) elaborated the theory for various mixture of Polya tree models and discussed computational approaches.

2. MIXTURES OF FINITE POLYA TREES

Using a relatively simple definition of Polya trees from Hanson (2006), we illustrate the process of generalizing a parametric family of distributions using the $N(\mu, \sigma^2)$ family, see Figure 1(a). Other parametric families are generalized similarly.

The generalization goes through a number of stages, say J. At each stage we introduce new parameters to generalize the previous stage. At the first stage, we split the real number line, that is, the support of the normal distribution, into two intervals divided by the median μ. We then allow changes in the probabilities of being below or above μ, but we retain the shape of the normal density both below μ and above μ. Figure 1(b) illustrates the density for the case when the probability of being below μ is 0.45.

The new parameters at the first stage are θ_{11}, the probability of being no greater than μ, and θ_{12}, the probability of being above μ. Formally, let X_1 have the first stage distribution, then

$$\theta_{11} \equiv \Pr[X_1 \leq \mu],$$

and

$$\theta_{12} \equiv \Pr[X_1 > \mu] = 1 - \theta_{11}.$$

Because we retain the shape of the normal on both sets, if $a \leq \mu$ and $Y \sim N(\mu, \sigma^2)$, conditionally we have

$$\Pr[X_1 \leq a | X_1 \leq \mu] = \frac{\Pr[Y \leq a]}{0.5} = 2\Phi[(a-\mu)/\sigma)],$$

where $\Phi(\cdot)$ is the cdf of a standard normal. Similarly, if $b > \mu$,

$$\Pr[X_1 > b | X_1 > \mu] = 2\Pr[Y > b] = 2\{1 - \Phi[(b-\mu)/\sigma)]\}.$$

Alternatively, we can write

$$\begin{aligned}\Pr[X_1 \leq a] &= \Pr[Y \leq a] 2\theta_{11} \\ \Pr[X_1 > b] &= \Pr[Y > b] 2\theta_{12}.\end{aligned}$$

With $I_A(x)$ the indicator function of A, the density of the stage 1 distribution is

$$f(x_1|\mu, \sigma^2, \theta_{11}, \theta_{12}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_1-\mu)^2/2\sigma^2} \times 2^1 \left[\theta_{11} I_{(-\infty,\mu]}(x_1) + \theta_{12} I_{(\mu,\infty)}(x_1)\right]. \quad (1)$$

In the first stage of the process, we split the real line at the median μ of the $N(\mu, \sigma^2)$ distribution. For the second stage, we split the line at the quartiles, say, q_1, μ, q_3 of the original distribution leading us to consider the sets $(-\infty, q_1]$, $(q_1, \mu]$, $(\mu, q_3]$, (q_3, ∞). For the normal, the quartiles are $q_1 \equiv \mu - 0.6745\sigma$, μ, and $q_3 \equiv \mu + 0.6745\sigma$. Under the original distribution, each of these sets has probability 0.25, but in stage 2 we allow the probabilities of the sets to change in a manner that is consistent with stage 1. An illustration of a stage 2 density is Figure 1(c).

Letting X_2 have the second stage distribution, we introduce new parameters, $\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}$, defined as conditional probabilities relative to the sets used in stage 1:

$$\begin{aligned}\theta_{21} &= \Pr[X_2 \leq q_1 | X_2 \leq \mu] \\ \theta_{22} &= \Pr[q_1 < X_2 \leq \mu | X_2 \leq \mu] \\ \theta_{23} &= \Pr[\mu < X_2 \leq q_3 | X_2 > \mu] \\ \theta_{24} &= \Pr[q_3 < X_2 | X_2 > \mu].\end{aligned}$$

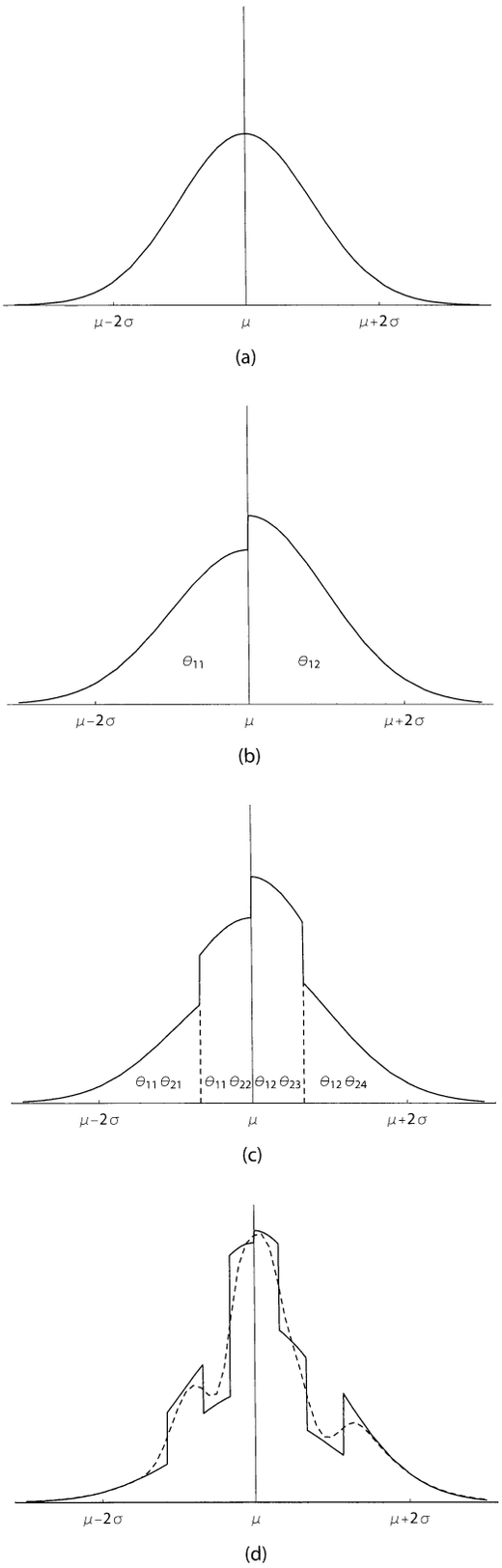

Figure 1. Finite PT densities. (a) Original $N(\mu, \sigma^2)$ centering density. (b) First stage: $\theta_{11} = 0.45$. (c) Second stage: $\theta_{21} = 0.4$, $\theta_{23} = 0.6$. (d) Third stage: $\theta_{31} = 0.3$, $\theta_{33} = 0.3$, $\theta_{35} = 0.6$, $\theta_{37} = 0.3$, $\mu \equiv \mu_0$, $\sigma^2 \equiv \sigma_0^2$; Third stage mixed over $\mu \sim N(\mu_0, (\sigma_0/10)^2)$, $\sigma \sim N(\sigma_0, (\sigma_0/10)^2)$.

Note that $\theta_{21} = 1 - \theta_{22}$ and $\theta_{23} = 1 - \theta_{24}$. Unconditionally, the four sets have the probabilities.

$$\begin{aligned} \Pr[X_2 \le q_1] &= \theta_{11}\theta_{21} \\ \Pr[q_1 < X_2 \le \mu] &= \theta_{11}\theta_{22} \\ \Pr[\mu < X_2 \le q_3] &= \theta_{12}\theta_{23} \\ \Pr[q_3 < X_2] &= \theta_{12}\theta_{24}. \end{aligned}$$

Within each set, we again use the shape of the original normal density so, for example, if $\mu < a < b \le q_3$ and $Y \sim N(\mu, \sigma^2)$,

$$\begin{aligned} &\Pr[a < X_2 \le b] \\ &= \Pr[a < X_2 \le b \mid \mu < X_2 \le q_3] \Pr[\mu < X_2 \le q_3] \\ &= \Pr[a < Y \le b \mid \mu < Y \le q_3] \Pr[\mu < X_2 \le q_3] \\ &= \frac{\Pr[a < Y \le b]}{\Pr[\mu < Y \le q_3]} \Pr[\mu < X_2 \le q_3] \\ &= \Pr[a < Y \le b] \frac{\theta_{12}\theta_{23}}{0.25}. \end{aligned}$$

In Figure 1(c) the density has $\theta_{11} = 0.45$, $\theta_{21} = 0.4$, $\theta_{23} = 0.6$. In general, the density of the stage 2 distribution is

$$\begin{aligned} &f(x_2 \mid \mu, \sigma^2, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_2-\mu)^2/2\sigma^2} 2^2 \left[\theta_{11}\theta_{21} I_{(-\infty, q_1]}(x_2) \right. \\ &\quad + \theta_{11}\theta_{22} I_{(q_1, \mu]}(x_2) + \theta_{12}\theta_{23} I_{(\mu, q_3]}(x_2) \\ &\quad \left. + \theta_{12}\theta_{24} I_{(q_3, \infty)}(x_2)\right]. \end{aligned} \tag{2}$$

Subsequent stages follow a similar pattern with stage 3 breaking the support of the $N(\mu, \sigma^2)$ distribution into eight sets based on the octiles so that each set has probability $1/8 = 1/2^3$ under the original parametric distribution. One introduces parameters $\theta_{31}, \ldots, \theta_{38}$ for the conditional probabilities of these sets given the stage 2 sets. Each parameter whose second subscript is even equals one minus the previous parameter, for example $\theta_{32} = 1 - \theta_{31}$ and $\theta_{38} = 1 - \theta_{37}$. Figure 1(d) illustrates stage 3 for $\theta_{31} = 0.3$, $\theta_{33} = 0.3$, $\theta_{35} = 0.6$, $\theta_{37} = 0.3$, and the previous θ_{js}'s. One continues these stages to a level J with 2^J sets that each have probability $1/2^J$ under the original parametric distribution and whose conditional probabilities given the stage $J-1$ sets are the parameters $\theta_{J1}, \ldots, \theta_{J,2^J}$ in which $\theta_{J,2k-1} = 1 - \theta_{J,2k}$, $k = 1, \ldots, 2^{J-1}$.

At the final stage J, the density at a point x_J depends on the string of sets from the various stages that contain x_J. For example, with $J = 3$, if x_3 is between the fifth and sixth octals, that is, if $\mu + 0.3186\sigma < x_3 \le \mu + 0.6745\sigma$, then x_3 is also in the sets $(\mu, \mu + 0.6745\sigma)$ and (μ, ∞). The corresponding θ parameters are $\theta_{36}, \theta_{23}, \theta_{12}$. Let $\Theta(x_J)$ be the collection of θ_{js}'s corresponding to the sets containing x_J. There are J such parameters. Define a step function similar to those in (1) and (2) that serves as a weighting factor

$$r(x_J \mid \mu, \sigma^2) = 2^J \prod_{\theta_{js} \in \Theta(x_J)} \theta_{js}.$$

For our x_3 between the fifth and sixth octals, $r(x_3) = 2^3\theta_{36}\theta_{23}\theta_{12}$. The density at stage J is just the product of the

weighting factor and the original parametric density, that is,

$$f(x_J|\mu,\sigma^2,\theta_{js}, j=1,\ldots,J, s=1,\ldots,2^j) = \psi(x_J|\mu,\sigma)r(x_J|\mu,\sigma),$$

where ψ denotes the normal density

$$\psi(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}.$$

Note that the θ_{js}'s appear only in the weight function $r(\cdot)$. Obviously, $\psi(\cdot)$ can be replaced by the density for any other parametric family, but corresponding changes in $r(\cdot)$ must be made. This is a highly flexible model because the θ_{js} parameters are numerous. There are $2^{J+1}-2$ of them, of which half are free parameters, so for $J=6$ there are 63 and for $J=8$ there are 255 free parameters. One often determines J as a function of the number of independent sampling units, say, n with corresponding $J \doteq \log_2(n)$.

To perform a Bayesian analysis with these sampling distributions, we need a joint prior distribution on the θ_{js} parameters. The θ_{js}'s are easily interpretable, so meaningful prior information may exist on many of them. An extreme case is choosing $\theta_{11}=\theta_{12}=.5$ with prior probability 1. This gives prior probability one to the median being μ for any J. Nonetheless, there are far too many parameters to choose a distribution that reflects meaningful prior information on all of the θ_{js}'s, so reference priors are also incorporated. Typically, meaningful prior information would be restricted to parameters from the first few stages j.

The parameters are all probabilities, so it is convenient to use beta priors. When the second subscript is odd, we assume for $j=1,\ldots,J, k=1,\ldots,2^{j-1}$ that

$$\theta_{j,2k-1} \sim \text{Beta}(\alpha_{j,2k-1},\alpha_{j,2k})$$

with $\theta_{j,2k}=1-\theta_{j,2k-1}$. In alternative notation, the consecutive pairs have Dirichlet distributions

$$(\theta_{j,2k-1},\theta_{j,2k}) \sim \text{Dirch}(\alpha_{j,2k-1},\alpha_{j,2k}).$$

We assume that all such pairs are independent in the prior.

For any parameters on which meaningful prior information is available, the α's are chosen to reflect that information. However, in terms of defining a workable prior for all of the θ_{js} parameters, we have not accomplished a great deal. We have reduced the problem of choosing a joint prior on $2^{J+1}-2$ parameters to choosing $2^{J+1}-2$ hyperparameters, the α_{js}'s. To make this more manageable, for parameters without meaningful prior information, and this may include all of them, we typically assume that $\alpha_{js}=c\rho(j)$ for some constant c and nondecreasing function $\rho(\cdot)$. For fixed $\rho(\cdot)$ the hyperparameter c indicates strength of prior beliefs in the original parametric family. Often we take $\rho(j)=j^2$.

One advantage of the $\alpha_{js}=c\rho(j)$ priors is that on average they give the same probabilities as the original parametric distributions. Thus, if $Y\sim N(\mu,\sigma^2)$, X_J has the stage J distribution, and A is any set,

$$\text{E}[\Pr(X_J\in A)] = \Pr(Y\in A),$$

where the expectation is over the prior on the θ_{js} parameters. To see this, first observe that by our construction, if one fixed $\theta_{js}=0.5$ for all j and s, then clearly $X_J\sim N(\mu,\sigma^2)$ and, in particular, the weight function is $r(x_J|\mu,\sigma^2)\equiv 1$. With these reference priors, for any x_J the θ_{js} parameters in $r(x_J)$ are independent with mean 0.5, so $\text{E}[r(x_J|\mu,\sigma^2)]=1$ and the average density is the normal density. Thus, these reference priors are particularly appropriate if we believe the data may come from the original parametric family but want to allow for other possibilities.

Correspondingly, prior or posterior distributions that focus high probability on regions around $\theta_{js}=0.5$ for all js will behave very much like normal distributions. This occurs whenever c is large in $\alpha_{js}=c\rho(j)$. With $\rho(j)$ increasing, large values of j imply that high prior probability is being placed on θ_{js} near 0.5. This is the property mentioned in the introduction that allows large numbers of parameters relative to the number of independent observations.

On the other hand, when c is small, the distribution is more "nonparametric." Let A_J be a set in the Jth level partition. When c is small, an observation in A_J has a large effect on all the posterior beta distributions of θ_{js}'s associated with A_J, thus causing high probability for A_J in the posterior distribution. Since A_J is a set in the finest partition considered, this causes jagged, approximating discrete, behavior in the posterior.

The Jth stage generalized sampling distribution, say G, depends on the θ_{js}'s. G, together with the prior on the θ_{js}'s, defines a finite Polya tree, which we write

$$G\sim PT_J(c,\rho,N(\mu,\sigma^2)).$$

A prior on (μ,σ) implies that the median μ, the quartiles, octiles, and so on, are random. This has the effect of smoothing out the abrupt jumps at these points that are noticeable in Figure 1. Figure 1(d) contains both a realization of a third stage Polya tree that is conditional on $\mu=\mu_0$, $\sigma^2=\sigma_0^2$, and the θ_{js}'s as well as a realization of a mixture of a third stage Polya tree that is integrated over μ and σ^2 but conditional on the θ_{js}'s. Specifically, $\mu\sim N(\mu_0,(\sigma/10)^2)$ and $\sigma\sim N(\sigma_0,(\sigma_0/10)^2)$.

The random sampling density obtained by generating a set of $\{\theta_{js}\}$ according to their prior, but averaged over a prior on (μ,σ) is termed a mixture of Polya trees, often written

$$G\sim\int PT_J(c,\rho,N(\mu,\sigma^2))p(\text{d}\mu,\text{d}\sigma^2).$$

Hanson (2006) showed that for typical priors the random MPT density is smooth. Polya trees and other versions of mixtures of Polya trees do not necessarily have this property; see Barron et al. (1999), Paddock (1999), and Berger and Guglielmi (2001).

2.1 Posterior Calculations

A critical advantage of using nonparametric Bayesian methods compared to parametric analyses is the ability to incorporate more uncertainty about the sampling distribution. However, this flexibility increases the computational complexity of the analysis. Much of the development of nonparametric Bayesian

models has been a direct result of advances in simulation-based computational methods, particularly MCMC methods; see, for example, Dey et al. (1998) and the references therein. The introduction of MCMC methods in the area began with the work of Escobar (1994) for Dirichlet process mixtures. In this section we discuss some computational aspects of posterior sampling. The model is given by

$$X_1, \ldots, X_n | G \overset{\text{iid}}{\sim} G,$$

$$G|c, \mu, \sigma \sim PT_J(c, \rho, N(\mu, \sigma^2)),$$

and

$$(\mu, \sigma^2) \sim p(\mu)p(\sigma^2).$$

Here c is fixed but it too can be treated as random.

Let $\mathbf{X} = (X_1, \ldots, X_n)'$. To explore the posterior distribution $p(G|\mathbf{X})$, a Gibbs sampling approach based on sampling from the appropriate full conditional distributions can be used. For illustration, let us consider $J = 2$, where the distribution G is fully characterized by μ, σ, θ_{11}, θ_{21}, and θ_{23}. Let $q_1 = q_1(\mu, \sigma^2)$, $q_2 = q_2(\mu, \sigma^2)$, and $q_3 = q_3(\mu, \sigma^2)$ be the quartiles of the $N(\mu, \sigma^2)$ distribution and let $n_{11} = \sum_{i=1}^n I_{(-\infty, q_2]}(X_i)$, $n_{12} = \sum_{i=1}^n I_{(q_2, \infty)}(X_i)$, $n_{21} = \sum_{i=1}^n I_{(-\infty, q_1]}(X_i)$, $n_{22} = \sum_{i=1}^n I_{(q_1, q_2]}(X_i)$, $n_{23} = \sum_{i=1}^n I_{(q_2, q_3]}(X_i)$, and $n_{24} = \sum_{i=1}^n I_{(q_3, \infty)}(X_i)$. A Markov chain for Gibbs sampling is described by the full conditionals

$$\begin{aligned} &p(\theta_{11}|\theta_{21}, \theta_{23}, \mu, \sigma^2, \mathbf{X}) \\ &\propto \left\{ \prod_{i=1}^n \theta_{11} I_{(-\infty, q_2]}(X_i) + \theta_{12} I_{(q_2, \infty)}(X_i) \right\} \theta_{11}^{c\rho(1)-1} \theta_{12}^{c\rho(1)-1} \\ &= \theta_{11}^{n_{11}+c\rho(1)-1} (1-\theta_{11})^{n_{12}+c\rho(1)-1}, \end{aligned} \tag{3}$$

$$\begin{aligned} &p(\theta_{21}|\theta_{11}, \theta_{23}, \mu, \sigma^2, \mathbf{X}) \\ &\propto \left\{ \prod_{i=1}^n \theta_{21} I_{(-\infty, q_1]}(X_i) + \theta_{22} I_{(q_1, q_2)}(X_i) \right\} \theta_{21}^{c\rho(2)-1} \theta_{22}^{c\rho(2)-1} \\ &= \theta_{21}^{n_{21}+c\rho(2)-1} (1-\theta_{21})^{n_{22}+c\rho(2)-1}, \end{aligned} \tag{4}$$

$$\begin{aligned} &p(\theta_{23}|\theta_{11}, \theta_{21}, \mu, \sigma^2, \mathbf{X}) \\ &\propto \left\{ \prod_{i=1}^n \theta_{23} I_{(q_2, q_3]}(X_i) + \theta_{24} I_{(q_3, \infty)}(X_i) \right\} \theta_{23}^{c\rho(2)-1} \theta_{24}^{c\rho(2)-1} \\ &= \theta_{23}^{n_{23}+c\rho(2)-1} (1-\theta_{23})^{n_{24}+c\rho(2)-1}, \end{aligned} \tag{5}$$

$$\begin{aligned} &p(\mu|\theta_{11}, \theta_{21}, \theta_{23}, \sigma^2, \mathbf{X}) \\ &\propto \left\{ \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(X_i-\mu)^2/2\sigma^2\} \right\} p(\mu) \\ &\quad \times \theta_{11}^{n_{11}+c\rho(1)-1} (1-\theta_{11})^{n_{12}+c\rho(1)-1} \\ &\quad \times \theta_{21}^{n_{21}+c\rho(2)-1} (1-\theta_{21})^{n_{22}+c\rho(2)-1} \\ &\quad \times \theta_{23}^{n_{23}+c\rho(2)-1} (1-\theta_{23})^{n_{24}+c\rho(2)-1}, \end{aligned} \tag{6}$$

and

$$\begin{aligned} &p(\sigma^2|\theta_{11}, \theta_{21}, \theta_{23}, \mu, \mathbf{X}) \\ &\propto \left\{ \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(X_i-\mu)^2/2\sigma^2\} \right\} p(\sigma^2) \\ &\quad \times \theta_{11}^{n_{11}+c\rho(1)-1} (1-\theta_{11})^{n_{12}+c\rho(1)-1} \\ &\quad \times \theta_{21}^{n_{21}+c\rho(2)-1} (1-\theta_{21})^{n_{22}+c\rho(2)-1} \\ &\quad \times \theta_{23}^{n_{23}+c\rho(2)-1} (1-\theta_{23})^{n_{24}+c\rho(2)-1}. \end{aligned} \tag{7}$$

Sampling from the full conditionals (3), (4), and (5) is straightforward. They are beta distributions with parameters $(n_{11} + c\rho(1), n_{12} + c\rho(1))$, $(n_{21} + c\rho(2), n_{22} + c\rho(2))$, and $(n_{23} + c\rho(2), n_{24} + c\rho(2))$, respectively.

Because the n_{js}'s depend on μ and σ^2, sampling from (6) and (7) is accomplished using a method that accepts or rejects candidate values μ^* and σ^{2*}. For Metropolis–Hastings useful candidate-generating distributions are $\mu^* \sim N(\mu, t_\mu \sigma^2/n)$ and $\sigma^{2*} \sim \text{LN}(\log(\sigma^2), t_\sigma)$, where $\text{LN}(a, s)$ refers to a log-normal distribution with location a and scale s, and t_μ and t_σ are appropriate positive parameters tuned to obtain good MCMC mixing. Alternatively, Jara et al. (2007) explored adaptive Metropolis algorithms that entirely automate the processes.

3. TWO EXAMPLES

The development in Section 2 provides a general class of distributions that can be used to model a variety of problems. As just illustrated, MPTs can be used for nonparametric analysis in one sample problems. MPTs can also be used to generalize the error distribution in linear models. In linear mixed models they can be used as an alternative to the standard assumption that random effects are normally distributed. In generalized linear mixed models, random effects are typically assumed to be normal, and our two data examples both involve generalizations of that assumption. In one case, the generalization is probably not needed, but causes little harm. In the other case, the data indicate that the random effects are not normal.

We now consider the Paraguayan and toenail examples in detail. All MPT computations were performed in R using the `DPpackage`. This package performs many routine analyses involving common Bayesian nonparametric priors, including generalized linear mixed models, Rasch IRT-type models, survival models, and others. General descriptions of the `DPpackage` syntax and design philosophy are in Jara (2007a,b). Our MPTs were all fitted with $\rho(j) = j^2$. For comparison we also fitted Bayesian normal theory models to the data. These used the same priors on the normal parameters as the MPTs.

3.1 Monkey Hunting

Our first example addresses an obvious question that arises when introducing many additional parameters into a parametric model, "What happens when all the extra machinery is not needed?" A useful aspect of Polya trees is that the centering family of distributions is included as a special case when $\theta_{js} = 0.5$ for all j and s. What is surprising is that the MPT

Table 1. Monkey hunting data from McMillan (2001). $Y_{i\bullet} = \sum_{j=1}^{N_i} Y_{ij}$ and $M_{i\bullet} = \sum_{j=1}^{N_i} M_{ij}$. The hunter's age in years is denoted a_i.

i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$
1	67	0	3	2	66	0	89	3	63	29	106	4	60	2	4
5	61	0	28	6	59	2	73	7	58	3	7	8	57	0	13
9	56	0	4	10	56	3	104	11	55	27	126	12	54	0	63
13	51	7	88	14	50	0	7	15	48	3	3	16	49	0	56
17	47	6	70	18	42	1	18	19	39	0	4	20	40	7	83
21	40	4	15	22	39	1	19	23	37	2	29	24	35	2	48
25	35	0	35	26	33	0	10	27	33	19	75	28	32	9	63
29	32	0	16	30	31	0	13	31	30	0	20	32	30	2	26
33	28	0	4	34	27	0	13	35	25	0	10	36	22	0	16
37	22	0	33	38	21	0	7	39	20	0	33	40	18	0	8
41	17	0	3	42	17	0	13	43	17	0	3	44	56	0	62
45	62	1	4	46	59	1	4	47	20	0	11				

often makes point predictions about as well as the parametric model, even when the data truly arise from the parametric model. However, the variability added from the unneeded $\{\theta_{js}\}$'s can inflate credible intervals for interesting parameters.

The Ache tribe of Paraguay are part-time hunter-gatherers who have been in contact with "modern civilization" only since the mid-1970s. McMillan (2001) spent a year living with the Ache, collecting data on many aspects of their life, including hunting. Part of Ache life is spent away from the village on extended forest treks, during which the Ache eat only food that they gather or hunt. This notably includes armadillos, large grubs, and the subject of our analysis, capuchin monkeys.

The monkeys have to be shot out of trees and are hard to kill unless there is a group of people coordinating activities. Often hunters split into two groups, one chasing a troop of monkeys towards the other group who shoot at them with bows and arrows. The monkeys are scared into flight by rattling vines and yelling; hunters try to keep monkeys moving so they can see them to shoot. As one might expect, monkey hunting is dangerous because arrows fired straight up fall back out of the trees.

A man's hunting ability changes over time. An interesting question is: after adjusting for individual heterogeneity, to what extent does the hunter's age affect his ability to hunt monkeys? And similarly, after adjusting for age, how heterogeneous is hunting ability?

McMillan (2001) collected data on the number of monkeys killed by $n = 47$ adult males from an Ache tribe over many forest treks of various lengths; these data are summarized in Table 1. Hunting success contributes to an Ache male's status within the group. It is of interest to quantify the association between a man's age and his monkey kill rate. One might wonder at what age monkey hunting prowess is greatest. We modeled the number of monkeys killed by an individual as Poisson. The log kill-rate is assumed to be a quadratic function of a man's age in years a_i plus a subject-specific random effect γ_i. The random effects account for the correlation of an individual's kills over multiple hunting treks. Several individuals embark on a trek together, for simplicity we ignore possible trek effects.

Let Y_{ij} be the number of monkeys killed by hunter i on his jth trek, $i = 1, \ldots, 47$, $j = 1, \ldots, N_i$. M_{ij} is the length in days of trek ij. Our basic stochastic model is

$$Y_{ij}|\lambda_i \sim \text{Pois}(\lambda_i M_{ij}),$$

so λ_i is the effect for hunter i. Our model for the monkey kill rate λ_i combines fixed effects capturing an overall population trend in hunting ability due to age and random effects that serve as a convenient surrogate for "innate" hunting ability:

$$\log \lambda_i = \beta_0 + \beta_1(a_i - 45) + \beta_2(a_i - 45)^2 + \gamma_i.$$

Commonly, the random effects would be assumed normal

$$\gamma_1, \ldots, \gamma_{47}|\sigma \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

We have incorporated an intercept into the model, so the random effects are centered at 0. Our experience is that this improves the MCMC mixing. The Bayesian analysis uses priors of $\sigma^{-2} \sim \Gamma(5/2, 3)$, that is, $E(\sigma^{-2}) = 5/6$ and $E(\sigma^2) = 2$, and independent $N(0, 100)$ distributions for the regression parameters. It gives posterior means, medians, standard deviations, and 95% highest posterior density (HPD) probability intervals (see Table 2).

For model comparisons we use two measures. DIC is the deviance information criterion of Spiegelhalter et al. (2002). LPML is the log pseudo marginal likelihood of Geisser and Eddy (1979) which is a leave-one-out cross-validatory measure based on predictive densities. For the parametric normal model, DIC $= 123.16$ and LPML $= -68.99$. The maximum of the quadratic kill-rate function is at $(\beta_2 90 - \beta_1)/2\beta_2$, the median of which is 48.3 years with a 95% credible interval of $(44.1, 60.4)$. The interval estimate is asymmetric which results from dividing by β_2, a number close to zero.

In examining hunting heterogeneity, note that there is no reason to believe, a priori, that hunting ability will follow a bell-shaped curve. There could be two well-separated groups of "good" and "bad" hunters. To allow leeway toward multimodal and skewed hunting ability distributions, consider the MPT generalization

$$\gamma_1, \ldots, \gamma_{47}|G \overset{\text{iid}}{\sim} G, \qquad G|c, \sigma \sim PT_5(c, \rho, N(0, \sigma^2)).$$

Given the mixing parameter σ, the random G is centered at a $N(0, \sigma^2)$ distribution. Further, we set the median of G to be zero

Table 2. Monkey hunting: Posterior summaries for the normal model.

				95% HPD	
Parameter	Mean	Median	Std. dev.	Lower	Upper
β_0	−2.549	−2.544	0.4756	−3.526	−1.671
β_1	0.04115	0.04020	0.02779	−0.01131	0.09762
β_2	−0.006083	−0.005877	0.002621	−0.01114	−0.001076

by fixing $\theta_{11} = \theta_{12} = 0.5$ as in Walker and Mallick (1999). We assume the hyperprior $c \sim \Gamma(1, 1)$ independent of the prior distributions already given for σ^{-2} and the regression parameters. The posterior gives the results shown in Table 3.

For the MPT model, DIC = 123.93 and LPML = −70.01. Both the DIC and LPML indicate slightly worse prediction from the Polya tree model. Figure 2 displaying the predictive distributions of a new random effect makes the normal assumption look reasonable. Thus, after adjusting for the hunter's age, innate monkey hunting ability is distributed roughly according to a normal distribution. The data do not indicate that well-separated groups of "good" and "bad" hunters exist. Generally, the means and medians are comparable to the parametric normal results but the standard deviations and intervals are larger or about the same. The MPT median apex of the kill-rate function is 48.6 years with a 95% interval of (43.7, 68.1), appreciably wider than from the parametric analysis.

3.2 Toenails

We consider data from a clinical trial in dermatology (De Backer et al. 1996). The data were obtained from a randomized, double-blind, parallel group, multicenter study for the comparison of the efficacy of two oral treatments for toenail dermatophyte onychomycosis infection: terbinafine and intraconazole. These are well-known, commercially available treatments. Onychomycosis, known popularly as toenail fungus, is a fairly common condition that not only can disfigure and sometimes destroy the nail but that also can lead to social and self-image issues for sufferers. It has been estimated that between 2% and 18% of people worldwide are afflicted by some form of the disease. Onychomycosis can be caused by several types of fungi known as dermatophytes, as well as by nondermatophytic yeasts or molds. Dermatophyte onychomycosis corresponds to the type caused by dermatophytes.

We consider data on a nasty side-effect of toenail fungus: the degree of separation of the nail plate from the nail bed, scored in four categories (0, absent; 1, mild; 2, moderate; 3, severe). For the $n = 294$ patients, the response was evaluated at seven visits (approximately on weeks 0, 4, 8, 12, 24, 36, and 48). A total of 937 measurements were made on the 146 intraconazole, $\text{Trt}_i = 0$, patients and 971 measurements were made on 148 terbinafine, $\text{Trt}_i = 1$, patients. The data are available at *http://www.blackwellpublishing.com/rss/Volumes/Cv50p3.htm*.

A typical approach to analyzing ordinal data is to fit a series of logistic models such as continuation ratios or cumulative logits; see Christensen (1997, sect. 4.6). Following Lesaffre and Spiessens (2001), we examine a logit mixed effects model. Specifically, let $Y_{ij} = 1$ if individual i has moderate or severe toenail separation at time j with $Y_{ij} = 0$ if toenail separation is absent or mild and consider the model,

$$\text{logit}\left\{\Pr\left(Y_{ij} = 1 \mid \boldsymbol{\beta}, \gamma_i\right)\right\} = \gamma_i + \beta_1 \text{Trt}_i + \beta_2 \text{Time}_{ij} + \beta_3 \text{Trt}_i \times \text{Time}_{ij},$$

$i = 1, \ldots, 294$, $j = 1, \ldots, N_i$. Here γ_i is a random effect for each subject, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ are regression parameters corresponding to Trt, the binary treatment indicator, Time, the visit time in four week periods, and the interaction Trt × Time. Typically, the γ_i's would be assumed independent $N(\mu, \sigma^2)$. We replace the normality assumption with a MPT prior,

$$\gamma_1, \ldots, \gamma_{294} | G \overset{\text{iid}}{\sim} G,$$

$$G|\mu, \sigma^2 \sim PT_8(c, j^2, N(\mu, \sigma^2)).$$

The methodology of our article allows the evaluation of the normality assumption of the random effects and the implications of a potential misspecification of the normal model. Different reference priors for the θ_{js}'s were considered using three values of c, $c = 0.1, 1, 10$, to reflect increasing degrees of belief in normality for the random effects. We also fitted the parametric normal model ($c \to \infty$). Earlier analyses based on normal assumptions displayed an inconsistency between marginal and

Table 3. Monkey hunting: Posterior summaries for the MPT model.

				95% HPD	
Parameter	Mean	Median	Std. dev.	Lower	Upper
β_0	−2.625	−2.583	0.5879	−3.934	−1.537
β_1	0.04245	0.04168	0.02738	−0.00769	0.1003
β_2	−0.005708	−0.005536	0.002711	−0.01157	−0.000981

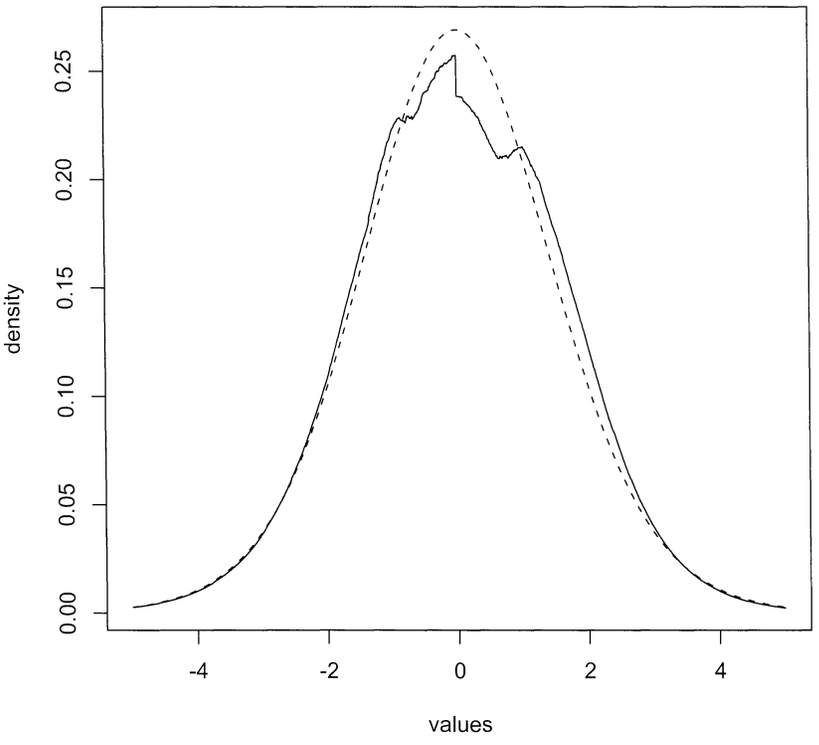

Figure 2. Monkey hunting: Predictive densities from the normal (dashed) and Polya tree (solid) models.

subject-specific effects that made us suspect the validity of the analysis under the normality assumption.

A Bayesian analysis requires priors for both the regression parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, and the parameters μ and σ^2 of the original normal family generalized by the MPT. We used independent conjugate normal—inverse gamma priors that display weak prior information. Specifically, $N_3(\mathbf{0}, 100 \times \boldsymbol{I}_3)$, $N(0, 100)$, and $\Gamma(2, 0.5)$ priors for $\boldsymbol{\beta}$, μ, and σ^{-2}, respectively. An alternative choice of $\sigma^{-2} \sim \Gamma(1.025, 0.01)$ gave similar results.

Figure 3 compares the predictive distributions of a new random effect, say γ_{295}, not in the observed dataset using the two best MPT models and the normal theory model. The plot clearly shows deviation from normality in such a way that the patients could be divided into two or three groups according to their resistance against infection and accompanying toenail separation. Note also the smooth appearance of the MPT predictive distributions. The fact that the density of the generalized distributions contain jumps typically washes out in the posterior analysis of a MPT.

Table 4 presents results from fitting the normal model and the three MPT models. The Polya trees outperform the normal model using either the LPML or DIC statistic, suggesting that the MPT model is better both for explaining the observed data and from a predictive viewpoint.

Table 4 also shows the effects of incorrectly assuming a normal distribution for the random intercept. Comparing the interval estimates to 0, none of the models shows a significant baseline effect (β_1) for treatments, as should be the case for a randomized experiment. All models show time effects (β_2) and that the treatment coded 1 works better over time (β_3). The posterior probability of $\Pr(\beta_3 > 0)$ was 2.03%, 2.18%, 3.56%, and 3.24% for $c = \infty$, $c = 10$, $c = 1$, and $c = 0.1$, respectively. Although, all models show evidence of β_3 effects, the two MPT models with weak normality assumptions show a reduction in the posterior evidence of (β_3) treatment effect. Note that $c = 1$ is the best fitting model. It fits better than $c = 0.1$ which is "more nonparametric."

4. APPLICATIONS AND FUTURE DIRECTIONS

Only recently have Polya trees and mixtures of Polya trees been used to seriously analyze data. Lavine (1994) used finite Polya trees to model errors in regression settings and performed the analysis of the hierarchical binomial model discussed by Berry and Christensen (1979). Finite Polya trees have been successfully applied in frailty models and GLMMs (Walker and Mallick 1997) and to model the error distribution in accelerated failure time models (Walker and Mallick 1999; Mallick and Walker 2003). Hanson (2006) provided computational strategies for obtaining inferences for mixtures of finite Polya tree models, describing algorithms for fitting binomial regression models with nonparametric link functions, random intercept models and a variety of survival models. These approaches have

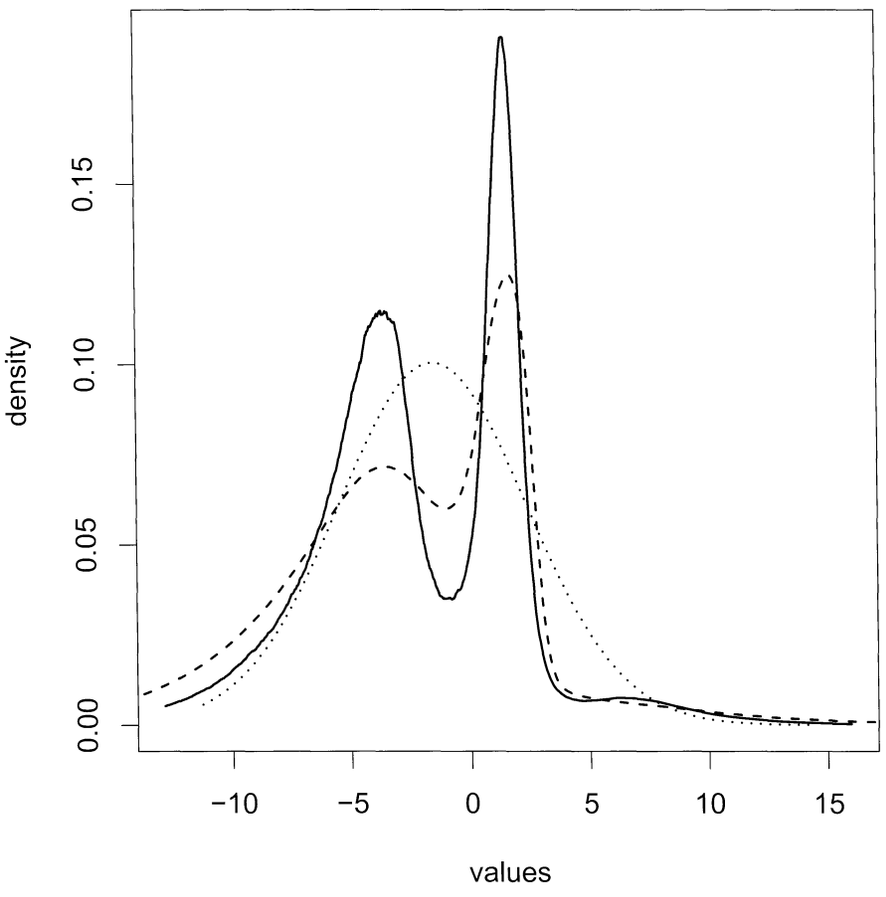

Figure 3. Toenail data estimated random effects distribution under the MPT ($c = 0.1$, solid line), MPT ($c = 1$, dashed line), and normal ($c \to \infty$, dotted line) models.

Table 4. Posterior means, model comparison criteria, and 95% HPD intervals for parameters of the toenail GLMM.

		MPT		
Parameter	Normal	$c = 10$	$c = 1$	$c = 0.1$
β_1(Trt)	−0.159	−0.051	0.292	0.364
β_2(Time)	−0.393	−0.390	−0.393	−0.376
β_3(Trt × Time)	−0.138	−0.136	−0.130	−0.129
μ	−1.604	−1.999	1.510	−0.181
σ^2	16.025	16.728	52.457	23.631
DIC	964.2	954.5	906.3	909.9
LPML	−484.0	−482.5	−465.3	−470.5
		Posterior Intervals		
β_1(Trt)	(−1.290,0.966)	(−1.145,1.090)	(−0.681,1.186)	(−0.486,1.198)
β_2(Time)	(−0.478,−0.305)	(−0.478,−0.304)	(−0.488,−0.307)	(−0.472,−0.289)
β_3(Trt × Time)	(−0.271,−0.005)	(−0.270,−0.003)	(−0.274,0.010)	(−0.266,0.007)
μ	(−2.473,−0.790)	(−3.544,−0.479)	(−3.402,3.884)	(−6.869,5.664)
σ^2	(10.445,22.053)	(9.562,24.867)	(8.541,116.273)	(1.436,42.767)

been successfully applied to density estimation with censored data (Yang, Hanson, and Christensen 2008), ROC curve estimation (Branscum et al. 2008; Hanson et al. 2008), meta-analysis (Branscum and Hanson 2008), and proportional odds models (Hanson and Yang 2007).

Interesting areas for further research are the extension of Polya tree constructions to multidimensional spaces, and extending Polya trees to define probability models over related distributions, that is, dependent Polya trees in time, space, or across covariate values. Paddock et al. (2003) and Hanson (2006) devised ways to generalize the univariate PT construction to higher dimensions. Paddock et al. (2003) provided a way to extend the Polya tree scheme to multi-dimensional simplexes by considering perpendicular splits of the axis. They also suggested the use of suitable univariate parametric CDFs to map the multivariate PT into another space. In a density estimation problem, Paddock et al. used the inverse of the uniform CDF on the range of the data. Defining PT priors on $\mathbb{R}^d$, Hanson (2006) used affine transformations of perpendicular splits of a grid. With this aim, Hanson (2006) considered a square symmetric decomposition of a given covariance matrix. Jara, Hanson, and Lesaffre (2007) discussed different, including random, decompositions of the covariance matrix and provide exact formulas for sampling functionals. The study of optimal partition schemes and the optimal number of levels in multidimensional spaces is a topic of interest. Locally adaptive partitioning schemes varying the number of levels throughout the regions of the space can be pursued in order to avoid the curse of dimensionality associated with high-dimensional problems.

Recently, applications of dependent distributions $G_{\mathbf{x}}$, where $\mathbf{x}$ is a predictor, time, space, or combinations of these three have been explored, most notably with stick-breaking priors (Griffin and Steel 2006; Dunson, Pillai, and Park 2007; Reich and Fuentes 2007; etc.) The Polya tree can similarly be adapted to produce dependent realizations suitable for modeling growth curve data, partially exchangeable random effects, or other applications requiring the complex evolution of a density in time, space, or with covariates.

[Received October 2007. Revised June 2008.]

REFERENCES

Barron, A., Schervish, M.J., and Wasserman, L. (1999), "Posterior Distributions in Nonparametric Problems," *The Annals of Statistics*, 27, 536–561.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996), "A New Perspective on Priors for Generalized Linear Models," *Journal of the American Statistical Association*, 91, 1450–1460.

Berger, J.O., and Guglielmi, A. (2001), "Bayesian and Conditional Frequentist Testing of a Parametric Model versus Nonparametric Alternatives," *Journal of the American Statistical Association*, 96, 174–184.

Berry, D., and Christensen, R. (1979), "Empirical Bayes Estimation of a Binomial Parameter via Mixtures of Dirichlet Processes," *The Annals of Statistics*, 7, 558–568.

Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions Via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Branscum, A., Johnson, W., Hanson, T., and Gardner, I. (2008), "Bayesian Semiparametric ROC Curve Estimation and Disease Risk Assessment," *Statistics in Medicine*, 27, 2474–2496.

Branscum, A., and Hanson, T. (2008), "Bayesian Nonparametric Meta-analysis Using Polya Tree Mixture Models," *Biometrics*, 64, 825–833.

Breslow, N.E., and Clayton, D. (1993), "Approximate Inference in Generalized Linear Models," *Journal of the American Statistical Association*, 88, 9–25.

Casella, G., and George, E.I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Christensen, R. (1997), *Log-Linear Models and Logistic Regression* (2nd ed.), New York: Springer-Verlag.

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126.

De Backer, M., De Keyser, P., De Vroey, C., and Lesaffre, E. (1996), "A 12-week Treatment for Dermatophyte Toe Onychomycosis: Terbinafine 250mg/day vs. Itraconazole 200mg/day—A Double-Blind Comparative Trial," *British Journal of Dermatology*, 134, 16–17.

Dey, D., Muller, P., and Sinha, D. (1998), *Practical Nonparametric and Semiparametric Bayesian Statistics*, New York: Springer.

Dubins, L.E., and Freedman, D.A. (1967), "Random Distribution Functions," *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, 2, 183–214.

Dunson, D.B., Pillai, N., and Park, J.-H. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society*, Series B, 69, 163–183.

Escobar, M. D. (1994), "Estimating Normal Means with a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277

Fabius, J. (1964), "Asymptotic Behavior of Bayes' Estimates," *The Annals of Mathematical Statistics*, 35, 846–856.

Ferguson, T.S. (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Freedman, D.A. (1963), "On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case," *The Annals of Mathematical Statistics*, 34, 1386–1403.

Geisser, S., and Eddy, W. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.

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

Hanson, T. (2006), "Inference for Mixtures of Finite Polya Tree Models," *Journal of the American Statistical Association*, 101, 1548–1565.

Hanson, T., Branscum, A., and Gardner, I. (2008), "Multivariate Mixtures of Polya Trees for Modelling ROC Data," *Statistical Modelling*, 8, 81–96.

Hanson, T., and Johnson, W.O. (2002), "Modeling Regression Error with a Mixture of Polya Trees," *Journal of the American Statistical Association*, 97, 1020–1033.

Hanson, T., and Yang, M. (2007), "Bayesian Semiparametric Proportional Odds Models," *Biometrics*, 63, 88–95.

Huzurbazar, A.V. (2005), *Flowgraph Models for Multistate Time-to-Event Data*, New York: Wiley.

Jara, A. (2007a), "DPpackage: Bayesian Nonparametric and Semiparametric Analysis," *User Manual Version 1.0-5, Biostatistical Centre, Catholic University of Leuven*. Available online at *http://cran.r-project.org/src/contrib/Descriptions/DPpackage.html*.

——— (2007b), "Applied Bayesian Non- and Semi-parametric Inference using DPpackage," *Rnews*, 7(3), 17–26.

Jara, A., Hanson T., and Lesaffre, E. (2007), "Robustifying Generalized Linear Mixed Models using Mixtures of Multivariate Polya Trees," *Technical Report, Biostatistical Centre, Catholic University of Leuven*.

Kraft, C.H. (1964), "A Class of Distribution Function Processes Which Have Derivatives," *Journal of Applied Probability*, 1, 385–388.

Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modeling," *The Annals of Statistics*, 20, 1222–1235.

——— (1994), "More Aspects of Polya Tree Distributions for Statistical Modeling," *The Annals of Statistics*, 22, 1161–1176.

Lesaffre, E., and Spiessens, B. (2001), "On the Effect of the Number of Quadrature Points in a Logistic Random-Effects Model: An Example," *Applied Statistics*, 50, 325–335.

Mallick, B.K., and Walker, S.G. (2003), "A Bayesian Semiparametric Transformation Model Incorporating Frailties," *Journal of Statistical Planning and Inference*, 112, 159–174.

Mauldin, R.D., and Williams, S.C. (1990), "Reinforced Random Walks and Random Distributions," in *Proceedings of the American Mathematical Society*, vol 110, American Mathematical Society, pp. 251–258.

Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992), "Polya Trees and Random Distributions," *The Annals of Statistics*, 20, 1203–1221.

McMillan, G.P. (2001), "Ache Residential Grouping and Social Foraging," PhD Dissertation, Dept. of Anthropology, The University of New Mexico, Albuquerque, NM.

Metivier, M. (1971), "On the Construction for Random Measures Almost Surely Absolutely Continuous with Respect to a Given Measure" (French: "Sur la construction de mesures aleatoires presque surement absolument continues par rapport a une mesure donnee"), *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 20, 332–334.

Paddock, S.M. (1999), "Randomized Polya Trees: Bayesian Nonparametrics for Multivariate Data Analysis," unpublished doctoral thesis, Institute of Statistics and Decision Sciences, Duke University.

Paddock, S.M., Ruggeri, F., Lavine M., and West, M. (2003), "Randomized Polya Tree Models for Nonparametric Bayesian Inference," *Statistica Sinica*, 13, 443–460.

Reich, B.J., and Fuentes, M. (2007), "A Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields," *Annals of Applied Statistics*, 1, 249–264.

Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society*, Series B, 64, 583–639.

Walker, S.G., and Mallick, B.K. (1997), "Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing," *Journal of the Royal Statistical Society*, Series B, 59, 845–860.

——— (1999), "Semiparametric Accelerated Life Time Model," *Biometrics*, 55, 477–483.

Yang, M., Hanson, T., and Christensen, R. (2008), "Nonparametric Bayesian Estimation of a Bivariate Density with Interval Censored Data," *Computational Statistics & Data Analysis*, 52, 5202–5214.