*The Annals of Statistics*
1994, Vol. 22, No. 3, 1161–1176

# MORE ASPECTS OF POLYA TREE DISTRIBUTIONS FOR STATISTICAL MODELLING[1]

BY MICHAEL LAVINE

*Duke University*

The definition and elementary properties of Polya tree distributions are reviewed. Two theorems are presented showing that Polya trees can be constructed to concentrate arbitrarily closely about any desired pdf, and that Polya tree priors can put positive mass in every relative entropy neighborhood of every positive density with finite entropy, thereby satisfying a consistency condition. Such theorems are false for Dirichlet processes. Models are constructed combining partially specified Polya trees with other information such as monotonicity or unimodality. It is shown how to compute bounds on posterior expectations over the class of all priors with the given specifications. A numerical example is given. A theorem of Diaconis and Freedman about Dirichlet processes is generalized to Polya trees, allowing Polya trees to be the models for errors in regression problems. Finally, empirical Bayes models using Dirichlet processes are generalized to Polya trees. An example from Berry and Christensen is reanalyzed with a Polya tree model.

**1. Introduction.** Polya trees form a class of distributions for a random probability measure $\mathcal{P}$ intermediate between Dirichlet processes [Ferguson (1973)] and tail-free processes [Freedman (1963) and Fabius (1964)]. Their advantage over Dirichlet processes is that they can be constructed to give probability 1 to the set of continuous or absolutely continuous probability measures, while their advantage over more general tail-free processes is their much greater tractability. The basic ideas of Polya trees can be found in Ferguson (1974), Mauldin, Sudderth and Williams (1992) and Lavine (1992). The rest of the introduction reviews the definitions and elementary properties of Polya trees. Section 2 contains two theorems showing the suitablility of Polya trees for statistical modelling. Section 3 combines robust and nonparametric Bayes ideas by showing the feasibility of modelling with partially specified Polya trees and incorporating other information such as shape constraints. A numerical example is given. Section 4 proposes models in which Polya trees are used to represent the errors in regression settings. The main result is a generalization of a theorem of Diaconis and Freedman (1986). Finally, Section 5 addresses the empirical Bayes problem, using Polya trees in place of Dirichlet processes. [See Antoniak (1974), Berry and Christensen (1979), Escobar (1994), Escobar and West (1990), Ferguson (1983), Kuo (1986), Lo (1984) and West (1990) for the use of Dirichlet processes.] Section 5 contains a reanalysis using Polya trees of a problem from Berry and Christensen (1979).

Received August 1991; revised November 1993.

[1] Supported in part by NSF Grant DMS-89-03842.

*AMS* 1991 *subject classifications.* Primary 62A15; secondary 62G07, 62G99.

*Key words and phrases.* Dirichlet processes, nonparametric regression, robust Bayes, tail-free processes.

Let $E = \{0,1\}$, $E^0 = \emptyset$; let $E^m$ be the $m$-fold product $E \times E \times \cdots \times E$ and $E^* = \cup_0^\infty E^m$; and let $E^N$ be the set of infinite sequences of elements of $E$. Let $\Omega$ be a separable measurable space, let $\pi_0 = \{\Omega\}$ and let $\Pi = \{\pi_m;\ m = 0, 1, \ldots\}$ be a separating binary tree of partitions of $\Omega$; that is, let $\pi_0, \pi_1, \ldots$ be a sequence of partitions such that $\cup_0^\infty \pi_m$ generates the measurable sets and such that every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\emptyset = \Omega$ and, for all $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two pieces into which $B_\varepsilon$ is split. Degenerate splits are permitted, for example, $B_\varepsilon = B_{\varepsilon 0} \cup \emptyset$. For every $\theta \in \Omega$, let $\varepsilon^m(\theta)$ be the element of $E^m$ such that $\theta \in B_{\varepsilon^m(\theta)}$. Note therefore that $\theta = \cap B_{\varepsilon^m(\theta)}$.

DEFINITION 1. A random probability measure $\mathcal{P}$ on $\Omega$ is said to have a Polya tree distribution, or a Polya tree prior, with parameter $(\Pi, \mathcal{A})$, written $\mathcal{P} \sim \mathrm{PT}(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\varepsilon\colon \varepsilon \in E^*\}$ and random variables $\mathcal{Y} = \{Y_\varepsilon\colon \varepsilon \in E^*\}$ such that the following hold:

(i) all the random variables in $\mathcal{Y}$ are independent;
(ii) for every $\varepsilon \in E^*$, $Y_\varepsilon$ has a Beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$;
(iii) for every $m = 1, 2, \ldots$ and every $\varepsilon \in E^m$,

$$\mathcal{P}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \left( \prod_{j=1;\, \varepsilon_j = 0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \right) \left( \prod_{j=1;\, \varepsilon_j = 1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}) \right),$$

where the first term in the products above is interpreted as $Y_\emptyset$ or $1 - Y_\emptyset$.

Degenerate Beta distributions are permitted, for example, $\alpha_{\varepsilon 0} = 0$ making the distribution degenerate at 0. The random variables $\Theta_1, \Theta_2, \ldots$ are said to be a sample from $\mathcal{P}$ if, given $\mathcal{P}$, they are i.i.d. with distribution $\mathcal{P}$. The $Y_\varepsilon$'s have the following interpretation: $Y_\emptyset$ and $1 - Y_\emptyset$ are, respectively, the probabilities that $\Theta_i \in B_0$ and $\Theta_i \in B_1$, and $Y_\varepsilon$ and $1 - Y_\varepsilon$ are the conditional probabilities that $\Theta_i \in B_{\varepsilon 0}$ and $\Theta_i \in B_{\varepsilon 1}$ given that $\Theta_i \in B_\varepsilon$.

Three important facts about Polya trees are the following:

1. They are conjugate. If $\mathcal{P}$ has a Polya tree distribution and $\Theta \mid \mathcal{P} \sim \mathcal{P}$, then $\mathcal{P} \mid \Theta$ has a Polya tree distribution [Ferguson (1974), Mauldin, Sudderth and Williams (1992)]. Updating a Polya tree after observing $\Theta_i$ is simple; for every $\varepsilon$ such that $\Theta_i \in B_\varepsilon$, add 1 to $\alpha_\varepsilon$. We call the new parameters $\mathcal{A} \mid \Theta$. Sometimes we will not have observed $\Theta_i$ exactly but will only know that $\Theta_i$ belongs to some set. If that set happens to be $B_\delta$ for some $\delta \in E^*$, then again the updating follows the same rule. The difference is that when $\Theta_i$ is observed exactly there are infinitely many $\alpha_\varepsilon$'s to update; when we see $\Theta_i \in B_\delta$, there are only finitely many.
2. Dirichlet processes are special cases of Polya trees. A Polya tree is a Dirichlet process if, for every $\varepsilon \in E^*$, $\alpha_\varepsilon = \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}$ [Ferguson (1974)]. The parameter of the Dirichlet process is $\alpha = mG_0$, where $m = \alpha_\emptyset$ and $G_0$ is determined by $G_0(B_\varepsilon) = \mathbb{E}[\mathcal{P}(B_\varepsilon)] = \Pr[\Theta_i \in B_\varepsilon]$.

3. Some Polya trees assign probability 1 to the set of continuous distributions, for example, when $\alpha_{\varepsilon_1,\dots,\varepsilon_m} = m^2$ as in Example 2 of Section 5.

## 2. Suitability of Polya trees for statistical modelling.

One occasionally wants to construct a nonparametric prior centered at and concentrated near a given distribution. Dalal and Hall (1980) show that Dirichlet processes can be so constructed if "near" is interpreted in the sense of weak convergence. Lavine (1992) gives a theorem showing that the cdf of $\mathcal{P}$, with a Polya tree distribution, can be made uniformly close to a given cdf, with arbitrarily high probability. The next theorem says that using Polya trees the pdf of $\mathcal{P}$ can be made close to a given pdf $q$. Such a theorem is impossible for Dirichlet processes because, under a Dirichlet process, $\mathcal{P}$ does not have a pdf; whereas Kraft (1964) and Metivier (1971) show that a Polya tree can be constructed so that $\mathcal{P}$ is almost surely absolutely continuous with respect to Lebesgue measure. See also Ferguson (1974) and Lavine (1992). Note that if degenerate Beta distributions are allowed, then Theorem 1 is true trivially: take the Polya tree to be degenerate at $Q$. The point of Theorem 1 is that the construction with nondegenerate Beta distributions still has full support in the set of all probability measures. Such constructions may be useful as priors in Bayesian analyses. Let $p$ be the random density of $\mathcal{P}$. Let ess sup denote the essential supremum, the supremum except perhaps on a null set.

THEOREM 1. *For a given probability measure $Q$ with density $q$, any positive number $k$ and any $\eta \in (0,1)$, there exists a Polya tree distribution for $\mathcal{P}$ such that $\mathcal{P}$ has a density $p$ satisfying* $\Pr[\operatorname{ess\,sup}_\theta |\log(p(\theta)/q(\theta))| < k] > \eta$.

PROOF. Construct the Polya tree so that

$$\mathbb{E}\left[\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]\right] = \frac{Q\left(B_{\varepsilon^m(\theta)}\right)}{Q\left(B_{\varepsilon^{m-1}(\theta)}\right)}.$$

Then,

$$\begin{aligned}
&\Pr\left[\operatorname*{ess\,sup}_\theta \left|\log\frac{p(\theta)}{q(\theta)}\right| < k\right] \\
&\quad = \Pr\left[\operatorname*{ess\,sup}_\theta \left|\sum_{m=1}^{\infty} \log\frac{\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]}{\mathbb{E}\left[\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]\right]}\right| < k\right] \\
&\quad \geq \Pr\left[\operatorname*{ess\,sup}_\theta \sum_{m=1}^{\infty} \left|\log\frac{\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]}{\mathbb{E}\left[\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]\right]}\right| < k\right] \\
&\quad \geq \Pr\left[\sum_{m=1}^{\infty} \operatorname*{ess\,sup}_\theta \left|\log\frac{\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]}{\mathbb{E}\left[\mathcal{P}\left[B_{\varepsilon^m(\theta)} \mid B_{\varepsilon^{m-1}(\theta)}\right]\right]}\right| < k\right] \\
&\quad = \Pr\left[\sum_{m=0}^{\infty} \sup_{\varepsilon \in E^m} \left(\left|\log\frac{Y_\varepsilon}{\mathbb{E}[Y_\varepsilon]}\right| \vee \left|\log\frac{1-Y_\varepsilon}{\mathbb{E}[1-Y_\varepsilon]}\right|\right) < k\right].
\end{aligned}$$

There are only finitely many elements in $E^m$ so, by choosing the $\alpha_\varepsilon$'s sufficiently large, each summand can be made arbitrarily small with arbitrarily large probability and the probability that the sum in less than $k$ can be made greater than $\eta$. □

For distributions $F_*$ and $F$ with densities $f_*$ and $f$, let $D(F_*,F)$ denote the relative entropy or Kullback–Leibler divergence $\int \log(f_*(\theta)/f(\theta))f_*(\theta)\,d\theta$ and let $N_\delta(F_*) = \{F\colon D(F_*,F) < \delta\}$ denote a relative entropy neighborhood. Barron (1986) remarks that if a prior $\mu$ satisfies $\mu(N_\delta(F_*)) > 0$ for all $\delta > 0$, then $\mu$ satisfies a consistency condition at $F_*$. It may therefore be desirable for a prior to put positive mass in every relative entropy neighborhood of a wide class of probability measures to ensure consistency over that class. The next theorem says that Polya trees can be constructed with positive mass in every relative entropy neighborhood of every distribution with an essentially positive density. Such a theorem is impossible for Dirichlet processes because they give probability 1 to the set of discrete distributions.

THEOREM 2. *A Polya tree prior $\mu$ can be constructed so that, for any distribution $F_*$ with density $f_*$ and having finite entropy, $\mu(N_\delta(F_*)) > 0$ for all $\delta > 0$.*

PROOF. We suppose that the $\mu$ to be constructed below almost surely has a density. Let $\delta$ be given and let $\lambda$ denote Lebesgue measure:

$$
\begin{aligned}
&D(F_*,F)\\
&\quad= \int \log\left(\lim_{m\to\infty}\left(\frac{F_*\big(B_{\varepsilon^m(\theta)}\big)/\lambda\big(B_{\varepsilon^m(\theta)}\big)}{F\big(B_{\varepsilon^m(\theta)}\big)/\lambda\big(B_{\varepsilon^m(\theta)}\big)}\right)\right) f_*(\theta)\,d\theta\\
&\quad= \int \log\left(\lim_{m\to\infty}\left(\prod_{j=1}^{m}\frac{F_*\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)/\lambda\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}{F\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)/\lambda\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}\right)\right) f_*(\theta)\,d\theta \qquad (1)\\
&\quad= \int \sum_{1}^{\infty}\log\left(\frac{F_*\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}{\lambda\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}\right) f_*(\theta)\,d\theta\\
&\qquad - \int \sum_{1}^{\infty}\log\left(\frac{F\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}{\lambda\big(B_{\varepsilon^j(\theta)}\,|\,B_{\varepsilon^{j-1}(\theta)}\big)}\right) f_*(\theta)\,d\theta.
\end{aligned}
$$

The first equality above is implicit in either Theorem 35.8 of Billingsley (1986) or Equation (1) of Lavine (1992). The first integral of the RHS of (1) converges by assumption. The second integral is

$$\begin{aligned}
&\int_{B_0} \sum_{1}^{\infty} \log\left(\frac{F\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}{\lambda\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}\right) f_*(\theta)\, d\theta \\
&\quad + \int_{B_1} \sum_{1}^{\infty} \log\left(\frac{F\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}{\lambda\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}\right) f_*(\theta)\, d\theta \\
&= F_*(B_0) \log\left(\frac{F(B_0)}{\lambda(B_0)}\right) + \int_{B_0} \sum_{2}^{\infty} \log\left(\frac{F\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}{\lambda\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}\right) f_*(\theta)\, d\theta \\
&\quad + F_*(B_1) \log\left(\frac{F(B_1)}{\lambda(B_1)}\right) + \int_{B_1} \sum_{2}^{\infty} \log\left(\frac{F\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}{\lambda\left(B_{\varepsilon^j(\theta)} \mid B_{\varepsilon^{j-1}(\theta)}\right)}\right) f_*(\theta)\, d\theta.
\end{aligned}$$

Similarly, by dividing the integral over $B_0$ into the sum of two integrals over $B_{00}$ and $B_{01}$ and likewise for the integral over $B_1$, and by continuing to subdivide integrals indefinitely, the previous expression becomes

$$\begin{aligned}
&\sum_{m=0}^{\infty} \sum_{\varepsilon \in E^m} F_*(B_{\varepsilon 0}) \log\left(\frac{F\left(B_{\varepsilon 0} \mid B_\varepsilon\right)}{\lambda\left(B_{\varepsilon 0} \mid B_\varepsilon\right)}\right) + F_*(B_{\varepsilon 1}) \log\left(\frac{F\left(B_{\varepsilon 1} \mid B_\varepsilon\right)}{\lambda\left(B_{\varepsilon 1} \mid B_\varepsilon\right)}\right) \\
(2) \qquad &= \sum_{m=0}^{\infty} \sum_{\varepsilon \in E^m} F_*(B_{\varepsilon 0}) \log\left(\frac{Y_\varepsilon}{\lambda\left(B_{\varepsilon 0} \mid B_\varepsilon\right)}\right) + F_*(B_{\varepsilon 1}) \log\left(\frac{1 - Y_\varepsilon}{\lambda\left(B_{\varepsilon 1} \mid B_\varepsilon\right)}\right) \\
&\leq \sum_{m=0}^{\infty} \max_{\varepsilon \in E^m} \left(\log\left(\frac{Y_\varepsilon}{\lambda\left(B_{\varepsilon 0} \mid B_\varepsilon\right)}\right) \vee \log\left(\frac{1 - Y_\varepsilon}{\lambda\left(B_{\varepsilon 1} \mid B_\varepsilon\right)}\right)\right).
\end{aligned}$$

The Polya tree parameter $\mathcal{A}$ can be chosen so that $\alpha_{\varepsilon 0}/(\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}) \equiv \mathbb{E}[Y_\varepsilon] = \lambda(B_{\varepsilon 0} \mid B_\varepsilon)$ and with the $\alpha_\varepsilon$'s increasing sufficiently rapidly with $m$ so that the sum in (2) converges with positive probability. However, if the sum in (2) converges, then the tail sum $\sum_M^\infty \cdots$ must be less than $\delta/2$ for some $M$. Thus,

$$0 < \Pr\left[(2) \text{ converges}\right] \leq \Pr\left[\sum_{1}^{\infty} \cdots < \frac{\delta}{2}\right] + \Pr\left[\sum_{2}^{\infty} \cdots < \frac{\delta}{2}\right] + \cdots,$$

which implies the existence of an $M$ such that $\Pr[\sum_M^\infty \cdots < \delta/2] > 0$. Finally, because the $Y_\varepsilon$'s have Beta distributions and hence have full support, there is positive probability that the sum $\sum_1^{M-1} \cdots$ differs from the first integral of (1) by no more than $\delta/2$. Therefore, there is positive probability that (1) is less than $\delta$. □

**3. Partially specified Polya trees.** It may appear initially that calculations with Polya tree models are impossible to perform exactly because of the need to update infinitely many parameters. That this is not necessarily so is demonstrated in Lavine (1992). Nonetheless, calculations and computer programs may be simplified if the Polya trees are updated only as far as a predetermined level. This section presents two scenarios under which it is sensible to stop updating below a given predetermined level and under which the error of approximation may be either estimated or bounded.

To give probability 1 to the set of absolutely continuous distributions, to model other beliefs in smoothness or to satisfy Theorems 1 and 2, Polya trees may be constructed so that the $\alpha$'s increase rapidly toward the bottom of the tree. A sample of $n$ observations then cannot affect the tree very strongly below the level at which the $\alpha$'s become large relative to $n$. To achieve, therefore, a specified accuracy in the computation of the predictive distribution, it is only necessary to update finitely many levels of the Polya tree. For a specific example, suppose that, for $\varepsilon \in E^{m-1}$, $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1} \geq m^2$ and that a sample of size $n$ has been observed. The updated predictive density $g$ at a point $\theta$ can be bounded above by

$$\begin{aligned}
\log g(\theta \mid \Theta_1, \dots, \Theta_n) &\leq \log g(\theta) + \sum_{m=1}^{\infty} \log \frac{2(\alpha_{\varepsilon^m(\theta)} + n)}{2\alpha_{\varepsilon^m(\theta)} + n} \\
&\leq \log g(\theta) + \sum_{m=1}^{\infty} \log\left(1 + \frac{n}{2\alpha_{\varepsilon^m(\theta)}}\right) \\
&\leq \log g(\theta) + \frac{n}{2} \sum_{m=1}^{\infty} \frac{1}{\alpha_{\varepsilon^m(\theta)}} \\
&\leq \log g(\theta) + \frac{n}{2} \sum_{m=1}^{\infty} \frac{1}{m^2}
\end{aligned}$$

and below by

$$\begin{aligned}
\log g(\theta \mid \Theta_1, \dots, \Theta_n) &\geq \log g(\theta) + \sum_{m=1}^{\infty} \log \frac{2(\alpha_{\varepsilon^m(\theta)})}{2\alpha_{\varepsilon^m(\theta)} + n} \\
&\geq \log g(\theta) + \sum_{m=1}^{\infty} \log\left(1 - \frac{n}{2\alpha_{\varepsilon^m(\theta)}}\right) \\
&\geq \log g(\theta) - \frac{n}{2} \sum_{m=1}^{\infty} \frac{1}{\alpha_{\varepsilon^m(\theta)}} \\
&\geq \log g(\theta) - \frac{n}{2} \sum_{m=1}^{\infty} \frac{1}{m^2}.
\end{aligned}$$

Therefore, $g(\theta \mid \Theta_1, \dots, \Theta_n)$ can be evaluated to within a factor of $\delta$ by updating the Polya tree as far as level $M$, where $\log \delta \geq (n/2)\Sigma_M^\infty m^{-2}$.

Another argument for considering only finitely many levels of a Polya tree arises out of robust Bayesian considerations. It is unreasonable to expect an elicitee to specify a single Polya tree, or even a mixture of Polya trees, that is the only reasonable representation of prior beliefs. However, it might be possible for the elicitee to specify finitely many parameters near the top of a tree, possibly along with other beliefs, such as shape constraints, about $\mathcal{P}$. There will then be a class $\Gamma$ of prior distributions consistent with the prior information, and one can search for upper and lower bounds on prior and posterior quantities of interest, over $\Gamma$.

Let $S$ be a finite subset of $E^*$ such that, for any $\varepsilon = \varepsilon_1 \cdots \varepsilon_m$ in $S$, the initial sequence $\varepsilon_1 \cdots \varepsilon_j$ for each $j < m$ is also in $S$, and suppose that the elicitee is willing to specify the parameters $\{B_{\varepsilon 0}, B_{\varepsilon 1}, \alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}\colon \varepsilon \in S\}$. Let $T_1 = \{\mathcal{P}(B_{\varepsilon 0}), \mathcal{P}(B_{\varepsilon 1})\colon \varepsilon \in S\}$ be the random probabilities assigned by the partially specified Polya tree, and let $T_2$ be the mass distribution of $\mathcal{P}$ conditional on $T_1$. Thus, $\mathcal{P} = (T_1, T_2)$ and $\mathcal{L}(\mathcal{P}) = \mathcal{L}(T_1) \times \mathcal{L}(T_2 \mid T_1)$.

DEFINITION 2. The random variable $T_1$ is said to have a finite Polya tree distribution with parameter $(\mathcal{B}^S, \mathcal{A}^S)$, written $T_1 \sim \mathrm{PT}(\mathcal{B}^S, \mathcal{A}^S)$, if there exist sets $\mathcal{B}^S = \{B_{\varepsilon 0}, B_{\varepsilon 1}\colon \varepsilon \in S\}$, numbers $\mathcal{A}^S = \{\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}\colon \varepsilon \in S\}$ and random variables $\mathcal{Y}^S = \{Y_\varepsilon\colon \varepsilon \in S\}$ such that the following hold:

(i) all the random variables in $\mathcal{Y}^S$ are independent;
(ii) for every $\varepsilon \in S$, $Y_\varepsilon$ has a Beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$;
(iii) for every $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in S$,

$$\mathcal{P}(B_{\varepsilon_1 \cdots \varepsilon_m 0}) = \left( \prod_{j=1;\, \varepsilon_j = 0}^{m+1} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \right) \left( \prod_{j=1;\, \varepsilon_j = 1}^{m+1} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}) \right).$$

Mauldin, Sudderth and Williams (1992) define finite Polya trees using a special case of the preceding definition. We suppose that the elicitee will be able to give a marginal prior for $T_1$, but not a conditional prior for $T_2$ given $T_1$. We suppose also that the elicitee will be able to name a set $\mathbf{P}$, a subset of the set of all possible distributions which will receive mass 1 under any reasonable prior. For example, $\mathbf{P}$ may be the set of all unimodal distributions. Let $E$ be the event $(T_1, T_2) \in \mathbf{P}$ and let $\nu$ be a measure on $\Omega$. The class of priors $\Gamma$ is the class of all priors, conditioned on $E$, such that $T_1 \sim \mathrm{PT}(\mathcal{B}^S, \mathcal{A}^S)$, such that $\Pr[\mathcal{P} \ll \nu] = \Pr[d\mathcal{P}/d\nu$ is a.e. positive$] = 1$ and such that $\mathcal{L}(T_2 \mid T_1)$ is otherwise arbitrary, that is, probabilities of the form $\Pr[\mathbf{P} \cap \{\mathcal{P} = (T_1, T_2)\colon T_1 = t\}]$ are governed by the finite Polya tree.

It is the unconditional distribution of $T_1$ that has the Polya tree distribution. Conditional on $E$, the distribution of $T_1$ might not be Polya tree. For example, let $\Omega = (0, 1]$ and $\mathcal{B} = \{B_1, B_2, B_3, B_4\}$, where $B_i = ((i-1)/4, i/4]$, and let $\mathbf{P}$ be the set of unimodal distributions. The distribution of $T_1$ given $E$ cannot be Polya tree because the event $[\mathcal{P}(B_1) > \mathcal{P}(B_2);\ \mathcal{P}(B_4) > \mathcal{P}(B_3)]$ has positive probability under any Polya tree but probability 0 conditional on $E$.

We want bounds on posterior expectations of interesting functions of $\mathcal{P}$. Two useful examples are the posterior predictive mean $\mathbb{E}[\int \theta \, d\mathcal{P}]$ and the posterior predictive probability of a set $\mathbb{E}[\mathcal{P}(C)]$. We concentrate on finding the upper bound, as finding the lower bound is similar.

The following theorem says that the class of posteriors $\Gamma_\Theta$ has the same form as the class of priors. Therefore finding bounds on posterior expectations can be reduced to the problem of finding bounds on prior expectations.

THEOREM 3. *Let $\nu$ be a measure on $\Omega$; let $\pi = \{B_1, \dots, B_k\}$ be a partition of $\Omega$; let $T_1 = (\mathcal{P}(B_1), \dots, \mathcal{P}(B_k))$ and $T_2 = (\mathcal{P} \mid T_1)$; let the set of priors $\Gamma$ be the set of all conditional distributions for $\mathcal{P} = (T_1, T_2)$ given $E$, where $\mathcal{L}(T_1) = \mathrm{PT}(\mathcal{B}^S, \mathcal{A}^S)$ is specified, $\Pr[\mathcal{P} \ll \nu] = \Pr[d\mathcal{P}/d\nu$ is a.e. positive$] = 1$ and where $\mathcal{L}(T_2 \mid T_1)$ is otherwise arbitrary. Let $\Theta$ be an observation from $\mathcal{P}$, and let $J$ be defined by $\Theta \in B_J$. Then the corresponding set of posteriors $\Gamma_\Theta$ is the set of all conditional distributions for $(T_1, T_2)$ given $E$, where $\mathcal{L}(T_1 \mid \Theta) = \mathrm{PT}(\mathcal{B}^S, \mathcal{A}^S \mid \Theta \in J)$, $\Pr[\mathcal{P} \ll \nu] = \Pr[d\mathcal{P}/d\nu$ is everywhere positive$] = 1$ and where $\mathcal{L}(T_2 \mid T_1)$ is otherwise arbitrary.*

PROOF. We prove Theorem 3 without restriction to $E$, then note that the set of posteriors corresponding to priors restricted to $E$ can be obtained by restricting to $E$ the set of posteriors of unrestricted priors. Note first that $J$ is sufficient for $T_1$, so $\mathcal{L}(T_1 \mid \Theta) = \mathrm{PT}(\mathcal{B}^S, \mathcal{A}^S \mid \Theta \in J)$ follows easily. Now we want to show that any probability measure $Q$ for $\mathcal{L}(T_2 \mid T_1)$ that is absolutely continuous with an almost everywhere positive density is the posterior $\mathcal{L}(T_2 \mid T_1, \Theta)$ for some prior $Q_0$. However, let $Q_0$ be proportional to $Q/(d\mathcal{P}/d\nu(\Theta))$ and the result follows. □

REMARK 1. The class $\Gamma$ changes into the class $\Gamma_\Theta$ only by updating the distribution for $T_1$, a finite set of probabilities and a more easily understood object than $\mathcal{P}$.

REMARK 2. Although Theorem 3 is proven for general distributions for $T_1$, we intend to use finite Polya tree distributions for $T_1$, because they are conjugate and easily updated.

REMARK 3. Given $T_1$, the classes $\Gamma$ and $\Gamma_\Theta$ are ordinary quantile classes, possibly with restrictions, so that known results for quantile classes may apply.

EXAMPLE 1. Let $\Omega = [0, 4)$. We present a class of finite Polya tree priors with shape constraints to model an opinion that the law of the data has a nonincreasing density that is approximately exponential with parameter 1, and we show how to bound expectations over that class. Let $S = \{\emptyset, 1, 11\}$. Define $B_0 = [0, \ln 4/3), B_{10} = [\ln 4/3, \ln 2)$ and $B_{110} = [\ln 2, \ln 4)$, so that $\Omega$ is the disjoint union $B_0 \cup B_{10} \cup B_{110} \cup B_{111}$. Let $\alpha_0 = 20$, $\alpha_1 = 60$, $\alpha_{10} = 80/3$, $\alpha_{11} = 160/3$ and $\alpha_{110} = \alpha_{111} = 40$. The opinion that $\Theta_1$ has roughly an Exponential(1) distribution is modelled as $Y = \mathcal{P}(B_0) \sim \mathrm{Beta}(20, 60)$, $Y_1 = \mathcal{P}(B_{10})/\mathcal{P}(B_1) \sim \mathrm{Beta}(80/3, 160/3)$

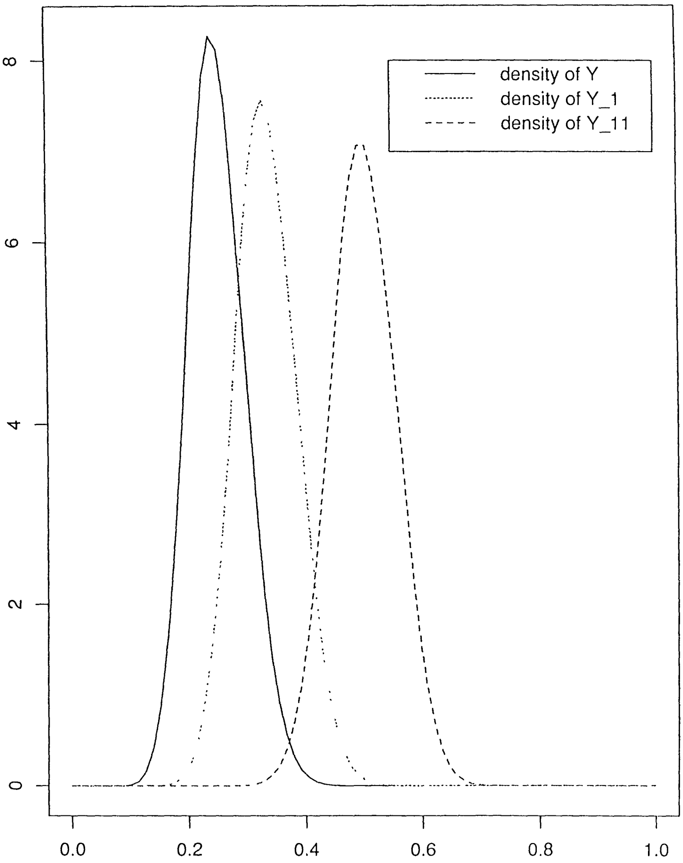

FIG. 1. *Densities of* $Y$, $Y_1$ *and* $Y_{11}$.

and $Y_{11} = \mathcal{P}(B_{110})/\mathcal{P}(B_{11}) \sim \text{Beta}(40, 40)$. Figure 1 shows the densities of $Y$, $Y_1$ and $Y_{11}$.

For any $Q \in \Gamma$, the expectation of $\Theta$ is $\int\int \mathbb{E}[\Theta \mid T_1, T_2]\, dQ(T_2 \mid T_1)\, dQ(T_1)$, which is maximized over $\Gamma$ by determining $Q(T_2 \mid T_1)$ separately for each value of $T_1$ to maximize $\mathbb{E}[\Theta]$ subject to monotonicity. The upper and lower bounds can be estimated by the following steps:

*Step 1.* Generate a sample $t_1, \dots, t_N$ from the known distribution of $T_1$.

*Step 2.* For each $t_j$ in the sample, let $q_{ji} = p_{ji}[\phi_i, \phi_{i+1})/(\phi_{i+1} - \phi_i)$, for $i \in \{0, \dots, 3\}$, be the average density assigned by $t_j$ to the interval $[\phi_i, \phi_{i+1})$.

*Step 3.* For each $j$, a nonincreasing density consistent with $t_j$ exists if and only if $q_{j0} \geq \cdots \geq q_{j3}$. If these inequalities do not hold, then drop the $j$th point from the sample. Let $N^*$ be the number of points remaining.

*Step 4.* Let $u_j$ and $l_j$ be the maximum and minimum expected values of $\Theta \mid t_j$, respectively.

*Step 5.* Let $\widehat{u} = \Sigma u_j/N^*$ and $\widehat{l} = \Sigma l_j/N^*$ be the estimates.

For Step 4 it is easy to see that $\widehat{u}$ is given by the density that is uniform and equal to $q_{ji}$ over the interval $[\phi_i, \phi_{i+1})$. The minimum, $\widehat{l}$, can be found by an optimization problem in a small number of variables [see Berger and O'Hagan (1988) for details of a similar optimization problem]. O'Hagan and Berger (1988) point out that the minimum can be bounded below using a density that is a step function with at most two values in each interval; the value of the bounding density in the interval $[\phi_i, \phi_{i+1})$ is $q_{j\,i-1}$ on the left side of the interval and $q_{j\,i+1}$ on the right side of the interval.

A sample of size $N = 1000$ was generated, of which $N^* = 818$ were acceptable according to Step 3. The estimates are $\widehat{u} = 1.09$ and $\widehat{l} = 0.80$. Figures 2 and 3 show how closely the model mimics the exp(1) distribution. Figure 2 shows boxplots of all 4000 of the $q_{ji}$'s grouped by interval. The solid diamonds are 0.25 divided by the interval length—a typical value for $q_{ji}$ according to the prior.

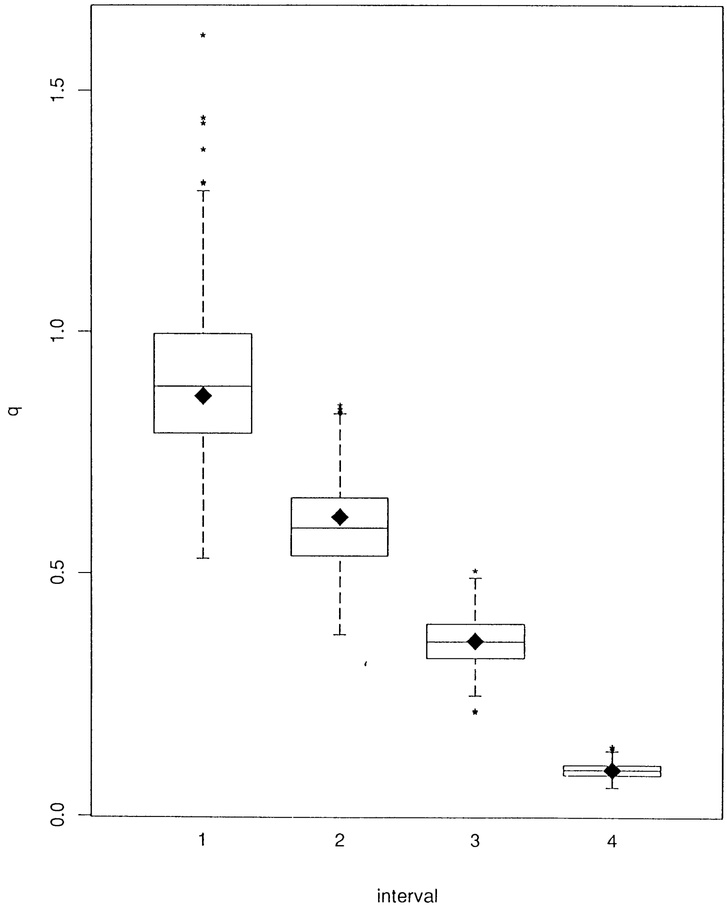

FIG. 2. *Boxplots of the $q_{ji}$ from Example* 1: *typical values according to the prior are indicated by diamonds.*

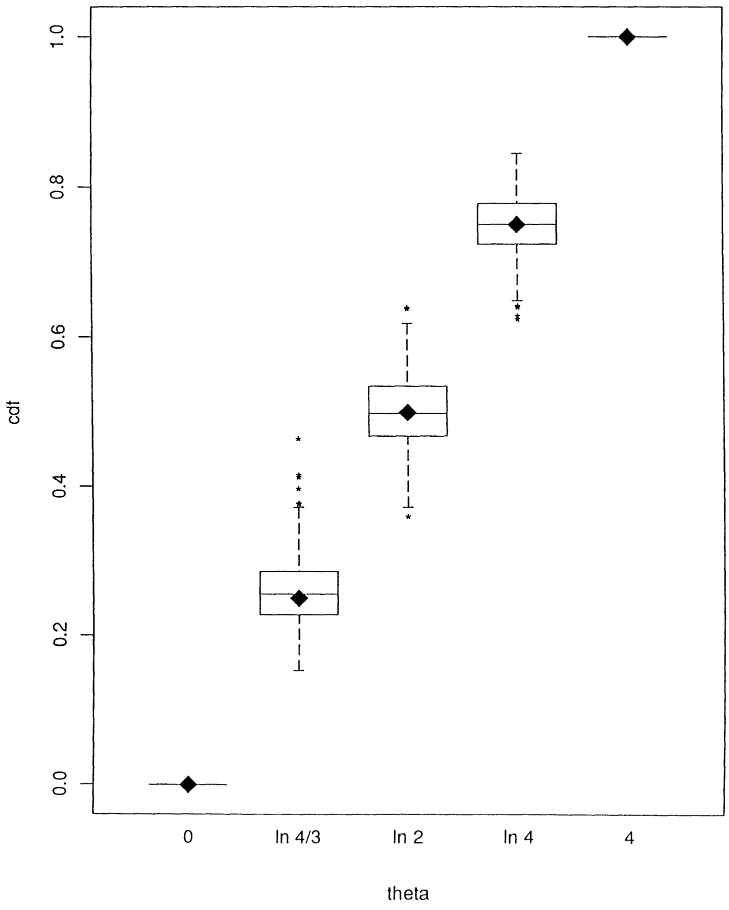

FIG. 3. *Boxplots of points on the random cdf's from Example* 1: *values from the* exp(1) *distribution are indicated by diamonds.*

Figure 3 shows boxplots of points on the 818 good cdf's. The solid diamonds are the values from the exp(1) cdf.

**4. Errors in regression.** This section describes the posterior distribution when Polya trees are used to model the errors in regression settings. The main result is a generalization of the following theorem which is implicit in Lemma 2.1 and Remark 1 of Diaconis and Freedman [(1986), page 71].

THEOREM DF. *Let $Y_i = \theta + \varepsilon_i$, where the $\varepsilon_i$ are independent with unknown distribution $\mathcal{P}$. With respect to the prior $Q$, let $\theta$ and $\mathcal{P}$ be independent, $\theta$ having density $f$ and $\mathcal{P}$ being Dirichlet with parameter measure $\alpha$ which is absolutely continuous; let $g = \alpha'/\|\alpha\|$, where $\|\alpha\|$ is the mass of $\alpha$. The posterior $Q_n$ can be* *characterized as follows*:

$$Q_n(d\theta) = C_n^{-1} f(\theta) \prod{}^{*} g(Y_i - \theta)\, d\theta,$$

$$Q_n\{d\mathcal{P} \mid \theta\} \text{ is } \mathcal{D}\left(\alpha + \sum_{i=1}^{n} \delta_{Y_i - \theta}\right),$$

*where*

$$C_n = \int_{-\infty}^{\infty} f(\theta) \prod{}^{*} g(Y_i - \theta)\, d\theta,$$

*and where* $\prod^*$ *is the product over distinct values in the sample.*

The first generalization is to drop the requirement that $\theta$ and $\mathcal{P}$ be independent with respect to the prior. Note that in the setting of Theorem DF, after a single observation $Y_1$, $\theta$ and $\mathcal{P}$ are no longer independent. The second generalization is from a single sample to a regression model: $\theta_i = r(X_i, \beta)$ where $r$ is a known regression function, $X_i$ is a known vector of covariates and $\beta$ is an unknown vector of parameters, with prior density $f$. Finally, the third generalization is from Dirichlet processes to Polya trees: $(\mathcal{P} \mid \beta) \sim \mathrm{PT}(\Pi_\beta, \mathcal{A}_\beta)$.

THEOREM 4. *Let* $Y_i = r(X_i, \beta) + \varepsilon_i$, *where the* $\varepsilon_i$ *are independent with unknown distribution* $\mathcal{P}$. *With respect to the prior* $Q$, *let* $\beta$ *have density* $f$ *and* $(\mathcal{P} \mid \beta) \sim \mathrm{PT}(\Pi_\beta, \mathcal{A}_\beta)$. *The posterior* $Q_n$ *can be characterized as follows*:

$$Q_n(d\beta) = C_n^{-1} f(\beta) \prod_{i=1}^{n} q(Y_i \mid \beta, Y_1, \dots, Y_{i-1})\, d\beta,$$

$$Q_n\{d\mathcal{P} \mid \beta\} \text{ is } \mathrm{PT}\big(\Pi_\beta, \mathcal{A}_\beta \mid Y_1 - r(X_1, \beta), \dots, Y_n - r(X_n, \beta)\big),$$

*where*

$$C_n = \int_B f(\beta) \prod_{i=1}^{n} q(Y_i \mid \beta, Y_1, \dots, Y_{i-1})\, d\beta$$

*and where* $q(Y_i \mid \beta, Y_1, \dots, Y_{i-1})$ *is the density of* $Y_i$ *given* $\beta, Y_1, \dots, Y_{i-1}$.

PROOF. The theorem follows by applying Bayes' theorem and the definition of conditional distribution. □

One reason for stating the theorem, apart from theoretical interest, is to point out that posterior densities for $\beta$ and for $Y_f$, a future observation with known covariate $X_f$, are computable. It is only necessary to evaluate $q(Y_i \mid \beta, Y_1, \dots, Y_{i-1})$, which is explained in Lavine (1992).

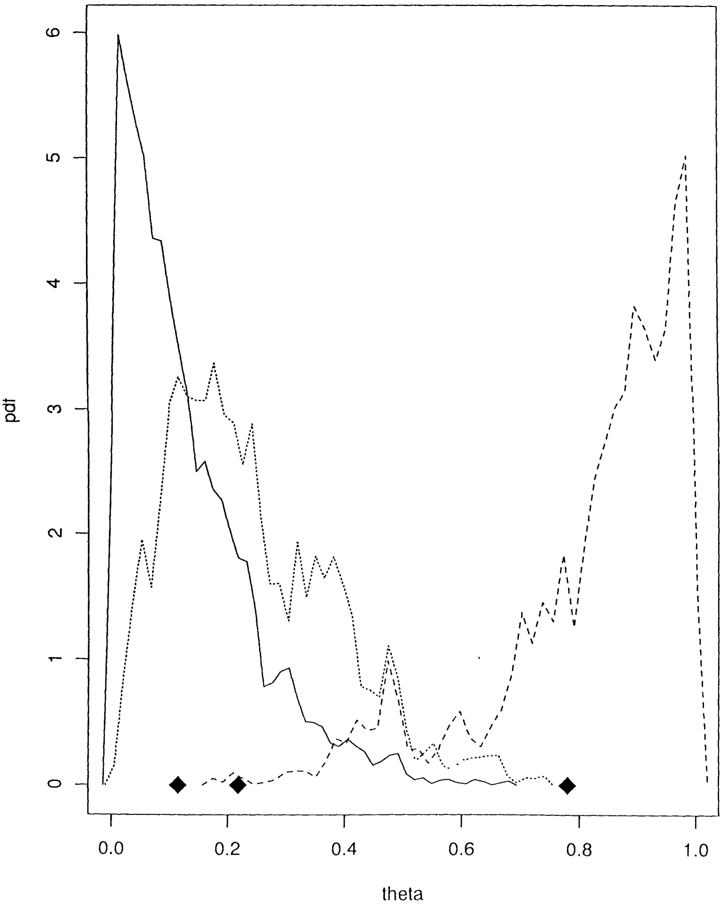

FIG. 4. *Density estimates for the* $\Theta_i$*'s from Example* 2: *the left curve is for* $\Theta_1$, $\Theta_3$, $\Theta_4$ *and* $\Theta_6$; *the middle curve is for* $\Theta_2$; *and the right curve is for* $\Theta_5$.

**5. Empirical Bayes problems.** The following empirical Bayes model was first stated by Antoniak (1974): $\mathcal{P}$, an unknown probability measure, has a prior that is a mixture of Dirichlet processes; conditional on $\mathcal{P}$, $\Theta_1, \dots, \Theta_n$ is a sample of size $n$ from $\mathcal{P}$; for each $i \in \{1, \dots, n\}$, conditional on $\mathcal{P}$, $\Theta_1, \dots, \Theta_n$, $X_i$ is a sample of size 1 from $F_{\Theta_i}$, independent of $\mathcal{P}$ and $\{X_j, \Theta_j: j \neq i\}$. When $\mathbf{X} = (X_1, \dots, X_n)$ is observed, but not $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_n)$, we may wish to estimate or find the posterior distributions of $\mathcal{P}$, $\Theta_i$, $\Theta_{n+1}$ or $X_{n+1}$.

The problem was further studied by Berry and Christensen (1979) in the case where $F_{\Theta_i}$ is the binomial distribution with parameter $\Theta_i$, by Ferguson (1983) in the case where $F_{\Theta_i}$ is the normal distribution with parameter $\Theta_i$ and by Lo (1984) in the general case. Kuo (1986) proposes a Monte Carlo method for computing the estimates. More recently, the model has appeared [Escobar (1994), West (1990), Escobar and West (1990)] with computations done by means of the Gibbs sampler [(Gelfand and Smith (1990)], an algorithm for drawing

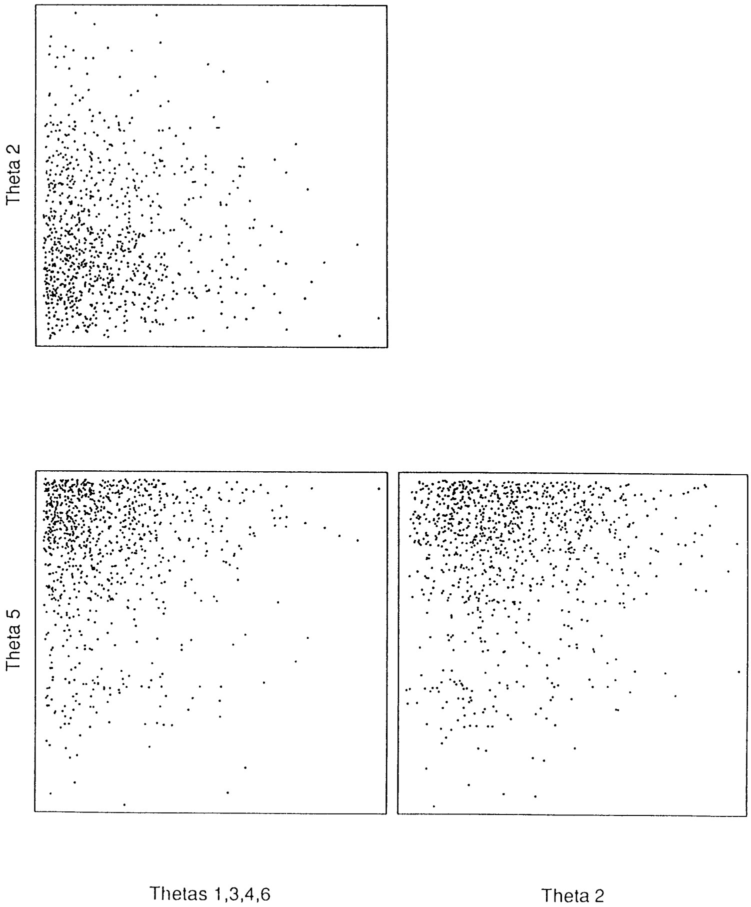

FIG. 5. *Pairwise plots of the Gibbs sampled points from Example* 2.

an approximate sample from a multivariate distribution when the conditional distribution of each variable given all the others can be sampled. Here we generalize the model so that $\mathcal{P}$ has a Polya tree prior, show how posteriors can be computed via the Gibbs sampler and discuss advantages provided by mixtures of Polya tree over mixtures of Dirichlet processes.

To generate a sample from the posterior distribution of $(\Theta)$ via the Gibbs sampler, it is required to generate an observation from the distribution of, say, $(\Theta_n \mid \mathbf{X}, \Theta_1, \dots, \Theta_{n-1})$, which is the same as the distribution of $(\Theta_n \mid X_n, \Theta_1, \dots, \Theta_{n-1})$. However, as described in Lavine (1992), the density of $(\Theta_n \mid \Theta_1, \dots, \Theta_{n-1})$ is a piecewise rescaled version of the prior density of $\Theta_n$, so the density of $(\Theta_n \mid X_n, \Theta_1, \dots, \Theta_{n-1})$ is a piecewise rescaled version of the density of $(\Theta_n \mid X_n)$. Therefore, if the density of $(\Theta_n \mid X_n)$ is available for sampling, then so is the density of $(\Theta_n \mid X_n, \Theta_1, \dots, \Theta_{n-1})$, at least to arbitrary accuracy. The sampling algorithm is particularly easy if only finitely many of the Polya tree parameters need be updated.

Algorithmically, the difference between a Dirichlet process prior and a Polya tree prior is that for the Dirichlet process the distribution of $(\Theta_n \mid X_n, \Theta_1, \dots, \Theta_{n-1})$ is a mixture of the distribution of $(\Theta_n \mid X_n)$ and the degenerate distributions $\delta_{\Theta_1}, \dots, \delta_{\Theta_{n-1}}$. Sampling from this mixture would typically be just as easy as the sampling required for the Polya tree model. However, the Dirichlet process model imposes features that may be deemed undesirable. For example, when $\mathcal{P}$ has a Dirichlet process prior, then with positive probability some of the $\Theta_i$'s are equal [Ferguson (1973)], and the law of the pattern of multiplicities has a specific form [Antoniak (1974)]. In contrast, the Polya tree prior can be constructed so that $\Theta_1, \dots, \Theta_n$ are distinct with probability 1.

EXAMPLE 2. An example from Berry and Christensen (1979) uses data from Martz and Lian (1974), who are quoted as saying, "The Portsmouth Naval Shipyard, Portsmouth, N.H. routinely must assess the quality of submitted lots of vendor produced material. The following data consist of the number of defects $x_i$ of a specified type in samples of size $n = 5$ from past lots of welding material. The past data are $(0, 1, 0, 0, 5)$ and in the current, i.e., sixth, lot, $x = 0$."

Berry and Christensen (1979) use a model in which $x_1, \dots, x_6$ are observations from binomial distributions with $n = 5$ and success probabilities $\Theta_1, \dots, \Theta_6$. The $\Theta$'s in turn are a sample from $\mathcal{P}$, which has a Dirichlet process prior with base measure Beta(1, 1). They estimate $\widehat{\Theta}_1 = \widehat{\Theta}_3 = \widehat{\Theta}_4 = \widehat{\Theta}_6 = 0.1143$, $\widehat{\Theta}_2 = 0.2178$ and $\widehat{\Theta}_5 = 0.7795$, the points shown as diamonds in Figure 4.

The data were reanalyzed with a similar model, but with $\mathcal{P}$ having a Polya tree prior, where $\pi_m$ is the collection of intervals $[i/2^m, (i+1)/2^m)$, for $i = 0, \dots, 2^m - 1$, and where $\alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 0} = \alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 1} = m^2$ to ensure the absolute continuity of $\mathcal{P}$. A sample of size 1000 from the posterior distribution of $\Theta_1, \dots, \Theta_6$ was obtained via the Gibbs sampler. Figure 4 shows univariate posterior density estimates. Figure 5 indicates the bivariate posterior distributions with plots of the sampled points. By symmetry, $\theta_1, \theta_3, \theta_4$ and $\theta_6$ have the same marginal posterior. However, because the Polya tree prior puts its mass on continuous $\mathcal{P}$, $\theta_1, \theta_3, \theta_4$ and $\theta_6$ are distinct with probability 1.

## REFERENCES

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

BARRON, A. (1986). Discussion of "On the consistency of Bayes estimates" by P. Diaconis and D. Freedman. *Ann. Statist.* **14** 26–30.

BERGER, J. O. and O'HAGAN, A. (1988). Ranges of posterior probabilities for unimodal priors with specified quantiles. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 45–65. Clarendon, Oxford.

BERRY, D. and CHRISTENSEN, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7** 558–568.

BILLINGSLEY, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

DALAL, S. and HALL, G. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8** 664–672.

DIACONIS, P. and FREEDMAN, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87.

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277.

ESCOBAR, M. D. and WEST, M. (1990). Bayesian prediction and density estimation. Discussion Paper 90-A16, Inst. Statistics and Decision Sciences, Duke Univ.

FABIUS, J. (1964). Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* **35** 846–856.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (H. Rizvi and J. Rustagi, eds.) 287–302. Academic, New York.

FREEDMAN, D. A. (1963). On the asymptotic behaviour of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1194–1216.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

KRAFT, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.* **1** 385–388.

KUO, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.* **7** 60–71.

LAVINE, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235.

LO, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12** 351–357.

MARTZ, H. F. and LIAN, M. G. (1974). Empirical Bayes estimation of the binomial parameter. *Biometrika* **61** 517–523.

MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* **20** 1203–1221.

METIVIER, M. (1971). Sur la construction de mesures aléatoires presque sûrement absolument continues par rapport à une mesure donnée. *Z. Wahrsch. Verw. Gebiete* **20** 332–344.

O'HAGAN, A. and BERGER, J. O. (1988). Ranges of posterior probabilities for quasiunimodal priors with specified quantiles. *J. Amer. Statist. Assoc.* **83** 503–508.

WEST, M. (1990). Bayesian kernel density estimation. Discussion Paper #90-A02, Inst. Statistics and Decision Sciences, Duke Univ.

INSTITUTE OF STATISTICS AND DECISION SCIENCES
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708-0251