Sankhyā : The Indian Journal of Statistics
2003, Volume 65, Part 3, pp 577-592

SOME FURTHER DEVELOPMENTS FOR STICK-BREAKING PRIORS: FINITE AND INFINITE CLUSTERING AND CLASSIFICATION

By HEMANT ISHWARAN
Cleveland Clinic Foundation, Cleveland, USA
and
LANCELOT F. JAMES
Hong Kong University of Science and Technology, Hong Kong

SUMMARY. The class of stick-breaking priors and their extensions are considered in classification and clustering problems in which the complexity, the number of possible models or clusters, can be either bounded or unbounded. A conjugacy property for the extended stick-breaking prior is established which allows for informative characterizations of the priors under i.i.d. sampling, and which further enables an informative characterization of the posterior in the classification model. Such characterizations show how to develop Monte Carlo algorithms for efficient posterior computing. One implication is that it is possible to estimate infinite complexity mixture models subject to arbitrary stick-breaking priors.

1. Introduction

A general classification problem can be described as follows. One observes i.i.d. observations $X_1, \ldots, X_n$ drawn from a mixture model whose true density is of the form

$$f_0(x) = \sum_{j=1}^{d} W_{j,0}\, k_j(x|Y_{j,0}), \tag{1}$$

where $k_j(x|Y)$ are known kernel densities in $x \in \mathcal{X}$ for each $Y \in \mathcal{Y}$. Given the data $\mathbf{X} = (X_1, \ldots, X_n)$, the problem is to classify each observation X_i

Paper received January 2001; revised August 2002.
AMS (2000) *subject classification.* Primary 62G99; secondary 62G09.
Keywords and phrases. Classification variables, Collapsed Gibbs sampling, Dirichlet process, Finite and infinite dimensional Dirichlet priors, Two-parameter Poisson-Dirichlet process, sequential importance sampling.

into one of the d possible models given that the weights $0 < W_{j,0} < 1$ and atoms $Y_{j,0} \in \mathcal{Y}$ for the mixture distribution are assumed unknown (note that $\sum_{j=1}^{d} W_{j,0} = 1$). We assume that d, the complexity of the true model, is finite and that it has some known bound N. The value of d may or may not be known to us. For example, applications in which $N = \infty$ correspond to an infinite collection of candidate kernels $k_1, k_2, \ldots$ of which only d are from the true model, but we do not know which ones. Here the bound for d is crude, $1 \leq d < N = \infty$. In other problems we may have more information so that our bound may be finite, $1 \leq d \leq N < \infty$. Finally, we might have a problem where d is known, so that our bound N is perfect, $N = d$. In a variation on the above problems, we will also consider the setting where all kernels are equal, $k_j = k_0$, in which case (1) corresponds to a finite mixture model. Here d plays the role of the dimension of the unknown mixing distribution and the focus shifts from classification, which is no longer meaningful with equal kernels, to the problem of clustering the data X_i (more discussion on this distinction will be forthcoming). Examples that have been considered from a Bayesian approach that fall into the above general framework include mixture models with bounded complexity (Ishwaran, James and Sun 2001), clustering and discrimination (Hartigan 1975, Binder 1978 and Lavine and West 1992) and change-point and switching regression problems (Chen and Liu 1996 and Frühwirth-Schnatter 2001). See also Brunner and Lo (1999) and Quintana and Iglesias (2003) for partition modelling approaches based on the Dirichlet process.

As mentioned, an important special case of (1) is the finite mixture model which corresponds to

$$f_0(x) = \int k_0(x|y)\, dQ_0(y),$$

where $Q_0(\cdot) = \sum_{j=1}^{d} W_{j,0}\delta_{Y_{j,0}}(\cdot)$ is the underlying unknown mixture distribution (δ_y denotes a discrete measure concentrated at y). A nice feature of this model, which will form the basis of our general development, is that it has an equivalent representation as a missing data model. By introducing hidden variables Y_i, such models can be written hierarchically as

$$(X_i|Y_i) \overset{\text{ind}}{\sim} k_0(X_i|Y_i), \quad (Y_i|Q_0) \overset{\text{iid}}{\sim} Q_0, \qquad i = 1, \ldots, n, \tag{2}$$

or equivalently, by writing Y_i as $Y_{K_i,0}$, as

$$(X_i|Y_{K_i,0}) \overset{\text{ind}}{\sim} k_0(X_i|Y_{K_i,0}), \quad (K_i|\mathbf{W}_0) \overset{\text{iid}}{\sim} \sum_{k=1}^{d} W_{k,0}\,\delta_k(\cdot), \qquad i = 1, \ldots, n, \tag{3}$$

where $\mathbf{W}_0 = (W_{1,0}, \dots, W_{d,0})$.

Exploiting such representations, Ishwaran and James (2001) (see also Ishwaran and Zarepour 2000) presented a general Bayesian approach for modeling semiparametric and nonparametric models, which can be applied to (3), using what they defined as the class of stick-breaking priors. Call a random measure $\mathcal{P}$ a *stick-breaking prior* if it has an expression of the form

$$\mathcal{P}(\cdot) = \sum_{k=1}^{N} W_k\, \delta_{Z_k}(\cdot), \tag{4}$$

where Z_k are i.i.d. values with a non-atomic distribution H over $(\mathcal{Y}, \mathcal{B})$, a measurable Polish space, and the Z_k are independent of W_k, which are random weights constructed to sum to one using what is often called a "stick-breaking construction" (Halmos 1944, Freedman 1963, Fabius 1964, Connor and Mosimann 1969, Patil and Taillie 1977). Specifically,

$$W_1 = V_1, \qquad W_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \geq 2, \tag{5}$$

where V_k are independent Beta(a_k, b_k) random variables with shape parameters $a_k, b_k > 0$. Such constructions hold for both the case when $N < \infty$ (in this case $V_N = 1$ and the prior is referred to as a *finite dimensional stick-breaking prior*) as well as the setting when $N = \infty$ (this involves constraints to a_k, b_k yielding priors referred to as *infinite dimensional stick-breaking priors*). More details and examples will be forthcoming in Section 2.

To model (2) (and hence (3)), Ishwaran and James (2001) proposed the use of stick-breaking prior $\mathcal{P}$ as a method for modelling Q_0. They proposed the use of the Bayesian nonparametric model

$$(X_i|Y_i) \overset{\text{ind}}{\sim} k_0(X_i|Y_i), \quad (Y_i|P) \overset{\text{iid}}{\sim} P, \quad P \sim \mathcal{P}, \qquad i = 1, \dots, n,$$

which by the identity $Y_i = Z_{K_i}$ was observed to be equivalent to

$$(X_i|Z_{K_i}) \overset{\text{ind}}{\sim} k_0(X_i|Z_{K_i}), \quad (K_i|\mathbf{W}) \overset{\text{iid}}{\sim} \sum_{k=1}^{N} W_k\, \delta_k(\cdot), \quad Z_k \overset{\text{iid}}{\sim} H, \quad \mathbf{W} \sim \pi_{\mathbf{W}}, \tag{6}$$

where $\mathbf{W}$ is the vector of random weights W_k (which can be either finite or infinite dimensional) whose prior $\pi_{\mathbf{W}}$ is defined by (5).

1.1. Classification using $K_1, \ldots, K_n$. Notice that the use of (6) in modelling (3) implicitly relies on the conditional distribution of X_i being fully specified in terms of $Y_i = Z_{K_i}$. The same method can also be applied to the general classification problem (1) if we further use K_i to identify the kernel corresponding to X_i. This points to a novel extension for analysing this problem. Given that we have different kernels, we assume that $Y_i = Z_{K_i}$, where Z_k are independent (but not necessarily identically distributed) with non-atomic distributions H_k over $(\mathcal{Y}, \mathcal{B})$. Thus we model Y_i as having a random distribution P drawn from a measure as in (4), but generalized to allow for different laws for Z_k; thus extending the notion of a stick-breaking prior. We call such measures *extended stick-breaking priors*. Now, allowing the conditioning on X_i to additionally include classification variables K_i as well as Z_{K_i}, the Bayesian model for (1) is

$$\begin{aligned}(X_i|Z_{K_i}, K_i) &\overset{\text{ind}}{\sim} \sum_{j=1}^{N} k_j(X_i|Z_{K_i}) I\{K_i = j\},\\ (K_i|\mathbf{W}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} W_k \delta_k(\cdot),\ Z_k \overset{\text{ind}}{\sim} H_k,\ \mathbf{W} \sim \pi_{\mathbf{W}}.\end{aligned} \tag{7}$$

The role of K_i in (7) is quite different than in the finite mixture model (3), where K_i is often included as a means for facilitating computations rather than for inference. In fact, because all kernels are the same in (3), one could bypass the use of K_i all together, working instead with Y_i, or at an even coarser level, the partition structure recording the clustering of Y_i; that is a partition of the integers $\{1, \ldots, n\}$. This latter approach was exploited by Brunner, Chan, James and Lo (2001) for analysis of mixture models subject to the Dirichlet process (see also Lo, Brunner and Chan 1996 and MacEachern, Clyde and Liu 1999), later being extended by Ishwaran and James (2003) to encompass mixture models for the class of all exchangeable urn distributions. One of the premises for basing such techniques on the partition structure for Y_i is that such information represents a minimal sufficient statistic for reducing the nonparametric mixture model into a collection of parametric models. Furthermore, such "collapsing" of the parameter space leads to Rao-Blackwell improvements in terms of Monte Carlo errors for computational algorithms. However, such techniques do not directly apply to (7) as the information contained in the partition structure is too "coarse" to classify observations into their models $\{k_1, k_2, \ldots\}$. In (7), the minimal amount of information necessary for classification is $K_1, \ldots, K_n$. Moreover, such partition based methods do not always readily apply to the mixture model (6), as the methods rely on explicit formulae for the prediction rule of the hidden Y_i variables, which for the general stick-breaking prior may not always be in a simple enough form for computational purposes. Recognizing this, Ishwaran and James (2001) and Ishwaran, James and Lo (2001) showed how to work with K_i to permit inference for (6) under the class of finite dimensional stick-breaking priors. Here, we will show how these methods can be extended to the general classification problem (7) under the class of extended stick-breaking priors, including both the finite as well as infinite dimensional cases.

In brief, then, the contributions of this paper will be to develop new surrounding theory for the hierarchical model (7) and show how these may be used to develop computational algorithms for computing posterior quantities. Our theoretical contributions include developing key properties for the class of extended stick-breaking measures, which includes establishing a conjugacy property of their random weights to i.i.d. sampling, and a characterization of the posterior for the extended stick-breaking prior under i.i.d. sampling. See Section 3. These properties then lead us in Section 4 to a general characterization for the posterior of (7). In Section 5 we outline a collapsed Gibbs sampling algorithm and an i.i.d. SIS (sequential importance sampling) algorithm that can be used for inference in (7). One important implication is our ability to fit the posterior of (6) subject to infinite dimensional stick-breaking measures. The paper begins with a brief discussion of stick-breaking priors in Section 2.

2. Some Examples of Stick-breaking Priors

Perhaps the best known example of an infinite dimensional stick-breaking prior is the Ferguson (1973,1974) Dirichlet process prior. Its stick-breaking construction was confirmed in Sethuraman (1994) with related work appearing in McCloskey (1965), Patil and Taillie (1977), Sethuraman and Tiwari (1982), Hoppe (1987), Donnelly and Joyce (1989), Perman, Pitman and Yor (1992) and Pitman (1996). This construction can also be viewed as a special case of the rich class of two-parameter Poisson-Dirichlet processes developed by Pitman and Yor (1997). Such measures can be expressed as infinite dimensional stick-breaking priors using $\text{Beta}(1-a, b+ka)$ laws for V_k where $0 \leq a < 1$ and $b > -a$. The Dirichlet process, written $\text{DP}(\alpha H)$, corresponds to the construction with $a = 0$ and $b = \alpha$. Another important class are the stable law processes with index $0 < \alpha < 1$, corresponding to $a = \alpha$ and $b = 0$. See Pitman and Yor (1997) and Pitman (1995, 1996) for further details. As discussed in Ishwaran and James (2001) an infinite dimensional stick-breaking prior is well defined if and only if its random weights satisfy $\sum_{k=1}^{\infty} E(\log(1-V_k)) = -\infty$ (a sufficient condition being $\sum_{k=1}^{\infty} \log(1+a_k/b_k) = +\infty$). Such conditions also apply to the class of extended stick-breaking priors.

A flexible class of finite dimensional stick-breaking priors are the *finite dimensional Dirichlet priors* discussed in Ishwaran and Zarepour (2002a). These are priors with random weights $\mathbf{W}$ whose law is a Dirichlet distribution, Dirichlet$(\alpha_1,\ldots,\alpha_N)$. The case where $\alpha_k = \alpha/N$, for $\alpha > 0$, has the important feature that it approximates the Dirichlet process, DP(αH) (see Ishwaran and Zarepour 2002b for further details). We denote its prior by DP$_N(\alpha H)$. In general, Dirichlet random weights $\mathbf{W}$ can be constructed as in (5) using independent Beta(a_k, b_k) random variables by setting $a_k = \alpha_k$, $b_k = \sum_{j=k+1}^{N} \alpha_j$ for $k = 1, 2, \ldots, N-1$, and by setting $V_N = 1$. The Dirichlet distribution is a special case of the generalized Dirichlet distribution (Connor and Mosimann 1969), the law arising from a finite stick-breaking construction (5) with $V_N = 1$. The class of finite dimensional Dirichlet priors can be extended to what we refer to as *infinite dimensional Dirichlet priors.* These are infinite dimensional stick-breaking priors in which $a_k = \alpha_k$ and $b_k = \sum_{j=k+1}^{\infty} \alpha_k$, for $\alpha_k > 0$ chosen so that $\sum_{k=1}^{\infty} \alpha_k < \infty$. We can verify directly that such a construction is well defined. Observe that since $a_k + b_k = b_{k-1}$,

$$E(W_k) = \frac{a_k}{a_k+b_k}\prod_{j=1}^{k-1}\frac{b_j}{a_j+b_j} = \frac{a_k}{a_1+b_1} = \frac{\alpha_k}{\sum_{k=1}^{\infty}\alpha_k},$$

and thus $\sum_{k=1}^{\infty} E(W_k) = 1$. From this it follows that $\sum_{k=1}^{\infty} W_k = 1$ almost surely. See Ishwaran and James (2001) for further discussion on stick-breaking priors.

3. Properties for Stick-breaking Priors and their Extensions

In this section we establish that the law of a stick-breaking random weight $\mathbf{W}$ is conjugate to i.i.d. sampling. That is, if $K_1, \ldots, K_n$ given $\mathbf{W}$ are i.i.d. values drawn from $\sum_{k=1}^{N} W_k\, \delta_k(\cdot)$, then the posterior for $\mathbf{W}$ given $K_1, \ldots, K_n$ has a stick-breaking representation (5). This important characterization will point to several key properties for the class of extended stick-breaking priors and their posteriors in the Bayesian classification model (7).

The key to establishing the conjugacy property can be based on the following lemma stating the posterior for $\mathbf{W}$ given the cluster variable K_1. In the setting where $N < \infty$ the result follows automatically from the conjugacy of the generalized Dirichlet distribution to multinomial sampling (Ishwaran and Zarepour 2000). However, it stands to reason that a similar result should also hold when $N = \infty$. This is stated as the following conjugacy result, which can be seen as a generalization of Lemma 4.2 of Sethuraman (1994) for the $\text{DP}(\alpha H)$ prior ($N = \infty$, $a_k = 1$ and $b_k = \alpha$).

LEMMA 1. *Suppose that K_1 given $\mathbf{W}$ has the law $\Pr\{K_1 \in \cdot|\mathbf{W}\} = \sum_{k=1}^N W_k\,\delta_k(\cdot)$, for $\mathbf{W}$ defined by the stick-breaking construction (5). Then, the law for $\mathbf{W}$ given K_1 is also defined by (5), where now V_k are independent Beta(a_k^*, b_k^*) random variables with $a_k^* = a_k + I\{K_1 = k\}$ and $b_k^* = b_k + \sum_{j=k+1}^N I\{K_1 = j\}$.*

PROOF. As mentioned above the result holds automatically for finite N, so we prove the case for $N = \infty$. We follow an argument similar to the proof of Lemma 4.2 of Sethuraman (1994) by noting that the posterior is identified by characterizing the joint distribution of K_1 and $V_1, \ldots, V_m$ for each positive integer m.

Let $A_1, \cdots, A_m$ be measurable sets over $[0,1]$. We have,

$$\begin{aligned}
&\Pr\{V_1 \in A_1, \ldots, V_m \in A_m, K_1 = j\}\\
&\quad = \int \prod_{k=1}^m I\{V_k \in A_k\} \Pr\{K_1 = j|V_1, V_2, \ldots\}\,\pi(dV_1, \ldots, dV_m)\\
&\quad = \int \prod_{k=1}^m I\{V_k \in A_k\}\,(1 - V_1)(1 - V_2)\cdots(1 - V_{j-1})V_j\,\pi(dV_1, \ldots, dV_m),
\end{aligned}$$

where the last line uses $\Pr\{K_1 = j|V_1, V_2 \ldots\} = W_j$. Deduce that $V_1, \ldots, V_m$ given K_1 are independent $\text{Beta}(a_k^*, b_k^*)$ random variables for $k = 1 \ldots, m$. This holds for all m, thus characterizing the posterior for $\mathbf{W}$. □

By applying Lemma 1 repeatedly, using conjugacy, we have the following important corollary.

COROLLARY 1. *Suppose that $K_1, \ldots, K_n$ given $\mathbf{W}$ are i.i.d. with law $\sum_{k=1}^N W_k\,\delta_k(\cdot)$. Then the law for $\mathbf{W}$ given $\mathbf{K} = (K_1, \ldots, K_n)$ is defined by (5), where V_k are independent Beta(a_k^*, b_k^*) random variables with $a_k^* = a_k + e_k$ and $b_k^* = b_k + \sum_{j=k+1}^N e_j$, where e_j denotes the number of K_i equaling j.*

3.1. Implications of conjugacy. The conjugacy of $\mathbf{W}$ leads to a posterior characterization for the stick-breaking measure under i.i.d. sampling. This can be used to show that if $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a sample from P with an extended stick-breaking law, then the posterior for P given $\mathbf{Y}$ and $\mathbf{K}$ has a representation expressible in terms of an extended stick-breaking measure. In what follows, let $K_1, \ldots, K_n$ denote classification variables as in Corollary 1. Furthermore, set $\mathbf{K}^* = \{K_1^*, \ldots, K_m^*\}$ to be the set of $m = n(\mathbf{K})$ unique values of $\mathbf{K}$.

THEOREM 1. *Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is an i.i.d. sample from an extended stick-breaking prior $\mathcal{P}$. That is, $(Y_i|P)$ are i.i.d. P where P has the law $\mathcal{P}$ described by (4) where Z_k are independent with laws H_k (equivalently, $Y_i = Z_{K_i}$). Then, the posterior of P given $\mathbf{Y}$ is characterized by*

$$\pi(dP|\mathbf{Y}) = \sum_{\mathbf{K}} \pi(dP|\mathbf{Y}, \mathbf{K}) \Pr(\mathbf{K}),$$

where the sum is over all $\mathbf{K}$,

$$\Pr(\mathbf{K}) = \Pr(K_1) \prod_{i=2}^{n} \Pr(K_i|K_1, \ldots, K_{i-1}) = E[W_{K_1}] \prod_{i=2}^{n} E\left[W_{K_i}|K_1, \ldots, K_{i-1}\right]$$

is the prior for $\mathbf{K}$ (which can be determined from Corollary 1) and the law of $(P|\mathbf{Y}, \mathbf{K})$ is a stick-breaking measure representable in terms of the random measure

$$P^*(\cdot) = \sum_{k \in \mathbf{K}^*} W_k \, \delta_{Y_k^*}(\cdot) + \sum_{k \in (\mathbf{K}^*)^c} W_k \, \delta_{Z_k}(\cdot), \tag{8}$$

where $Y_1^*, \ldots, Y_m^*$ are the unique values of $Y_1, \ldots, Y_n$, and W_k are the random weights with law $\pi(\mathbf{W}|\mathbf{K})$ described in Corollary 1 and are indenendent of Z_k.

PROOF. The first two displayed equations follow by definition, which leaves us to prove (8). The law for P is determined by the law for $\mathbf{W}$ and $\mathbf{Z}$, where $\mathbf{Z}$ is the vector of Z_k variables. We have that

$$\mathcal{L}(\mathbf{W}, \mathbf{Z}|\mathbf{Y}, \mathbf{K}) = \mathcal{L}(\mathbf{Z}|\mathbf{W}, \mathbf{Y}, \mathbf{K}) \times \mathcal{L}(\mathbf{W}|\mathbf{Y}, \mathbf{K}) = \mathcal{L}(\mathbf{Z}|\mathbf{Y}, \mathbf{K}) \times \mathcal{L}(\mathbf{W}|\mathbf{K}).$$

The law for the second term is defined by Corollary 1, while the first term is easily determined using $Y_i = Z_{K_i}$. Deduce that,

$$\mathcal{L}(P|\mathbf{Y}, \mathbf{K}) = \mathcal{L}\left(\sum_{k \in \mathbf{K}^*} W_k \, \delta_{Y_k^*} + \sum_{k \in (\mathbf{K}^*)^c} W_k \, \delta_{Z_k} \right),$$

for W_k defined by $\pi(\mathbf{W}|\mathbf{K})$. □

REMARK 1. Note that the representation (8) for P^* involves components which are not independent. An alternate representation which may be useful in some settings is the following. Let ℓ be the largest component of $\mathbf{K}$, i.e. $\ell = \max\{K_1, \ldots, K_n\}$. Then, (8) can be rewritten as

$$P^*(\cdot) = \sum_{k \in \mathbf{K}^*} W_k\, \delta_{Y_k^*}(\cdot) + \sum_{k \in \mathcal{M} - \mathbf{K}^*} W_k\, \delta_{Z_k}(\cdot) + \left[1 - \sum_{k=1}^{\ell} W_k\right] P_\ell(\cdot), \tag{9}$$

where $\mathcal{M} = \{1, \ldots, \ell\}$ and P_ℓ is an independent stick-breaking prior defined by (assuming $N = \infty$)

$$\begin{aligned} P_\ell(\cdot) &= \sum_{k=\ell+1}^{\infty} \frac{W_k}{[1 - \sum_{j=1}^{\ell} W_j]} \delta_{Z_k}(\cdot) \\ &= V_{\ell+1}\delta_{Z_{\ell+1}}(\cdot) + \sum_{k=2}^{\infty} \left[\prod_{j=1}^{k-1} (1 - V_{\ell+j}) V_{\ell+k}\right] \delta_{Z_{\ell+k}}(\cdot) \\ &= \sum_{k=1}^{\infty} \overline{W}_k\, \delta_{Z_{\ell+k}}(\cdot), \end{aligned} \tag{10}$$

where $\overline{W}_1 = V_{\ell+1}$ and $\overline{W}_k = (1 - V_{\ell+1})(1 - V_{\ell+2}) \cdots (1 - V_{\ell+k-1}) V_{\ell+k}$. Observe by Corollary 1 that $V_{\ell+j}$ are independent $\text{Beta}(a_{\ell+j}, b_{\ell+j})$ random variables.

REMARK 2. A nice simplification occurs in computing $\Pr(\mathbf{K})$ for the class of finite and infinite dimensional Dirichlet priors. From their property that $a_k + b_k = b_{k-1}$ (and hence $a_k^* + b_k^* = b_{k-1}^*$), it follows from Corollary 1 that

$$E[W_k|\mathbf{K}] = \frac{a_k^*}{a_k^* + b_k^*} \prod_{j=1}^{k-1} \frac{b_j^*}{a_j^* + b_j^*} = \frac{\alpha_k + e_k}{\sum_{k=1}^{N} \alpha_k + n}.$$

Thus, deduce that

$$\Pr(\mathbf{K}) = \frac{\alpha_{K_1}}{\sum_{k=1}^{N} \alpha_k} \prod_{i=2}^{n} \frac{\alpha_{K_i} + e_i^*}{\sum_{k=1}^{N} \alpha_k + i - 1},$$

where e_i^* equals the number of K_j for $j = 1, \ldots, i-1$ equal to K_i. Thus, for example, for the $\text{DP}_N(\alpha H)$ measure,

$$\Pr(\mathbf{K}) = \frac{1}{N} \prod_{i=2}^{n} \frac{\alpha/N + e_i^*}{\alpha + i - 1}.$$

REMARK 3. Although a simple closed form expression for $\Pr(\mathbf{K})$ can sometimes be useful, it is important to note that this is not necessary to fit the Bayesian classification model. As we will see in Section 5, it will be enough that we can compute the expression $E\left[W_j|K_1,\ldots,K_{i-1}\right]$, which is of course readily available from Corollary 1.

4. Posterior Characterization for the Classification Model

The importance of $\mathbf{K}$ in the classification problem is now revealed by the following posterior characterization for (7), which builds on Theorem 1. The result is analogous to the posterior characterizations with respect to the partition structure appearing in Lo, Brunner and Chan (1996), Brunner, Chan, James and Lo (2001) and Ishwaran and James (2003). See also Lo (1984). As before, let $\mathbf{K}^*$ denote the unique set of values of $\mathbf{K}$.

THEOREM 2. *Let $\mathbf{Y} = (Y_1,\ldots,Y_n)$ be an i.i.d. sample from an extended stick-breaking prior as in Theorem 1. Then the conditional distribution of $\mathbf{Y}$ given $\mathbf{K}$ and $\mathbf{X}$ from (7) is such that the sequence $Y_1,\ldots,Y_n$ consists of $m = n(\mathbf{K})$ unique values $Y_1^*,\ldots,Y_m^*$, where $Y_j^* = Z_{K_j^*}$, such that each value Y_j^* is conditionally independent, with conditional distribution determined from*

$$\prod_{j=1}^{m}\pi(dY_j^*|\mathbf{K},\mathbf{X}) = \prod_{j\in\mathbf{K}^*}\pi(dZ_j|\mathbf{K},\mathbf{X}) = \frac{\prod_{j\in\mathbf{K}^*} H_j(dZ_j) \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)}{\prod_{j\in\mathbf{K}^*}\int_{\mathcal{Y}} H_j(dZ_j) \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)}. \tag{11}$$

Moreover, the posterior for $\mathbf{K}$ given $\mathbf{X}$ equals

$$\pi(\mathbf{K}|\mathbf{X}) = \frac{\Pr(\mathbf{K})\prod_{j\in\mathbf{K}^*}\int_{\mathcal{Y}} H_j(dZ_j)\prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)}{\sum_{\mathbf{K}}\Pr(\mathbf{K})\prod_{j\in\mathbf{K}^*}\int_{\mathcal{Y}} H_j(dZ_j)\prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)}.$$

Furthermore, the posterior for $\mathcal{P}$, the law of the extended stick-breaking prior for P, is characterized by

$$\begin{aligned}\pi(dP|\mathbf{X}) &= \sum_{\mathbf{K}}\int_{\mathcal{Y}^n}\mathcal{P}(dP|\mathbf{Y},\mathbf{K},\mathbf{X})\,\pi(d\mathbf{Y}|\mathbf{K},\mathbf{X})\,\pi(\mathbf{K}|\mathbf{X})\\ &= \sum_{\mathbf{K}}\int_{\mathcal{Y}^m}\mathcal{P}(dP|Y_1^*,\ldots,Y_m^*,\mathbf{K})\prod_{j=1}^{m}\pi(dY_j^*|\mathbf{K},\mathbf{X})\,\pi(\mathbf{K}|\mathbf{X}),\end{aligned}$$

where $\mathcal{P}(dP|Y_1^*,\ldots,Y_m^*,\mathbf{K})$ is defined by (8).

Theorem 2 reveals the importance of $\mathbf{K}$ in the classification problem. The values K_i not only tell us the assignment of observations X_i to their models, but additionally how to update the selected models given the data. That is, given $\mathbf{K}$ and $\mathbf{X}$, the updated value for Y in a model k_{j^*}, where $j^* = K_j^*$, is interpreted as the posterior distribution of Y given the observations $\{X_i : K_i = j^*\}$ where *a priori* $(X_i|Y)$ are i.i.d. $k_{j^*}(X_i|Y)$ and Y has prior H_{j^*}.

PROOF OF THEOREM 2. By Bayes rule and Theorem 1,

$$\pi(dP|\mathbf{X}) = \sum_{\mathbf{K}} \int_{\mathcal{Y}^m} \mathcal{P}(dP|Y_1^*,\dots,Y_m^*,\mathbf{K})\,\pi(d\mathbf{Y}|\mathbf{K},\mathbf{X})\,\pi(\mathbf{K}|\mathbf{X})$$

where $\mathcal{P}(dP|Y_1^*,\dots,Y_m^*,\mathbf{K})$ is defined by (8). To work out $\pi(d\mathbf{Y}|\mathbf{K},\mathbf{X})$, note that the joint distribution for $(\mathbf{Y},\mathbf{K},\mathbf{X})$ equals

$$\begin{aligned} &\Pr(\mathbf{K})\,\pi(d\mathbf{Y}|\mathbf{K})\,f(\mathbf{X}|\mathbf{K},\mathbf{Y}) \\ &\quad = \Pr(\mathbf{K}) \times \pi(dZ_{K_1},\dots,dZ_{K_n}|\mathbf{K}) \times \left(\prod_{j\in\mathbf{K}^*} \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j) \right) \\ &\quad = \Pr(\mathbf{K}) \prod_{j\in\mathbf{K}^*} H_j(dZ_j) \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j), \qquad (12) \end{aligned}$$

where we have used the fact that $Y_i = Z_{K_i}$. Thus,

$$\pi(d\mathbf{Y}|\mathbf{K},\mathbf{X}) = \frac{\prod_{j\in\mathbf{K}^*} H_j(dZ_j) \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)}{\prod_{j\in\mathbf{K}^*} \int_{\mathcal{Y}} H_j(dZ_j) \prod_{\{i:\,K_i=j\}} k_j(X_i|Z_j)},$$

from which the conditional independence for Y_j^* also follows. Now to work out $\pi(\mathbf{K}|\mathbf{X})$, by Bayes rule we have

$$\pi(\mathbf{K}|\mathbf{X}) = \frac{\Pr(\mathbf{K}) f(\mathbf{X}|\mathbf{K})}{\sum_{\mathbf{K}} \Pr(\mathbf{K}) f(\mathbf{X}|\mathbf{K})},$$

where $\Pr(\mathbf{K}) f(\mathbf{X}|\mathbf{K})$ is determined by integrating (12) with respect to $\mathbf{Y}$. □

5. Monte Carlo Algorithms

We now outline two different Monte Carlo procedures for drawing posterior values for $\mathbf{K}$ from the Bayesian classification model (7), these being: (i) a new collapsed Gibbs sampler and (ii) an SIS based technique. We begin in Section 5.1 with the proposed collapsed Gibbs sampler. The use of the term "collapsing" here refers to the fact that the technique involves Gibbs sampling classification variables $\mathbf{K}$ subject to an extended stick-breaking prior, and thus is a form of collapsing of the stick-breaking prior. Seen this way, the method can also be viewed as a collapsed version of the "blocked Gibbs sampler" of Ishwaran and James (2001). See MacEachern (1994) for a collapsed Gibbs sampler for Dirichlet process mixture models. As an almost immediate consequence of the Markov transitions used in our Gibbs algorithm we are able to readily identify an SIS procedure for drawing importance values for $\mathbf{K}$ (see Remark 4 of Section 5.1). This procedure is an extension of the SIS algorithm presented in Ishwaran, James and Lo (2001) for mixture models subject to finite dimensional stick-breaking priors. For a general discussion of SIS techniques see Kong, Liu and Wong (1994). Section 5.2 ends by outlining the use of our algorithms in infinite complexity mixture models.

5.1. Algorithms. We begin by describing the collapsed Gibbs sampler. Let $\mathbf{K}_{-i}$ denote the $(i-1)$-dimensional classification vector formed by removing the i-th coordinate K_i from $\mathbf{K}$. The collapsed Gibbs sampler works by cycling through draws $\pi(\mathbf{K}|\mathbf{K}_{-i},\mathbf{X})$ for $i=1,\ldots,n$. Note that the i-th conditional draw for $\mathbf{K}$ involves only updating the i-th coordinate K_i. Repeating this n-cycle draw many times produces a Markov chain with stationary distribution $\pi(\mathbf{K}|\mathbf{X})$ and our posterior draw for $\mathbf{K}$.

Theorem 2 can be used to work out the conditional draw for $\mathbf{K}$ from $\pi(\mathbf{K}|\mathbf{K}_{-i},\mathbf{X})$. Let $\mathbf{K}^*_{-i}$ denote the unique set of values for $\mathbf{K}_{-i}$. To draw $\mathbf{K}$ we update the value for its i-th coordinate K_i as follows. Assuming $(\mathbf{K}^*_{-i})^c \neq \emptyset$, we choose a new value $K_i = j_c \in (\mathbf{K}^*_{-i})^c$ with probability

$$\frac{E[W_{j_c}|\mathbf{K}_{-i}]}{\lambda(i)} \times \int_{\mathcal{Y}} k_{j_c}(X_i|Z)\,H_{j_c}(dZ), \tag{13}$$

or we select a value already used, $K_i = j \in \mathbf{K}^*_{-i}$, with probability

$$\frac{E[W_j|\mathbf{K}_{-i}]}{\lambda(i)} \times \int_{\mathcal{Y}} k_j(X_i|Z_j)\,\pi(dZ_j|\mathbf{K}_{-i},\mathbf{X}), \tag{14}$$

where $\lambda(i)$ is the appropriate normalizing constant and

$$\pi(dZ_j|\mathbf{K}_{-i},\mathbf{X}) = \frac{H_j(dZ_j)\prod_{\{l\neq i:\,K_l=j\}} k_j(X_l|Z_j)}{\int_{\mathcal{Y}} H_j(dZ_j)\prod_{\{l\neq i:\,K_l=j\}} k_j(X_l|Z_j)}.$$

Remark 4. Observe that to implement the draws from (13) and (14) we need to compute $E[W_j|\mathbf{K}_{-i}]$ for $j = 1,\ldots,N$, but this follows straightforwardly from Corollary 1. For example, consider the case $i = n$ without loss of generality (by exchangeability the values $E[W_j|\mathbf{K}_{-i}]$ have the same form for each $i = 1,\ldots,n$). Replacing n with $n-1$ in Corollary 1 now shows how to readily compute $E[W_j|\mathbf{K}_{-n}]$.

Remark 5. The Markov transitions worked out in our collapsed Gibbs sampler can be used to extend the SIS algorithm of Ishwaran, James and Lo (2001) for finite mixture models to the setting here of unequal kernels and general stick-breaking priors. Briefly the method works as follows. To generate an importance draw for $\mathbf{K}$ we build up a sequence of classification vectors $\mathbf{K}_1, \mathbf{K}_2\ldots,\mathbf{K}_n$, where the eventual draw for $\mathbf{K}$ is $\mathbf{K}_n$. To start, assign label j to $\mathbf{K}_1$ with probability $E(W_j)$. Next, we build up $\mathbf{K}_{r+1}$ sequentially for $r = 2,\ldots,n-1$, where the generation of $\mathbf{K}_{r+1} = (\mathbf{K}_r, j)$ is based on the current value $\mathbf{K}_r$ using a posterior update rule similar to (13) and (14). The simplest way to define the update rule is by considering the case $r = n-1$ (with obvious modifications for general r). In this case, $\mathbf{K}_n = (\mathbf{K}_{n-1}, j)$, where $j \in (\mathbf{K}^*_{-n})^c$ is selected with probability (13) or $j \in \mathbf{K}^*_{-n}$ is selected with probability (14).

5.2. Infinite complexity mixture models. Several simplifications occur in the finite mixture model (6) because kernels are equal (i.e. $k_j = k_0$) and the variables Z_j used in our priors are identically distributed (i.e. $H_j = H$). This points the way to a novel estimation procedure for mixture models with infinite complexity $N = \infty$. We describe this new method in the context of our collapsed Gibbs algorithm, but the method outlined can be modified in an obvious way for the SIS algorithm.

First note that in the finite mixture setting, the normalizing constant equals

$$\begin{aligned}\lambda(i) &= \Big[1 - \sum_{j\in\mathbf{K}^*_{-i}} E[W_j|\mathbf{K}_{-i}]\Big]\int_{\mathcal{Y}} k_0(X_i|Z)\,H(dZ)\\ &\qquad + \sum_{j\in\mathbf{K}^*_{-i}} E[W_j|\mathbf{K}_{-i}]\int_{\mathcal{Y}} k_0(X_i|Z_j)\,\pi(dZ_j|\mathbf{K}_{-i},\mathbf{X}),\end{aligned}$$

which can be computed even when $N = \infty$ since it involves a finite number of terms. Now all that remains is to draw (13) and (14). The draw from (14) is straightforward since it involves only a finite number of terms. The draw (13) is also possible since it essentially requires simulating a discrete variable from the countable discrete distribution

$$\Pi_i(\cdot) = \sum_{j \in (\mathbf{K}^*_{-i})^c} \Gamma_j \delta_j(\cdot), \qquad \text{where } \Gamma_j = \frac{E[W_j|\mathbf{K}_{-i}]}{1 - \sum_{j \in \mathbf{K}^*_{-i}} E[W_j|\mathbf{K}_{-i}]}.$$

One method for drawing from Π_i uses rejection sampling. Order the indices in $(\mathbf{K}^*_{-i})^c$ as $i_1 < i_2 < \cdots$ and draw U from a Uniform $[0,1]$ distribution. A draw J from Π_i is the value $J = i_{j_0}$ where j_0 is such that

$$\sum_{j=1}^{j_0 - 1} \Gamma_{i_j} \le U < \sum_{j=1}^{j_0} \Gamma_{i_j}.$$

Observe that this requires only a finite number of computations. See Doss (1994, page 1768) for a similar technique in survival analysis problems subject to Dirichlet process priors.

REMARK 6. The draws for $\mathbf{K}$ can be used to estimate the hidden variables Y_i and the unknown mixing distribution Q_0 of (2). For example, to estimate the posterior mean of a function $t(\mathbf{Y})$ note that by the double-expectation rule

$$E[t(\mathbf{Y})|\mathbf{X}] = E\{E[t(\mathbf{Y})|\mathbf{K}, \mathbf{X}]\}.$$

So to estimate $E[t(\mathbf{Y})|\mathbf{X}]$, draw $\mathbf{K}$ from the Gibbs sampler, followed by a draw for the unique values $Y_1^*, \dots, Y_m^*$ from (11) in Theorem 2, where $m = n(\mathbf{K})$. This gives us a draw for $\mathbf{Y}$ from $\pi(\mathbf{Y}|\mathbf{K}, \mathbf{X})$. Given B such draws, say, $\mathbf{K}^{(1)}, \dots, \mathbf{K}^{(B)}$, and corresponding values $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(B)}$, estimate $E[t(\mathbf{Y})|\mathbf{X}]$ by $\sum_{b=1}^{B} t(\mathbf{Y}^{(b)})/B$.

To estimate Q_0 we draw from the law of $\mathcal{P}(dP|Y_1^*, \dots, Y_m^*, \mathbf{K})$ described in Theorem 1. For a given value for $\mathbf{K}$ and $Y_1^*, \dots, Y_m^*$, this can be computed conveniently by the random measure P^* defined in (9). Thus, to approximate the law for a functional $t(Q_0) = \int g(Y) Q_0(dY)$, use $\sum_{b=1}^{B} I\{t(P^{*(b)}) \in \cdot\}/B$, where for a given $\mathbf{K}$ and $Y_1^*, \dots, Y_m^*$,

$$t(P^*) = \sum_{k \in \mathbf{K}^*} W_k\, g(Y_k^*) + \sum_{k \in \mathcal{M} - \mathbf{K}^*} W_k\, g(Z_k) + \Big[1 - \sum_{k=1}^{\ell} W_k\Big] P_\ell(g),$$

where W_k are the random weights with law $\pi(\mathbf{W}|\mathbf{K})$, and P_ℓ is the independent random measure defined by (10) where ℓ is the largest component of $\mathbf{K}$. In practice, P_ℓ will need to be approximated, but this can be done easily enough using a truncation argument.

References

Binder, D.A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31-38.

Brunner, L.J., Chan, A.T., James, L.F. and Lo, A.Y. (2001). Weighted Chinese restaurant processes and Bayesian mixture models. Manuscript.

Brunner, L.J. and Lo, A.Y. (1999). Bayesian classification. Research Report, Dept. of Statistics, University of Toronto.

Chen, R. and Liu, J.S. (1996). Predictive updating methods with application to Bayesian classification. *J.R. Statist. Soc., Ser. B* **58**, 397-415.

Connor, R.J. and Mosimann, J.E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **64**, 194-206.

Donnelly, P. and Joyce, P. (1989). Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Process. Appl.* **31**, 89-103.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22**, 1763-1786.

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.

Freedman, D.A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Statist.* **34**, 1386-1403.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96**, 194-209.

Halmos, P. (1944). Random alms. *Ann. Math. Stat.* **15**, 182-189.

Hartigan, J.A. (1975). *Clustering Algorithms.* Wiley, New York.

Hoppe, F.M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biology* **25**, 123-159.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161-173.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture model. *Statistica Sinica* **13**, 1211-1235.

Ishwaran H., James, L.F. and Lo, A.Y. (2001). Generalized weighted Chinese restaurant and SIS stick-breaking algorithms for semiparametric models. Manuscript.

Ishwaran H., James, L.F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96**, 1316-1332.

Ishwaran, H and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371-390.

Ishwaran, H and Zarepour, M. (2002a). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12**, 941-963.

Ishwaran, H. and Zarepour, M. (2002b). Exact and approximate sum-representations for the Dirichlet process. *Canad. J. Statist.* **30**, 269-283.

Kong, A., Liu, J.S. and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89**, 278-288.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canad. J. Statist.* **20**, 451-461.

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

Lo, A.Y., Brunner, L.J. and Chan, A.T. (1996). Weighted Chinese restaurant processes and Bayesian mixture models. Research Report 1, Hong Kong University of Science and Technology.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727-741.

MacEachern, S.N., Clyde, M. and Liu, J.S. (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canad. J. Statist.* **27**, 251-267.

McCloskey, J.W. (1965). A model for the distribution of individuals by species in an environment. Unpublished Ph.D. thesis, Michigan State University.

Patil, G.P. and Taillie C. (1977). Diversity as a concept and its implications for random communities. *Bull. Int. Stat. Inst.* **XLVII**, 497-515.

Perman, M., Pitman, J. and Yor M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21-39.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102**, 145-158.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, T.S. Ferguson, L.S. Shapley and J.B. MacQueen, eds., 245-267. IMS Lecture Notes-Monograph series, Vol 30.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855-900.

Quintana, F.A. and Iglesia, P.L. (2003). Nonparametric Bayesian clustering and product partition models. *J.R. Statist. Soc., Ser. B* **65**, 557-574.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-650.

Sethuraman, J. and Tiwari, R.C. (1982). Convergence of Dirichlet measures and the interpretation of their parameters. *Statistical Decision Theory and Related Topics III* **2**, 305-315.

HEMANT ISHWARAN
CLEVELAND CLINIC FOUNDATION
DEPT. OF BIOSTATISTICS / WB4
9500 EUCLID AVENUE
CLEVELAND, USA
E-mail: ishwaran@bio.ri.ccf.org

LANCELOT F. JAMES
HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPT. OF INFORMATION SYSTEMS AND MANAGEMENT
CLEAR WATER BAY
KOWLOON, HONG KONG
E-mail: lancelot@ust.hk