Some Theory and Applications of Stick-Breaking Priors to Bayes Nonparametrics

Hemant Ishwaran
Cleveland Clinic Foundation, Department of Biostatistics
9500 Euclid Avenue, OH 44195
Cleveland, USA
ishwaran@bio.ri.ccf.org

1. Introduction

Ishwaran and James (2001) recently introduced a rich class of random probability measures, which they coined stick-breaking priors, and discussed how the simple constructive form of such measures make them readily applicable as well as computationally feasible as priors in Bayesian non- and semiparametric problems. Call $\mathcal{P}$ a *stick-breaking prior* if it can be expressed in the form

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

where Z_k are iid values with a non-atomic distribution H over a measurable Polish space $(\mathcal{Y}, \mathcal{B})$ 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". Specifically,

$$W_1 = V_1, \qquad W_k = (1-V_1)(1-V_2)\cdots(1-V_{k-1})V_k, \quad k = 2, 3, \ldots, \tag{2}$$

where V_k are independent beta(a_k, b_k) random variables. Stick-breaking priors can be finite dimensional, corresponding to the case when $N < \infty$, or they can be infinite dimensional in the case when $N = \infty$. The properties and computational approaches for these two sub-classes are quite different and it is worth taking the time to clearly delineate between them.

In the finite dimensional case, the random weights $\mathbf{W} = (W_1, \ldots, W_N)$ are constructed as in (**??**) but with $V_N = 1$ to ensure that $W_1 + \cdots + W_N = 1$. This follows since the stick-breaking construction (**??**) implies that

$$1 - \sum_{k=1}^{K} W_k = (1-V_1)(1-V_2)\cdots(1-V_{K-1}), \quad \text{for each } K = 2, \ldots, N.$$

An important feature in the finite case is that the law for $\mathbf{W}$ is a generalized Dirichlet distribution [Connor and Mosimann (1969)], written as $\mathbf{W} \sim \mathcal{GD}(\mathbf{a}, \mathbf{b})$, where $\mathbf{a} = (a_1, \ldots, a_{N-1})$ and $\mathbf{b} = (b_1, \ldots, b_{N-1})$. This is important for the computational algorithms to be discussed which will exploit the key fact that the generalized Dirichlet distribution is conjugate to the multinomial. An important class of finite dimensional stick-breaking priors are the class of

finite dimensional Dirichlet priors [Ishwaran and Zarepour (2000)], which are defined using a weight vector $\mathbf{W}$ having a Dirichlet$(\alpha_1, \ldots, \alpha_N)$ law.

Infinite dimensional stick-breaking measures are defined under $\sum_{k=1}^{\infty} E[\log(1-V_k)] = -\infty$, which is necessary and sufficient condition for the random weights to sum to one [Ishwaran and James (2001)]. An important class of such measures are the class of two-parameter Poisson-Dirichlet processes developed by Pitman and Yor (1997). This class corresponds to measures (**??**) with random weight vector $\mathbf{W} = (W_1, W_2, \ldots)$ constructed using beta(a_k, b_k) random variables where

$$a_k = 1 - a, \qquad b_k = b + ka, \qquad \text{where } 0 \leq a < 1 \text{ and } b > -a.$$

See Pitman (1995, 1996) and Pitman and Yor (1997). A well known example of the two-parameter Poisson-Dirichlet process is the Ferguson Dirichlet process (1973). The Dirichlet process with finite measure parameter αH, written as DP(αH), corresponds to the selection of parameters with $a = 0$ and $b = \alpha > 0$.

2. Hierarchical Models and their Posterior Characterizations

A rich application of the stick-breaking prior is in analysis of Bayesian semiparametric models. These are models in which the data $\mathbf{X} = (X_1, \ldots, X_n)$ is derived from a hierarchical model of the form

$$\begin{aligned} (X_i|Y_i, \theta) &\overset{\text{ind}}{\sim} k(X_i|Y_i, \theta), \quad i = 1, \ldots, n \\ (Y_i|P) &\overset{\text{iid}}{\sim} P, \qquad P \sim \mathcal{P} \\ \theta &\sim \pi(d\theta), \end{aligned} \tag{3}$$

where $k(X_i|Y_i, \theta)$ is a kernel density in X_i for given values for Y_i and θ, with $Y_i \in \mathcal{Y}$ drawn from P, a random probability measure with a stick-breaking prior $\mathcal{P}$, and θ is a hyperparameter drawn from $\pi(\theta)$, a prior over a finite dimensional space Θ (usually Θ is a subset of $\Re^q$).

A key to analysis and computational algorithms for (**??**) is to rewrite the model as

$$\begin{aligned} (X_i|\mathbf{Z}, \mathbf{K}, \theta) &\overset{\text{ind}}{\sim} k(X_i|Z_{K_i}, \theta), \quad i = 1, \ldots, n \\ (K_i|\mathbf{W}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} W_k \, \delta_k(\cdot), \quad 1 \leq N \leq \infty \\ \theta &\sim \pi(\theta), \end{aligned} \tag{4}$$

where we use the notation $\mathbf{Z}$ for the vector of Z_k variables (which could be either finite or infinite dimensional depending on N) and $\mathbf{K} = (K_1, \ldots, K_n)$. Note that the key identity to (**??**) is $Y_i = Z_{K_i}$.

The following theorem characterizes the posterior in semiparametric models subject to stick-breaking priors. Its proof follows from arguments in Ishwaran and James (2001) and the

use of disintegrations for joint product measures which are known to always exist on Polish spaces [see Le Cam (1986)].

Theorem 1. *For both the finite and infinite dimensional cases where $N < \infty$ and $N = \infty$, the posterior for P from (??) is characterized by*

$$\mathcal{P}(dP|\mathbf{X}) = \sum_{\mathbf{K}\in\mathcal{K}} \iiint \mathcal{P}(dP|\mathbf{W},\mathbf{Z})\,\pi(d\mathbf{W}|\mathbf{K})\,\pi(d\mathbf{Z}|\mathbf{K},\theta,\mathbf{X})\,\pi(d\theta|\mathbf{K},\mathbf{X})\,\Pr(\mathbf{K}|\mathbf{X}),$$

where the sum is over all $\mathbf{K}$ *in* $\mathcal{K} = \{1,\ldots,N\}^n$,

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

and

$$f(\mathbf{X}|\mathbf{K}) = \int_\Theta \pi(d\theta)\left[\prod_{j\in\mathbf{K}^*}\int_{\mathcal{Y}} H(dZ)\prod_{\{i:K_i=j\}} k(X_i|Z,\theta)\right],$$

where $\mathbf{K}^* = \{K_1^*,\ldots,K_m^*\}$ *denotes the unique set of* K_i *values and* $\Pr(\mathbf{K})$ *is the prior for* $\mathbf{K}$ *defined by*

$$\Pr(K_1 = k_1,\ldots,K_n = k_n) = E\left(\prod_{i=1}^n W_{k_i}\right).$$

Although Theorem 1 may appear too formidable to be applicable to computational algorithms, in fact, the posterior characterization points to several powerful Monte Carlo methods for fitting semiparametric models. In each of these cases, the key is to recognize that the above decomposition becomes simple to work with once we have a method for drawing posterior values for $\mathbf{K}$, as once we know $\mathbf{K}$ we know how the data decomposes into a group of finite parametric models. For example, the law for $\pi(d\mathbf{Z}|\mathbf{K},\theta,\mathbf{X})$ is the joint distribution defined where Z_k are iid H for $k \in \mathcal{K} - \mathbf{K}^*$ and where Z_j, for $j \in \mathbf{K}^*$, are independent with law

$$\pi(dZ_j|\mathbf{K},\theta,\mathbf{X}) = \frac{H(dZ_j)\prod_{\{i:K_i=j\}} k(X_i|Z_j,\theta)}{\int_{\mathcal{Y}} H(dZ_j)\prod_{\{i:K_i=j\}} k(X_i|Z_j,\theta)}. \tag{5}$$

Of course, trying to actually draw all the Z_k variables for $k \in \mathcal{K} - \mathbf{K}^*$ in the infinite dimensional case is not computationally feasible, and moreover it is not clear how one could in general obtain the law for $\pi(\mathbf{W}|\mathbf{K})$ in such cases. However, these are not limitations in the case where $N < \infty$. The special conjugacy feature of the generalized Dirichlet distribution tells us immediately that $(\mathbf{W}|\mathbf{K}) \sim \mathcal{GD}(\mathbf{a}^*,\mathbf{b}^*)$, where $a_k^* = a_k + e_k$, $b_k^* = b_k + \sum_{l=k+1}^N e_l$ and e_k is the number of K_i values equal to k.

The conjugacy of the generalized Dirichlet distribution, along with the key idea that knowing $\mathbf{K}$ effectively decomposes the semiparametric model into a finite number of parametric models, was exploited in Ishwaran, James and Lo (2001) in a sequential importance sampling (SIS) technique for approximating posterior laws for functionals of P. The same idea is also

implicitly used in Ishwaran and James (2001) to develop a Gibbs sampling technique, called the *blocked Gibbs sampler* for drawing posterior values for P directly. We give a brief outline of the procedure now.

To draw P, iteratively draw values from the conditional distributions of the following blocked variables:

$$(\mathbf{K}|\mathbf{Z},\mathbf{W},\theta,\mathbf{X}), \quad (\theta|\mathbf{Z},\mathbf{K},\mathbf{X}), \quad (\mathbf{Z}|\mathbf{K},\theta,\mathbf{X}), \quad (\mathbf{W}|\mathbf{K}).$$

Doing so eventually produces values drawn from the distribution of $(\mathbf{K},\theta,\mathbf{Z},\mathbf{W}|\mathbf{X})$. Thus, each draw $(\mathbf{K}^*,\theta^*,\mathbf{Z}^*,\mathbf{W}^*)$ defines a random probability measure $P^*(\cdot)=\sum_{k=1}^N W_k^*\,\delta_{Z_k^*}(\cdot)$, which (eventually) gives us the draw from the posterior $\mathcal{P}(\cdot|\mathbf{X})$ that we are after. Note carefully that each of the above draws are based on usual parametric techniques (see for example (**??**)) and at least in the cases of $\mathbf{W}$ and $\mathbf{K}$ are drawn from easy to simulate multivariate distributions.

REFERENCES

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.

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

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

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

Ishwaran, H. and Zarepour, M. (2000). Dirichlet prior sieves in finite normal mixtures. Conditionally accepted by *Statistica Sinica.*

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer-Verlag.*

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Prob. 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. Prob.*, **25**, 855-900.

RESUME

The class of stick-breaking priors is defined and their application to semiparametric models is considered. The characterization of the resulting posterior is given and points to the underlying technique used in various Monte Carlo methods, including SIS and blocked Gibbs sampling, for approximating posterior quantities.