Chapter 7

Species Sampling Models

7.1. Introduction

One of the reasons for the widespread popularity of the DP prior model is the computational simplicity of posterior simulation and posterior predictive inference. This simplicity is in part due to the almost sure discrete nature of a random probability measure G with DP prior.

Recall from the discussion in §2.1 that the discrete nature of G naturally induces a prior on random partitions, in the following sense. Consider a random sample, $x_i \mid G \sim G$, $i = 1, \ldots, n$, with $G \sim \mathsf{DP}(M, G_0)$. The discreteness of G implies a positive probability of ties among the x_i. We can use these ties to partition experimental units into clusters defined by the unique values. Let $x_j^\star$, $j = 1, \ldots, k$, define the $k \leq n$ unique values among the x_i and define clusters $S_j = \{i : x_i = x_j^\star\}$. Also, let $n_j = |S_j|$ denote the size of the j-th cluster and let $\boldsymbol{n}_n = (n_{1n}, \ldots, n_{kn})$. When the number n of experimental units is understood from the context we drop the index $_n$.

Finally, recall posterior predictive inference under i.i.d. sampling from a DP random measure:

$$p(x_{n+1} \mid x_1, \ldots, x_n) = \begin{cases} \delta_{x_j^\star}(x_{n+1}) & \text{w. prob } \frac{n_j}{n+\alpha} \equiv p_j(\boldsymbol{n}) \\ G_0(x_{n+1}) & \text{w. prob } \frac{\alpha}{n+\alpha} \equiv p_{k+1}(\boldsymbol{n}). \end{cases} \tag{7.1}$$

In anticipation of the upcoming discussion, the posterior predictive distribution can also be written as

$$x_{n+1} \mid x_1, \ldots, x_n \sim \sum_{j=1}^{k} p_j(\boldsymbol{n}) \delta_{x_j^\star} + p_{k+1}(\boldsymbol{n}), \tag{7.2}$$

with weights $p_j(\boldsymbol{n})$. This is known as the predictive probability function (PPF). The first important feature to see in (7.1) is what is not seen. The posterior predictive is integrated with respect to the random probability measure G. This is important for computation. It would be impossible to keep an infinite dimensional quantity G in computer memory. In other words, the Pólya urn (7.1) characterizes the marginal distribution $p(x_1, \ldots, x_n)$, simply by multiplying the posterior predictive $p(x_{i+1} \mid x_1, \ldots, x_i)$ for $i = 1, \ldots, n-1$ and the marginal $p(x_1) = G_0(x_1)$. The second important feature to note is that the weights $p_j(\cdot)$ of the clusters are a function of only the cluster size n_j. Neither the actual observations $x_i \in S_j$, nor the number and sizes of other clusters enter the expression.

7.2. Predictive Probability Functions

We introduced the PPF in (7.2) as the predictive rule that is implied by i.i.d. sampling from a probability measure with DP prior. However, also the opposite is

true; the PPF characterizes the DP prior and is a defining property of the process. This is not immediately obvious; to formally state this we introduce the notion of a species sampling sequence.

An *exchangeable* sequence of random variables $x_1, x_2, \ldots$, is called a species sampling sequence (SSS) if

$$x_{n+1} \mid x_1, \ldots, x_n \sim \sum_{j=1}^{k_n} p_j(\boldsymbol{n}_n)\delta_{x_j^\star} + p_{k_n+1}(\boldsymbol{n}_n)G_0,$$

with weights depending on the data only indirectly through the cluster sizes n_j.

The family of functions $\{p_j(\boldsymbol{n}_n)\}$ is called the predictive probability function (PPF). As the weights in the posterior predictive distribution any PPF needs to satisfy

$$p_j(\boldsymbol{n}_n) \geq 0, \qquad \sum_{j=1}^{k_n+1} p_j(\boldsymbol{n}_n) = 1, \tag{7.3}$$

for all $\boldsymbol{n}_n$. In the definition of the SSS and PPF, the restriction to exchangeable sequences $(x_n)_{n\geq 1}$ is important. Not every family of functions $\{p_j(\mathbf{n})\}$ that satisfies (7.3) is a PPF. In fact, most such families are not.

The expression for the weights $p_j(\cdot)$ in the Pólya urn for the DP appear particularly simple with $p_j \propto n_j$. It can be shown (Gnedin and Pitman, 2006; Lee *et al.*, 2008) that any PPF with weights that are functions of the cluster size only must be of essentially that form. More specifically, if $p_j = f(n_j)$ for some function of the cluster size, $j = 1, \ldots, k_n$ and $p_{k_n+1} = \theta$, then $f(n_j) = an_j$ for some $a > 0$. Actually, for finite exchangeable sequences $x_1, x_2, \ldots, x_{n+1}$ of categorical random variables, i.e., possible outcomes $x_i \in \{1, \ldots, t\}$, the same result is known as Johnson's sufficientness postulate (Zabell, 1982).

We are now ready to state how the predictive rule of a SSS characterizes a random probability measure. An exchangeable sequence of random variables (x_n) is a SSS if and only if $x_i \sim G$, i.i.d., for some random distribution G that admits a representation of the form

$$G = \sum_{h=1}^{\infty} p_h \delta_{m_h} + RG_0,$$

with $m_h \sim G_0$, i.i.d., and some sequence of positive random weights p_h such that $\sum_{h=1}^{\infty} p_h \leq 1$ (Pitman, 1996, Proposition 11). The random probability measure G is called the species sampling model (SSM) of the SSS (x_n).

A SSM can be defined directly by specifying a prior for the weights p_h, and a distribution for the point masses m_h. The only constraint is the positivity of the p_h and the constraint on the sum of the weights. Alternatively an SSM can be (implicitly) defined through its PPF. The characterization is very useful for computational purposes, but of little use to construct an SSM, because of the difficult constraint that the implied sequence x_n be exchangeable.

A third characterization of an SSM is through the implied prior on the sequence of random partitions. A sequence of discrete random variables (x_n) defines a partition of $\{1, \ldots, n\}$ into clusters $S_j = \{i : \ x_i = x_j^\star\}$ of tied observations. Thus the SSM indirectly defines a sequence of priors for partitions. As before, let $\boldsymbol{n}_n = (n_1, \ldots, n_{k_n})$ denote the cluster sizes of the partition of $\{1, \ldots, n\}$. Since the

sequence x_n is exchangeable it suffices to specify the probability of $\boldsymbol{n}_n$. The probability for any two partitions with the same cluster sizes $\boldsymbol{n}_n$ must be the same. The implied prior $p(\boldsymbol{n}_n)$ is known as the exchangeable partition probability function (EPPF). Let $\mathsf{N}^\star = \cup_{k=1}^{\infty} \mathsf{N}^k$ and let $\boldsymbol{n}^{j+}$ denote $\boldsymbol{n}$ with the the j-th cluster size incremented by 1. Formally an EPPF $p(\cdot)$ is a symmetric function $p : \mathsf{N}^\star \to [0, 1]$ with

$$p(\boldsymbol{n}) = \sum_{j=1}^{k_n+1} p(\boldsymbol{n}^{j+}) \text{ for all } \boldsymbol{n} \in \mathsf{N}^\star \tag{7.4}$$

and $p(1) = 1$. The condition simply formalizes coherence across sample sizes. The probability of partitions for the first n elements of a SSS must match the appropriate marginal of the probabilities for partitions of the first $n + 1$ elements.

The converse is also true. For any function that could be interpreted as an EPPF, i.e., that satisfies the above condition, there is a SSS that gives rise to it (Pitman, 1996, Proposition 13). Again, similar to the characterization of a SSM by PPF, the definition through the EPPF is of little practical use. It is difficult to directly elicit and specify a legitimate EPPF that satisfies (7.4).

Finally, there is an obvious link between the EPPF and the PPF. Every EPPF defines a PPF through

$$p_j(\boldsymbol{n}) \equiv p(\boldsymbol{n}^{j+})/p(\boldsymbol{n}).$$

7.3. More SSMs

We used the DP prior to introduce the notion of the PPF. Some other examples of SSMs are the Pitman Yor (PY) process, the normalized inverse Gaussian (NIG) and Gibbs type priors.

Pitman-Yor Process

The PY process (Pitman, 1995; Pitman and Yor, 1997) is more easily introduced as a stick breaking prior. A random probability measure $G = \sum_h w_h \delta_{\theta_h}$ has a $\mathsf{PY}(\sigma, \alpha, G_0)$ prior if $w_h = \prod_{\ell<h}(1-v_\ell)v_h$ for $v_h \sim \mathsf{Be}(1-\sigma, \alpha+h\sigma)$, independently with $0 \le \sigma < 1$ and $\alpha > -\sigma$, and the locations θ_h are a random sample from the base measure, $\theta_h \sim G_0$. See Ishwaran and James (2001) for a discussion of this construction and a larger class of random probability measures defined by similar stick breaking algorithms. The PPF implied by the PY process is simply

$$x_{n+1} \mid x_1, \ldots, x_n \sim \sum_{j=1}^{k_n} \frac{n_j - \sigma}{n + \theta} \delta_{x_j^\star}(x_{n+1}) + \frac{\theta + k_n\sigma}{n + \theta} G_0(x_{n+1}),$$

while the EPPF reduces to

$$p(\boldsymbol{n}) = \frac{\Gamma(\theta + 1)}{(\theta + k_n\sigma)\Gamma(\theta + n)} \prod_{j=1}^{k_n} \left\{ (\theta + j\sigma) \frac{\Gamma(n_j - \sigma)}{\Gamma(1 - \sigma)} \right\}.$$

Homogeneous NRMIs

Many more SSMs exist. Any homogeneous NRMI is a SSM. Recall from §1.2.6 the construction of NRMI's as normalized CRM, which in turn can be constructed with

a Poisson process with intensity $\nu(x,s)$ on $X \times \Re^+$. An NRMI is called homogeneous when the intensity factors as $\nu(x,s) = \rho(s)G_0(x)$. While all homogeneous NRMIs define a SSM, most do not allow a closed form expression for the weights in the PPF. The most prominent exception is the DP. Another is the normalized inverse Gaussian process (NIG) that was already briefly introduced in §1.2.6. The NIG is in many ways similar to the DP prior. Recall the characterization of a DP for G as assigning a Dirichlet prior to $(G(A_1), \ldots, G(A_k))$ for any partition $\{A_1, \ldots, A_k\}$ of the sample space. Similarly a NIG prior for random probability measure G can be defined by requiring a normalized inverse Gaussian distribution for $(G(A_1), \ldots, G(A_k))$. See §1.2.6 for a statement of the normalized inverse Gaussian distribution. Like the DP the NIG allows closed form expressions for the PPF. See, for example, Lijoi *et al.* (2007b) (using $\sigma = 1/2$) or Lijoi *et al.* (2005). The NIG is a special case of the more general normalized generalized gamma (NGG) process.

Gibbs Type Priors

Another large class of SSMs are Gibbs type priors (Gnedin and Pitman, 2006). Gibbs type priors can be defined by the EPPF. Let $a_k = a(a+1)\cdot\ldots\cdot(a+k-1)$ define a rising factorial. A Gibbs type prior is a prior for a discrete random probability measure with EPPF

$$p(\boldsymbol{n}) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}.$$

with $\sigma < 1$, $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n - k\sigma) + V_{n+1,k+1}$. The last condition is simply (7.4). The implied weights in the PPF are

$$p_j(\boldsymbol{n}) \propto V_{n+1,k}(n_j - \sigma), \qquad p_{k+1}(\boldsymbol{n}) \propto V_{n+1,k+1}.$$

Lijoi *et al.* (2007a) discuss some results about the predictive distribution under this model.