14

Discrete Random Structures

Random, exchangeable partitions of a finite or countable set arise naturally through the pattern of tied observations when sampling repeatedly from a random discrete measure. They may be used as prior models for clustering, or more generally for sharing of features. The Chinese restaurant process is a distinguished example of a random partition, and arises from the Dirichlet process. After presenting general theory on random partitions, we discuss many other examples of species sampling processes, such as Gibbs processes, and the Poisson-Kingman and Pitman-Yor processes. These random discrete distributions can be viewed as generalizations of the Dirichlet process. Normalized completely random measures are another generalization, with the normalized inverse-Gaussian process as a particularly tractable example. Random discrete distributions with atoms or locations that depend on a covariate can be used as priors for conditional distributions. We discuss some constructions that lead to tractable processes. Finally we discuss a prior distribution on infinite binary matrices, known as the Indian buffet process, as a model for overlapping clusters and for sharing a potentially unlimited number of features.

14.1 Exchangeable Partitions

A *partition* $\{A_1, \ldots, A_k\}$ of the finite set $\mathbb{N}_n = \{1, \ldots, n\}$ is a decomposition of this set in disjoint (nonempty) subsets: $\mathbb{N}_n = \cup_i A_i$ and $A_i \cap A_j = \varnothing$ for $i \neq j$. In this section we discuss probability measures on the collection of all partitions of $\mathbb{N}_n$, where $n \in \mathbb{N}$ may vary. Exchangeable probability measures that can be defined "consistently" across n are of special interest. These probability measures may serve as prior distributions in methods for clustering data.

The cardinalities $n_i = |A_i|$ of the sets in a partition of $\mathbb{N}_n$ are said to form a *partition of* n: an unordered set $\{n_1, \ldots, n_k\}$ of natural numbers such that $n = \sum_{i=1}^k n_i$. The sets A_i in a partition are considered to be unordered, unless specified otherwise, but if the sets are listed in a particular order, then so are their cardinalities. An "ordered partition" $(n_1, \ldots, n_k)$ of n is called a *composition* of n, and the set of all compositions of n is denoted by $\mathcal{C}_n$. The particular order by the sizes of the minimal elements of the sets (so $\min A_1 < \min A_2 < \cdots < \min A_k$) is called the *order of appearance*; thus in this case A_1 contains the element 1, A_2 contains the smallest number from $\mathbb{N}_n$ not in A_1, etc.

A *random partition* of $\mathbb{N}_n$ is a random element defined on some probability space taking values in the set of all partitions of $\mathbb{N}_n$. (Measurability is understood relative to the discrete σ-field on the (finite) set of all partitions.) Its induced distribution is a probability measure on the set of all partitions of $\mathbb{N}_n$. We shall be interested in "exchangeable partitions."

Write $\sigma(A) = \{\sigma(i): i \in A\}$ for the image of a subset $A \subset \mathbb{N}_n$ under a permutation σ of $\mathbb{N}_n$.

Definition 14.1 (Exchangeable partition) A random partition $\mathcal{P}_n$ of $\mathbb{N}_n$ is called *exchangeable* if its distribution is invariant under the action of any permutation $\sigma: \mathbb{N}_n \mapsto \mathbb{N}_n$, i.e. for every partition $\{A_1, \ldots, A_k\}$ of $\mathbb{N}_n$ the probability $\mathrm{P}(\mathcal{P}_n = \{\sigma(A_1), \ldots, \sigma(A_k)\})$ is the same for every permutation σ of $\mathbb{N}_n$. Equivalently, a random partition $\mathcal{P}_n$ of $\mathbb{N}_n$ is *exchangeable* if there exists a symmetric function $p: \mathcal{C}_n \to [0, 1]$ such that, for every partition $\{A_1, \ldots, A_k\}$ of $\mathbb{N}_n$,

$$\mathrm{P}(\mathcal{P}_n = \{A_1, \ldots, A_k\}) = p(|A_1|, \ldots, |A_k|). \tag{14.1}$$

The function p is called the *exchangeable partition probability function* (EPPF) of $\mathcal{P}_n$.

The EPPF is defined as a function on compositions $(n_1, \ldots, n_k)$, but, as it is necessarily symmetric, could have been defined on the corresponding *partitions* $\{n_1, \ldots, n_k\}$. Its value $p(n_1, \ldots, n_k)$ is the probability of a *particular* partition $\{A_1, \ldots, A_k\}$ with $|A_i| = n_i$ for every i, and its number of arguments varies from 1 to n.[1]

Example 14.2 (Product partition model) An EPPF is said to lead to a *product partition model* if the function p in (14.1) factorizes as $p(|A_1|, \ldots, |A_k|) = V_{n,k} \prod_{i=1}^{k} \rho(|A_i|)$. The function $\rho: \mathbb{N} \to [0, 1]$ is then called the *cohesion function* of the product partition model.

Example 14.3 (Ties in exchangeable vector) Given an ordered list $(x_1, \ldots, x_n)$ of n arbitrary elements, the equivalence classes for the equivalence relation $i \sim j$ if and only if $x_i = x_j$ form a partition of $\mathbb{N}_n$. The partition defined in this way by an exchangeable random vector $(X_1, \ldots, X_n)$ is exchangeable.

Example 14.4 (General finite partition) Any exchangeable random partition of $\mathbb{N}_n$ can be obtained by the following scheme.[2] First generate a composition $(n_1, \ldots, n_k)$ from a given but arbitrary distribution on the set of all compositions of n. Next, given $(n_1, \ldots, n_k)$, form a random vector $(X_1, \ldots, X_n)$ by randomly permuting[3] the set of n symbols

$$(\underbrace{1, \ldots, 1}_{n_1}, \underbrace{2, \ldots, 2}_{n_2}, \ldots, \underbrace{k, \ldots, k}_{n_k}).$$

Finally define a partition of $\mathbb{N}_n$ by the equivalence classes under the relationship $i \sim j$ if and only if $X_i = X_j$.

An alternative way of coding the information in a partition $\{n_1, \ldots, n_k\}$ of n generated by a partition $\{A_1, \ldots, A_k\}$ of $\mathbb{N}_n$ is to list the corresponding *multiplicity class* $(m_1, \ldots, m_n)$,

[1] For instance $p(1, 2)$ is the probability of the partition of $\mathbb{N}_3$ into $\{1\}$ and $\{2, 3\}$, and also of the partition in $\{2\}$ and $\{1, 3\}$; it is equal to $p(2, 1)$, which is the probability of the partition into e.g. $\{1, 3\}$ and $\{2\}$; $p(1, 1, 1)$ is the probability of the partition $\{\{1\}, \{2\}, \{3\}\}$; $p(3)$ is the probability of $\{\{1, 2, 3\}\}$.

[2] See Section 11 of Aldous (1985) for this and further discussion. We shall not use this characterization in the following.

[3] This is to say by a permutation drawn from the uniform distribution on the set of all $n!$ permutations of $\mathbb{N}_n$.

where m_i is defined as the number of sets A_j of cardinality i in the partition. The entries m_i of a multiplicity vector are nonnegative integers satisfying $\sum_{i=1}^{n} i m_i = n$ (with necessarily many zero m_i).

This coding was used already in the statement of Ewens's sampling formula given in Proposition 4.11. The following proposition generalizes this formula to general partitions.

Proposition 14.5 (Sampling formula) *The probability that a random exchangeable partition of $\mathbb{N}_n$ consists of m_i sets of cardinality i, for $i = 1, 2, \ldots, n$, is equal to, for any composition $(n_1, \ldots, n_k)$ compatible with the multiplicity class $(m_1, \ldots, m_n)$,*

$$\frac{n!}{\prod_{i=1}^{n} m_i!(i!)^{m_i}}\, p(n_1, \ldots, n_k).$$

Furthermore, the probability that a random exchangeable partition of $\mathbb{N}_n$ consists of k sets, for $k = 1, \ldots n$, is equal to, with the sum over all compositions $(n_1, \ldots, n_k)$ of n in k elements,

$$\sum_{(n_1,\ldots,n_k)} \frac{1}{k!} \binom{n}{n_1 \cdots n_k}\, p(n_1, \ldots, n_k).$$

Proof The EPPF gives the probability of a particular partition. The factor preceding it is the cardinality of the number of partitions giving the multiplicity class. See the proof of Proposition 4.10 for details.

For the proof of the second formula it is helpful to create an *ordered* random partition of $\mathbb{N}_n$ by randomly ordering the sets of an ordinary (i.e. unordered) random partition. The probability that the partition contains k sets is the same, whether the partition is ordered or not. For a given composition $(n_1, \ldots, n_k)$ of n, the probability of a particular ordered partition with set sizes $n_1, \ldots, n_k$ is $p(n_1, \ldots, n_k)/k!$. We can construct all ordered partitions of $\mathbb{N}_n$ with composition $(n_1, \ldots, n_k)$ by first ordering the elements $1, 2, \ldots, n$ in every possible way and next defining the first set to consist of the first n_1 elements, the second set of the next n_2 elements, etc. There are $n!$ possible orderings of $1, 2, \ldots, n$, but permuting the first n_1 elements, the next n_2 elements, etc. gives the same ordered partition. Thus there are $n!/\prod_i n_i!$ ordered partitions with composition $(n_1, \ldots, n_k)$. By exchangeability they all have the same probability, $p(n_1, \ldots, n_k)/k!$. □

Rather than elaborating on exchangeable partitions of a finite set, we turn to *partition structures*, which link partitions across n. We want these structures to be *consistent* in the following sense. From a given partition $\{A_1, \ldots A_k\}$ of $\mathbb{N}_n$, we can obtain a partition of $\mathbb{N}_m$ for a given $m < n$ by removing the elements $\{m + 1, \ldots, n\}$ from the sets A_i, and removing possible empty sets thus created. If this is applied to an exchangeable partition $\mathcal{P}_n$, then the resulting partition will be an exchangeable partition of $\mathbb{N}_m$, since a permutation of just the elements of $\mathbb{N}_m$ can be identified with the permutation of $\mathbb{N}_n$ that leaves $m + 1, \ldots, n$ invariant.

Definition 14.6 (Exchangeable partition) An *infinite exchangeable random partition* (or *exchangeable random partition of* $\mathbb{N}$) is a sequence $(\mathcal{P}_n : n \in \mathbb{N})$ of exchangeable random partitions of $\mathbb{N}_n$ that are consistent in the sense that $\mathcal{P}_{n-1}$ is equal to the partition obtained

from $\mathcal{P}_n$ by leaving out the element n, almost surely, for every n. The function $p: \cup_{n=1}^{\infty} \mathcal{C}_n \to [0, 1]$ whose restriction to $\mathcal{C}_n$ is equal to the EPPF of $\mathcal{P}_n$ is called the *exchangeable partition probability function* (EPPF) of $(\mathcal{P}_n : n \in \mathbb{N})$.

That the definition does not refer directly to infinite partitions, but focuses on consistent finite partitions instead, avoids measurability issues. However, a partition of $\mathbb{N}$ is implied as it can be ascertained from $\mathcal{P}_{i \vee j}$ whether i and j belong to the same set.

An infinite exchangeable random partition may be obtained from an (infinite) exchangeable sequence of random variables $X_1, X_2, \ldots$, by defining $\mathcal{P}_n$ by the pattern of ties in the collection $(X_1, \ldots, X_n)$, for every n, as in Example 14.3. The sequence $X_1, X_2, \ldots$ is said to *generate* the infinite random partition in this case. The following theorem shows that *all* infinite exchangeable random partitions are generated by an exchangeable sequence. It is similar to de Finetti's representation of exchangeable sequences, to which it can also be reduced.

Theorem 14.7 (Kingman's representation) *For any infinite exchangeable random partition $(\mathcal{P}_n : n \in \mathbb{N})$ defined on a probability space that is rich enough to support an independent i.i.d. sequence of uniform variables, there exists a random probability measure P on $[0, 1]$ and a sequence of random variables $X_1, X_2, \ldots$ defined on the same probability space with $X_1, X_2, \ldots \mid P \overset{iid}{\sim} P$ that generates $(\mathcal{P}_n : n \in \mathbb{N})$. Furthermore, the size $N_{(j),n}$ of the jth largest set in $\mathcal{P}_n$ satisfies $n^{-1} N_{(j),n} \to W_{(j)}$ a.s. as $n \to \infty$, for $W_{(1)} \geq W_{(2)} \geq \cdots$ the sizes of the atoms of P ordered in decreasing size.*

Proof Let $\xi_1, \xi_2, \ldots$ be an i.i.d. sequence of uniform random variables defined on the same probability space as $(\mathcal{P}_n)$ and independent of it. By deleting a null set it can be ensured that these variables "surely" assume different values. For any $i \in \mathbb{N}$ let $j(i)$ be the smallest number in the partitioning set to which i belongs and define $X_i = \xi_{j(i)}$. (As $i \in \mathbb{N}_i$ and $(\mathcal{P}_n)$ is consistent, it suffices to consider the partition $\mathcal{P}_i$, and $j(i) \leq i$.) Then $X_1, X_2, \ldots$ generates the partition $(\mathcal{P}_n)$. We can write $X_i = g_i((\xi_j), (\mathcal{P}_n))$, for some measurable map g_i, and then have $X_{\sigma(i)} = g_i((\xi_{\sigma(j)}), (\sigma\langle\mathcal{P}_n\rangle))$, for any permutation σ of $\mathbb{N}$ that permutes only a finite set of coordinates, where $(\sigma\langle\mathcal{P}_n\rangle)$ is the sequence of partitions of $\mathbb{N}_n$ obtained from $\sigma\langle\mathcal{P}_n\rangle$ by swapping the elements of $\mathbb{N}$ according to σ. Because the distribution of $((\xi_{\sigma(j)}), (\sigma\langle\mathcal{P}_n\rangle)$ is invariant under the permutation σ, the sequence $X_1, X_2, \ldots$ is exchangeable. By de Finetti's theorem there exists a random measure P such that given P it is an i.i.d. sequence.

To prove the last assertion we note that the classes of the partition $\mathcal{P}_n$ generated by $X_1, \ldots, X_n$ are among the sets $\{i \in \mathbb{N}_n : X_i = x\}$, when x varies (most of which are empty), and hence the numbers $N_{(j),n}$ are the first n elements of the numbers $N_n(x) := \#\{i \in \mathbb{N}_n : X_i = x\}$ ordered in decreasing size (hence with the uncountably many zeros at the end). By the (ergodic) law of large numbers $n^{-1} N_n(x) \to \mathrm{P}(X_1 = x \mid P)$, a.s., which implies the result. □

The unit interval and the random probability measure P in the theorem are rather arbitrary. Only the sizes (W_i) of the atoms of P play a role for the pattern of equal X_i; their locations merely serve as labels. The atomless part of P is even more arbitrary: it

produces labels that are used only once and hence give rise to singletons in the induced partition. This will motivate to make in the next section a convenient choice of P, in the form of a "species sampling model," which gives a random measure on an arbitrary Polish space.

By marginalization, the EPPFs of the random exchangeable partitions of $\mathcal{P}_{n-1}$ and $\mathcal{P}_n$ obtained from an infinite exchangeable random partition satisfy the relationships

$$p(n_1, \ldots, n_k) = \sum_{j=1}^{k} p(n_1, \ldots, n_j + 1, \ldots, n_k) + p(n_1, \ldots, n_k, 1). \tag{14.2}$$

Here we abuse notation by using the same letter p for both EPPFs, with p on the left side referring to $\mathcal{P}_{n-1}$ and every p on the right side to $\mathcal{P}_n$. This is customary and should not cause confusion as the sum of the arguments reveals the set $\mathbb{N}_n$ to which it refers.[4] If for $\boldsymbol{n} = (n_1, \ldots, n_k)$ we introduce the notations $\boldsymbol{n}^{j+} = (n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k)$, for $1 \le j \le k$, and $\boldsymbol{n}^{(k+1)+} = (n_1, \ldots, n_k, 1)$, then the preceding formula can also be written as $p(\boldsymbol{n}) = \sum_{j=1}^{k+1} p(\boldsymbol{n}^{j+})$. The $j+$ refers to the set A_j from which n was deleted when reducing $\mathcal{P}_n$ to $\mathcal{P}_{n-1}$, or perhaps better added when extending the latter to the former.

Equation (14.2) can be viewed as expressing consistency of the partitions in a distributional sense, which is a consequence of the almost sure consistency that is baked into Definition 14.6. As we shall be mainly interested in distributions and not exact sample space constructions, this distributional consistency will be good enough for our purposes. In fact, given a sequence of symmetric functions $p: \cup_{n=1}^{\infty} \mathcal{C}_n \to [0, 1]$ that satisfy the marginalization equation (14.2), one can always construct an infinite exchangeable partition that has p as its EPPF. To see this we may first construct a joint distribution of a sequence $(\mathcal{P}_n : n \in \mathbb{N})$ of partitions of $\mathbb{N}_n$ by recursively extending $\mathcal{P}_{n-1}$ to $\mathcal{P}_n$, for $n = 2, 3, \ldots$, as indicated in Figure 14.1. Equation (14.2) ensures that this construction leads to a proper probability distribution on the set of all partitions of $\mathbb{N}_n$. Symmetry of p in its arguments implies the exchangeability of the partitions. Finally we may appeal to an abstract theorem by Ionescu-Tulcea on the existence of a joint distribution given consistent marginal distributions to establish the distribution of the whole infinite sequence $\{\mathcal{P}_n\}$.

Instead of as a sequence of partitions of $\mathbb{N}_n$ it is convenient to think of an infinite exchangeable partition as a sequential process that adds the points $2, 3, \ldots$ in turn to the existing partition formed by the preceding points. The (conditional) probability that the point $n + 1$ is added to the jth set in the partition of $\mathcal{P}_n$ of $\mathbb{N}_n$ with composition $\boldsymbol{n} = (n_1, \ldots, n_k)$ is equal to

$$p_j(\boldsymbol{n}) = \frac{p(\boldsymbol{n}^{j+})}{p(\boldsymbol{n})}, \qquad j = 1, \ldots, k + 1. \tag{14.3}$$

The collection of functions $p_j: \cup_n \mathcal{C}_n \to [0, 1]$ is called the *prediction probability function* (PPF). Clearly $(p_1(\boldsymbol{n}), \ldots, p_{k+1}(\boldsymbol{n}))$ is a probability vector for every composition $\boldsymbol{n} = (n_1, \ldots, n_k)$. In Figure 14.1 these probability vectors are placed as weights on the branches emanating from the node $\boldsymbol{n}$, so that the probabilities of the terminal nodes are

[4] For instance $p(1, 2)$ refers to a (particular) partition of $\mathbb{N}_3$ in two sets of sizes 1 and 2; by the preceding formula it is equal to $p(2, 2) + p(1, 3) + p(1, 2, 1)$, whose three terms refer to (particular) partitions of $\mathbb{N}_4$.

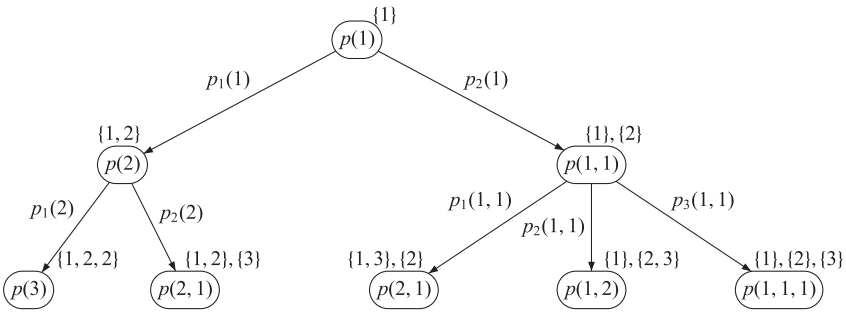

Figure 14.1 Recursive partitioning. The three levels show all partitions of $\mathbb{N}_1$, $\mathbb{N}_2$ and $\mathbb{N}_3$, respectively. The partitions at each consecutive level are constructed by adding the next number (2 and 3 for the levels shown, 4 for the next level not shown) to every of the sets in a partition of the preceding level, thus splitting a node into as many branches as sets present in the partition at the node plus 1. Equation (14.2) ensures that the probability of a node splits along the branches. In the tree as shown exchangeability of the partition means (only) that the middle three leaves (the partitions $\{1, 2\}, \{3\}$ and $\{1, 3\}, \{2\}$ and $\{1\}, \{2\}, \{3\}$ of $\mathbb{N}_3$) have equal probability.

the products of the weights on the branches along the path connecting them to the root of the tree.

Rather than deriving a PPF from an EPPF as in (14.3), we might start from a collection of functions p_j, and sequentially generate a sequence of partitions. To obtain an infinite exchangeable partition, the functions p_j must correspond to an EPPF p. The following lemma gives necessary and sufficient conditions that the implied p is an EPPF.

Lemma 14.8 *A collection of functions p_j is a PPF of an infinite exchangeable partition if and only if $(p_1(\boldsymbol{n}), \ldots, p_{k+1}(\boldsymbol{n}))$ is a probability vector, for every composition $\boldsymbol{n} = (n_1, \ldots, n_k)$, and, for all $i, j = 1, 2, \ldots,$*

$$p_i(\boldsymbol{n})\, p_j(\boldsymbol{n}^{i+}) = p_j(\boldsymbol{n})\, p_i(\boldsymbol{n}^{j+}), \tag{14.4}$$

and for every permutation σ of $\mathbb{N}_k$, the following permutation equivariance condition holds:

$$p_i(n_1, \ldots, n_k) = p_{\sigma^{-1}(i)}(n_{\sigma(1)}, \ldots, n_{\sigma(k)}). \tag{14.5}$$

Proof The necessity of (14.4) follows upon substituting (14.3) for all four occurrences of a p_j, which reduces both sides of the equation to $p((\boldsymbol{n}^{i+})^{j+})/p(\boldsymbol{n})$.

Conversely, given the p_j we may use (14.3) to define p recursively by $p(1) = 1$, and $p(\boldsymbol{n}^{j+}) = p_j(\boldsymbol{n})p(\boldsymbol{n})$. Because a given composition $\boldsymbol{n}$ can be reached by many different paths starting at 1 and augmenting the composition by steps of 1, it must be checked that this definition is well posed, and the function so obtained satisfies the consistency condition (14.2) of an EPPF. Both are guaranteed by (14.4).

Equation (14.5) can be seen to be necessary (e.g. by (14.3)) and sufficient for the function p so constructed to be permutation symmetric. □

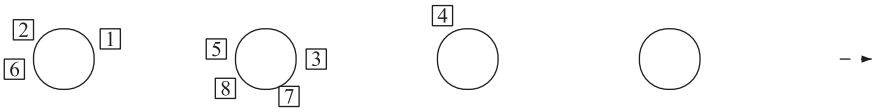

Figure 14.2 Chinese restaurant process. The ninth customer will sit at the tables from left to right with probabilities $3/(M+8)$, $4/(M+8)$, $1/(M+8)$ and $M/(M+8)$. There is an infinite list of empty tables.

14.1.1 The Chinese Restaurant Process

The infinite exchangeable random partition generated by a sample from the Dirichlet process $\mathrm{DP}(MG)$ with atomless center measure G, as in Section 4.1.4, is called the *Chinese restaurant process* (CRP) with parameter M. By Proposition 4.11 its EPPF is given by

$$p(n_1, \ldots, n_k) = \frac{M^k \prod_{i=1}^k (n_i - 1)!}{M^{[n]}}, \tag{14.6}$$

where $a^{[k]} = a(a+1)\cdots(a+k-1)$ stands for the ascending factorial. This is a product partition model with cohesion function given by $\rho(n_i) = M(n_i - 1)!$. The PPF of the process is given by the generalized Pólya urn (or Blackwell-MacQueen scheme):

$$p_j(n_1, \ldots, n_k) = \frac{n_j}{M+n}, \quad j = 1, \ldots, k, \qquad p_{k+1}(n_1, \ldots, n_k) = \frac{M}{M+n}.$$

Thus the probability of adding a next element to an existing set in the partition is proportional to the number of elements already in the partition, and a new set is "opened" with probability proportional to M.

The name derives from the following metaphor. Suppose that customers arrive sequentially in a Chinese restaurant with an infinite number of tables, each with infinite seating capacity. The first customer chooses an arbitrary table. The second customer has two options, sit at the table opened by customer 1 or open a new table, between which he decides with probabilities $1/(M+1)$ and $M/(M+1)$. More generally, the $(n+1)st$ customer finds n customers seated at k tables in groups of $n_1, \ldots, n_k$, where $\sum_{j=1}^k n_j = n$, and chooses to sit at the jth open table with probability $n_j/(M+n)$, or open a new table with probability $M/(M+n)$. The gravitational effect of this scheme – more massive tables, apparently with more known faces, attract a newcomer with a higher probability – is valuable for clustering variables together in groups.

The following result shows that the simple form of the PPF of the Chinese restaurant process as proportional to a function of the respective cardinalities, characterizes this process.

Proposition 14.9 *The Chinese restaurant process is the only exchangeable random partition with PPF* $(p_j : j \in \mathbb{N})$ *of the form, for some function* $f : \mathbb{N} \to (0, \infty)$ *not depending on* j*, some* $M > 0$*, and every composition* $(n_1, \ldots, n_k)$*,*

$$p_j(n_1, \ldots, n_k) \propto \begin{cases} f(n_j), & j = 1, \ldots, k, \\ M, & j = k+1. \end{cases}$$

Proof By equation (14.4) applied with $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $i \neq j \in \{1, \ldots, k\}$, the number

$$\frac{f(n_i)}{\sum_{l=1}^{k} f(n_l) + M} \times \frac{f(n_j)}{\sum_{l=1:l\neq i}^{k} f(n_l) + f(n_i + 1) + M}$$

is invariant under swapping i and j. This is possible only if $f(n_j) + f(n_i + 1)$ is invariant, so that $f(n_i + 1) - f(n_i)$ is constant in i. Hence f must be of the form $f(n) = an + b$ for some real constants a, b.

The same reasoning with $i = 1, \ldots, k$ and $j = k + 1$ leads to the relation $f(n_i + 1) = f(n_i) + f(1)$, whence we must have $b = 0$, that is $f(n) = an$. Thus (p_j) is the PPF of a Dirichlet process. □

14.1.2 The Chinese Restaurant Franchise Process

The *Chinese restaurant franchise* process is a hierarchical partition structure based on the hierarchical Dirichlet process (HDP) introduced in Example 5.12. The hierarchical Dirichlet process consists of a "global" random probability measure G generated from a $\mathrm{DP}(M_0 G_0)$-distribution, "local" random probability measures $G_1, G_2, \ldots$ generated i.i.d. from $\mathrm{DP}(MG)$ for given G, and a random sample $X_{i,1}, X_{i,2}, \ldots$ from every G_i independent across i.

The measure G can be written in a stick-breaking representation $G = \sum_{j=1}^{\infty} W_j \delta_{\theta_j}$, for an i.i.d. sequence $\theta_1, \theta_2, \ldots$ from G_0. Similarly every of the measures G_i has a stick-breaking representation $G_i = \sum_{j=1}^{\infty} W_{i,j} \delta_{\theta_{j,i}}$, but with the support points $\theta_{1,i}, \theta_{2,i}, \ldots$ now chosen from G, which means that they are subsets of the variables $\theta_1, \theta_2, \ldots$ Finally the observations $X_{i,1}, X_{i,2}, \ldots$ are subsets of the sequences $\theta_{1,i}, \theta_{2,i}, \ldots$ As a result, observations from different samples are linked through the hierarchical structure, and when the samples $X_{i,1}, X_{i,2}, \ldots$ are used as before to induce a partitioning by ties, then this partition structure will be dependent across the samples.

This structure becomes better visible in the predictive distributions of the samples. Given G the variables $X_{i,1}, X_{i,2}, \ldots$ are a sample from the Dirichlet process $\mathrm{DP}(MG)$ in the sense of Section 4.1.4, and can be generated by the Pólya urn scheme

$$X_{i,j} \,|\, X_{i,1}, \cdots, X_{i,j-1}, G \sim \sum_{t=1}^{K_{i,j}} \frac{N_{i,t}}{M + j - 1} \delta_{\tilde{X}_{i,t}} + \frac{M}{M + j - 1} G, \tag{14.7}$$

where $\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,K_{i,j}}$ are the distinct values in $X_{i,1}, \ldots, X_{i,j-1}$ and $N_{i,1}, \ldots, N_{i,K_{i,j}}$ their multiplicities. The G_i do not appear in this scheme, but "have been integrated out." The variable $X_{i,j}$ generated in the display is either a previous value $\tilde{X}_{i,t}$ from the same sample, or a "new" value from G. The latter measure is itself a realization from the $\mathrm{DP}(M_0 G_0)$-process. The same realization is used in the sequential scheme for every sample $X_{i,1}, X_{i,2}, \ldots$, for $i = 1, 2, \ldots$, but the sampling in the display is independent across i. Every time the variable $X_{i,j}$ must be "new," then this new value must be independently drawn from G, both within a sample of observations $X_{i,1}, X_{i,2}, \ldots$ as across i. Because only a sample from G is needed, rather than realizing the complete measure we can appeal a second time to the Pólya urn scheme of Section 4.1.4, and generate a sample from G, in the form

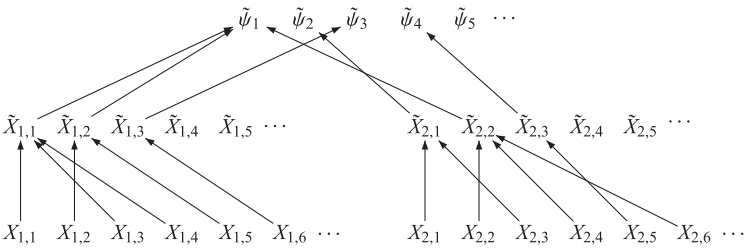

Figure 14.3 Chinese restaurant franchise process. At the top level are the distinct values in a sample $\psi_1, \psi_2, \ldots$ from the DP(MG_0) process. The middle level shows the distinct values in two samples of observations shown at the bottom level. The arrows show a possible dependency structure in a single realization: the variable at the bottom is a copy from the variable at the top.

$$\psi_k | \, \psi_1, \ldots, \psi_{k-1} \sim \sum_{s=1}^{L_k} \frac{P_{k,s}}{M_0 + k - 1} \delta_{\tilde{\psi}_s} + \frac{M_0}{M_0 + k - 1} G_0,$$

where $\tilde{\psi}_1, \ldots, \tilde{\psi}_{L_k}$ are the distinct values in $\psi_1, \ldots, \psi_{k-1}$ and $P_{k,s}$ are their multiplicities. Only a single such sequential scheme is needed to generate all samples $X_{i,1}, X_{i,2}, \ldots$: every time it is required to sample a new value (from G) in (14.7), we substitute the first value ψ_k that has not been used before. In practice we shall generate the sample values $X_{i,j}$ in a given order and generate the next observation ψ_k whenever it is needed to extend one of the samples $X_{i,1}, X_{i,2}, \ldots$ This next observation will then be equal to an existing value $\tilde{\psi}_s$ or a new value generated from G_0, as in a Chinese restaurant process. In the end every $X_{i,j}$ will be one of the distinct values $\tilde{\psi}_s$ in the sequence $\psi_1, \psi_2, \ldots$, which may well occur multiple times in the sample $X_{i,1}, X_{1,2}, \ldots$, as well as in the other samples. This practical implementation presupposes an ordering in which the observations $X_{i,j}$ are sampled. This is a bit arbitrary, but unimportant for the resulting clustering. Figure 14.3 illustrates the scheme.

The following culinary metaphor explains the name "Chinese restaurant franchise process." Consider a number of Chinese restaurants that serve a common "franchise-wide" menu. Customers arrive and are seated in each restaurant according to independent Chinese restaurant processes with parameter M. Their overall order of arrivals is arbitrary, but fixed. All customers at a given table eat the same dish, which is chosen by the first customer at the table from the dishes of the previously opened tables in the franchise or from the franchise-wide menu with probabilities proportional to the total number of tables eating this dish in the franchise and M_0 for the new dish, respectively. In the preceding notation $X_{i,1}, X_{i,2}, \ldots$ are the dishes eaten by the customers in restaurant i, $\psi_1, \psi_2, \ldots$ are dishes that are new to a restaurant, and $\tilde{\psi}_1, \tilde{\psi}_2, \ldots$ are dishes that are new to the franchise. The resulting process induces a random partition by table in each restaurant, which is nested within a partition by dish across the franchise.

14.2 Species Sampling Processes

As noted following its statement only the sizes (W_i) of the atoms of the directing measure P in Kingman's representation theorem, Theorem 14.7, play a role for the pattern of equal X_i. It is convenient to represent P in the special form of a *species sampling model*.

Definition 14.10 (Species sampling model) A *species sampling model* (SSM) is a pair of a sequence (X_i) of random variables and a random measure P such that $X_1, X_2, \ldots \mid P \overset{\text{iid}}{\sim} P$ and P takes the form

$$P = \sum_{j=1}^{\infty} W_j \delta_{\theta_j} + (1 - \sum_{j=1}^{\infty} W_j) G, \tag{14.8}$$

for $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$ for an atomless probability distribution G on a Polish space $\mathfrak{X}$ and an independent random subprobability vector (W_j). The random distribution P in a SSM is called a *species sampling process* (SSP). If $\sum_{j=1}^{\infty} W_j = 1$, the SSP is called *proper*.

Because G is atomless, the "labels" θ_j are all different (almost surely), and hence the species sampling process P in (14.8) has atoms (W_j), as the random measure P in Kingman's theorem. Thus the partitions generated by the output $X_1, X_2, \ldots$ of a species sampling model can be viewed as a general model for an infinite exchangeable partition.

One advantage of this special construction is that the predictive distributions of $X_1, X_2, \ldots$ take a simple form. Write $\tilde{X}_1, \tilde{X}_2, \ldots$ for the distinct values in the sequence of variables $X_1, X_2, \ldots$ in *order of appearance*, and for each n let $N_{j,n}$ be the multiplicity of $\tilde{X}_j$ in $X_1, \ldots, X_n$, and K_n the total number of distinct values in the latter set.[5] Write $N_n = (N_{1,n}, \ldots, N_{K_n,n})$ for the vector of counts of distinct values.

Lemma 14.11 *The predictive distributions of variables $X_1, X_2, \ldots$ in the species sampling model* (14.8) *take the form $X_1 \sim G$ and, for $n \geq 1$,*

$$X_{n+1} \mid X_1, \ldots, X_n \sim \sum_{j=1}^{K_n} p_j(N_n) \delta_{\tilde{X}_j} + p_{K_n+1}(N_n) G, \tag{14.9}$$

where the functions $p_j : \cup_n \mathcal{C}_n \to [0,1]$ are the PPF of the infinite exchangeable random partition generated by $X_1, X_2, \ldots$ (cf. (14.3)*). A sequence $X_1, X_2, \ldots$ with the same distribution as the original sequence can be generated by first generating the infinite random exchangeable partition defined by this PPF and next attaching to the partitioning sets in order of appearance the values of an independent i.i.d. sequence $\tilde{X}_1, \tilde{X}_2, \ldots$ from G.*

Proof The variable X_1 is either a θ_j or generated from G. As $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$ it follows that marginally $X_1 \sim G$.

For the proof of the general formula we introduce latent variables (I_i) that indicate the partitioning classes of the (X_i). We generate random variables hierarchically in the order: $(W_j)_{j \geq 0}$, where $W_0 = 1 - \sum_j W_j$, next $(I_i)_{i \geq 1} \mid (W_j) \overset{\text{iid}}{\sim} (W_j)$ on $\mathbb{N} \cup \{0\}$, and $(\theta_j)_{j \geq 1} \overset{\text{iid}}{\sim} G$ and $(\theta'_i)_{i \geq 1} \overset{\text{iid}}{\sim} G$ independently from each other and from (W_j) and (I_i). Finally we set $X_i = \theta_{I_i}$ if $I_i \geq 1$ and $X_i = \theta'_i$ if $I_i = 0$. Then, for $X_{1:n}$ abbreviating $X_1, \ldots, X_n$,

[5] So $\tilde{X}_1 = X_1$, $\tilde{X}_2 = X_{\hat{\imath}(2)}$, for $\hat{\imath}(2) = \min\{i \geq 2 : X_i \neq X_1\}$, etc. In the statement of the following lemma the ordering of the distinct values is actually irrelevant, but it is made explicit to unify notation.

$$\mathrm{P}(X_{n+1} \in B \,|\, X_{1:n}) = \sum_{l=1}^{\infty} \mathrm{P}(\theta_l \in B \,|\, X_{1:n}, I_{n+1} = l)\mathrm{P}(I_{n+1} = l \,|\, X_{1:n}) + \mathrm{P}(\theta'_{n+1} \in B \,|\, X_{1:n}, I_{n+1} = 0)\mathrm{P}(I_{n+1} = 0 \,|\, X_{1:n}).$$

Since θ'_{n+1} is independent of everything the leading probability in the second term on the right can be reduced to $G(B)$. The corresponding probabilities in the first term can be decomposed as the expectation

$$\mathrm{E}(\mathrm{P}(\theta_l \in B \,|\, X_{1:n}, I_1, \ldots, I_n, I_{n+1} = l) \,|\, X_{1:n}, I_{n+1} = l).$$

If $I_i = 0$, then $X_i = \theta'_i$ and is independent of everything and can be removed from the inner conditioning. If $I_i = k \geq 1$ and $k \neq l$, then $X_i = \theta_k$ and can be removed from the inner conditioning as well. Then either $l \notin \{I_1, \ldots, I_n\}$ and all $X_1, \ldots, X_n$ can be removed from the inner conditioning and the display reduces to $G(B)$, or $l = I_i$ for some i and then the display reduces to $\mathbb{1}\{X_i \in B\}$. For $\tilde{I}_1, \tilde{I}_2, \ldots$ the sequence $I_1, I_2, \ldots$ with nonzero duplicates removed the latter indicator can also be written $\sum_{j=1}^{K_n} \mathbb{1}\{\tilde{I}_j = l, \tilde{X}_j \in B\}$. Multiplying this and the indicator of $l \notin \{I_1, \ldots, I_n\}$ by $\mathrm{P}(I_{n+1} = l \,|\, X_{1:n})$ and summing over l, we obtain that $\mathrm{P}(X_{n+1} \in B \,|\, X_{1:n})$ is equal to

$$\sum_{l=1}^{\infty} \Big[\sum_{j=1}^{K_n} \mathbb{1}\{\tilde{I}_j = l, \tilde{X}_j \in B\} + G(B)\mathbb{1}\{l \notin \{I_1, \ldots, I_n\}\} \Big] + G(B)\mathrm{P}(I_{n+1} = 0 \,|\, X_{1:n})$$
$$= \sum_{j=1}^{K_n} \mathbb{1}\{\tilde{X}_j \in B\}\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, X_{1:n}) + G(B)\mathrm{P}(I_{n+1} \notin \{\tilde{I}_1, \ldots, \tilde{I}_{K_n}\} \,|\, X_{1:n}).$$

The events $\{I_{n+1} = \tilde{I}_j\}$ describe that the point $n+1$ is added to the jth set when enlarging the partition generated by $X_1, X_2, \ldots$ from $\mathbb{N}_n$ to $\mathbb{N}_{n+1}$ if $j \leq K_n$, or opens a new set if $j = K_n + 1$. The probabilities $\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, X_{1:n})$ can therefore be interpreted as the transition probabilities p_j of the transition $\mathcal{P}_n$ to $\mathcal{P}_{n+1}$. We must still show that they depend on $X_{1:n}$ through N_n only.

The conditioning variable $X_{1:n}$ is equal in information to the partition $\mathcal{P}_n$ of $\mathbb{N}_n$, the common values $\tilde{X}_j$ shared by the variables in the partitioning sets, and the ordering of these values repeated by their multiplicities. By the exchangeability of $X_1, X_2, \ldots$ the ordering is random and not informative on the events $\{I_{n+1} = \tilde{I}_j\}$. The common values are not informative either, as they are i.i.d. variables with law G and independent of the partition. In the case of a proper species sampling model this follows, as for any partition π of $\mathbb{N}_n$ with composition $(n_1, \ldots, n_k)$ and any measurable sets $B_1, \ldots, B_k$,

$$\mathrm{P}(\mathcal{P}_n = \pi, \tilde{X}_1 \in B_1, \ldots, \tilde{X}_k \in B_k) = \mathrm{E} \sum_{1 \leq i_1 \neq \cdots \neq i_k < \infty} \prod_{j=1}^{k} W_{i_j}^{n_j} \prod_{j=1}^{k} \mathbb{1}\{\theta_{i_j} \in B_j\}.$$

(The indices $i_1, \ldots, i_k$ represent the indices of the k different θ_i chosen.) As the weights (W_j) and locations (θ_j) are independent, the expectation factorizes, and the right side can be seen to be equal to $\mathrm{P}(\mathcal{P}_n = \pi) \prod_{j=1}^{k} G(B_j)$. This argument can be extended to a possibly improper species sampling process. We conclude that the conditioning information $X_{1:n}$ can

be reduced to the partition $\mathcal{P}_n$; equivalently by exchangeability, to the multiplicities of the partitioning sets.[6]

That $X_1, X_2, \ldots$ can be generated as in the final assertion of the lemma follows from the preceding. □

Species sampling processes take their name from a probabilistic model for discovery of new types of objects, or *species*. Interpret W_j as the relative frequency of the jth species, which carries label θ_j, in a population, and suppose we explore the population by sequentially sampling individuals. After sampling n individuals we record the number K_n of distinct species discovered, along with their multiplicities $N_n = (N_{1,n}, \ldots, N_{K_n,n})$ in the sample, and may wish to predict the type of the next individual, or estimate the frequency distribution of the species in the population. The predictive probabilities p_j (see (14.9)) are the solution to the first question, with $p_1(N_n), \ldots, p_{K_n}(N_n)$ describing the probabilities of sampling one of the K_n previously discovered species, and $p_{K_n+1}(N_n)$ the probability of finding a thus far undiscovered species.

The preceding lemma expresses the close link between a species sampling model and the (exchangeable) partition generated by the observed species. In accordance, a species sampling model can be characterized through any of the following three pairs of objects:

(i) G and the distribution of $(W_1, W_2, \ldots)$;
(ii) G and the EPPF;
(iii) G and the PPF.

The measure G is the mean measure $\mathrm{E}(P) = G$ of the random measure (see Section 3.4.2), and is also called the *center measure* of the species sampling process.

A random (sub)probability measure $(W_1, W_2, \ldots)$ on $\mathbb{N}$ can easily be obtained by the methods of Section 3.3. For the definition of the species sampling model, the ordering of the weights W_j is irrelevant, but for calculations one ordering may be more convenient than another. An easy way to identify the weights is to put them in decreasing order, as in Kingman's representation, Theorem 14.7. However, this may lead to a complicated description of their joint distribution. An alternative is a description in "order of appearance," which is a limit of the predictive distributions (14.9), and gives the weights in *size-biased order*.

Definition 14.12 (Size-biased permutation) The *size-biased permutation* of a probability distribution (w_j) on $\mathbb{N}$ is the random vector $(\tilde{w}_1, \tilde{w}_2, \ldots)$ for $\tilde{w}_j = w_{\tilde{I}_j}$ and $\tilde{I}_1, \tilde{I}_2, \ldots$ the distinct values in an i.i.d. sequence $I_1, I_2, \ldots$ from (w_j), in order of appearance. A random probability distribution $(\tilde{W}_j)$ on $\mathbb{N}$ is said to be *invariant under size-biased permutation* (ISBP), or to be in *size-biased order*, if its distribution is invariant under size-biased permutation by a sequence $I_1, I_2, \ldots$ that given $(\tilde{W}_j)$ is i.i.d. from $(\tilde{W}_j)$.

Because size-biased permutation takes a full probability vector (w_j) on $\mathbb{N}$ as input, listed in arbitrary order, and the size-biased version $(\tilde{w}_j)$ is a (random) permutation of the original sequence, a second independent size-biased permutation of a deterministic sequence does not change its distribution. It follows that any random sequence (W_j) can be changed in an

[6] Hansen and Pitman (2000) prove that for *any* exchangeable $X_1, X_2, \ldots$ with predictive distributions of the form (14.9) with p_j possibly dependent on $X_{1:n}$, these p_j necessarily depend on the partition only.

ISBP sequence $(\tilde{W}_j)$ by a single (independent) size-biased permutation, and its distribution is the same irrespective of the ordering of the original sequence (W_j).

If one thinks of the (w_j) as the relative frequencies of species in a population (labeled $1, 2, \ldots$), and $I_1, I_2, \ldots$ as a random sample of individuals, then $\tilde{I}_1, \tilde{I}_2, \ldots$ are the distinct species in order of appearance. The name "size-biased" expresses that more frequent species (with bigger w_i) will tend to show up earlier.

Rather than as the distinct values in an i.i.d. sequence, a sequence $\tilde{I}_1, \tilde{I}_2, \ldots$ with the same distribution can also be generated sequentially by the algorithm:

- Choose $\tilde{I}_1$ from $\mathbb{N}$ according to the probability vector (w_j).
- Given $\tilde{I}_1, \ldots, \tilde{I}_{j-1}$, choose $\tilde{I}_j$ from $\mathbb{N} \setminus \{\tilde{I}_1, \ldots, \tilde{I}_{j-1}\}$ according to the probability vector (w_j) conditioned to the latter set.

The following theorem reveals the role of size-biasing. Recall the notation $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ for the distinct species in $X_1, \ldots, X_n$, and $N_{1,n}, \ldots N_{K_n,n}$ for their multiplicities.

Theorem 14.13 (Pitman's representation) *If $X_1, X_2, \ldots$ follow the species sampling model* (14.8)*, then for every j the sequence $N_{j,n}/n$ tends a.s. surely to a limit $\tilde{W}_j$, and the predictive distributions* (14.9) *tend a.s. in total variation norm to $\tilde{P} = \sum_{j=1}^{\infty} \tilde{W}_j \delta_{\tilde{X}_j} + (1 - \sum_{i=1}^{\infty} \tilde{W}_j)G$. Here $\tilde{X}_1, \tilde{X}_2, \ldots \overset{iid}{\sim} G$ are independent of $(\tilde{W}_j)$, and for every $n \geq 1$,*

$$X_{n+1} | \, X_1, \ldots, X_n, \tilde{W}_1, \tilde{W}_2, \ldots \sim \sum_{j=1}^{K_n} \tilde{W}_j \delta_{\tilde{X}_j} + (1 - \sum_{j=1}^{K_n} \tilde{W}_j)G. \tag{14.10}$$

If the species sampling model is proper, then $(\tilde{W}_j)$ is a size-biased permutation of (W_j).

Proof Let $I_1, I_2, \ldots$ denote the indices of the components δ_{θ_j} from which $X_1, X_2, \ldots$ is chosen, with $I_i = 0$ if X_i is chosen from the continuous component G (and $X_i = \theta_{I_i}$ otherwise). Furthermore, let $\tilde{I}_1, \tilde{I}_2, \ldots$ be the distinct values in this sequence, in order of appearance.

Let $\hat{j}$ be the time of appearance of the jth species, so that $X_{\hat{j}} = \tilde{X}_j$. From the fact that the variables $X_1, X_2, \ldots$ are conditionally i.i.d. given P, it can be seen that the variables $Y_1, Y_2, \ldots$, for $Y_i = X_{\hat{j}+i}$, are conditionally i.i.d. given P and $X_1, \ldots, X_{\hat{j}}$. Since $N_{j.n} = 1 + \sum_{i=1}^{n-\hat{j}} \mathbb{1}\{Y_i = \tilde{X}_j\}$, the law of large numbers gives that $n^{-1} N_{j,n} \to W_{\tilde{I}_j} =: \tilde{W}_j$, almost surely. If the species sampling model is proper, then $I_1, I_2, \ldots$ is an i.i.d. sequence from (W_j) and hence the sequence $(\tilde{W}_j)$ is a size-biased permutation of the sequence (W_j).

In the proof of Lemma 14.11 the distinct values $\tilde{X}_1, \tilde{X}_2, \ldots$ were seen to be independent and identically distributed according to G and independent of the partition generated by $X_1, X_2, \ldots$. Thus they are also independent of the variables N_n and hence of the limits $\tilde{W}_j$.

In the proof of Lemma 14.11 the predictive probabilities were expressed as $p_j(N_n) = \mathrm{P}(I_{n+1} = \tilde{I}_j | \, I_{1:n})$, for $j \leq K_n$, where $I_{1:n} = (I_1, \ldots I_n)$ can also be replaced by N_n in the conditioning. Consider first the conditional probabilities $\mathrm{P}(I_{n+1} = \tilde{I}_j | \, I_{1:n}, N_m)$, for $m > n$. The extra conditioning information indicates that $N_{j,m} - N_{j,n}$ variables out of $X_{n+1}, \ldots, X_m$ must be realized as equal to $\tilde{X}_j$, to make the cardinality of the jth species

grow from $N_{j,n}$ to $N_{j,m}$. By exchangeability the indices of these variables are equally distributed over $n+1, \ldots, m$, so that $\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, I_{1:n}, N_m) = (N_{j,m} - N_{j,n})/(m-n)$. The latter variable has the same limit as $N_{j,m}/m$ if $m \to \infty$ and n is fixed, which was identified as $\tilde{W}_j$ in the preceding paragraph. Applying the dominated convergence theorem to the conditional expectations $\mathrm{E}(\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, I_{1:n}, N_m) \,|\, I_{1:n})$, we conclude that $p_j(N_n) = \mathrm{E}(\tilde{W}_j \,|\, I_{1:n})$. Finally, martingale convergence gives that $p_j(N_n) \to \tilde{W}_j$, almost surely as $n \to \infty$.

If we condense the continuous part G of the predictive distributions to an additional atom x_0, then the predictive distributions and $\tilde{P}$ concentrate on the countable set $\{x_0, \tilde{X}_1, \tilde{X}_2, \ldots\}$ and the sizes of the atoms of the predictive distributions on all but the first element of this set have been shown to tend to those of $\tilde{P}$. For any subsequence along which the sizes of the atoms at x_0 also converge, Scheffe's theorem gives convergence in total variation. Because the predictive distributions and the limit $\tilde{P}$ are all probability measures, the whole sequence must converge in total variation to $\tilde{P}$.

As limits of the sequence $N_{j,n}/n$, the variables $\tilde{W}_j$ are clearly measurable relative to $\mathcal{F} := \cap_m \mathcal{F}_m$, for $\mathcal{F}_m = \sigma\langle N_m, N_{m+1}, \ldots\rangle$. By the Markov property the conditional probability $\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, I_{1:n}, N_m)$, for $m > n$, does not change if we add $\mathcal{F}_m$ to the conditioning, and the reverse submartingale thus obtained tends almost surely to $\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, I_{1:n}, \mathcal{F})$, as $m \to \infty$, for fixed j and n. As the limit was already seen to be equal to $\tilde{W}_j$, it follows that taking the conditional expectation given the smaller σ-field generated by $I_{1:n}$ and the full sequence $(\tilde{W}_i)$ does not change the limit, whence $\mathrm{P}(I_{n+1} = \tilde{I}_j \,|\, I_{1:n}, (\tilde{W}_i)) = \tilde{W}_j$. This shows that the partition structure is generated according to equation (14.10) with the X_i replaced by the I_i. Together with the independence of the partition structure and the values $\tilde{X}_j$ attached to the partitioning sets, this implies the latter equation. □

The constructive part of the theorem is the displayed equation (14.10), which shows how to generate the observations $X_1, X_2, \ldots$ sequentially, given the sequence $(\tilde{W}_j)$. The $(n+1)$st observation X_{n+1} is with probability $\tilde{W}_j$ equal to the existing value $\tilde{X}_j$ (for $j = 1, \ldots, K_n$), and newly generated from G otherwise.

The observations $X_1, X_2, \ldots$ generate a sequence of partitions, which grows likewise. The first partition is $\mathcal{P}_1 = \{1\}$. Given the partition $\mathcal{P}_n = \{\mathcal{P}_{1,n}, \ldots, \mathcal{P}_{K_n,n}\}$ of $\mathbb{N}_n$ generated by $(X_1, \ldots, X_n)$, listed in *order of appearance*, the partition $\mathcal{P}_{n+1}$ is created by

(i) attaching the element $n+1$ to $\mathcal{P}_{j,n}$, with probability $\tilde{W}_j$, for $j = 1, \ldots, K_n$;
(ii) opening the new set $\mathcal{P}_{K_{n+1},n+1} = \{n+1\}$, with probability $1 - \sum_{j=1}^{K_n} \tilde{W}_j$.

Once the sequence of partitions $\mathcal{P}_n$ is created, the values $X_1, X_2, \ldots$ can be created by attaching an i.i.d. sequence $\tilde{X}_1, \tilde{X}_2, \ldots$ from G to the partitioning sets, one $\tilde{X}_i$ to each newly opened set, and equating every X_i to the $\tilde{X}_j$ of its partitioning set. This was already seen in Lemma 14.11.

The weights $\tilde{W}_j$ in the theorem are the almost sure limits of the multiplicities of the ties in $X_1, X_2, \ldots$, but if we are interested only in distributions (and not "strong" definitions as maps on a probability space), then we may turn the construction around. Given a proper sequence (W_j), we determine (the distribution of) its size-biased permutation $(\tilde{W}_j)$, and

generate the partitions and variables by the preceding scheme. This gives a realization from the species sampling model, as the distribution of the right side of (14.10) is determined by the *distribution* of $(\tilde{W}_j)$ only.

The size-bias of the sequence $(\tilde{W}_j)$ is crucial for this representation. In the predictive formula (14.9) the weights $p_j(N_n)$ change with the generation of every new variable, whereas the size-biased weights $\tilde{W}_j$ in (14.10) are fixed at the beginning. Intuitively, they can be fixed precisely because they are size-biased. Importantly, to generate the beginning $X_1, \ldots, X_n$ of the sequence of observations, only the initial weights $\tilde{W}_1, \ldots \tilde{W}_{K_n}$ are needed.

As a first illustration of the usefulness of the size-biased representation, the following corollary expresses the EPPF in the weights $(\tilde{W}_j)$. The formula may be contrasted with the formula using arbitrary weights, given in Problem 14.1, which involves infinite sums.

Corollary 14.14 *The EPPF of an infinite exchangeable partition given by a weight sequence $(\tilde{W}_j)$ listed in size-biased order satisfies, for every composition $(n_1, \ldots, n_k)$,*

$$p(n_1, \ldots, n_k) = \mathrm{E}\Big[\prod_{j=1}^{k} \tilde{W}_j^{n_j - 1} \prod_{j=2}^{k} \Big(1 - \sum_{i<j} \tilde{W}_i\Big)\Big]. \tag{14.11}$$

Proof In view of Theorem 14.13, a random partition $\mathcal{P}_n$ of $\mathbb{N}_n$ in the order of appearance is obtained by following the scheme (i)–(ii). The specification of an event $\{\mathcal{P}_n = \{A_1, \ldots, A_k\}\}$ for a partition $\{A_1, \ldots, A_k\}$ in order of appearance completely determines the steps (i)–(ii) by which the elements $2, 3, \ldots, n$ are added to the existing partition: the partition of $\mathbb{N}_1$ is $\{1\}$; next if $2 \in A_1$, the element 2 was attached to $\{1\}$, which has probability $\tilde{W}_1$, or otherwise $2 \in A_2$, which has probability $1 - \tilde{W}_1$; next there are three possibilities for 3: $3 \in A_1$, which has probability $\tilde{W}_1$, $3 \in A_2$, which has probability $\tilde{W}_2$, or otherwise $3 \in A_3$, which has probability $1 - \tilde{W}_1 - \tilde{W}_2$; etc. In total, we attach $n_j - 1$ times a new element to the jth opened set, and open $k - 1$ times a new set. Multiplying the probabilities gives the probability that $\mathcal{P}_n$ is this particular partition. Because $(\mathcal{P}_n)$ is assumed exchangeable, this probability is equal to the probability $p(n_1, \ldots, n_k)$ of any partition with the same composition. □

The construction (i)–(ii) can also be applied with an *arbitrary* sub-probability sequence (W_j), but the resulting partition is then not necessarily exchangeable: it leads to the wider class of *partially exchangeable random partitions*, defined as infinite random partitions $(\mathcal{P}_n)$ such that for any partition $\{A_1, \ldots, A_k\}$ of $\mathbb{N}_n$, in the *order of appearance*,

$$\mathrm{P}(\mathcal{P}_n = \{A_1, \ldots, A_k\}) = p(|A_1|, \ldots, |A_k|),$$

for some function $p: \cup_{n=1}^{\infty} \mathcal{C}_n \to [0, 1]$. By the proof of the preceding lemma this function p will then be given by (14.11); the partition is exchangeable if and only if this function is symmetric in its arguments. This leads to the following characterization of size-biasedness.

Lemma 14.15 (Size-bias) *A random probability vector $(\tilde{W}_j)$ is in size-biased order, if $\tilde{W}_1 > 0$ a.s. and the measure $A \mapsto \mathrm{E}[\mathbb{1}_A(\tilde{W}_1, \ldots, \tilde{W}_k) \prod_{j=2}^{k} (1 - \sum_{i<j} \tilde{W}_i)]$ on $\mathbb{R}^k$ is symmetric, for every $k \in \mathbb{N}$.*

Proof If the given measure is symmetric, then the function p defined by (14.11) in Corollary 14.14 is symmetric. The preceding scheme (i)–(ii) then generates an infinite exchangeable partition. Comparison of (i)–(ii) with Theorem 14.13 shows that $(\tilde{W}_j)$ from (i)–(ii) must be equal in distribution to the sequence $(\tilde{W}_j)$ in this theorem. The theorem asserts that this sequence is size-biased if the corresponding species sampling model is proper. If this would not be the case, then the first observation would with positive probability be generated from the atomless component G. On this event, the first set in the partition would remain a singleton, whence $n^{-1}N_{1,n} = n^{-1}$ for every n. By Theorem 14.13 this sequence tends almost surely to $\tilde{W}_1$, and hence this event is excluded by the assumption that $\tilde{W}_1 > 0$, almost surely. □

Example 14.16 (Dirichlet process) A random sample from a Dirichlet process $\mathrm{DP}(MG)$ is a proper species sampling model with center measure G. This follows from the description of a Dirichlet process as a random discrete measure in Theorem 4.12.

The distribution of the weights (W_j) in decreasing magnitude $W_1 \geq W_2 \geq \cdots$ is known as the *Poisson-Dirichlet distribution*. The stick-breaking weights in Theorem 4.12 are the size-biased permutation of this sequence, as will be seen in greater generality in Theorem 14.33. Its distribution is also known as the *GEM distribution*, named after Griffiths, Engen and McClosky.

Example 14.17 (Fisher process) The *Fisher process* is the species sampling process with weight sequence given by $(W_1, \ldots, W_m) \sim \mathrm{Dir}(m; -\sigma, \ldots, -\sigma)$, for some $\sigma < 0$, and $W_j = 0$, for $j > m$. The corresponding model gives partitions in at most m sets. Using Kingman's formula (see Problem 14.1) and Corollary G.4 its EPPF can be computed as, for $k \leq m$,

$$\begin{aligned} p(n_1, \ldots, n_k) &= \frac{m!}{(m-k)!} \frac{\Gamma(-m\sigma)}{\Gamma(-m\sigma + n)} \prod_{j=1}^{k} \frac{\Gamma(-\sigma + n_j)}{\Gamma(-\sigma)} \\ &= \frac{\prod_{i=1}^{k-1}(-m\sigma + i\sigma)}{(-m\sigma + 1)^{[n-1]}} \prod_{j=1}^{k} (1-\sigma)^{[n_j - 1]}. \end{aligned}$$

The second rendering exhibits the model as a Pitman-Yor process, discussed in Section 14.4. This connection is also the motivation to write the (positive!) parameter as $-\sigma$.

The weight vector $(W_1, \ldots, W_m)$ is exchangeable, and not size-biased. In Theorem 14.33 it will be seen, by an application of Corollary 14.14 and Lemma 14.15, that the size-biased weights are given by the stick-breaking scheme $\tilde{W}_j = V_j \prod_{i<j}(1 - V_i)$, for $V_1, \ldots, V_{m-1}$ independent with $V_j \sim \mathrm{Be}(1 - \sigma, -(m - j)\sigma)$, for $j = 1, \ldots, m - 1$, and $\tilde{W}_m = 1 - \sum_{j<m} \tilde{W}_j$. (This is *not* the stick-breaking representation of the $\mathrm{Dir}(m; -\sigma, \ldots, -\sigma)$-distribution.)

14.2.1 Posterior Distribution

A species sampling process P may be used as a prior for the distribution of a random sample $X_1, \ldots, X_n$ of observations. As always the posterior mean is given by the predictive

distribution, which for a species sampling process has an attractive expression in terms of the PPF: $\mathrm{E}(P \mid X_1, \ldots, X_n)$ is equal to the distribution on the right side of (14.9). The full posterior distribution is a modified species sampling process. The following theorem shows that it has an easy description in terms of the weight sequence in size-biased order.

Theorem 14.18 (Posterior distribution) *The posterior distribution of P in the model with observations $X_1, \ldots, X_n \mid P \overset{iid}{\sim} P$ with P following a proper species sampling prior* (14.8) *with the weight sequence $(W_j) = (\tilde{W}_j)$ in size-biased order is the distribution of*

$$\sum_{j=1}^{K_n} \widehat{W}_j \delta_{\tilde{X}_j} + \sum_{j=K_n+1}^{\infty} \widehat{W}_j \delta_{\widehat{X}_j},$$

where $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ are the distinct values in $X_1, \ldots, X_n$ in order of appearance, the variables $\widehat{X}_{K_n+1}, \widehat{X}_{K_n+2}, \ldots \overset{iid}{\sim} G$, and $\widehat{W} = (\widehat{W}_j)$ is an independent vector with distribution given by, for every bounded measurable function f,

$$\mathrm{E}(f(\widehat{W}) \mid X_1, \ldots, X_n) = \frac{1}{p(N_n)} \mathrm{E}\Big[f(\tilde{W}) \prod_{j=1}^{K_n} \tilde{W}_j^{N_{j,n}-1} \prod_{j=2}^{K_n} \Big(1 - \sum_{i<j} \tilde{W}_i\Big)\Big],$$

where p is the EPPF of the species sampling model, and the expectation on the right side is under the prior distribution of $(\tilde{W}_i)$.

Proof Generate $X_1, \ldots, X_n$ by the scheme of Theorem 14.13: first sequentially generate a random partition of $\mathbb{N}_n$ by the scheme (i)–(ii) described following the theorem (also see the proof of Corollary 14.14) and next attach to each partitioning set a different value from an i.i.d. sequence $\tilde{X}_1, \tilde{X}_2, \ldots$ from G. Now suppose that we have observed $X_1, \ldots, X_n$. Then the value of K_n and the values $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ can be recovered exactly from the observations $X_1, \ldots, X_n$, whereas the values $\tilde{X}_{K_n+1}, \tilde{X}_{K_n+2}, \ldots$ remain i.i.d. from G, also conditionally given these observations. By (the proof of) Corollary 14.14 the likelihood of the observed partition given the weight sequence $(\tilde{W}_j)$ is given by $\prod_{j=1}^{K_n} \tilde{W}_j^{n_j-1} \prod_{j=2}^{K_n} (1 - \sum_{i<j} \tilde{W}_i)$. By Bayes's rule the posterior distribution of the weight sequence is obtained by reweighting its prior with the likelihood, and renormalizing. The normalizing constant is $p(N_n)$, by Corollary 14.14. □

Under the mild conditions of Lemma 3.5 on the weight sequence (W_j) a proper species sampling process has full weak support in the set of all probability measures on $\mathfrak{X}$. This might suggest that it leads to posterior consistency under the weak topology when used as a prior on the distribution of i.i.d. data. This is *not* true in general. The following theorem gives (somewhat abstract) sufficient conditions for posterior consistency. We shall encounter several examples where these conditions are not met and the posterior distribution is inconsistent. As the prior is neither tail-free, nor possesses the Kullback-Leibler property (as the family of distributions under consideration is not even dominated), it should not be a complete surprise that examples of inconsistency are abundant.

Theorem 14.19 (Consistency) *Let S be the support of the discrete part P_0^d of the probability measure $P_0 = P_0^c + P_0^d$. If P follows a species sampling process prior with PPF (p_j) satisfying, for nonnegative numbers $\alpha_n = O(1)$ and numbers $\delta_n = O(1)$,*

$$\sum_{j=1:\tilde{X}_j \in S}^{K_n} \left| p_j(N_n) - \frac{\alpha_n N_{j,n} + \delta_n}{n} \right| \to 0, \qquad \text{a.s. } [P_0^\infty], \tag{14.12}$$

$$\sum_{i=1}^{K_n} \sum_{j=1}^{K_n} \left| p_i(N_n) p_j(N_n^{i+}) - p_i(N_n) p_j(N_n) \right| \to 0, \qquad \text{a.s. } [P_0^\infty], \tag{14.13}$$

then the posterior distribution of P in the model $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ is strongly consistent at P_0 relative to the topology of pointwise convergence on bounded measurable functions if both $\alpha_n \to 1$ and $p_{K_n+1}(N_n) \to 0$ a.s. The latter two conditions are necessary if $P_0^d \neq 0$ or $G \neq P_0^c / \|P_0^c\|$, respectively; and equivalent if P_0 is discrete. Furthermore, if $\alpha_n \to \alpha$ and $p_{K_n+1}(N_n) \to \gamma$ and either $p_{K_n+2}(N_n^{(K_n+1)+}) \to \gamma$ or $\gamma = 0$, then the posterior distribution tends to $\alpha P_0^d + \beta P_0^c + \gamma G$, for $\beta = (1 - \alpha \|P_0^d\| - \gamma)/\|P_0^c\|$.

Proof For posterior consistency it is necessary and sufficient that the posterior mean of Pf tends to $P_0 f$ and the posterior variance to zero, for every bounded measurable function f (see Lemma 6.4). By conditioning on P and using the towering property of conditional expectation, the posterior mean and second moment can be expressed in the predictive distributions as $\mathrm{E}(Pf | X_{1:n}) = \mathrm{E}(f(X_{n+1}) | X_{1:n})$ and $\mathrm{E}((Pf)^2 | X_{1:n}) = \mathrm{E}(f(X_{n+2}) f(X_{n+1}) | X_{1:n})$. In conjunction with the predictive formula (14.9) this permits to express posterior mean and variance in the PPF of the species sampling model.

Observations X_i not in the support S of the discrete part P_0^d of P_0 appear only once. Thus they contribute distinct values $\tilde{X}_j$ of multiplicity $N_{j,n} = 1$ in $X_1, \ldots, X_n$, and can be viewed as sampled i.i.d. from the continuous part P_0^c of P_0 conditioned to S^c. The predictive probabilities $p_j(N_n)$ corresponding to these distinct values are all the same. Indeed by symmetry of the EPPF, the value of $p_j(\boldsymbol{n}) = p(\boldsymbol{n}^{j+})/p(\boldsymbol{n})$ (see (14.3)) depends on $\boldsymbol{n}$, but is the same for every j with $n_j = 1$: it does not matter which 1 is raised to 2 to obtain $\boldsymbol{n}^{j+}$.

For every given m, the fraction of values $X_1, \ldots, X_n$ landing on atoms θ_j with $j > m$ tends to $\sum_{j>m} W_j$ almost surely, by the law of large numbers. Therefore the number K_n^d of *distinct values* in $X_1, \ldots, X_n$ that belong to S is bounded above by $m + n \sum_{j>m} W_i (1 + o(1))$. As this is true for every m, it follows that $K_n^d = o(n)$ almost surely.

We have that $f(\tilde{X}_j) N_{j,n}$ is equal to $\sum_{i=1}^n f(X_i) \mathbb{1}\{X_i = \tilde{X}_j\}$, which sums to $n\mathbb{P}_n(f \mathbb{1}_S)$ when summed over j with $\tilde{X}_j \in S$, for $\mathbb{P}_n$ the empirical distribution of $X_1, \ldots, X_n$. Using this after applying (14.12), we see, with $\beta_n \geq 0$ the common value of $n p_j(N_n)$ for j such that $\tilde{X}_j \notin S$, and $\gamma_n = p_{K_n+1}(N_n)$,

$$\begin{aligned} \mathrm{E}[Pf | X_{1:n}] &= \sum_{j=1:\tilde{X}_j \in S}^{K_n} p_j(N_n) f(\tilde{X}_j) + \sum_{j=1:\tilde{X}_j \notin S}^{K_n} p_j(N_n) f(\tilde{X}_j) + p_{K_n+1}(N_n) G f \\ &= \alpha_n \mathbb{P}_n(f \mathbb{1}_S) + \delta_n o(1) + o(1) + \beta_n \mathbb{P}_n(f \mathbb{1}_{S^c}) + \gamma_n G f. \end{aligned}$$

Up to the $o(1)$ terms the right side is a mixture of three (sub)probability measures (where $\beta_n > 1$ is not a priori excluded). By the law of large numbers $\mathbb{P}_n(f\mathbb{1}_S) \to P_0^d f$ and $\mathbb{P}_n(f\mathbb{1}_{S^c}) \to P_0^c f$, almost surely, and hence the right side of the display reduces to $\alpha_n P_0^d f + \beta_n P_0^c f + \gamma_n Gf + o(1) + o(\beta_n)$, almost surely. If $\alpha_n \to 1$ and $\gamma_n \to 0$, then the first term tends to $P_0^d f$ and the third term to zero. Choosing $f \equiv 1$ in the display shows that the total mass of the measure remains one, whence $\beta_n \to 1$. We conclude that the limit is $P_0^d f + P_0^c f = P_0 f$, and hence consistency pertains.

Since only the first of the three measures can give rise to P_0^d, for consistency in the case that the latter measure is nonzero, it is necessary that $\alpha_n \to 1$. Similarly if G is not proportional to P_0^c, then consistency requires the third measure $\gamma_n G$ to disappear and hence $\gamma_n \to 0$.

If $\alpha_n \to \alpha$ and $\gamma_n \to \gamma$, then the total masses of the first and third measures tend to limits. Consequently, if $P_0(S^c) > 0$, then β_n must also converge to a limit; in the other case β_n is irrelevant for the limit of the mixture, as $\mathbb{P}_n(f\mathbb{1}_{S^c})$ vanishes for every n. The limit of the mixture takes the form $\alpha P_0^d + \beta P_0^c + \gamma G$, where the weight β can be identified from the fact that the mixture is a probability measure.

It remains to consider the posterior variance of Pf. The posterior second moment satisfies

$$\mathrm{E}[(Pf)^2 | X_{1:n}] = \int \mathrm{E}(f(X_{n+2})f(x) | X_{n+1} = x, X_{1:n})\, dP^{X_{n+1}|X_{1:n}}(x).$$

We use (14.9) twice, the second time with $n+1$ instead of n, to replace first the integrating measure and next the integrand by a mixture. We use that K_{n+1} is equal to K_n if $X_{n+1} = \tilde{X}_j$ for some j and equal to $K_n + 1$ otherwise, and apply (14.4) to combine the terms with coefficients $p_i(N_n)p_j(N_n^{i+})$ and $p_j(N_n)p_i(N_n^{j+})$ into a single one. Subtracting the square of the posterior mean we finally find that

$$\begin{aligned}
\mathrm{var}[Pf | X_{1:n}] = &\sum_{i=1}^{K_n}\sum_{j=1}^{K_n}\Big(p_i(N_n)p_j(N_n^{i+}) - p_i(N_n)p_j(N_n)\Big)f(\tilde{X}_i)f(\tilde{X}_j) \\
&+ 2\sum_{i=1}^{K_n}\Big(p_i(N_n)p_{K_n+1}(N_n^{i+}) - p_i(N_n)p_{K_n+1}(N_n)\Big)f(\tilde{X}_i)Gf \\
&+ \Big(p_{K_n+1}(N_n)p_{K_n+2}(N_n^{(K_n+1)+}) - p_{K_n+1}(N_n)^2\Big)(Gf)^2 \\
&+ p_{K_n+1}(N_n)p_{K_n+1}(N_n^{(K_n+1)+})G(f^2).
\end{aligned}$$

The first term on the right tends to zero by assumption (14.13). The second one also, as is seen by rewriting the probabilities $p_{k+1}(\boldsymbol{n})$ as $1 - \sum_{j=1}^k p_j(\boldsymbol{n})$. The third and fourth terms tend to zero if $p_{K_n+1}(N_n) \to 0$.

The third term also tends to zero if $p_{K_n+1}(N_n)$ and $p_{K_n+2}(N_n^{(K_n+1)+})$ have the same limit. By evaluating the display with $f \equiv 1$, we see that the fourth term can be written as minus the sum of the first three terms for $f = 1$. This shows that it also tends to zero. □

The product $p_i(N_n)p_j(N_n^{i+})$ in (14.13) is the probability that X_{n+1} is added to the ith set and next X_{n+2} to the jth set in the partitions generated by $X_1, X_2, \ldots$ Thus condition (14.13) requires that for large n two consecutive steps of the Markov chain of partitions are

almost independent. This seems reasonable as the change in the partition from time n to $n+1$ is minor.

Condition (14.12) with $\alpha_n \approx 1$ requires that the predictive probabilities settle down on the empirical frequencies. This seems natural, also in view of Theorem 14.13, which shows that this is true under the prior law. However, presently the convergence must take place under the "true" distribution of $X_1, X_2, \ldots$, whence the condition is independent of the latter theorem, and certainly not automatic. The condition is empty if P_0 is continuous.

The theorem shows that a species sampling process can be an appropriate prior (in that leads to posterior consistency) only if $p_{K_n+1}(N_n) \to 0$, except in the accidentally lucky case that the continuous component P_0^c is proportional to the prior center measure G. The convergence $p_{K_n+1}(N_n) \to 0$ requires that eventually new observations should not open new sets in the partition. In view of the form of the predictive distribution (or posterior mean) this is intuitively natural, as otherwise the prior center measure G does not wash out. In many examples $p_{K_n+1}(N_n)$ converges to a positive limit γ (or even $\gamma = 1$) if P_0 has a nonzero absolutely continuous component and hence leads to inconsistency (see the next sections and Problems 14.2 and 14.3). This includes the Pitman-Yor process, discussed in Section 14.4, whenever this does not coincide with the Dirichlet process.

14.2.2 Species Sampling Process Mixtures

Even though inconsistency can easily occur when using a species sampling process prior directly on the distribution of the observations, putting the prior on the mixing distribution of a kernel mixture will typically work. For a species sampling process with large weak support the Kullback-Leibler support of the corresponding kernel mixture prior contains many true densities, by Theorem 7.2. Hence a kernel mixture of a species sampling process prior will lead to posterior consistency for the weak topology, and under entropy conditions also to consistent density estimation, just as for the Dirichlet mixture process prior, considered in Chapter 5. Consistency under $\mathbb{L}_1$-norm can also be obtained if more structure is imposed; see Problem 14.16.

A generalization of the Gibbs sampling scheme for Dirichlet process mixtures, considered in Chapter 5, may be used for computing the posterior distribution. For families of probability densities $x \mapsto \psi_i(x;\theta_i)$ indexed by parameters (or latent variables) θ_i, we assume that P follows a species sampling process and

$$X_i \,|\, \theta_i \overset{\text{ind}}{\sim} \psi_i(\cdot;\theta_i), \qquad \theta_i \,|\, P \overset{\text{iid}}{\sim} P, \qquad i = 1, 2, \ldots$$

Assume that the species sampling process has an atomless center measure G (as usual) and PPF $\{p_j\}$. Then as in Theorem 5.3, which specializes to the Dirichlet process, the posterior distribution of the latent variables θ_i satisfies

$$\theta_i \,|\, \theta_{-i}, X_1, \ldots, X_n \sim \sum_{j=1}^{K_{-i}} q_{i,j}\delta_{\theta^*_{-i,j}} + q_{i,0}G_{b,i}, \tag{14.14}$$

where $\theta^*_{-i,j}$ is the jth distinct value in $\{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$, K_{-i} is the total number of distinct values and

$$q_{i,j} = cp_j(N_{-i,1}, \ldots, N_{-i,K_{-i}})\psi_i(X_i; \theta^*_{-i,j}), \qquad j \neq 0,$$
$$q_{i,0} = cp_{K_{-i}+1}(N_{-i,1}, \ldots, N_{-i,K_{-i}}) \int \psi_i(X_i; \theta)\, dG(\theta), \tag{14.15}$$

where $N_{-i,1}, \ldots, N_{-i,K_{-i}}$ are the multiplicities of the distinct values $\theta^*_{-i,j}$, c is chosen to satisfy $q_{i,0} + \sum_{j \neq i} q_{i,j} = 1$, and $G_{b,i}$ is the "baseline posterior measure" proportional to $\psi_i(X_i; \theta)\, dG(\theta)$.

14.3 Gibbs Processes

Partition models in which the EPPF depends multiplicatively on the sizes of the partitioning sets are attractive for their simplicity. Lemma 14.21 below shows that these are all of Gibbs type, as defined in the following definition.

Definition 14.20 (Gibbs process) A *Gibbs partition* of *type* $\sigma \in (-\infty, 1)$ is an infinite exchangeable random partition with EPPF of the form[7]

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)^{[n_j-1]}, \qquad n = \sum_{i=1}^{k} n_i, \tag{14.16}$$

where $V_{n,k}$, for $n = 1, 2, \ldots, k = 1, \ldots, n$, are nonnegative numbers satisfying the backward recurrence relation

$$V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \qquad V_{1,1} = 1. \tag{14.17}$$

The species sampling process corresponding to a Gibbs partition and an atomless measure G is called a *Gibbs process*.

For a given type σ, multiple arrays $(V_{n,k})$ will satisfy relation (14.17), and define different Gibbs processes. The Dirichlet processes form a family of Gibbs processes of type $\sigma = 0$, indexed by the precision parameter M (note that $1^{[n]} = n!$ and see (14.6) and/or apply the following lemma). Other prominent examples of Gibbs processes are the Pitman-Yor processes, discussed in Section 14.4. Mixtures of Gibbs processes of the same type are Gibbs processes of again the same type.

The recurrence relation (14.17) arises naturally, as it ensures that p given by (14.16) satisfies the consistency equation (14.2), and thus that p is indeed an EPPF. At first sight the defining relation (14.16) is odd, but it turns out to be the only possible form of a product partition model, apart from trivial partition models.

Lemma 14.21 (Product partition) *Any infinite exchangeable partition with EPPF of the form* $p(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} \rho(n_i)$ *for every composition* $(n_1, \ldots, n_k)$ *of* n*, for a given array of constants* $V_{n,k}$ *and a strictly positive, nonconstant function* ρ*, is a Gibbs partition.*

Proof Equation (14.2) applied to the product representation $p(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} \rho(n_i)$ can be rearranged to give $V_{n,k} = V_{n+1,k} \sum_{j=1}^{k} \rho(n_j + 1)/\rho(n_j) +$

[7] Recall that $a^{[n]} = a(a+1)\cdots(a+n-1)$ denotes the ascending factorial, and $a^{[0]} = 1$.

$\rho(1)V_{n+1,k+1}$. Set $f(n) = \rho(n+1)/\rho(n)$ and apply this equation with $k = 2$ to find that $V_{n,2} = V_{n+1,2}\big[f(n_1)+f(n_2)\big]+\rho(1)V_{n+1,3}$, for every partition $n = n_1+n_2$. If $V_{n+1,2} > 0$ for every n, then we can conclude that the sum $f(n_1) + f(n_2)$ depends on $n_1 + n_2$ only, whence $f(n+1)-f(n) = f(m+1)-f(m)$, for every m, n. This implies that $f(n) = bn-a$, for some $b > 0$ and $a < b$ (as ρ is strictly positive and nonconstant), and hence

$$\rho(n) = \rho(1)\prod_{i=1}^{n-1} f(i) = \rho(1)(b-a)(2b-a)\cdots((n-1)b-a) = \rho(1)b^{n-1}(1-a/b)^{[n-1]}.$$

Substituting this in the product formula for $p(n_1, \ldots, n_k)$ we see that this is of Gibbs form, with $V_{n,k}$ taken equal to $V_{n,k}b^{n-k}\rho(1)^k$ and $\sigma = a/b$.

If $V_{n,2} = 0$ for some $n \geq 2$, then $p(n_1, n_2) = 0$ for every partition $n = n_1 + n_2$ of n. This implies that a partition of $\mathbb{N}_n$ in more than one block is impossible. The consistency of the exchangeable partition shows then this is true for every $n \geq 1$, which indicates that the species sampling model can have one species only. This is the Gibbs partition with $V_{n,k} = 0$ for $k \geq 2$, and $V_{n,1} = (1-\sigma)^{[n-1]}$, for any $\sigma < 1$. □

From (14.3) the PPF of a Gibbs process is obtained as

$$p_j(n_1, \ldots, n_k) = \begin{cases} \dfrac{V_{n+1,k}}{V_{n,k}}(n_j - \sigma), & j = 1, \ldots, k, \\ \dfrac{V_{n+1,k+1}}{V_{n,k}}, & j = k+1. \end{cases} \tag{14.18}$$

Thus the probability of discovering a new species after observing K_n different species in n individuals is $V_{n+1,K_n+1}/V_{n,K_n}$. This depends on n and the total number K_n of observed species, but not on their multiplicities $N_{1,n}, \ldots, N_{K_n,n}$. Ignoring the multiplicities and bringing to bear only K_n may be considered a deficiency of the Gibbs model, but also a reasonable compromise between flexibility and tractability. In terms of flexibility it is still a step up from the Dirichlet process DP(MG) (which is a Gibbs process of type $\sigma = 0$) for which the probability of observing a new species is $M/(M+n)$, and hence depends only on n. The following proposition shows that the latter property characterizes the Dirichlet process within the Gibbs processes.

Proposition 14.22 *In a nontrivial Gibbs process, the probability $p_{k+1}(n_1, \ldots, n_k)$ of discovering a new species depends only on the total number of individuals $n = \sum_i n_i$, for all $n = 1, 2, \ldots$ and $k = 1, 2, \ldots n$, if and only if the partition corresponds to the Dirichlet process.*

Proof For a Dirichlet process, the probability of observing a new species is $M/(M+n)$ and hence depends on n only. Conversely, by (14.18) (or (14.17)) the probability of a new species in a Gibbs partition satisfies $V_{n+1,k+1}/V_{n,k} = 1-(n-\sigma k)V_{n+1,k}/V_{n,k}$ and hence does not depend on k if and only if $V_{n+1,k}/V_{n,k} = c_n/(n-\sigma k)$ for some sequence c_n depending only on n, in which case $V_{n+1,k+1}/V_{n,k} = 1-c_n$. Applying the latter two equations, also with n replaced by $n+1$ and k by $k+1$, we can rewrite the obvious identity

$$\frac{V_{n+2,k+1}}{V_{n+1,k+1}}\frac{V_{n+1,k+1}}{V_{n,k}} = \frac{V_{n+2,k+1}}{V_{n+1,k}}\frac{V_{n+1,k}}{V_{n,k}}$$

in the form $\sigma\big[k(d_{n+1}-d_n)+d_{n+1}\big] = (n+1)d_{n+1} - nd_n$, for $d_n = (1-c_n)/c_n$. This can be true for all $k = 1, \ldots n$ and $n = 1, 2, \ldots$ only if either $d_{n+1} = d_n$ or $\sigma = 0$. In the first case the equation implies $\sigma = 1$, which is excluded by definition, or $d_n = 0$ for every n, which implies $V_{n,k} = 0$ for all $n, k \geq 2$, so that the partition is the trivial one in one block. In the second case we find $d_{n+1} = d_n n/(n+1)$. Iterating the latter equation shows that $d_n = d_1/n$ or $c_n = n/(d_1+n)$. Thus the PPF of the partition is that of a Dirichlet process with $M = d_1$, and the proposition follows by application of Proposition 14.9. □

The following proposition shows that, given the number of sets K_n in a Gibbs partition, the distribution of the composition of the partition is free of the $V_{n,k}$ parameters, but it does depend on the type parameter σ. In other words, in a Gibbs process of known type the number of sets K_n is a sufficient statistic.

The first formula in the following proposition generalizes (4.16) for a Dirichlet process.

Proposition 14.23 (Number of species) *The distribution of the number K_n of sets in a Gibbs partition $\mathcal{P}_n$ and the partition probabilities conditioned on this number are given by*

$$\mathrm{P}(K_n = k) = B_{n,k} V_{n,k},$$

$$\mathrm{P}(\mathcal{P}_n = \{A_1, \ldots, A_k\} | \, K_n = k) = \frac{1}{B_{n,k}} \prod_{j=1}^{k} (1-\sigma)^{[|A_j|-1]},$$

where $\sigma^k B_{n,k}$ is a generalized factorial coefficient*, given by, with the sum over all compositions $(n_1, \ldots, n_k)$ of n of size k,*

$$B_{n,k} = \sum_{(n_1,\ldots,n_k)} \frac{n!}{k! \prod_{j=1}^{k} n_j!} \prod_{j=1}^{k} (1-\sigma)^{[n_j-1]} = \frac{1}{\sigma^k k!} \sum_{j=1}^{k} (-1)^j \binom{k}{j} (-j\sigma)^{[n]}.$$

Proof The second formula follows from the formula for conditional probability, the first formula, and expression (14.16) for the EPPF. Proving the first formula is equivalent to verifying the formula for $B_{n,k}$. The first expression for $B_{n,k}$ follows from substituting (14.16) in the general formula for $\mathrm{P}(K_n = k)$ given in the second part of Proposition 14.5. To see that this is identical to the second expression we consider the generating function of the sequence $\sigma^k B_{n,k}/n!$, for $n = 1, 2, \ldots$, for a fixed value of k. Since $k \leq n$, the first $k-1$ values of this sequence vanish, and hence the generating function at argument s is given by

$$\frac{(-1)^k}{k!} \sum_{n=k}^{\infty} s^n \sum_{n_1+\cdots+n_k=n} \prod_{j=1}^{k} \frac{(-\sigma)^{[n_j]}}{n_j!} = \frac{(-1)^k}{k!} \prod_{j=1}^{k} \sum_{n_j=1}^{\infty} \frac{(-\sigma)^{[n_j]} s^{n_j}}{n_j!}.$$

In view of the identity $1 - (1-s)^t = -\sum_{m=1}^{\infty} (-t)^{[m]} s^m/m!$, the last display is equal to

$$\frac{1}{k!}\big[1-(1-s)^{\sigma}\big]^k = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (1-s)^{j\sigma} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \sum_{n=0}^{\infty} \frac{(-j\sigma)^{[n]} s^n}{n!},$$

by the binomial formula and a second application of the identity, in the other direction. Exchanging the order of summation and comparing the coefficient of s^n to $\sigma^k B_{n,k}/n!$ gives the desired identity. □

The type parameter σ has a key role in determining the structure of Gibbs processes. One may discern three different regimes.

Proposition 14.24 *The following characterizations of Gibbs processes hold.*

(i) *Gibbs partitions of type* $\sigma < 0$ *are mixtures over* m *of finite-dimensional Dirichlet* $\mathrm{Dir}(m; -\sigma, \dots, -\sigma)$ *partitions (also called "Fisher" or "two-parameter partitions"* $(\sigma, -\sigma m)$ *and discussed in Example 14.17 and Section 14.4).*
(ii) *Gibbs partitions of type* $\sigma = 0$ *are mixtures over* M *of Dirichlet* $\mathrm{DP}(MG)$ *partitions.*
(iii) *Gibbs partitions of type* $0 < \sigma < 1$ *are Poisson-Kingman* $\mathrm{PK}(\rho_\sigma, \eta)$ *partitions for* ρ_σ *a* σ*-stable Lévy measure (equivalently mixtures over* y *of* $\mathrm{PK}(\rho_\sigma \,|\, y)$ *partitions), as discussed in Section 14.5 (and Example 14.43).*

Proof See Gnedin and Pitman (2006), Theorem 12. □

Here, "mixture" refers to a decomposition $p(\boldsymbol{n}) = \int p(\boldsymbol{n}|\, m)\, \pi(dm)$ of the EPPF p of the partition into the EPPFs $p(\cdot|\, m)$ of the basic partitions, relative to some mixing measure π. Equivalently, the representing species sampling process (in particular its weight sequence) is a mixture of the species sampling processes of the basic partitions. In (i) and (ii) "Dirichlet partition" is short for the partition generated by a random sample $X_1, X_2, \dots$ from the random Dirichlet distribution or process. In (ii) the partition is the Chinese restaurant process, whereas in (i) the Dirichlet distribution $\mathrm{Dir}(m; -\sigma, \dots, -\sigma)$ is understood as a model for a random probability distribution on the finite set $\{1, 2, \dots, m\}$ and $X_1, X_2, \dots$ are variables with values in this finite set. This model can also be framed as the species sampling model with the infinite weight sequence whose first m elements are distributed according to $\mathrm{Dir}(m; -\sigma, \dots, -\sigma)$ and whose remaining elements are zero. More formally, a Gibbs partition of type (i) can be described as the random partition generated by a species sampling model with weight sequence $(W_1, W_2, \dots)$ satisfying

$$(W_1, \dots, W_m, W_{m+1}, \dots)|\, m \sim \mathrm{Dir}(m; -\sigma, \dots, -\sigma) \times \delta_0^\infty, \qquad m \sim \pi. \tag{14.19}$$

Here π can be any probability measure on $\mathbb{N}$.

The three types can be differentiated by the distribution of the number of distinct species K_n among the first n individuals. In the basic type (i) the total number of species is m and hence K_n is bounded; mixing over m relieves this restriction. The number of species in a type (ii) Gibbs processes (i.e. the number of tables in the Chinese restaurant process) is of the order $\log n$ by Proposition 4.8. Finally, the number of distinct species in a Gibbs partition of type (iii) is of the order n^σ (see Theorem 14.50). This has obvious consequences for modeling purposes. If the number of clusters is a priori thought to be large, then type (iii) partitions may be more useful than types (i)–(ii). In contrast, type (i) is natural as a model for discovery of a finite number of species, where moderate mixing over m allows an a priori unknown total number.[8]

[8] For instance, for a type (i) Gibbs process with mixing measure $\pi(j) = \gamma(1-\gamma)^{[j-1]}/j!$, for $j = 1, 2, \dots$ and some $\gamma \in (0, 1)$, the total number of species K_n remains finite a.s., but its limit has infinite expectation; see Gnedin (2010).

Gibbs processes of positive type admit an explicit stick-breaking representation with *dependent* stick-breaking variables.

Theorem 14.25 (Stick-breaking) *A Gibbs process of type $0 < \sigma < 1$ and atomless center measure G can be represented as $P = \sum_{i=1}^{\infty} \tilde{W}_i \delta_{\theta_i}$, for $\theta_1, \theta_2, \ldots \overset{iid}{\sim} G$, and $\tilde{W}_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, where $V_i \,|\, V_1 = v_1, \ldots, V_{i-1} = v_{i-1}$ has density function on $(0, 1)$ given by*

$$v_i \mapsto \frac{\sigma/\Gamma(1-\sigma)}{v_i^{\sigma} \prod_{j=1}^{i-1}(1-v_j)^{\sigma}} \frac{\int_0^{\infty} y^{-i\sigma} f_{\sigma}(y \prod_{j=1}^{i}(1-v_j))/f_{\sigma}(y)\, d\eta(y)}{\int_0^{\infty} y^{-(i-1)\sigma} f_{\sigma}(y \prod_{j=1}^{i-1}(1-v_j))/f_{\sigma}(y)\, d\eta(y)}.$$

Here f_{σ} is the density of a positive σ-stable random variable (with Laplace transform equal to $e^{-\lambda^{\sigma}}$), and η is the mixing measure appearing in Proposition 14.24 (iii).

Proof In view of Proposition 14.24, a Gibbs process with $0 < \sigma < 1$ is a Poisson-Kingman $\mathrm{PK}(\rho_{\sigma}, \eta)$ process, for ρ_{σ} a σ-stable Lévy measure (see Example 14.43). The result is a special case of Theorem 14.49. □

Gibbs process priors, and species sampling models in general, are infinite series priors, which were seen to have full weak support under mild conditions in Lemma 3.5. These conditions are satisfied by most Gibbs processes. In the following theorem we assume that the sample space is a Polish space.

Theorem 14.26 (Support) *The weak support of a Gibbs process prior with type-parameter $\sigma \geq 0$ and center measure G is equal to $\{P \in \mathfrak{M}: \mathrm{supp}(P) \subset \mathrm{supp}(G)\}$. The same is true for type $\sigma < 0$, provided the mixing distribution π in* (14.19) *has unbounded support in $\mathbb{N}$.*

Proof Without loss of generality we may assume that G has full support; otherwise we redefine the sample space. By Proposition 14.24 the Gibbs prior is a mixture. In the cases $\sigma = 0$ and $0 < \sigma < 1$ the mixture components are the Dirichlet process and the Poisson-Kingman processes $\mathrm{PK}(\rho_{\sigma} \,|\, y)$, for ρ_{σ} a normalized σ-stable measure. These can each be seen to have full support by Lemma 3.5, and this is then inherited by the mixture. If $\sigma < 0$ the Gibbs process is a mixture of finite-dimensional Dirichlet distributions of varying dimension. The full support follows again from Lemma 3.5, provided this dimension is unbounded under the mixing distribution (the prior π (14.19)). □

The following theorem specializes Theorem 14.19 for posterior consistency to Gibbs processes. The main requirement that the probability of a new species tends to zero translates into convergence of the sequence $V_{n+1,K_n+1}/V_{n,K_n}$ to zero. The posterior limit is further dependent on the limit of K_n/n, which is the total mass of the continuous part of P_0 and can be anywhere in the interval $[0, 1]$.

Theorem 14.27 (Consistency) *If P follows a Gibbs process prior with coefficients such that both $V_{n+1,K_n+1}/V_{n,K_n} \to \gamma$ and $V_{n+2,K_n+2}/V_{n+1,K_n+1} \to \gamma$ almost surely, for some $0 < \gamma \leq 1$, or $V_{n+1,K_n+1}/V_{n,K_n} \to 0$, then the posterior distribution of P in the*

model $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ *converges almost surely under* P_0 *relative to the weak topology to* $\alpha P_0^d + \beta P_0^c + \gamma G$, *for* $\alpha = (1-\gamma)/(1-\sigma\|P_0^c\|)$ *and* $\beta = (1-\gamma)(1-\sigma)/(1-\sigma\|P_0^c\|)$. *In particular, unless* P_0^c *is proportional to* G, *the posterior distribution is consistent if and only if* $\gamma = 0$ *and one of the three possibilities is true:* $\sigma = 0$ *or* P_0 *is discrete or* P_0 *is atomless.*

Proof Observations $X_1, \ldots, X_n$ not in the support of the discrete part of P_0 occur only once, and hence contribute a count of 1 to the total number of distinct values. As shown in the proof of Theorem 14.19 observations from the discrete part contribute $o(n)$ distinct values. Therefore $K_n/n \to \|P_0^c\| =: \xi$, almost surely.

By (14.18) $p_{K_n+1}(N_n) = V_{n+1,K_n+1}/V_{n,K_n}$, which tends to γ by assumption. Furthermore $p_j(N_n) = (\alpha_n N_{j,n} + \delta_n)/n$, for $\alpha_n = nV_{n+1,K_n}/V_{n,K_n}$ and $\delta_n = -\sigma\alpha_n$, where by (14.17) (or the fact that the p_j add up to 1) $(n - \sigma K_n)V_{n+1,K_n}/V_{n,K_n} = 1 - V_{n+1,K_n+1}/V_{n,K_n} \to 1 - \gamma$, so that $\alpha_n \to (1-\sigma\xi)^{-1}(1-\gamma)$. We conclude that condition (14.12) of Theorem 14.19 is satisfied, trivially with the right side equal to 0.

In view of (14.18), the left side of (14.13) is equal to

$$\sum_{i=1}^{K_n} p_i(N_n) \sum_{j=1}^{K_n} \left| \frac{V_{n+2,K_n}}{V_{n+1,K_n}}(N_{j,n} + \mathbb{1}\{i=j\} - \sigma) - \frac{V_{n+1,K_n}}{V_{n,K_n}}(N_{j,n} - \sigma) \right|$$
$$\leq \left| \frac{V_{n+2,K_n}}{V_{n+1,K_n}} - \frac{V_{n+1,K_n}}{V_{n,K_n}} \right| (n + K_n\sigma) + \frac{V_{n+2,K_n}}{V_{n+1,K_n}}.$$

As noted the assumptions imply that the three quotients in the right side are asymptotically equivalent to $(1-\gamma)/(n-\sigma K_n)$, where $n - \sigma K_n \sim n(1-\sigma\xi)$, almost surely. It follows that the expression tends to zero.

For the final assertion we solve the equation $\alpha P_0^d + \beta P_0^c + \gamma G = P_0$. If the measure G, which is continuous by assumption, is not proportional to P_0^c, then this implies $\gamma = 0$. Next if both the discrete and continuous parts of P_0 are nonzero, the equation is valid if and only if $\alpha = \beta = 1$, which is true if and only if $\sigma = 0$. If P_0^d vanishes, then $\xi = 1$ and $\beta = 1$, and the equation is valid. Finally, if P_0^c vanishes, then $\xi = 0$ and $\alpha = 1$ and the equation is valid. □

It is remarkable that if P_0 possesses both a discrete and a continuous component, then consistency under the conditions of the preceding theorem pertains only if $\sigma = 0$. Otherwise (if $\gamma = 0$) the limit measure is a mixture of the discrete and continuous components of P_0, but with incorrect weights.

The following result specializes to the case of Gibbs process of type $\sigma < 0$. Consistency then depends on the mixing distribution π in (14.19). For most distributions the probabilities $\pi(j)$ are decreasing for sufficiently large j; this is sufficient for consistency at discrete distributions P_0. Faster decrease guarantees consistency also at continuous distributions.

Lemma 14.28 (Consistency, negative type) *If* P *follows a Gibbs process prior of type* $\sigma < 0$ *with fully supported mixing measure* π *as in* (14.19), *then the posterior distribution based on* $X_1 \ldots, X_n | P \overset{iid}{\sim} P$ *is consistent relative to the weak topology at any discrete* P_0 *if*

$\pi(j+1) \le \pi(j)$, for all sufficiently large j. Furthermore, it is consistent at any atomless P_0 if $\pi(j+1) \lesssim j^{-1}\pi(j)$, for all sufficiently large j.

Proof It suffices to show that the limit γ in Theorem 14.27 is equal to 0. The coefficients of the Gibbs processes corresponding to sampling from a $\mathrm{Dir}(m; -\sigma, \ldots, -\sigma)$-distribution can be written in the form,

$$V_{n,k}^m = |\sigma|^{k-n} \frac{(m-1)(m-2)\ldots(m-k+1)}{(m+s)(m+2s)\cdots(m+(n-1)s)},$$

where $s = -1/\sigma$. The coefficients with $k > m$ are understood to vanish (as also follows from the formula). The parameter m in the expression for the corresponding EPPF is attached only to $V_{n,k}^m$, and hence the mixture over m relative to the prior π is Gibbs with coefficients $V_{n,k} = \sum_{m \ge k} \pi(m) V_{n,k}^m$. We need to consider the quotient

$$\frac{V_{n+1,k+1}}{V_{n,k}} = \frac{\sum_{m \ge k+1} \pi(m) V_{n+1,k+1}^m}{\sum_{m \ge k} \pi(m) V_{n,k}^m} = \frac{\sum_{m \ge k} \pi(m+1) V_{n+1,k+1}^{m+1}}{\sum_{m \ge k} \pi(m) V_{n,k}^m}.$$

By a little algebra

$$\frac{V_{n+1,k+1}^{m+1}}{V_{n,k}^m} = \frac{m}{m+1+ns} \prod_{i=1}^{n-1} \frac{m+is}{m+1+is}.$$

This quotient is smaller than 1 for every m, as all terms of the product are, and bounded by $c/(c+s)$ if $m \le cn$.

If $\pi(m+1) \lesssim m^{-1}\pi(m)$, then we bound $\pi(m+1)V_{n+1,k+1}^{m+1}$ by $m^{-1}\pi(m)V_{n,k}^m$, and it immediately follows that $V_{n+1,k+1}/V_{n,k} \lesssim 1/k$. In the case of an atomless P_0, the number of distinct observations K_n is equal to n and tends to infinity, whence the proof of the second assertion of the lemma is complete.

If only $\pi(m+1) \lesssim \pi(m)$, then we can still replace the terms $\pi(m+1)V_{n+1,k+1}^{m+1}$ for $m \le cn$ by $c/(c+s)$ times $\pi(m)V_{n,k}^m$, where $c/(c+s)$ can be made arbitrarily small by choice of a small c. Arguing similarly, we see that the first part of the series can be made arbitrarily small and it only remains to prove that $\sum_{m \ge cn} \pi(m) V_{n,K_n}^m / V_{n,K_n} \to 0$, for every $c > 0$. Now $V_{n,k} \ge \pi(k) V_{n,k}^k$ and, again by some algebra

$$\frac{V_{n,k}^m}{V_{n,k}^k} = \binom{m}{k} \prod_{i=0}^{n-1} \frac{k+is}{m+is}.$$

The binomial coefficient is bounded by $(me/k)^k$. The terms of the product are increasing in i, and bounded by 1 for $m \ge k$. Hence the product of the first $2k$ terms is bounded above by $(k+2ks)^{2k}/(m+2ks)^{2k}$, which we combine with the bound on the binomial coefficient, and the product of the remaining $n-2k$ terms is bounded above by the product of any number of these terms. Under the assumption that $n \ge 4k-1$, we can use the two remaining terms given by $i = 3k$ and $i = 4k$, and arrive at the bound

$$\sum_{m \ge cn} \pi(m) \frac{V_{n,k}^m}{V_{n,k}} \lesssim \sum_{m \ge cn} \frac{V_{n,k}^m}{V_{n,k}^k} \le \sum_{m \ge cn} \left[\frac{me(k+2ks)^2}{k(m+2ks)^2}\right]^k \frac{k+3ks}{m+3ks} \frac{k+4ks}{m+4ks}.$$

If $k = o(n)$, then the expression in square brackets tends to zero, uniformly in $m \geq cn$, and the right side can be bounded by a multiple of $r^k k^2 \sum_{m \geq cn} m^{-2}$, for some $r < 1$, which tends to zero as $n \to \infty$. If P_0 is discrete, then this suffices, as $K_n/n \to 0$ almost surely, in this case. □

Example 14.29 (Geometric) For the geometric distribution π conditioned to be positive the sequence $\pi(j+1)/\pi(j)$ tends to a positive limit smaller than 1. This ensures consistency at every discrete P_0, but not at continuous P_0. By directly using Theorem 14.27 the posterior can be shown to be inconsistent at atomless true distributions (see Problem 14.3).

Example 14.30 (Poisson) The Poisson distribution π conditioned to be positive (i.e. $\pi(j) = (1-e^{-\lambda})^{-1} e^{-\lambda} \lambda^j / j!$, for $j = 1, 2, \ldots$) satisfies $\pi(j+1)/\pi(j) = \lambda/(j+1) \lesssim j^{-1}$. This ensures consistency at every discrete or continuous P_0.

14.4 Pitman-Yor Process

The Pitman-Yor processes form an important subclass of Gibbs processes, which share and generalize essential properties of the Dirichlet process.

Definition 14.31 (Pitman-Yor process, two-parameter family) The *Pitman-Yor partition* or *two-parameter family* is the Gibbs process of type σ with

$$V_{n,k} = \frac{\prod_{i=1}^{k-1} (M + i\sigma)}{(M+1)^{[n-1]}}. \tag{14.20}$$

The parameters σ and M are restricted to either ($\sigma < 0$ and $M \in \{-2\sigma, -3\sigma, \ldots\}$) or ($\sigma \in [0, 1)$ and $M > -\sigma$). The distribution of the weight sequence $(W_1, W_2, \ldots)$ of the corresponding species sampling model, listed in size-biased order, is known as the *Pitman-Yor distribution* and the corresponding random measure as the *Pitman-Yor process* $\mathrm{PY}(\sigma, M, G)$, where G is the center measure.

The relation (14.20) may seem odd. It turns out to be the only possibility so that the numbers $V_{n,k}$ factorize over n and k.

Lemma 14.32 *Any array of nonnegative numbers $V_{n,k}$ that satisfies the backward recurrence relation* (14.17) *with $V_{1,1} = 1$ and factorize as $V_{n,k} = V_k / c_n$, for a nonnegative sequence $V_1, V_2, \ldots$ with $V_2 > 0$ and a sequence of positive numbers $c_1, c_2, \ldots$, is of the form* (14.20) *for numbers $\sigma < 1$ and M such that either ($\sigma < 0$ and $M \in \{-2\sigma, -3\sigma, \ldots\}$) or ($\sigma \in [0, 1)$ and $M > -\sigma$).*

Proof For $V_{n,k}$ that factorize as given, relation (14.17) can be written in the form $V_{k+1} = V_k(c_{n+1}/c_n - n + \sigma k)$, for $k \leq n$ and $n = 1, 2, \ldots$ This shows that either $V_k > 0$ for all k or there exists k_0 with $V_k > 0$ for $k \leq k_0$ and $V_k = 0$ for $k > k_0$. The assumption $V_{1,1} = 1$ implies $V_1 > 0$, which shows that k_0 solves $c_{n+1}/c_n - n + \sigma k_0 = 0$. In both cases the equation can be rearranged to $V_{k+1}/V_k - \sigma k = c_{n+1}/c_n - n$, for $k \leq k_0 \leq n$, where the left side is constant in n and the right side constant in k. Thus the two sides have a common value

M, and hence $V_{k+1} = V_k(M+\sigma)$ and $c_{n+1} = c_n(M+n)$, for $k \le k_0 \le n$ and $n = 1, 2, \ldots$ Iterating these equations gives $V_{k+1} = V_1 \prod_{i=1}^{k}(M+i\sigma)$ and $c_{n+1} = c_1(M+1)^{[n]}$, and hence relation (14.17), since $V_1/c_1 = V_{1,1} = 1$, by assumption. If $V_k > 0$ for all $k \ge 1$, then $M + i\sigma > 0$ for every $i \in \mathbb{N}$, which forces $\sigma \ge 0$ and $M > -\sigma$. In the other case, that there exists k_0 with $V_k > 0$ for $k \le k_0$ and $V_k = 0$ for $k > k_0$, we have already seen that $M = -\sigma k_0$. The assumption $V_2 > 0$ forces $k_0 \ge 2$ and $M + \sigma > 0$ and hence $\sigma < 0$, as $M + \sigma = \sigma(1 - k_0)$. □

Besides "two-parameter" and "Pitman-Yor," various other names are attached to the process. First the process is also referred to as the *two-parameter Poisson-Dirichlet model* and as the *Ewens-Pitman model*, with as special cases:

(i) For $\sigma < 0$ the model is also known as the *Fisher model*. For $M = -m\sigma$ the exchangeable random partition is then generated by a random sample $X_1, X_2, \ldots$ from the $\mathrm{Dir}(m; -\sigma, \ldots, -\sigma)$ distribution, and hence eventually will consist of m sets (or "species").

(ii) For $\sigma = 0$ the model is generated by a sample from the Dirichlet process $\mathrm{DP}(MG)$, whence the underlying partition is the Chinese restaurant process, which is also known as the *Ewens model*.

Proofs of these statements can be based on Theorem 14.33 below, or deduced from the resulting expression for the EPPF. Second the Pitman-Yor model is also a Poisson-Kingman model obtained from a σ-stable subordinator or generalized gamma process; see Section 14.5.

The EPPF and PPF of this model are given by, with $n = \sum_{j=1}^{k} n_j$,

$$p(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1}(M+i\sigma)}{(M+1)^{[n-1]}} \prod_{j=1}^{k} (1-\sigma)^{[n_j-1]},$$

$$p_j(n_1, \ldots, n_k) = \begin{cases} (n_j - \sigma)/(M+n), & j = 1, \ldots, k, \\ (M + k\sigma)/(M+n), & j = k+1. \end{cases}$$

This follows from the definition and the expression (14.18) for the PPF of a Gibbs process. For $\sigma = 0$ we recognize the PPF of the Dirichlet process.

The following theorem describes the (size-biased) weight sequence $(\tilde{W}_1, \tilde{W}_2, \ldots)$ of a Pitman-Yor process explicitly through stick breaking (3.2). It generalizes Theorem 4.12 for the Dirichlet process (which is the special case $\sigma = 0$). The stick-breaking algorithm is also called a *residual allocation model* (RAM) in this context.[9]

Theorem 14.33 (Stick-breaking) *Let* $V_j \overset{ind}{\sim} \mathrm{Be}(1-\sigma, M+j\sigma)$ *and set* $\tilde{W}_j = V_j \prod_{l=1}^{j-1}(1-V_l)$, *for* $j = 1, 2, \ldots$

(i) *If* $\sigma \ge 0$ *and* $M > -\sigma$, *then the sequence* $(\tilde{W}_1, \tilde{W}_2, \ldots)$ *is in size-biased order and possesses the Pitman-Yor distribution.*

[9] The independence of the allocations distinguishes the Pitman-Yor process from other species sampling processes: the Pitman-Yor process is the only species sampling process for which the size-biased random permutation of the atoms admit independent relative stick lengths; see Pitman (1996a).

(ii) *If $\sigma < 0$ and $M = -m\sigma$, then the vector $(\tilde{W}_1, \ldots, \tilde{W}_{m-1}, 1 - \sum_{j<m} \tilde{W}_j)$ is in size-biased order and this vector augmented with zeros possesses the Pitman-Yor distribution.*

Proof (i). Since $\log(1-x) \le -x$, for $x < 1$, we have $\sum_j \log \mathrm{E}(1 - V_j) \le -\sum_j (1 - \sigma)/(M + 1 + (j-1)\sigma) = -\infty$. Therefore, the sequence $(\tilde{W}_j)$ is a proper probability vector, by Lemma 3.4.

To show that the sequence $(\tilde{W}_j)$ is in size-biased order it suffices to verify that the measure λ_k given by $\lambda_k(A) = \mathrm{E}\mathbb{1}_A(\tilde{W}_1, \ldots, \tilde{W}_k) \prod_{j=2}^k (1 - \sum_{i<j} \tilde{W}_i)$, given in Lemma 14.15, is permutation symmetric, for every k. Since $(\tilde{W}_1, \ldots, \tilde{W}_k)$ is a transformation of $(V_1, \ldots, V_k)$, we can express the expectation in the definition of λ_k as an integral relative to the density g_k of $(V_1, \ldots, V_k)$ and next transform into an integral over $(w_1, \ldots, w_k) \in A$ by substituting the inverse stick-breaking map $v_1 = w_1$, $v_2 = w_2/(1 - v_1)$, $v_3 = w_3/(1 - w_1 - w_2)$, etc. (Note that the remaining stick lengths satisfy $1 - \sum_{i<j} w_i = \prod_{i<j}(1 - v_i)$.) The Jacobian $|\partial v/\partial w|$ of this transformation cancels the factor $\prod_{j=2}^k (1 - \sum_{i<j} w_j)$ and hence

$$\lambda_k(A) = \int_A g_k\Big(w_1, \frac{w_2}{1 - w_1}, \ldots, \frac{w_k}{1 - \sum_{i<k} w_i}\Big)\, dw_1 \cdots dw_k.$$

Evaluating g_k as a product of the given beta densities reduces the integrand to a multiple of $(w_1 w_2 \cdots w_k)^{-\sigma}(1 - w_1 - w_2 - \cdots - w_k)^{M + k\sigma - 1}$, which indeed is symmetric in $(w_1, \ldots, w_k)$. Thus λ_k is symmetric and hence the distribution of $(\tilde{W}_j)$ is in size-biased order, by Lemma 14.15.

Using Corollary 14.14 we can now compute the EPPF as

$$p(n_1, \ldots, n_k) = \mathrm{E} \prod_{j=1}^k \tilde{W}_j^{n_j - 1} \prod_{j=2}^k \Big(1 - \sum_{i<j} \tilde{W}_i\Big) = \prod_{l=1}^k \mathrm{E}\big[(1 - V_l)^{\sum_{j=l+1}^k n_l} V_l^{n_l - 1}\big].$$

The right side can be evaluated with the help of Corollary G.4 to give (after a little algebra on the factorials), the EPPF in the Gibbs form of Definition 14.31.

(ii). Again it suffices to show that the measure λ_k is symmetric for all k. For $k > m$ the measure vanishes, as $1 - \sum_{i=1}^m \tilde{W}_i = 0$, by construction. For $k < m$ the proof given under (i) is valid. Only the case $k = m$ needs further consideration. For any measurable sets $A_1, \ldots, A_m$,

$$\lambda_k(A_1 \times \cdots \times A_m) = \int_{\substack{w_i \in A_i, i<m \\ 1 - \sum_{i<m} w_i \in A_m}} g_{m-1}\Big(w_1, \frac{w_2}{1 - w_1}, \ldots, \frac{w_{m-1}}{1 - \sum_{i<m-1} w_i}\Big) \times (1 - \sum_{i<m} w_i)\, dw_1 \cdots dw_{m-1}.$$

Unlike in the preceding proof, the factor $1 - \sum_{i<m} w_i$ from the definition of λ_m has not cancelled, because the integral only involves $m - 1$ variables and the inverse Jacobian is $\prod_{j=2}^{m-1}(1 - \sum_{i<j} w_j)$. However, since $M = -m\sigma$, the density g_{m-1} evaluated at the relative stick lengths now takes the form $(w_1 w_2 \cdots w_{m-1})^{-\sigma}(1 - w_1 - w_2 - \cdots - w_{m-1})^{-\sigma - 1}$, and

hence the integrand reduces to $(w_1 w_2 \cdots w_{m-1} w_m)^{-\sigma}$, for $w_m = 1 - \sum_{i<m} w_i$. Thus λ_m is the $\mathrm{Dir}(m; 1-\sigma, \ldots, 1-\sigma)$-distribution and hence is symmetric.

The EPPF can be derived as under (i). □

The following propositions describe the mean and covariances and an integral transform of integrals $\int \psi \, dP$ of functions with respect to a random probability measure P distributed according to a Pitman-Yor process. In particular, they give expressions for means and covariances of probabilities of sets. The assertions of the first proposition reduce to those given for the Dirichlet process in Propositions 4.2 and 4.3 as $\sigma \to 0$.

Proposition 14.34 (Moments) *If $P \sim \mathrm{PY}(\sigma, M, G)$, then for any real-valued, bounded, measurable functions ϕ, ψ,*

$$\begin{aligned} \mathrm{E}[P(\psi)] &= G(\psi), \\ \mathrm{E}[P(\psi)P(\phi)] &= \frac{1-\sigma}{M+1} G(\psi\phi) + \frac{M+\sigma}{M+1} G(\psi)G(\phi), \\ cov(P(\psi), P(\phi)) &= \frac{1-\sigma}{M+1}[G(\psi\phi) - G(\psi)G(\phi)]. \end{aligned}$$

Proof The first assertion holds as $\mathrm{PY}(\sigma, M, G)$ is a species sampling process with center measure G, and the third assertion follows from combining the first and the second.

To prove the second assertion write $\mathrm{E}[P(\psi)P(\phi)] = \mathrm{E}\big[\psi(X_1)\phi(X_2)\big]$ for $X_1, X_2 | P \overset{\mathrm{iid}}{\sim} P$, and use (14.9) to obtain

$$\mathrm{E}[\phi(X_2) | X_1] = \frac{1-\sigma}{M+1}\phi(X_1) + \frac{M+\sigma}{M+1} G(\phi).$$

Next multiply both sides by $\psi(X_1)$ and integrate with respect to the marginal distribution $X_1 \sim G$, to find the formula. □

Proposition 14.35 (Cauchy-Stieltjes transform) *If $P \sim \mathrm{PY}(\sigma, M, G)$ for $\sigma > 0$ and $M > 0$, then for any nonnegative, measurable function ψ with $\int \psi^\sigma \, dG < \infty$,*

$$\Big(\mathrm{E}\Big[(1 + \int \psi \, dP)^{-M}\Big]\Big)^{-1/M} = \Big(\int (1+\psi)^\sigma \, dG\Big)^{1/\sigma}.$$

Consequently, if $P_i \overset{ind}{\sim} \mathrm{PY}(\sigma, M_i, G)$ are independent of $(W_1, \ldots, W_k) \sim \mathrm{Dir}(k; M_1, \ldots, M_k)$, then $\sum_{i=1}^k W_i P_i \sim \mathrm{PY}(\sigma, \sum_i M_i, G)$.

Proof For the derivation of the formula for $M \neq 0$ we use the characterization of the Pitman-Yor process as a Poisson-Kingman process, as derived in Example 14.47: for (Y_j) a σ-stable Poisson point process, with total mass $Y = \sum_j Y_j$, and $\Phi = \sum_j Y_j \delta_{\theta_j}$ for $\theta_j \overset{\mathrm{iid}}{\sim} G$, the $\mathrm{PY}(\sigma, M, G)$-process is equal in distribution to the mixture of the conditional distribution of Φ/Y given $Y = y$ over the distribution with density $y \mapsto c_{\sigma,M} y^{-M} f_\sigma(y)$, where $c_{\sigma,M} = \sigma\Gamma(M)/\Gamma(M/\sigma)$ is the norming constant and f_σ is the density of Y. Given $M > 0$ the $(-M)$th power of the left side of the formula is

$$\int \mathrm{E}\Big[\Big(1+\frac{\Phi[\psi]}{Y}\Big)^{-M} \mid Y=y\Big]\,\frac{c_{\sigma,M}}{y^M} f_\sigma(y)\,dy = \mathrm{E}\Big[(Y+\Phi[\psi])^{-M}\Big] c_{\sigma,M}$$
$$= \frac{c_{\sigma,M}}{\Gamma(M)} \mathrm{E}\Big[\int_0^\infty \lambda^{M-1} e^{-\lambda(Y+\Phi[\psi])}\,d\lambda\Big],$$

since $y^{-M} = \int_0^\infty \lambda^{M-1} e^{-y\lambda}\,d\lambda/\,\Gamma(M)$. Exchanging the expectation and integral and applying formula (J.6) for the Laplace transform of a completely random measure to $Y+\Phi[\psi] = \int (1+\psi)\,d\Phi$, we see that the preceding display is equal to

$$\frac{c_{\sigma,M}}{\Gamma(M)} \int_0^\infty \lambda^{M-1} e^{-\int\int (1-e^{s\lambda(1+\psi(x))}) s^{-\sigma-1}\,ds\,dG(x)\,\sigma/\,\Gamma(\sigma)}\,d\lambda$$
$$= \frac{c_{\sigma,M}}{\Gamma(M)} \int_0^\infty \lambda^{M-1} e^{-\lambda^\sigma \int (1+\psi)^\sigma\,dG}\,d\lambda,$$

by evaluation of the integral over s with the help of partial integration on $\sigma s^{-\sigma-1}\,ds = ds^{-\sigma}$. The integral on the right side can be reduced to gamma form by changing the variable λ^σ, and then transforms into to the right side of the proposition to the power $-M$.

For the proof of the last assertion, set $V_i \overset{\text{ind}}{\sim} \mathrm{Ga}(M_i, 1)$, so that $V = \sum_i V_i \sim \mathrm{Ga}(\sum_i M_i, 1)$ and independent of $(W_1, \dots, W_k) \sim (V_1, \dots, V_k)/V$, by Proposition G.2. By the formula for the Laplace transform of a gamma variable and the formula of the present proposition, for $\lambda \geq 0$,

$$\mathrm{E}\big[e^{-\lambda \sum_i V_i P_i[\psi_i]}\big] = \prod_i \mathrm{E}\Big[\Big(1+\lambda P_i[\psi_i]\Big)^{-M_i}\Big] = \Big(\int (1+\lambda\psi)^\sigma\,dG\Big)^{-\sum_i M_i/\sigma}.$$

By the same calculation (but without the product) it follows that the right side is the Laplace transform of the variable $VP[\psi]$, for $P \sim \mathrm{PY}(\sigma, \sum_i M_i, G)$. By uniqueness of Laplace transforms (Feller 1971, Theorem XIII.1.1), it follows that $VP[\psi] \sim \sum_i V_i P_i[\psi] = V \sum_i W_i P_i[\psi]$. Here the variable V at the right is independent of $\sum_i W_i P_i$, Since the Fourier transform of $\log V$ is never zero, the logarithm of this distributional equation can be deconvolved, and the last assertion of the proposition follows. □

Example 14.36 (Mean functional) In terms of the probability density function $h_{M,\sigma,G,\psi}$ of $P(\psi)$ Proposition 14.35 gives the identity

$$\int (1+\lambda x)^{-M} h_{M,\sigma,G,\psi}(x)\,dx = \Big[\int (1+\lambda\psi)^\sigma\,dG\Big]^{-M/\sigma}.$$

The left-hand side is the generalized *Cauchy-Stieltjes transform* of order M of the density function $h_{M,\sigma,G,\psi}$, which may be inverted to get $h_{M,\sigma,G,\psi}$. In the limit $\sigma \to 0$ this becomes the *Cifarelli-Regazzini identity* or *Markov-Krein identity*

$$\int (1+\lambda x)^{-M} h_{M,0,G,\psi}(x)\,dx = \exp\Big[-M \int \log(1+\lambda\psi)\,dG\Big].$$

Here $h_{M,0,G,\psi}$ is the density of a mean functional of the Dirichlet process, and the formula is a close analog of (4.26). Another special case is obtained by letting $M \to 0$ when $\sigma \in (0,1)$ is fixed, leading to

$$\exp\Big[-\int \log(1+\lambda x) h_{0,\sigma,G,\psi}(x)\,dx\Big] = \Big[\int (1+\lambda\psi)^{\sigma}\,dG\Big]^{-1/\sigma}.$$

The left side of this equation coincides with the right side of the preceding equation when $M = 1$, $\psi = \iota$ is the identity, and G is replaced by the probability measure $H_{\sigma,0,G,\psi}$ with density $h_{\sigma,0,G,\psi}$. In particular, the generalized Cauchy-Stieltjes transform of order 1 of $h_{0,\sigma,G,\psi}$ coincides with that of $h_{1,0,\sigma,H_{\sigma,0,G,\psi},\iota}$. In view of the uniqueness of the inversion of generalized Cauchy-Stieltjes transform of order 1, this implies the curious distributional equality: $P(\psi) =_d Q(\iota)$ if $P \sim \mathrm{PY}(1,\sigma,G)$ and $Q \sim \mathrm{DP}(H_{\sigma,0,G,\psi})$.

Unlike the Dirichlet process, the Pitman-Yor process is not conjugate under observing independent samples. However, the posterior distribution still has a neat characterization, in terms of Pitman-Yor process with updated parameters.

Theorem 14.37 (Posterior distribution) *If $P \sim \mathrm{PY}(\sigma, M, G)$ for $\sigma \geq 0$, then the posterior distribution of P based on observations $X_1, \ldots, X_n | P \sim P$ is the distribution of the random measure*

$$R_n \sum_{j=1}^{K_n} \hat{W}_j \delta_{\tilde{X}_j} + (1 - R_n) Q_n, \tag{14.21}$$

where $R_n \sim \mathrm{Be}(n - K_n\sigma, M + K_n\sigma)$, $(\hat{W}_1, \ldots, \hat{W}_{K_n}) \sim \mathrm{Dir}(K_n; N_{1,n} - \sigma, \ldots, N_{K_n,n} - \sigma)$, and $Q_n \sim \mathrm{PY}(\sigma, M + \sigma K_n, G)$, all independently distributed. Here $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ are the distinct values in $X_1, \ldots, X_n$ and $N_{1,n}, \ldots, N_{K_n,n}$ their multiplicities.

Proof Theorem 14.33 gives the size-biased weight sequence of the Pitman-Yor process in the stick-breaking form $\tilde{W}_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$, for $V_l \overset{\mathrm{ind}}{\sim} \mathrm{Be}(1-\sigma, M + l\sigma)$, $l = 1, 2, \ldots$ Since the remaining stick lengths satisfy $1 - \sum_{i=1}^{l} \tilde{W}_i = \prod_{l=1}^{j}(1 - V_l)$, on the event $K_n = k$ and $N_{1,n} = n_1, \ldots, N_{k,n} = n_k$

$$\prod_{j=1}^{k} \tilde{W}_j^{n_j - 1} \prod_{j=2}^{k} \Big(1 - \sum_{i<j} \tilde{W}_i\Big) = \prod_{l=1}^{k} (1 - V_l)^{\sum_{j=l+1}^{k} n_l} V_l^{n_l - 1}.$$

By Theorem 14.18 the posterior distribution is a species sampling model whose weight sequence $(\hat{W}_j)$ has density relative to the prior of $(\tilde{W}_j)$ proportional to this expression. The stick-breaking variables $V_{k+1}, V_{k+2}, \ldots$ do not enter in this expression and are independent of the earlier variables. Thus their posterior distribution is the same as their prior distribution and they remain independent of $(V_1, \ldots, V_k)$. To obtain the posterior distribution of the latter vector we multiply the display by its prior density, which is proportional to $\prod_{l=1}^{k} V_l^{-\sigma}(1 - V_l)^{M + \sigma l - 1}$. The resulting product factorizes as a product of beta densities, up to the normalizing constant, whence we conclude that the weights $V_1, \ldots, V_k$ are again independent under the posterior, with $V_j \sim \mathrm{Be}(n_j - \sigma, M + j\sigma + \sum_{i=j+1}^{k} n_i)$, for $j \leq k$.

The posterior weights $\hat{W}_j$ for $j > k$ can be decomposed as $\hat{W}_j = (1 - R_n) \prod_{l=k+1}^{j-1}(1 - V_l) V_j$, for $R_n = 1 - \prod_{l=1}^{k}(1 - V_l) = \sum_{i=1}^{k} \hat{W}_i$. Thus $\sum_{j>k} \hat{W}_j \delta_{\hat{X}_j}$ is equal to $1 - R_n$ times a species sampling process with weights $\prod_{l=k+1}^{j-1}(1 - V_l) V_j$ for $j > k$. This entails a shift

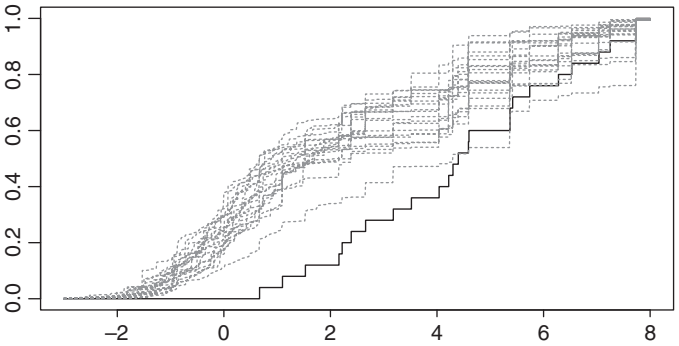

Figure 14.4 Pitman-Yor process. Realization (solid curve) of the empirical distribution of a sample of size 25 from the normal distribution with mean 4 and variance 4, and 20 realizations (dashed curves) from the corresponding posterior distribution relative to the Pitman-Yor prior process with $\sigma = 1/2$, $M = 1$ and standard normal base measure. The posterior inconsistency is clearly visible.

by k in the parameter j of the beta distributions of these variables, which can be viewed as shift by $k\sigma$ in the parameter M. In view of Theorem 14.33 the weights $(\hat{W}_j)_{j>k}$ are those of a $\mathrm{PY}(\sigma, M + k\sigma)$ process.

The vector $(\hat{W}_1, \ldots, \hat{W}_k, 1 - R_n)$ possesses a $\mathrm{Dir}(k+1; n_1 - \sigma, \ldots, n_k - \sigma, M + k\sigma)$-distribution in view of of the stick-breaking representation of a discrete Dirichlet process, as given in Corollary G.5 with $k+1$ instead of k, $\alpha_j = n_j - \sigma$, for $j = 1, \ldots, k$, and $\alpha_{k+1} = M + k\sigma$. Equivalently, by the aggregation property of Dirichlet distributions, given in Proposition G.3, the vector $(\hat{W}_1, \ldots, \hat{W}_k)/R_n$ possesses a $\mathrm{Dir}(k; n_1 - \sigma, \ldots, n_k - \sigma)$-distribution, and is independent of R_n, which has a $\mathrm{Be}(n - k\sigma, M + k\sigma)$-distribution. □

The Pitman-Yor process prior inherits the (in)consistency of the general Gibbs processes as a prior for estimating a distribution P of a sample of observations. It is consistent at discrete distributions, but inconsistent for distributions with a continuous component, except when the center measure is (accidentally) proportional to the continuous part of the true distribution, or in the case $\sigma = 0$ that the Pitman-Yor process coincides with the Dirichlet process.

Theorem 14.38 (Consistency) *If P follows a Pitman-Yor $\mathrm{PY}(\sigma, M, G)$ process, then the posterior distribution of P in the model $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ converges under P_0 relative to the weak topology to $P_0^d + (1-\sigma)P_0^c + \sigma\xi G$, for $\xi = \|P_0^c\|$. In particular, the posterior distribution is consistent if and only if P_0 is discrete or G is proportional to P_0^c or $\sigma = 0$.*

Proof A $\mathrm{PY}(\sigma, M, G)$ process is Gibbs, with probability of a new observation equal to $V_{n+1,K_n+1}/V_{n,K_n} = (M + K_n\sigma)/(M+n)$ and $V_{n+2,K_n+2}/V_{n+1,K_n+1} = (M + K_n\sigma + \sigma)/(M+n+1)$. As seen in the proof of Theorem 14.27 $K_n/n \to \xi$, almost surely, whence both quotients converge to $\gamma = \sigma\xi$. The theorem is a consequence of Theorem 14.27. □

In particular, in the case of an atomless P_0 the Pitman-Yor posterior contracts to the (typically wrong) distribution $(1-\sigma)P_0 + \sigma G$. The following lemma of Bernstein–von Mises type gives a preciser description of the asymptotics in this case.

Lemma 14.39 (Bernstein–von Mises) *If P follows a $\mathrm{PY}(\sigma, M, G)$ process with $\sigma \geq 0$, then the posterior distribution of P in the model $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ satisfies*

$$\sqrt{n}(P - (1-\sigma)\mathbb{P}_n - \sigma G) | X_1, \ldots, X_n \rightsquigarrow \sqrt{1-\sigma}\,\mathbb{G}_{P_0} + \sqrt{\sigma(1-\sigma)}\,\mathbb{G}_G + \sqrt{\sigma(1-\sigma)}\,Z(P_0 - G)$$

in $\mathfrak{L}_\infty(\mathcal{F})$ a.s. $[P_0^\infty]$, for every atomless measure P_0 and for every P_0-Donsker class of functions $\mathcal{F}$ for which the $\mathrm{PY}(\sigma, \sigma, G)$-process satisfies the central limit theorem in $\mathfrak{L}_\infty(\mathcal{F})$. Here $\mathbb{G}_{P_0}$ and $\mathbb{G}_G$ are independent P_0- and G-Brownian bridges independent of the variable $Z \sim \mathrm{Nor}(0,1)$. In particular this is is true for every sufficiently measurable class of functions $\mathcal{F}$ with finite P_0- and G-bracketing integrals or a finite uniform entropy integral and envelope function that is square integrable under both P_0 and G.

Proof The Pitman-Yor process for $\sigma = 0$ is the Dirichlet process and hence the theorem follows from Theorem 12.2. We may assume that $\sigma > 0$.

Because P_0 is atomless, all observations $X_1, \ldots, X_n$ are distinct, whence $K_n = n$, the multiplicities $N_{j,n}$ are all 1, and $\tilde{X}_1, \ldots, \tilde{X}_n$ are the original observations. Therefore, by Theorem 14.37, the posterior distribution can be represented as $R_n\tilde{\mathbb{P}}_n + (1-R_n)Q_n$, for $\tilde{\mathbb{P}}_n = \sum_{i=1}^n W_{n,i}\delta_{X_i}$, where $R_n \sim \mathrm{Be}(n - n\sigma, M + n\sigma)$, $W_n = (W_{n,1}, \ldots, W_{n,n}) \sim \mathrm{Dir}(n; 1-\sigma, \ldots, 1-\sigma)$ and $Q_n \sim \mathrm{PY}(\sigma, M + \sigma n, G)$ are independent variables.

By representing R_n as a quotient of gamma variables, as in Proposition G.2, and applying the delta-method and the central limit theorem, we see that $\sqrt{n}(R_n - 1 + \sigma) \rightsquigarrow \sqrt{\sigma(1-\sigma)}Z$.

The vector nW_n can be represented as $(Y_1, \ldots, Y_n)/\bar{Y}_n$ for $Y_i \overset{\mathrm{iid}}{\sim} \mathrm{Ga}(1-\sigma, 1)$, and hence is exchangeable as in Example 3.7.9 of van der Vaart and Wellner (1996) with $c = (1-\sigma)^{-1/2}$. Consequently $\sqrt{n}(\tilde{\mathbb{P}}_n - \mathbb{P}_n) | X_1, X_2, \ldots \rightsquigarrow c\mathbb{G}_{P_0}$, almost surely, by the exchangeable bootstrap theorem, Theorem 3.7.13 of the same reference.

Applying the delta-method we find that

$$\sqrt{n}(R_n\tilde{\mathbb{P}}_n - (1-\sigma)\mathbb{P}_n) = \sqrt{n}(R_n - 1 + \sigma)P_0 + (1-\sigma)\sqrt{n}(\tilde{\mathbb{P}}_n - \mathbb{P}_n) + o_P(1),$$

where the remainder term tends to zero given $X_1, X_2, \ldots$ almost surely. If we can show that $\sqrt{n}(Q_n - G) \rightsquigarrow \sqrt{(1-\sigma)/\sigma}\,\mathbb{G}_G$, then the same reasoning gives

$$\sqrt{n}((1-R_n)Q_n - \sigma G) = -\sqrt{n}(R_n - 1 + \sigma)G + \sigma\sqrt{n}(Q_n - G) + o_P(1).$$

The lemma follows by adding the last two displays.

By the last assertion of Proposition 14.35 the Pitman-Yor process Q_n can be represented as $\sum_{i=0}^n W_{n,i}P_i$, for $(W_{n,0}, \ldots, W_{n,n}) \sim \mathrm{Dir}(n+1; M, \sigma, \ldots, \sigma)$ independent of the Pitman-Yor processes $P_0, \ldots, P_n$ with parameters (σ, M, G) for $i = 0$ and (σ, σ, G) for $i > 0$. Since $Q_n - G = W_{n,0}(P_0 - G) + (1 - W_{n,0})\sum_{i=1}^n \tilde{W}_{n,i}(P_i - G)$, where $\sqrt{n}W_{n,0} \to 0$ in probability and the variables $\tilde{W}_{n,i} = W_{n,i}/(1 - W_{n,0})$ form a Dirichlet vector of dimension n with parameters $(\sigma, \ldots, \sigma)$, the term for $i = 0$ is negligible, and the limit distribution of $\sqrt{n}(Q_n - G)$ is as if the term for $i = 0$ did not exist. For simplicity of notation, assume the latter. Represent the remaining Dirichlet coefficients as V_i/V, for $V_i \overset{\mathrm{iid}}{\sim} \mathrm{Ga}(\sigma, 1)$, for $i = 1, \ldots, n$, and $V = \sum_{i=1}^n V_i$.

For a given $f \in \mathbb{L}_2(G)$, the sequence of variables $n^{-1/2}\sum_{i=1}^n V_i(P_i f - Gf)$ converges by the univariate central limit theorem in distribution to a normal distribution with

mean zero and variance $\operatorname{var}(V_1(P_1 f - Gf)) = \mathrm{E}[V_1^2](1-\sigma)/(\sigma+1)\operatorname{var}\big[\mathbb{G}_G(f)\big]$, by Proposition 14.34. By Slutsky's lemma the sequence $\sqrt{n}\sum_{i=1}^n (V_i/V)(P_i f - Gf)$ converges in distribution to $1/\mathrm{E}[V_1]$ times the same limiting process. Here $\mathrm{E}[V_1^2]/(\mathrm{E}[V_1])^2 = (1+\sigma)/\sigma$, giving the final variance equal to $(1-\sigma)/\sigma \operatorname{var}\big[\mathbb{G}_G(f)\big]$.

This shows that the processes $\sqrt{n}\sum_{i=1}^n (V_i/V)(P_i - G)$ converge marginally in distribution to the process $((1-\sigma)/\sigma)^{1/2}\mathbb{G}_G$. We finish by showing that this sequence of processes, or equivalently the processes $n^{-1/2}\sum_{i=1}^n V_i(P_i - G)$, are asymptotically tight (i.e. equicontinuous) in $\ell^\infty(\mathcal{F})$. The assumption that the PY(σ, σ, G)-process satisfies the central limit theorem in $\ell^\infty(\mathcal{F})$ entails that the sequence $n^{-1/2}\sum_{i=1}^n (P_i - G)$ converges to a tight Gaussian limit and hence it is asymptotically tight. By Lemma 2.9.1 (also see the proof of Theorem 2.9.2) of van der Vaart and Wellner (1996), this asymptotic tightness is inherited by the sequence of processes $n^{-1/2}\sum_{i=1}^n V_i(P_i - G)$.

For the final assertion of the theorem, it suffices to show that the sequence $n^{-1/2}\sum_{i=1}^n (P_i - G)$ converges in distribution to a tight limit in $\ell^\infty(\mathcal{F})$. This follows because the processes $Z_i = P_i$ satisfy the conditions of Theorems 2.11.9 or 2.11.1 of van der Vaart and Wellner (1996), where Lemma 2.11.6 in the same reference can be used to verify the main condition of the second theorem. □

Samples from Pitman-Yor processes of different types $\sigma \geq 0$ cluster in different manners. The numbers K_n of species forms a Markov chain, which increases by jumps of size one if a new species is added. Since a new species is added with probability $(M + K_n\sigma)/(M+n)$,

$$\mathrm{E}(K_{n+1} \mid X_1, \ldots, X_n) = K_n + \frac{M + K_n\sigma}{M+n} = a_n K_n + b_n,$$

for $a_n = 1 + \sigma/(M+n)$ and $b_n = M/(M+n)$. It follows that the process

$$\frac{K_n}{\prod_{i=1}^{n-1}[1+\sigma/(M+i)]} - \sum_{i=1}^{n-1} \frac{M/(M+i)}{\prod_{j=1}^{i}[1+\sigma/(M+j]}$$

is a martingale. If $\sigma > 0$, then the numbers $\prod_{i=1}^{n-1}[1+\sigma/(M+i)]$ are of the order n^σ, while the sum in the display is bounded in n. It follows that the mean number of species $\mu_n = \mathrm{E}(K_n)$ is of the order n^σ. By using the martingale convergence theorem it can further be derived that K_n/n^σ converges almost surely to a finite random variable, as $n \to \infty$. The limit random variable, known as the *σ-diversity*, can be shown to have a positive Lebesgue density. Thus the number of clusters formed by a Pitman-Yor process is of the order n^σ, which is much higher than the logarithmic order of the number of clusters formed by the Dirichlet process, established in Proposition 4.8. For this reason, in some applications, the Pitman-Yor process model is favored over the Dirichlet process model. For further discussion, within the more general context of Poisson-Kingman partitions, see Theorem 14.50.

14.5 Poisson-Kingman Processes

A *Poisson point process* on $(0,\infty)$ with intensity (or Lévy) measure ρ can be defined as a collection $\{Y_j : j \in \mathbb{N}\}$ of random variables with values in $(0,\infty)$ such that the numbers of variables (or "points") $\#\{j : Y_j \in A_i\}$ in disjoint Borel sets A_i are independent Poisson variables with means $\rho(A_i)$. Consider intensity measures ρ such that

$$\int (s \wedge 1)\, \rho(ds) < \infty, \qquad \rho(0, \infty) = \infty. \tag{14.22}$$

Then the number of points is infinite (only clustering at zero) and their sum $Y := \sum_{i=1}^{\infty} Y_i$ is finite almost surely. The points Y_i can be viewed as the "jumps" in the series Y. By placing these jumps at (possibly random) locations in $(0, \infty)$ one obtains a process with independent increments or "completely random measure." See Appendix J for background on these processes.

By Lemma J.2 the Laplace transform of $Y = \sum_{j=1}^{\infty} Y_j$ is given by, for f the probability density function of Y,

$$\int e^{-\lambda y} f(y)\, dy = e^{-\psi(\lambda)}, \qquad \psi(\lambda) = \int (1 - e^{-\lambda s})\, \rho(ds).$$

The function ψ is called the *Laplace exponent* of Y or the point process (or also *Lévy exponent*).

Definition 14.40 (Poisson-Kingman process) The (special) *Poisson-Kingman distribution* $\mathrm{PK}(\rho)$ with intensity measure (or Lévy measure) ρ is the distribution of $(Y_{(1)}, Y_{(2)}, \ldots)/Y$ for $Y_{(1)} \geq Y_{(2)} \geq \cdots$ the ranked values of a Poisson point process on $(0, \infty)$ with intensity measure ρ and $Y = \sum_{i=1}^{\infty} Y_i$ (which is assumed finite). The *Poisson-Kingman distribution* $\mathrm{PK}(\rho|\, y)$ is the conditional distribution of the same vector given $Y = y$. Given a probability measure η on $(0, \infty)$, the *Poisson-Kingman distribution* $\mathrm{PK}(\rho, \eta)$ is the mixture $\int \mathrm{PK}(\rho|\, y)\, d\eta(y)$. The corresponding species sampling models $P = \sum_{i=1}^{\infty} (Y_i/Y)\delta_{\theta_i}$, for an independent random sample $\theta_1, \theta_2, \ldots$ from an atomless probability distribution G, are said to follow *Poisson-Kingman processes.*

Example 14.41 (Scaling) By the renormalization of the vector (Y_j) by its sum, the scale of the variables Y_j is irrelevant in the definition of the Poisson-Kingman distribution. Thus any measure of the scale family $\rho_\tau(A) = \rho(A/\tau)$ of a given intensity measure gives the same Poisson-Kingman distribution. (On the other hand, multiplying the intensity measure with a constant, to obtain $\tau\rho$, gives a different process.)

Example 14.42 (Gamma process) The sum Y of all points of the point process with intensity measure given by $\rho(ds) = Ms^{-1}e^{-s}\, ds$ possesses a gamma distribution with shape parameter M and scale 1. For this process, the ranked normalized jump sizes $(Y_{(1)}, Y_{(2)}, \ldots)/Y$ are independent of the sum Y (see below). The Poisson-Kingman distributions $\mathrm{PK}(\rho|\, y)$ and $\mathrm{PK}(\rho, \eta)$ are therefore free of y and η, and hence coincide with $\mathrm{PK}(\rho)$. This turns out to be the distribution of the ranked jump sizes of a Dirichlet process, so that the corresponding species sampling process can be identified with the Dirichlet prior. In the present context, the distribution of the ranked normalized jump sizes is often called the *one-parameter Poisson-Dirichlet distribution* with parameter M (which should not be confused with $\mathrm{PK}(\rho)$ with a single but general parameter ρ).

The present point process (Y_j) and its intensity measure ρ can be obtained by marginalization to the second coordinate of the Poisson point process $\{(X_j, Y_j) : j \in \mathbb{N}\}$ with

intensity $s^{-1}e^{-s}\,dx\,ds$ on the space $(0, M] \times (0, \infty)$. The process $t \mapsto \sum_j Y_j \mathbb{1}\{X_j \le t\}$ is the gamma process, which has marginal distributions $\mathrm{Ga}(t, 1)$ and independent increments (see Example J.14). The ranked jump sizes $(Y_{(1)}, Y_{(2)}, \ldots)$ can be seen to be the coordinatewise limits of the ranked increments over increasingly finer uniform partitions of $[0, M]$. This may be combined with the independence of the sum and ratios of independent gamma variables (Proposition G.2), to show that the ranked normalized jump sizes $(Y_{(1)}, Y_{(2)}, \ldots)/Y$ are independent of the sum Y. The connection to the gamma process also shows that the normalized jumps are those of the Dirichlet process (see Section 4.2.3).

Example 14.43 (Stable process) The Laplace exponent $\psi_\sigma(\lambda) = \int (1 - e^{-\lambda s})\,\rho_\sigma(ds)$ of the process with intensity measure $\rho_\sigma(ds) = (\sigma/\Gamma(1-\sigma))\,s^{-1-\sigma}\,ds$, for $\sigma \in (0, 1)$, takes the simple form $\psi_\sigma(\lambda) = \lambda^\sigma$. (One way to see this is to compute the derivative as $\psi'_\sigma(\lambda) = \sigma\lambda^{\sigma-1}$, after differentiating under the integral.) It follows that the density f_σ of Y has Laplace transform $e^{-\lambda^\sigma}$, and hence Y possesses the σ-stable distribution. The corresponding process with independent increments generated by the intensity function $\rho_\sigma(ds)\,dx$ is known as the *σ-stable process* (see Example J.13).

The EPPFs of Poisson-Kingman processes can be expressed in the intensity measure ρ. The formulas are somewhat involved, but lead to simple representations for several examples. The density f in the following theorem is the density of the sum Y of all weights.

Theorem 14.44 (EPPF) *Let $\psi(\lambda) = \int_0^\infty (1 - e^{-\lambda s})\,\rho(ds)$ be the Laplace exponent of an absolutely continuous measure ρ on $(0, \infty)$ satisfying* (14.22)*, let $\psi^{(n)}$ be its nth derivative, and let f be the probability density with Laplace transform $e^{-\psi}$. Then the EPPF of the Poisson-Kingman processes* $\mathrm{PK}(\rho)$ *and* $\mathrm{PK}(\rho|\,y)$ *are given by, respectively, for* $n = \sum_{j=1}^k n_j$,

$$p(n_1, \ldots, n_k) = \frac{(-1)^{n-k}}{\Gamma(n)} \int_0^\infty \lambda^{n-1} e^{-\psi(\lambda)} \prod_{j=1}^k \psi^{(n_j)}(\lambda)\,d\lambda,$$

$$p(n_1, \ldots, n_k|\,y) = \int\limits_{\sum_{j=1}^k y_j < y} \cdots \int \frac{f(y - \sum_{j=1}^k y_j)}{y^n\, f(y)} \prod_{j=1}^k y_j^{n_j} \prod_{j=1}^k \rho(dy_j).$$

Proof By Lemma J.2, f is the density of the sum $Y = \sum_i Y_i$ of the jumps of the Poisson process. If $\mathcal{P}_n$ is the partition generated by the species sampling model defined by the renormalized sequence $W_i = Y_i/Y$, and $\tilde{W}_i = \tilde{Y}_i/Y$ is this sequence in size-biased order, then for any partition $A_1, \ldots, A_k$ of $\mathbb{N}_n$ in order of appearance,

$$\begin{aligned}&\mathrm{P}(\mathcal{P}_n = \{A_1, \ldots, A_k\}, \tilde{Y}_1 \in B_1, \ldots, \tilde{Y}_k \in B_k, Y \in C)\\ &\quad = \int_{B_1} \cdots \int_{B_k} \int_C y^{-n} f\Big(y - \sum_{j=1}^k y_j\Big)\,dy \prod_{j=1}^k y_j^{|A_j|} \prod_{j=1}^k \rho(dy_j).\end{aligned}$$

An intuitive argument for this formula is as follows. (For a longer but precise proof, see the proof of Lemma 14.62.) Form the partition by covering $(0, Y)$ with contiguous intervals of lengths Y_i, generating an independent sample $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{Unif}(0, Y)$, and letting i and j belong to the same partitioning set if U_i and U_j belong to the same interval. Then $\prod_{j=1}^{k} \rho(dy_j)$ is the probability that $\tilde{Y}_j = y_j$ are the jumps in the Poisson process corresponding to the k sets in the partition, $f(y - \sum_{j=1}^{k} y_j)\, dt$ is the conditional probability that the sum of all jumps is $Y = y$ given these points, and $\prod_{j=1}^{k} (y_j/y)^{|A_j|}$ is the probability that the U_i fall in the intervals in the manner that produces the partition.

The EPPFs follow by marginalizing out the jumps $\tilde{Y}_i$, and marginalizing out (for p) or conditioning on $Y = y$ (for $p(\cdot \,|\, y)$), respectively. To obtain the formula for p we further write $y^{-n} = \int_0^\infty \lambda^{n-1} e^{-\lambda y}\, d\lambda / \Gamma(n)$, apply Fubini's theorem, make the change of variables $t = y - \sum_{j=1}^{k} y_j$, and substitute the identities $\int e^{-\lambda t} f(t)\, dt = e^{-\psi(\lambda)}$ and $\psi^{(n)}(\lambda) = (-1)^{n-1} \int_0^\infty s^n e^{-\lambda s}\, \rho(ds)$. □

Because the Poisson-Kingman distribution $\text{PK}(\rho, \eta)$ is a mixture of $\text{PK}(\rho \,|\, y)$ distributions, so is its EPPF. Thus the formulas in the preceding theorem extend to general Poisson-Kingman processes. In the following two examples, which involve polynomial or exponential *tilting* of the intensity measure, the resulting integrations can be reduced to a simple form.

Example 14.45 (Polynomial tilting) For a given absolutely continuous infinite Lévy measure ρ, and given $M \geq 0$, let f_M be the probability density of the form $f_M(y) = c_M y^{-M} f(y)$, for f the density of $Y = \sum_{j=1}^{\infty} Y_j$ associated with ρ. Then the EPPF of the $\text{PK}(\rho, f_M)$ process is given by

$$p(n_1, \ldots, n_k) = \frac{c_M (-1)^{n-k}}{\Gamma(M+n)} \int_0^\infty \lambda^{M+n-1} e^{-\psi(\lambda)} \prod_{j=1}^{k} \psi^{(n_j)}(\lambda)\, d\lambda.$$

For $M = 0$ this reduces to the formula of the EPPF of the $\text{PK}(\rho)$ process, given in Theorem 14.44. The formula can be derived in exactly the same manner as the latter formula, by computing $\int p(n_1, \ldots, n_k \,|\, y) f_M(y)\, dy$.

Example 14.46 (Exponential tilting) For a given absolutely continuous Lévy measure ρ, and given $\tau \geq 0$, the measure $\rho_\tau(ds) = e^{-\tau s} \rho(ds)$ is another Lévy measure. It turns out that the associated Poisson-Kingman distributions $\text{PK}(\rho_\tau \,|\, t)$ are identical to the $\text{PK}(\rho \,|\, t)$-distribution, for every t. As a consequence, the Poisson-Kingman distributions $\text{PK}(\rho_\tau, \eta)$ are also independent of this *exponential tilting* of the intensity measure.

On the other hand, the special Poisson-Kingman distributions $\text{PK}(\rho_\tau)$ do depend on τ, in general, as the associated mixing distribution (the distribution of Y) does.

To verify the claim, first note that the Laplace exponent of the tilted intensity measure satisfies $\psi_\tau(\lambda) = \psi_0(\lambda + \tau) - \psi_0(\tau)$, and hence the density of Y satisfies $f_\tau(y) = f_0(y) e^{\psi_0(\tau) - \tau y}$. When substituting these expressions in the formula for $p(\cdot \,|\, t)$ in Theorem 14.44, the tilting terms can be seen to cancel out.

Another, perhaps more insightful argument, is to note that the laws P_τ of the Poisson processes N_τ with intensities ρ_τ are mutually absolutely continuous, with the density of P_τ with respect to P_0 equal to $dP_\tau/dP_0 = e^{\psi_0(\tau)-\tau Y}$, for $Y = \sum_{i=1}^{\infty} Y_i$, as before. This can be seen from the identity, for any bounded measurable function f,

$$\mathrm{E}e^{-\int f\,dN_0} e^{-\tau Y} = \mathrm{E}e^{-\int (f(s)+\tau s)\,N(ds)} = e^{-\int (1-e^{f(s)+\tau s})\,\rho(ds)} = e^{-\psi_0(\tau)} \mathrm{E}e^{-\int f\,dN_\tau}.$$

This identity follows by two applications of the formula for the Laplace transform of a Poisson process in Lemma J.2. The fact that the density of P_τ depends on the process only through the sum Y of its points means that the latter variable is statistically sufficient for the parameter τ, and hence the conditional law of the process given Y is free of τ.

Example 14.47 (Pitman-Yor) The Pitman-Yor process with parameter $(\sigma, M) \in (0,1) \times (0,\infty)$ is identical to the Poisson-Kingman distribution $\mathrm{PK}(\rho_\sigma, \eta_{\sigma,M})$ with intensity and mixing measures given by

$$\rho_\sigma(ds) = \frac{\sigma s^{-1-\sigma}}{\Gamma(1-\sigma)}\,ds, \qquad \eta_{\sigma,M}(dy) = \frac{\sigma\Gamma(M)}{\Gamma(M/\sigma)}\,y^{-M} f_\sigma(y)\,dy,$$

where f_σ is the probability density (associated with ρ_σ) with Laplace transform $e^{-\lambda^\sigma}$.

To see this, note that the intensity measure is the σ-stable measure and has Laplace exponent $\psi_\sigma(\lambda) = \lambda^\sigma$ (see Example 14.43) and the measure $\eta_{\sigma,M}$ in the display is a polynomial tilting of the associated density f_σ. Hence the EPPF can be computed as in Example 14.45, with $\psi_\sigma^{(n)}(\lambda) = \sigma(-1)^{n-1}(1-\sigma)^{[n-1]}\lambda^{\sigma-n}$.

As noted in Example 14.46 the Poisson-Kingman distributions $\mathrm{PK}(\rho|t)$ are invariant under exponential tilting of the intensity measure, and hence the Pitman-Yor process can also be obtained as a mixture of a tilted stable process, i.e. a generalized gamma process. See (iv) of Example 14.48 below.

Example 14.48 (Generalized gamma) The *generalized gamma process* has intensity measure $\rho(ds) = (\sigma/\Gamma(1-\sigma))\,s^{-1-\sigma}e^{-\tau s}\,ds$, where $\sigma \in (0,1)$ and $\tau > 0$. The process is an exponentially tilted version of the σ-stable process. By the invariance of the Poisson-Kingman distribution under the scaling $s \mapsto s/\tau$, the exponential term can without loss of generality be simplified to the unit form e^{-s} if at the same time the full intensity measure is multiplied by τ^σ, giving $(\sigma/\Gamma(1-\sigma))\tau^\sigma\, s^{-1-\sigma}e^{-s}\,ds$.

The EPPF of the Poisson-Kingman process based on a generalized gamma process can be found with the help of Theorem 14.44. First we compute the Laplace exponent as $\psi(\lambda) = \int(1-e^{-\lambda s})\,\rho(ds) = (\lambda+\tau)^\sigma - \tau^\sigma$. (Write $1-e^{-\lambda s}$ as the integral of its derivative and apply Fubini's theorem.) This leads to the derivatives $\psi^{(n)}(\lambda) = \sigma(\sigma-1)\cdots(\sigma-n+1)(\lambda+\tau)^{\sigma-n} = \sigma(-1)^{n-1}(1-\sigma)^{[n]}(\lambda+\tau)^{\sigma-n}$, and hence to the EPPF

$$p(n_1,\ldots,n_k) = \int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n)}(\lambda+\tau)^{k\sigma-n}e^{-\psi(\lambda)}\,d\lambda\,\sigma^k(-1)^{n-k}\prod_{j=1}^{k}(1-\sigma)^{[n_j]}.$$

It follows that the special Poisson-Kingman process based on a generalized gamma process (i.e. a normalized generalized gamma process) is a Gibbs process of type σ with coefficients

$$V_{n,k} = \sigma^k(-1)^{n-k}\int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n)}(\lambda+\tau)^{k\sigma-n}e^{-\psi(\lambda)}\,d\lambda$$
$$= \frac{e^{\tau^\sigma}\sigma^{k-1}(-1)^{n-k}}{\Gamma(n)}\sum_{j=0}^{n-1}\binom{n-1}{j}(-\tau)^j\Gamma(k-j/\sigma;\tau^\sigma),$$

where $\Gamma(n;x) = \int_x^\infty s^{n-1}e^{-s}ds$ is the incomplete gamma function, and the second expression follows after some algebra.

The class of generalized gamma processes contains several interesting special cases.

(i) For $\tau \to 0$ the process reduces to the σ-stable process.
(ii) For $\sigma \to 0$ the process with intensity measure $\sigma^{-1}\rho$ tends to the intensity measure of the gamma process, and hence the resulting Poisson-Kingman process reduces to the Dirichlet process.
(iii) The choice $\sigma = 1/2$ leads to tractable marginal distributions of the normalized process; it is called the *normalized inverse-Gaussian process* and is discussed in Section 14.6.
(iv) The Pitman-Yor process $\mathrm{PY}(\sigma, M, G)$ is a mixture over ξ of the generalized gamma processes with parameters σ and $\tau = \xi^{1/\sigma}$, with a $\mathrm{Ga}(M/\sigma, 1)$-mixing distribution on ξ. (See Problem 14.9.) By scaling and normalization these generalized gamma processes are equivalent to the processes with intensity measure $\xi(\sigma/\Gamma(1-\sigma))\,s^{-1-\sigma}e^{-s}\,ds$.

One reason to be interested in Poisson-Kingman distributions is that the weight sequence in size-biased order is relatively easy to characterize. In the following theorem let $(\tilde{Y}_j)$ be the random permutation of the point process (Y_j) such that $(\tilde{Y}_1, \tilde{Y}_2, \ldots)$ is in size-biased order. Then $\tilde{W}_j = \tilde{Y}_j/Y$ are the normalized weights of the Poisson-Kingman species sampling process in size-biased order. The theorem gives the joint density of the stick-breaking weights (V_j) corresponding to this sequence: the variables such that $\tilde{W}_i = V_i\prod_{l=1}^{i-1}(1-V_l)$, for $i = 1, 2, \ldots$.

Theorem 14.49 (Residual allocation, stick-breaking) *If ρ is absolutely continuous, and $\tilde{H}(ds\,|\,t) = (s/t)\,f(t-s)/f(t)\,\rho(ds)$, for f the density of $Y = \sum_{j=1}^\infty Y_j$, then the points of the Poisson-Kingman process in size-biased order satisfy*

$$\tilde{Y}_i\,|\,Y, \tilde{Y}_1, \ldots, \tilde{Y}_{i-1} \sim \tilde{H}\Big(\cdot\,|\sum_{j\ge i}\tilde{Y}_j\Big), \qquad i = 1, 2, \ldots \tag{14.23}$$

Furthermore, the joint density of $(Y, V_1, \ldots, V_i)$ of Y and the stick-breaking variables (V_j) is given by, for r the density of ρ,

$$(y, v_1, \ldots, v_i) \mapsto \prod_{j=1}^{i}\Big[v_j\prod_{l=1}^{j-1}(1-v_l)\,r\Big(yv_j\prod_{l=1}^{j-1}(1-v_l)\Big)\Big]\,y^i\,f\Big(y\prod_{j=1}^{i}(1-v_j)\Big).$$

Consequently, in the $\mathrm{PK}(\rho, \eta)$ *process the conditional density of* V_i *given* $(V_1, \dots, V_{i-1}) = (v_1, \dots, v_{i-1})$ *is*

$$v_i \mapsto \frac{v_i \prod_{j=1}^{i-1}(1-v_j) \int \prod_{j=1}^{i} r\big(y v_j \prod_{l=1}^{j-1}(1-v_l)\big)\, y^i\, f\big(y \prod_{j=1}^{i}(1-v_j)\big)/f(y)\, d\eta(y)}{\int \prod_{j=1}^{i-1} r\big(y v_j \prod_{l=1}^{j-1}(1-v_l)\big)\, y^{i-1}\, f\big(y \prod_{j=1}^{i-1}(1-v_j)\big)/f(y)\, d\eta(y)}.$$

Proof As noted in the proof of Theorem 14.44, for any measurable sets B and C,

$$\mathrm{P}(\tilde{Y}_1 \in B, Y \in C) = \int_B \int_C f(y - y_1) \frac{y_1}{y}\, \rho(dy_1)\, dy.$$

Since f is the density of Y, the identity can also be obtained from Lemma J.4(i), which implies that (Y_1, Y) has density $(y_1, y) \mapsto f(y - y_1) y_1 / y$ relative to the product of ρ and Lebesgue measure. The equation readily gives the conditional density of $\tilde{Y}_1$ given Y and proves (14.23) for $i = 1$.

Define N_1 as the point process obtained from the point process $N = (Y_j)$ by removing the point $\tilde{Y}_1$, and let $T_1 = Y - \tilde{Y}_1 = \sum_{j \geq 2} \tilde{Y}_j$ be its total mass. By general properties of Poisson processes (see Lemma J.4) the conditional distribution of N_1 given $(T_1 = t_1, \tilde{Y}_1)$ is identical to the conditional distribution of N given $Y = t_1$. Since the point $\tilde{Y}_2$ can be viewed as the first element of N_1 in order of appearance, we can repeat the conclusion of the preceding paragraph to obtain (14.23) for $i = 2$. By further induction we obtain the identity for every i.

Equation (14.23) shows that the vector $(\tilde{Y}_1, \dots, \tilde{Y}_i)$ has conditional density given $Y = y$ of the form, for $\tilde{h}$ the density of $\tilde{H}$,

$$\prod_{j=1}^{i} \tilde{h}\Big(y_j \,|\, y - \sum_{l<j} y_l\Big) = \prod_{j=1}^{i} \Big[\frac{y_j}{y - \sum_{l<j} y_l} r(y_j) \frac{f(y - \sum_{l \leq j} y_l)}{f(y - \sum_{l<j} y_l)}\Big].$$

Given $Y = y$ the weights $(\tilde{W}_1, \dots, \tilde{W}_i)$ are a simple scaling of $(\tilde{Y}_1, \dots, \tilde{Y}_i)$ by y, and the stick-breaking variables satisfy $V_j = \tilde{W}_j / (1 - \sum_{l<j} \tilde{W}_j) = \tilde{Y}_j / (y - \sum_{l<j} \tilde{Y}_l)$, since $1 - \sum_{l<j} \tilde{W}_l = \prod_{l<j}(1 - V_l)$ is the remaining stick length. Thus the joint density of $(V_1, \dots, V_i)$ given $Y = y$ can be obtained from the joint density of $(\tilde{Y}_1, \dots, \tilde{Y}_i)$ given $Y = y$ by the change of variables $v_j = y_j / (y - \sum_{l<j} y_l)$, which has inverse $y_j = y v_j \prod_{l<j}(1 - v_l)$, since $y - \sum_{l<j} y_j = y(1 - \sum_{l<j} w_l)$. The Jacobian of the transformation is $|\partial y / \partial v| = y^i \prod_{j=1}^{i} \prod_{l<j}(1 - v_l)$ and hence the joint density of $(V_1, \dots, V_i)$ given $Y = y$ is given by

$$\prod_{j=1}^{i} \Big[v_j r\Big(y v_j \prod_{l<j}(1 - v_l)\Big) \frac{f(y \prod_{l \leq j}(1 - v_l))}{f(y \prod_{l<j}(1 - v_l))}\Big]\, y^i \prod_{j=1}^{i} \prod_{l<j} (1 - v_l).$$

The product of quotients involving f telescopes out to a simple quotient, with $f(y)$ in the denominator. The latter factor cancels upon multiplying with the density of Y to obtain the joint density of $(V_1, \dots, V_i, Y)$, as given in the theorem.

The $\mathrm{PK}(\rho, \eta)$ distribution is a mixture over $y \sim \eta$ of the $\mathrm{PK}(\rho \,|\, y)$-distributions. The density of its stick-breaking weights can be obtained by mixing the corresponding densities of

the PK($\rho|\, y$)-distribution. We obtain the marginal density of $(V_1, \ldots, V_i)$ by returning to the conditional density of this vector given $Y = y$ (hence divide by $f(y)$), and next integrating the variable Y out relative to its marginal distribution η. Finally the conditional density of V_i is the quotient of the joint densities of $(V_1, \ldots, V_i)$ and $(V_1, \ldots, V_{i-1})$. □

Although Theorem 14.49 gives an explicit expression for the joint density of the stick-breaking proportions (V_j), it is clear from (14.23) that a description of the Poisson-Kingman process, in size-biased order, in terms of the unnormalized weights is simpler. The conditioning variable $\sum_{i\geq j} \tilde{Y}_i = Y - \sum_{i<j} \tilde{Y}_i$ in the right side of (14.23) is the "remaining (unnormalized) mass" at stage j. The conditioning variables $Y, \tilde{Y}_1, \ldots, \tilde{Y}_{i-1}$ on the left side of the equation jointly contain the same information as the collection of remaining lengths $Y = \sum_{j\geq 1} \tilde{Y}_j, \sum_{j\geq 2} \tilde{Y}_j, \ldots, \sum_{j\geq i} \tilde{Y}_j$, and the equation says that only the last variable is relevant for the distribution of $\tilde{Y}_i$. Equation (14.23) expresses that the masses $\tilde{Y}_1, \tilde{Y}_2, \ldots$ of the Poisson-Kingman species sampling process, in size-biased order, are allocated consecutively by chopping them off the remaining mass, where the remaining mass decreases with every new allocation, but the distribution for allocating the next mass keeps the same general form. (An equivalent way of formulating the equation is that the remaining masses form a stationary Markov chain. See Problem 14.8.)

The parameter σ in a stable or generalized gamma process controls the distribution of the number K_n of distinct species when sampling n individuals from the process. A partition model is said to possess *σ-diversity* S if S is a finite strictly positive random variable such that

$$\frac{K_n}{n^\sigma} \to S, \qquad \text{a.s.} \tag{14.24}$$

The Dirichlet process is outside the range of this definition, since it has $K_n/\log n$ converging to a constant. In contrast, by the following theorem stable or generalized gamma partitions with parameter $\sigma > 0$ have diversity equal to $U^{-\sigma}$, for U the mixing variable. If this mixing variable is chosen heavier-tailed (such as is true for the default choice of the special Poisson-Kingman distribution), then the distribution of K_n will be more spread out. This is attractive for modeling clustering in some applications, where the Dirichlet process gives a sufficient number of clusters only if the precision parameter M is sufficiently large, which makes the distribution relatively concentrated and/or makes the analysis sensitive to the prior on this parameter.

Theorem 14.50 (Diversity) *The* PK(ρ_σ, η) *partition for ρ_σ the σ-stable intensity for $0 < \sigma < 1$, or equivalently the generalized gamma intensity, has σ-diversity equal to $U^{-\sigma}$, for $U \sim \eta$. In particular, the* PK($\rho_\sigma|\, y$) *partition has σ-diversity equal to the constant $y^{-\sigma}$.*

A proof of the theorem can be based on a result of Karlin (1967), which connects the σ-diversity for sampling from a (random) distribution on $\mathbb{N}$ with the limiting behavior of the ranked probabilities $p_{(1)} \geq p_{(2)} \geq \cdots$. It can be shown that $K_n \sim Sn^\sigma$ if $p_{(i)} \sim (S/\Gamma(1-\sigma)i)^{1/\sigma}$. For the process PK($\rho_\sigma|\, y$) this can be shown for $S = y^{-\sigma}$ (see Kingman 1975).

14.6 Normalized Inverse-Gaussian Process

The normalized inverse-Gaussian process is both a special Poisson-Kingman process, and a specific Gibbs process of type $1/2$. It is of special interest, because the marginal distributions $(P(A_1), \ldots, P(A_k))$ of its species sampling process P admit explicit expressions. This is similar to the Dirichlet process, and allows to define the process in the same way as a Dirichlet process.

Appendix H gives an introduction to finite-dimensional inverse-Gaussian distributions.

Definition 14.51 (Normalized inverse-Gaussian process) A random probability measure P on a Polish space $(\mathfrak{X}, \mathscr{X})$ is said to follow a *normalized inverse-Gaussian process* with base measure α if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{NIGau}(k; \alpha(A_1), \ldots, \alpha(A_k)).$$

In the definition α is a finite Borel measure on $(\mathfrak{X}, \mathscr{X})$. By Proposition H.5 its normalization $\bar{\alpha} = \alpha/|\alpha|$, where $|\alpha| = \alpha(\mathfrak{X})$ is the total mass, is the mean measure of the normalized inverse-Gaussian process; it is called the *center measure*.

To see that the definition is well posed and the random measure exists we may employ Theorem 3.1 and consistency properties of the normalized inverse-Gaussian distribution. Alternatively, in the following theorem the process is constructed as a species sampling process.

Theorem 14.52 *The Poisson-Kingman species sampling process with intensity measure* $\rho(ds) = \frac{1}{2} M \pi^{-1/2} s^{-3/2} e^{-s}\, ds$ *and mean measure* G *is the normalized inverse-Gaussian process with* $\alpha = MG$.

Proof Both the normalized inverse-Gaussian distribution and the Poisson-Kingman distribution arise by normalization. Therefore it suffices to show that a multiple of the unnormalized Poisson-Kingman process $\Phi := \sum_{i=1}^{\infty} Y_i \delta_{\theta_i}$ is an inverse-Gaussian process. We shall show that this is true for 2Φ; it suffices to verify that $2\Phi(A_i) \overset{\text{ind}}{\sim} \mathrm{IGau}(\alpha(A_i), 1)$, for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$.

The marginal distribution of $\Phi(A)$ is the sum of all points in the process obtained by thinning the process $(Y_1, Y_2, \ldots)$ by the indicator variables $(\mathbb{1}\{\theta_1 \in A\}, \mathbb{1}\{\theta_2 \in A\}, \ldots)$. Since these indicators are independent Bernoulli variables with parameter $G(A)$, the thinned process is a Poisson process with intensity measure $G(A)\rho$. Hence by Lemma J.2 the log-Laplace transform of the sum $\Phi(A)$ of its points is given by

$$\log \mathrm{E} e^{-\lambda \Phi(A)} = -G(A) \int (1 - e^{-\lambda s})\, \rho(ds) = -\frac{\alpha(A)}{2\sqrt{\pi}} \int_0^{\infty} (1 - e^{-\lambda s}) s^{-3/2} e^{-s}\, ds.$$

The remaining integral can be evaluated by first taking its derivative with respect to λ, which is $\int_0^{\infty} e^{-(\lambda+1)s} s^{-1/2}\, ds = \Gamma(1/2)(\lambda+1)^{-1/2}$, and next integrating this with respect to λ, resulting in $2\Gamma(1/2)((\lambda+1)^{1/2} - 1)$. Thus the preceding display with λ replaced by 2λ evaluates to $-\alpha(A)((2\lambda+1)^{1/2} - 1)$, which is the exponent of the $\mathrm{IGau}(\alpha(A), 1)$-distribution, by Proposition H.4.

The variables $\Phi(A_1), \ldots, \Phi(A_k)$ are the sums of all points in the k Poisson processes obtained by separating (or thinning) the process (Y_j) in k processes using the independent multinomial vectors $(\mathbb{1}\{\theta_j \in A_1\}, \ldots, \mathbb{1}\{\theta_j \in A_k\})$. By general properties of Poisson processes these k processes are independent, and hence so are the sums of their points. □

The intensity measure ρ in Theorem 14.52 corresponds to the generalized gamma process with $\sigma = 1/2$ and $\tau = M^2$, considered in Example 14.48. In particular, the normalized inverse-Gaussian process is Gibbs of type $1/2$. Its EPPF and PPF admit explicit, albeit a little complicated, expressions. They follow the general forms for Gibbs processes, given in (14.16) and (14.18), with the coefficients $V_{n,k}$ given in Example 14.48.

The number of distinct species in a sample of size n from a normalized inverse-Gaussian process is of the order $\sqrt{n}$, as for every generalized gamma process of type $1/2$. The following proposition specializes the formula for the exact distribution of the number of species, given in Proposition 14.23 for general Gibbs processes, to the normalized inverse-Gaussian process.

Proposition 14.53 *If P follows a normalized inverse-Gaussian process with atomless base measure α, then the number K_n of distinct values in a sample $X_1, \ldots, X_n \mid P \overset{iid}{\sim} P$ satisfies, for $k = 1, \ldots, n$,*

$$\mathrm{P}(K_n = k) = \binom{2n-k-1}{n-1} \frac{e^{|\alpha|}(-|\alpha|)^{2n-2}}{2^{2n-k-1}\Gamma(k)} \sum_{r=0}^{n-1} \binom{n-1}{r} (-1)^r |\alpha|^{-2r} \Gamma(k+2+2r-2n; |\alpha|).$$

Being a Poisson-Kingman process, the normalized inverse-Gaussian process allows an easy characterization through residual allocation. The ensuing stick-breaking algorithm given in Theorem 14.49 can be described in terms of auxiliary variables, with standard (albeit somewhat exotic) distributions, as in the following proposition.[10]

Proposition 14.54 (Stick-breaking) *Let $\xi_1 \sim \mathrm{GIGau}(M^2, 1, -1/2)$ be independent of the sequence $\zeta_1^{-1}, \zeta_2^{-1}, \ldots \overset{iid}{\sim} \mathrm{Ga}(1/2, 1/2)$, and for $i = 2, 3, \ldots$,*

$$\xi_i \mid \xi_1, \ldots, \xi_{i-1}, \zeta_1, \ldots, \zeta_{i-1} \sim \mathrm{GIGau}\Big(M^2 \prod_{j=1}^{i-1} \Big(1 + \frac{\xi_j}{\zeta_j}\Big), 1, -\frac{i}{2}\Big).$$

Then $\tilde{W}_j = V_j \prod_{l=1}^{j-1}(1-V_l)$ for $V_j = \xi_j/(\xi_j+\zeta_j)$ give the weight sequence of the Poisson-Kingman process with generalized gamma intensity measure with parameters $\sigma = 1/2$ and $\tau = M^2$. Consequently, the process $\sum_{j=1}^{\infty} \tilde{W}_j \delta_{\theta_j}$, with $\theta_j \overset{iid}{\sim} G$, for $j = 1, 2, \ldots$, follows a normalized inverse-Gaussian process with base measure $\alpha = MG$.

The normalized inverse-Gaussian process provides an alternative to the Dirichlet process as a prior on mixing distributions. Empirical studies show that compared with the Dirichlet process the resulting procedure is less sensitive to the choice of the total mass parameter $|\alpha|$.

[10] See Definition H.6 for the "generalized inverse-Gaussian distribution." The positive $1/2$-stable distribution is the distribution of of $1/Z$ for $Z \sim \mathrm{Ga}(1/2, b/2)$. It has density $z \mapsto (2\pi)^{-1/2} b^{1/2} z^{-3/2} e^{-b/(2z)}$ on $(0, \infty)$. For details of the derivation, see Favaro et al. (2012b).

As for the Dirichlet process, the prior and posterior distributions of the mean of a normalized inverse-Gaussian process can be obtained explicitly; see Problems 14.13 and 14.14.

14.7 Normalized Completely Random Measures

A completely random measure (CRM) on a Polish space $\mathfrak{X}$ is a random element Φ in the space $(\mathfrak{M}_\infty, \mathscr{M}_\infty)$ of Borel measures on $\mathfrak{X}$ such that the variables $\Phi(A_1), \ldots, \Phi(A_k)$ are independent, for every measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$. Appendix J gives a review of CRMs, and shows that, apart from a deterministic component, a CRM can be represented as

$$\Phi(A) = \int_A \int_0^\infty s\, N^c(dx, ds) + \sum_j \Phi_j \delta_{a_j}(A),$$

for a Poisson process N^c on $\mathfrak{X} \times (0, \infty]$, arbitrary independent, nonnegative random variables $\Phi_1, \Phi_2, \ldots$, and given elements $a_1, a_2, \ldots$ of $\mathfrak{X}$, called "fixed atoms" or "fixed jumps." For N^d the point process consisting of the points (a_j, Φ_j) and $N = N^c + N^d$, the equation can be written succinctly as $\Phi(A) = \int_A \int_0^\infty s\, N(dx, ds)$, but N is not a Poisson process, but called an "extended Poisson process." The distribution of N, and hence of Φ, is determined by its intensity measure ν on $\mathfrak{X} \times [0, \infty)$, which splits as $\nu = \nu^c + \nu^d$, where ν^c is the mean measure of N^c and gives mass 0 to every strip $\{x\} \times [0, \infty]$, and ν^d is concentrated on $\cup_j \{a_j\} \times [0, \infty]$, with $\nu^d(\{a_j\} \times B) = \mathrm{P}(\Phi_j \in B)$, for every j. The representation shows in particular that the realizations of a CRM are discrete measures.

If $0 < \Phi(\mathfrak{X}) < \infty$ a.s., then the CRM can be renormalized to the random probability measure $\Phi/\Phi(\mathfrak{X})$, which is also discrete.

Definition 14.55 (Normalized completely random measure) The random probability measure $\Phi/\Phi(\mathfrak{X})$, for a given completely random measure Φ with $0 < \Phi(\mathfrak{X}) < \infty$ a.s., is called a *normalized completely random measure* (NCRM), and for $\mathfrak{X}$ an interval in $\mathbb{R}$, also a *normalized random measure with independent increments* (NRMI).

The "independent increments" in the second name refers to the situation that a CRM can be identified with a process with independent increments through its cumulative distribution function.

Since there is a one-to-one correspondence between CRMs and their intensity measures, every intensity measure defines an NCRM. However, the normalization to a probability measure cancels the scale of the jump sizes so that the measure $\nu(dx, c\, ds)$ gives the same NCRM as $\nu(dx, ds)$, for every $c > 0$.

For a prior specification we shall usually choose an intensity measure without discrete component (i.e. $\nu^d = 0$), but "fixed jumps" will appear in the posterior distribution at the observations, as shown in the theorem below.

A completely random measure without fixed atoms is said to be *homogeneous* if its intensity measure is equal to a product measure $\nu = G \times \rho$, for an atomless probability measure G on $\mathfrak{X}$. Such a measure can be generated by independently laying down point masses $(Y_1, Y_2, \ldots)$ according to a Poisson point process on $(0, \infty)$ with intensity measure ρ and

locations $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$ in $\mathfrak{X}$, and next forming $\Phi = \sum_i Y_i \delta_{\theta_i}$ (the corresponding Poisson process sits at the points (θ_i, Y_i); see Example J.8). The normalized completely random measure is then $\sum_i (Y_i/Y)\delta_{\theta_i}$, for $Y = \sum_i Y_i$, which is exactly the species sampling model corresponding to the PK(ρ)-distribution (see Definition 14.40). Therefore, homogeneous NCRM do not add beyond the special Poisson-Kingman distribution, and are proper species sampling models. On the other hand, non-homogeneous normalized completely random measures are not species sampling models.

Although for constructing a prior distribution we shall usually choose a homogeneous intensity measure, the perspective of NCRMs leads to a different representation of the posterior distribution. Even though the NCRM does not have a structural conjugacy property, except in the special case of the Dirichlet process.[11] the posterior can be described as a *mixture* of NCRMs. The representation in the following theorem allows simulating the posterior distribution by the general *Ferguson-Klass algorithm* for simulating a completely random measure.[12]

As before let $\tilde{X}_1, \tilde{X}_2, \ldots$ be the distinct values in $X_1, X_2, \ldots$, and let $(N_{1,n}, \ldots, N_{K_n,n})$ give the number of times they appear in $D_n := (X_1, \ldots, X_n)$.

Theorem 14.56 (Posterior distribution) *If P follows a completely random measure with intensity measure $\rho(ds|x)\,\alpha(dx)$ for $x \mapsto \rho(ds|x)$ weakly continuous and α atomless, then the posterior distribution of P in the model $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ is a mixture over λ of the distributions of the NCRM $\Phi_\lambda/\Phi_\lambda(\mathfrak{X})$ with intensity measures*

$$\nu^c_{\Phi|\lambda,D_n}(dx, ds|\lambda) = e^{-\lambda s}\,\rho(ds|x)\,\alpha(dx),$$
$$\nu^d_{\Phi|\lambda,D_n}(\{\tilde{X}_j\}, ds|\lambda) \propto s^{N_{j,n}} e^{-\lambda s}\,\rho(ds|\tilde{X}_j),$$

where $\nu^d_{\Phi|\lambda,D_n}(\{x\}, \cdot|\lambda)$ are probability measures on $(0,\infty)$, with mixing density proportional to, with $\psi(\lambda) = \int\int (1 - e^{-\lambda s})\,\rho(ds|x)\,\alpha(dx)$,

$$\lambda \mapsto \lambda^{n-1} e^{-\psi(\lambda)} \prod_{j=1}^{K_n} \int_0^\infty s^{N_{j,n}} e^{-\lambda s} \rho(ds|\tilde{X}_j).$$

Proof For a given partition $A_1, \ldots, A_k$ of the sample space the vector $M = (M_1, \ldots, M_k)$ of counts $M_j = \#\{i : X_i \in A_j\}$ possesses a multinomial distribution with cell probabilities $\Phi(A_j)/\Phi(\mathfrak{X})$ and hence has likelihood $L_M(\Phi) \propto \Phi(\mathfrak{X})^{-n} \prod_{j=1}^k \Phi(A_j)^{M_j}$. By Bayes's rule, for any bounded function f,

$$\mathrm{E}(e^{-\int f\,d\Phi}|M) = \frac{\mathrm{E}_\Phi(e^{-\int f\,d\Phi} L_M(\Phi))}{\mathrm{E}_\Phi(L_M(\Phi))}.$$

Here the expectations on the right side are taken relative to the prior of Φ. We shall explicitly evaluate this expression, and then take the limit along partitions that become finer and finer and whose union generate the Borel σ-field of $\mathfrak{X}$. The martingale convergence theorem then

[11] The Dirichlet process prior is the only NCRM whose posterior distribution is again an NCRM; see Theorem 1 of James et al. (2006).

[12] See Barrios et al. (2013) and Favaro and Teh (2013) for descriptions.

implies that the left side tends to $\mathrm{E}(e^{-\int f\,d\Phi} | X_1, \ldots, X_n)$, and the posterior distribution of Φ can be identified from the limit of the right side.

We start by rewriting the term $\Phi(\mathfrak{X})^{-n}$ in the likelihood as $\int \lambda^{n-1} \prod_j e^{-\lambda\Phi(A_j)}\,d\lambda/\,\Gamma(n)$. Using Fubini's theorem and the independence of Φ on the disjoint sets A_j, we see that on the event $M = m$,

$$\mathrm{E}_\Phi(e^{-\int f\,d\Phi} L_M(\Phi)) = \int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n)} \prod_j \mathrm{E}\Big[e^{-\int_{A_j}(f+\lambda)\,d\Phi}\Phi(A_j)^{m_j}\Big]\,d\lambda.$$

The expected values in the product take the general form

$$\frac{d^m}{d\lambda^m}\mathrm{E}\Big[e^{-\int_A (f+\lambda)\,d\Phi}\Big](-1)^m = \frac{d^m}{d\lambda^m} e^{-\int_A \int (1-e^{-s(f(x)+\lambda)})\,\nu(dx,ds)}(-1)^m,$$

by the expression for the Laplace transform of a CRM, given in Proposition J.6. The derivative on the right side can be computed, and be written in the form

$$e^{-\int_A \int (1-e^{-s(f(x)+\lambda)})\,\nu(dx,ds)} \sum_{(i_1,\ldots,i_r)} a_{i_1,\ldots,i_r} \prod_{l=1}^r \int_A \int e^{-s(f(x)+\lambda)} s^{i_l}\,\nu(dx,ds),$$

where the sum is over all partitions $(i_1, \ldots, i_r)$ of m and is to be read as 1 if $m = 0$. As $\alpha(A) \to 0$, all the integrals $\int_A \int e^{-s(f(x)+\lambda)} s^i\,\nu(dx,ds)$ tend to zero, and hence the sum is dominated by the term for which the product contains only one term ($r = 1$), which corresponds to the partition of m in a single set, and has coefficient $a_m = (-1)^m$.

Only finitely many of the partitioning sets contain data points and have $m_j > 0$. If the partition is fine enough so that the distinct values $\tilde{X}_i$ are in different sets, then the nonzero m_i are equal to the multiplicities $N_{i,n}$, and if f is uniformly continuous, then $f(x) \to f(\tilde{X}_i)$ along the sequence of partitions uniformly in x in the partitioning set $\tilde{A}_i$ that contains $\tilde{X}_i$. We conclude that, as the partitions become finer, the numerator $\mathrm{E}_\Phi(e^{-\int f\,d\Phi} L_M(\Phi))$ is asymptotic to

$$\int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n)} e^{-\int\int (1-e^{-s(f(x)+\lambda)})\,\nu(dx,ds)} \prod_i \Big[\int e^{-s(f(\tilde{X}_i)+\lambda)} s^{N_{i,n}}\,\rho(ds|\,\tilde{X}_i)\alpha(\tilde{A}_i)\Big]\,d\lambda.$$

When divided by $\mathrm{E}_\Phi L_M(\Phi)$, the factors $\alpha(\tilde{A}_i)$ cancel and both numerator and denominator tend to a limit. It remains to write this in the form corresponding to the claim of the theorem.

In view of (J.10), the product over i corresponds to fixed jumps, where the norming factors $\int e^{-s\lambda} s^{N_{i,n}}\,\rho(ds|\,\tilde{X}_i)$ of the integrals can be compensated in the integral over λ. The leading exponential can be written as $e^{-\psi(\lambda)} \exp\big[-\int\int (1-e^{-sf(x)})\,e^{-\lambda s}\,\nu(dx,ds)\big]$, and contributes a factor $e^{-\psi(\lambda)}$ to the integral over λ and the continuous part of the CRM. The denominator provides the norming factor to the integral over λ. □

Example 14.57 (Normalized generalized gamma CRM) If P follows the normalized process of the generalized gamma CRM described in Example 14.48, the continuous part $\nu^c(\cdot|\,\lambda)$ of the CRM Φ_λ in the theorem has intensity measure

$$\tau^\sigma\,(\sigma/\,\Gamma(1-\sigma))\,s^{-(1+\sigma)}\,e^{-(1+\lambda)s}\alpha(dx)\,ds,$$

again of generalized gamma form; and the distributions in the discrete part are $\mathrm{Ga}(N_{i,n} - \sigma, \lambda + 1)$. Furthermore, the mixing density is proportional to $\lambda \mapsto \lambda^{n-1}(1+\lambda)^{K_n\sigma - n} e^{-|\alpha|(1+\lambda)^\sigma}$.

The normalized generalized gamma CRM is the only normalized random measure of Gibbs type (see Lijoi et al. 2008, Proposition 2).

Example 14.58 (Normalized extended gamma CRM) An example of a non-homogeneous CRM is given by the extended gamma process defined in Example J.15. This has intensity measure given by $s^{-1}e^{-s/b(x)}\alpha(dx)\,ds$, for a positive measurable function b and a σ-finite measure α with $\int \log(1+b)\,d\alpha < \infty$. The corresponding normalized measure is called an *extended gamma CRM*.

Corollary 14.59 (Predictive distribution) *If $X_1, X_2, \ldots \mid P$ are a random sample from a normalized random measure P with intensity $\nu(dx, ds) = \rho(ds \mid x)\,\alpha(dx)$ satisfying the conditions of Theorem 14.56, then*

$$X_{n+1} \mid X_1, \ldots, X_n \sim \sum_{j=1}^{K_n} p_j(X_1, \ldots, X_n)\delta_{\tilde{X}_j} + p_{K_n+1}(X_1, \ldots, X_n \mid x)\,\bar{\alpha},$$

for, with $\tau_n(\lambda \mid x) = \int s^n e^{-\lambda s}\,\rho(ds \mid x)$,

$$p_j(X_1, \ldots, X_n) = \frac{\int_0^\infty \lambda^n e^{-\psi(\lambda)} \tau_{N_{j,n}+1}(\lambda \mid \tilde{X}_i) \prod_{i=1, i\neq j}^{K_n} \tau_{N_{i,n}}(\lambda \mid \tilde{X}_i)\,d\lambda}{n \int_0^\infty \lambda^{n-1} e^{-\psi(\lambda)} \prod_{i=1}^{K_n} \tau_{N_{i,n}}(\lambda \mid \tilde{X}_i)\,d\lambda},$$

$$p_{K_n+1}(X_1, \ldots, X_n \mid x) = \frac{\int_0^\infty \lambda^n e^{-\psi(\lambda)} \tau_1(\lambda \mid x) \prod_{i=1}^{K_n} \tau_{N_{i,n}}(\lambda \mid \tilde{X}_i)\,d\lambda}{n \int_0^\infty \lambda^{n-1} e^{-\psi(\lambda)} \prod_{i=1}^{K_n} \tau_{N_{i,n}}(\lambda \mid \tilde{X}_i)\,d\lambda}.$$

Proof The predictive distribution is the mean $\mathrm{E}\big[P \mid X_1, \ldots, X_n\big]$ of the posterior distribution of P. By Theorem 14.56 the posterior distribution of P is a mixture of normalized completely random measures, and hence the posterior mean is a mixture of means of normalized completely random measures. For the latter means, we apply the general formula given in Lemma 14.60, with the intensity measure equal to the one given in Theorem 14.56. The Laplace exponent of the latter measure is $\psi_{\Phi|\lambda, D_n}(\theta) = \int\int (1 - e^{-\theta s}) e^{-\lambda s} \rho(ds \mid x)\,\alpha(dx) = \psi(\theta + \lambda) - \psi(\lambda)$, for ψ the Laplace exponent of the prior intensity measure, while for $k = 0, 1$,

$$\int s^k e^{-\theta s}\, d\nu^d_{\Phi|\lambda, D_n}(\{\tilde{X}_i\}, ds) = \frac{\tau_{N_{i,n}+k}(\theta + \lambda \mid \tilde{X}_i)}{\tau_{N_{i,n}}(\lambda \mid \tilde{X}_i)}.$$

We substitute these expressions in the formula of Lemma 14.60, taking the integration variable there equal to θ, multiply with the density of λ from Theorem 14.56, which is proportional to $\lambda^{n-1} e^{-\psi(\lambda)} \prod_i \tau_{N_{i,n}}(\lambda \mid X_i)$, and integrate over λ. In the resulting double integral with respect to $(\lambda, \theta) > 0$, we make the change of variables $\theta + \lambda = u$, next integrate over $\lambda \in (0, u)$, giving the factor u^n/n. The sum over the fixed jump points then gives the formulas for p_j and $j \leq K_n$, while the continuous part gives the formula for p_{K_n+1}, after an application of Fubini's theorem. □

Lemma 14.60 (Mean) *The mean measure of the normalized completely random measure $P = \Phi/\Phi(\mathfrak{X})$ with intensity measure ν and strictly positive fixed jumps $\Phi(\{a_j\})$ at the points (a_j) is given by, for $\psi(\lambda) = \int\int (1 - e^{\lambda s})\, \nu^c(dx, ds)$,*

$$\mathrm{E}\Big[\frac{\Phi(A)}{\Phi(\mathfrak{X})}\Big] = \int e^{-\psi(\lambda)} \int_A \int s e^{-\lambda s}\, \nu^c(dx, ds) \prod_j \int e^{-\lambda s}\, \nu^d(\{a_j\}, ds)\, d\lambda$$
$$+ \sum_{j:a_j \in A} \int e^{-\psi(\lambda)} \int s e^{-\lambda s}\, \nu^d(\{a_j\}, ds) \prod_{i \neq j} \int e^{-\lambda s}\, \nu^d(\{a_i\}, ds)\, d\lambda.$$

Proof Applying the identity $y^{-1} = \int e^{-\lambda y}\, d\lambda$ to $y = \Phi(\mathfrak{X})$ and Fubini's theorem, we can write the left side in the form $\int \mathrm{E}\big[\Phi(A) e^{-\lambda \Phi(\mathfrak{X})}\big]\, d\lambda$. Next we decompose $\Phi(\mathfrak{X}) = \Phi(A) + \Phi(A^c)$ and use the independence of the two variables on the right to further rewrite the left side of the lemma as $\int \mathrm{E}\big[\Phi(A) e^{-\lambda \Phi(A)}\big]\, \mathrm{E}\big[e^{-\lambda \Phi(A^c)}\big]\, d\lambda$. By combining (J.7) and (J.10), we have

$$\mathrm{E}\big[e^{-\lambda \Phi(A^c)}\big] = e^{-\int_{A^c} \int (1 - e^{-\lambda s})\, \nu^c(dx, ds)} \prod_{j:a_j \in A^c} \int e^{-\lambda s}\, \nu^d(\{a_j\}, ds).$$

Applying equations (J.7) and (J.10) a second time, we also find

$$\mathrm{E}\big[\Phi(A) e^{-\lambda \Phi(A)}\big] = -\frac{d}{d\lambda}\Big[e^{-\int_A \int (1 - e^{-\lambda s})\, \nu^c(dx, ds)} \prod_{j:a_j \in A} \int e^{-\lambda s}\, \nu^d(\{a_j\}, ds)\Big].$$

We substitute these expressions in the integral, first for a set A without fixed jumps, and next for $A = \{a_j\}$, and simplify the resulting expression to the one given in the lemma. □

The structure of the predictive distribution is similar to that obtained in species sampling models, in Lemma 14.11. A difference is that the probabilities of redrawing a preceding observation or generating a new value depend on the *values* of the preceding observations, next to the partition they generate. This is due to the fact that for a general normalized completely random measure (when the intensity measure is not homogeneous) the value of an observation and the probability that it is drawn are confounded.

Even though a NCRM is in general not a species sampling model, a random sample $X_1, X_2, \ldots$ from an NCRM is still exchangeable and hence generates an infinite exchangeable partition. The following theorem shows that its EPPF has exactly the same form as that of a special Poisson-Kingman process, given in Theorem 14.44.

Theorem 14.61 (EPPF) *The EPPF of the exchangeable partition generated by a random sample from an NCRM with intensity measure ν without fixed atoms and Laplace exponent $\psi(\lambda) = \int\int (1 - e^{-\lambda s})\, \nu(dx, ds)$ is given by, for $n = \sum_{j=1}^{k} n_j$,*

$$p(n_1, \ldots, n_k) = \frac{(-1)^{n-k}}{\Gamma(n)} \int \lambda^{n-1} e^{-\psi(\lambda)} \prod_{j=1}^{k} \psi^{(n_j)}(\lambda)\, d\lambda.$$

Proof Because ν does not have fixed atoms, the corresponding (extended) Poisson process has at most one point on every half line $\{x\} \times (0, \infty)$. Sampling an observation X from a NCRM $\Phi/\Phi(\mathfrak{X})$ given Φ is therefore equivalent to choosing one of the points (x_j, s_j) from the realization of Φ. The pattern of ties in a random sample $X_1, X_2, \ldots$ arises by the observations X_i choosing the same or different points. The probability of choosing a point (x_j, s_j) is equal to $s_j / \sum_k s_k$, for $\sum_k s_k$ equal to the realization of $\Phi(\mathfrak{X})$; it does not depend on the value of x_j. It follows that the mechanism of forming ties depends on the weights s_j only, which form a Poisson process on $(0, \infty)$ with intensity measure the marginal measure of ν given by $\rho_2(D) = \nu(\mathfrak{X} \times D)$. The Laplace exponent of ρ_2 is equal to $\int (1 - e^{-\lambda s})\, \rho_2(ds) = \int\int (1 - e^{-\lambda s})\, \nu(dx, ds) = \psi(\lambda)$. The theorem is a corollary of Theorem 14.44. □

If we write $\kappa_n(\lambda) = \int\int s^n e^{-\lambda s}\, \nu(dx, ds)$, then the theorem shows that given λ the EPPF is essentially of the product form $\prod_{j=1}^{k} \kappa_{n_j}(\lambda)$, and hence the exchangeable partition generated by a normalized completely random measure is conditionally Gibbs, by Lemma 14.21. By reweighting we can write, for any positive probability density g,

$$p(n_1, \ldots, n_k) = \int p(n_1, \ldots, n_k | \lambda)\, g(\lambda)\, d\lambda, \quad p(n_1, \ldots, n_k | \lambda) = \frac{\lambda^{n-1} e^{-\psi(\lambda)}}{\Gamma(n) g(\lambda)} \prod_{j=1}^{k} \kappa_{n_j}(\lambda).$$

The Gibbs form is convenient, as general results on Gibbs processes become available, conditional on λ.

In a species sampling model the distinct values $\tilde{X}_1, \tilde{X}_2, \ldots$ can be viewed as generated independently from the center measure G and attached to the partitioning sets as the common values of the observations (see Theorem 14.13). For general normalized random measures the partitioning and the distinct values are dependent. Their joint distribution is given in the following lemma, a version of which was used in the proof of Theorem 14.44.

Lemma 14.62 *The distinct values $\tilde{X}_1, \tilde{X}_2, \ldots$ and partitions $\mathcal{P}_n$ generated by a random sample from a normalized completely random measure with intensity measure $\nu(dx, ds) = \rho(ds | x)\, \alpha(dx)$ without fixed atoms and Laplace exponent $\psi(\lambda) = \int\int (1 - e^{-\lambda s})\, \nu(dx, ds)$ satisfy, for every measurable sets $B_1, \ldots, B_k$ and partition $A_1, \ldots, A_k$ of $\{1, 2, \ldots, n\}$, with $\tau_n(\lambda | x) = \int s^n e^{-\lambda s}\, \rho(ds | x)$,*

$$\begin{aligned} &\mathrm{P}(\tilde{X}_1 \in B_1, \ldots, \tilde{X}_k \in B_k, \mathcal{P}_n = \{A_1, \ldots, A_k\}) \\ &\quad = \int_{B_1} \cdots \int_{B_k} \int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n)} e^{-\psi(\lambda)} \prod_{j=1}^{k} \tau_{n_j}(\lambda | x_j)\, d\lambda\, \alpha(dx_1) \cdots \alpha(dx_k). \end{aligned}$$

Proof For notational convenience we give the proof in the case that $k = 2$ only. The partition structure $\mathcal{P}_n = \{A_1, A_2\}$ expresses that the observations $X_1, \ldots, X_n$ are obtained in

a particular order in two groups, where two different points (x_1, s_1) and (x_2, s_2) are chosen from the point process N on $\mathfrak{X} \times (0, \infty)$ that generates the NCRM, by two particular observations (say with indices $\min A_1$ and $\min A_2$), setting these equal to x_1 and x_2, and all other observations choose either (x_1, s_1) or (x_2, s_2), depending on whether their index is in A_1 or in A_2. Given N the observations are independent and choose a point (x, s) with probability s/S, for $S = \int\int s\, N(dx, ds)$ the "total weight" in N. It follows that, for $(n_1, n_2) = (|A_1|, |A_2|)$,

$$\begin{aligned}\mathrm{P}(\tilde{X}_1 \in B_1, \tilde{X}_2 \in B_2, \mathcal{P}_n = \{A_1, A_2\} \,|\, N) \\ = \int_{B_1} \int \Big(\frac{s_1}{S}\Big)^{n_1 - 1} \int_{B_2} \int \Big(\frac{s_2}{S}\Big)^{n_2 - 1} \frac{s_2}{S}\, (N - \{(x_1, s_1)\})(dx_2, ds_2)\, \frac{s_1}{S}\, N(dx_1, ds_1).\end{aligned}$$

Here $N - \{(x_1, s_1)\}$ is the point process N with the point (x_1, s_1) removed; this arises because the point (x_2, s_2) must be different from the point (x_1, s_1). The two factors $(s_i/S)^{n_i - 1}$ are the (conditional) probabilities that $n_i - 1$ points choose the point (x_i, s_i), and $(s_1/S)\, N(dx_1, ds_1)$ and $(s_2/S)\, (N - \{(x_1, s_1)\})(dx_2, ds_2)$ are the law of the first distinct point, and the conditional law of the second distinct point given that it is different from the first, respectively.

The probability in the lemma is the expectation over the preceding display relative to N. We next proceed by exchanging the order of expectation and integration twice, with the help of formula (J.1), which is based on Palm theory for the Poisson process N. Application of the formula on $N(dx_1, ds_1)$, with the other, inner integral treated as a fixed function of N, gives that the expectation of the preceding display is equal to

$$\int_{B_1} \int \mathrm{E}\Big(\frac{s_1}{S + s_1}\Big)^{n_1} \int_{B_2} \int \Big(\frac{s_2}{S + s_1}\Big)^{n_2} N(dx_2, ds_2)\, \nu(dx_1, ds_1).$$

To see this, note that (J.1) demands to replace N by the process $N \cup \{(x_1, s_1)\}$, thus changing $S = \int\int s\, N(dx, ds)$ into $S + s_1$, and $N - \{(x_1, s_1)\}$ into N. A second application of formula (J.1), now on $N(dx_2, ds_2)$, allows to rewrite this further as

$$\int_{B_1} \int \int_{B_2} \int \mathrm{E}\Big(\frac{s_1}{S + s_1 + s_2}\Big)^{n_1} \Big(\frac{s_2}{S + s_1 + s_2}\Big)^{n_2} \nu(dx_2, ds_2)\, \nu(dx_1, ds_1).$$

The remaining expectation refers only to the variable S, which is the total weight of the Poisson process with intensity measure $\nu(\mathfrak{X}, ds)$ on $(0, \infty)$ and has density f with Laplace transform $e^{-\psi(\lambda)}$. Thus the preceding display is equal to

$$\int_{B_1} \int \int_{B_2} \int \int \Big(\frac{s_1}{t + s_1 + s_2}\Big)^{n_1} \Big(\frac{s_2}{t + s_1 + s_2}\Big)^{n_2} f(t)\, dt\, \nu(dx_2, ds_2)\, \nu(dx_1, ds_1).$$

To arrive at the expression given by the lemma we express the factor $(t + s_1 + s_2)^{-n}$, as $\Gamma(n)^{-1} \int_0^\infty \lambda^{n-1} e^{-\lambda(t + s_1 + s_2)}\, d\lambda$, rearrange by several applications of Fubini's theorem, and use that $\int e^{-\lambda t} f(t)\, dt = e^{-\psi(\lambda)}$. □

14.8 Relations between Classes of Discrete Random Probability Measures

Figure 14.5 pictures the relations between the various classes of discrete random probability measures developed in the preceding sections. Except for the non-homogeneous normalized

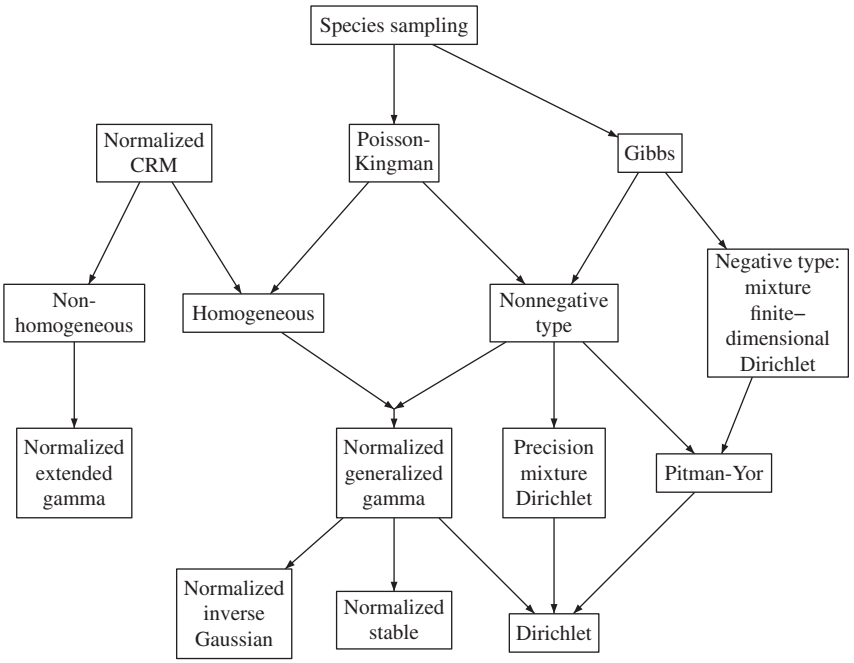

Figure 14.5 Relations between classes of random probability measures. An arrow indicates that a target is a special (or limiting) case of the origin of the arrow.

completely random measures, all classes are species sampling models, and characterized by the distribution of an infinite probability vector. In a Poisson-Kingman process this distribution is induced by normalizing the points of a Poisson process in $(0, \infty)$ by its sum. This distribution may be conditioned on the sum, or mixed over the sum, using either the intrinsic distribution of the sum or an arbitrary distribution. Mixtures with the intrinsic distribution give exactly the same class of random distributions as the homogeneous normalized completely random measures. Mixtures with an arbitrary distribution and the Poisson process equal to a σ-stable process or generalized gamma process give all Gibbs processes of type $0 < \sigma < 1$. Gibbs processes are precisely those species sampling processes for which the exchangeable partition probability function factorizes (and the probability of discovering a new species depends only on the sample size and the number of distinct elements). A Gibbs process of type $\sigma = 0$ is a mixture of Dirichlet DP(MG)-distributions over M, while negative type Gibbs processes are mixtures of finite-dimensional Dirichlet Dir$(m; -\sigma, \ldots, -\sigma)$-distributions over m. Within the Gibbs processes the Pitman-Yor processes are prominent, and contain the Dirichlet process as a special case. They can be obtained as Poisson-Kingman processes with σ-stable intensity and mixing distribution a polynomially tilted positive σ-stable distribution. All Pitman-Yor processes admit a stick-breaking representation with independent beta variables. Normalized completely random measures with generalized gamma intensity form the intersection of Gibbs processes and normalized completely random measures. This class parameterized by (σ, τ) contains the Dirichlet process in the limit $\sigma \to 0$, the normalized inverse-Gaussian process for $\sigma = 1/2$, and the normalized σ-stable process in the limit $\tau \to 0$.

14.9 Dependent Random Discrete Distributions

Many applications, especially in spatial statistics, involve a collection of distributions P_z on a sample space $\mathfrak{X}$, indexed by a parameter z belonging to some "covariate" space $\mathfrak{Z}$. A typical example is that P_z is the conditional distribution of a variable X given a covariate $Z = z$; *density regression* then focuses on estimating the conditional density of X given Z. A useful prior distribution on the P_z should not treat these measures separately, but view them as related quantities. As an extreme case all P_z could be taken equal, and a prior distribution be imposed on the common measure, but such total dependence is usually too restrictive. More typical would be to incorporate some form of continuity of the map $z \mapsto P_z$ into the prior specification.

One possibility is to model the variables x and z simultaneously, for instance through a stochastic process $(\xi(x, z): (x, z) \in \mathfrak{X} \times \mathfrak{Z})$ or a set of basis functions on the product space $\mathfrak{X} \times \mathfrak{Z}$. For instance, a prior distribution on the conditional density of X given Z can be constructed from a Gaussian process ξ and a link function Ψ through the relation

$$p(x \mid z) = \frac{\Phi(\xi(x, z))}{\int \Phi(\xi(y, z))\, dy}.$$

Alternatively, a random series in (x, z) may replace the Gaussian process. Continuity in z is then expressed in the stochastic process or basis functions, where it may be sensible to treat the arguments x and z asymmetrically.

In this section we focus on priors of the form

$$P_z = \sum_{j=1}^{\infty} W_j(z)\delta_{\theta_j(z)}, \tag{14.25}$$

for $\mathfrak{Z}$-indexed stochastic processes of "weights" $(W_j(z): z \in \mathfrak{Z})$ and "locations" $(\theta_j(z): z \in \mathfrak{Z})$. The processes θ_j may be i.i.d. copies of some $\mathfrak{X}$-valued process; a Gaussian process may be convenient; an infinite-dimensional exponential family, or a random series are alternatives. The weights W_j could be constructed using stick breaking:

$$W_j(z) = V_j(z) \prod_{l=1}^{j-1} (1 - V_l(z)). \tag{14.26}$$

The processes $(V_j(z): z \in \mathfrak{Z})$ can be taken independent (or even i.i.d.), and should then satisfy $\sum_{j=1}^{\infty} \log \mathrm{E}\big[1 - V_j(z)\big] = -\infty$ (cf. Lemma 3.4) to ensure that the P_z are proper probability distributions (and in addition a regularity condition in z such as separability to ensure that all P_z are simultaneously proper almost surely).

A process of random measures $(P_z: z \in \mathfrak{Z})$ defined by the combination of (14.25) and (14.26) is called a *dependent stick-breaking process* (DSBP).

In a dependent stick-breaking process closeness of $V_j(z)$ and $V_j(z')$, and $\theta_j(z)$ and $\theta_j(z')$, for $j \in \mathbb{N}$ and $z, z' \in \mathfrak{Z}$, implies closeness of P_z and $P_{z'}$. One of the processes V_j or θ_j may be chosen constant in z without losing too much flexibiity. The extent of sharing of atoms across different locations might be controlled by an auxiliary variable.

An advantage of stick-breaking weights is that the series can be truncated to a finite number of terms, thus effectively reducing computations to finitely many variables. For posterior computation, the Metropolis-Hastings algorithm is applicable, but more specialized algorithms, such as a block-Gibbs sampler (cf. Ishwaran and James 2001), often work better.

Measurability issues may be resolved by the result that a map of two arguments that is continuous in the first argument and measurable in the second, is measurable provided the spaces involved are nice, such as Polish. In the present case P_z is a measurable random element in $\mathfrak{M}$, for every z, and $z \mapsto P_z$ is weakly continuous if the processes V_j and θ_j have weakly continuous sample paths, since $x \mapsto \delta_x$ is weakly continuous.

In view of the success of the Dirichlet process as a prior distribution, it is reasonable to make each marginal prior P_z a Dirichlet process. By the stick-breaking representation of a Dirichlet process, the specifications $\theta_j(z) \overset{\text{iid}}{\sim} G_z$ and $V_j(z) \overset{\text{iid}}{\sim} \text{Be}(1, M_z)$, for $j = 1, 2, \ldots$ and a probability measure G_z on $\mathfrak{X}$ and some $M_z > 0$, ensure that $P_z \sim \text{DP}(M_z G_z)$, for every fixed $z \in \mathfrak{Z}$. A process satisfying the latter requirement is referred to as a *dependent Dirichlet process* (DDP). We discuss some specific constructions later in this section.

Instead of Dirac measures $\delta_{\theta_j}(z)$ in (14.25) one might use general random probability distributions. Choosing these independent of z leads to priors of the form

$$P_z = \sum_{j=1}^{\infty} W_j(z) G_j. \tag{14.27}$$

The measures G_j might be i.i.d. from a prior on the set of probability distributions, such as a Dirichlet process.

The density regression problem requires a prior on functions rather than on measures, and an extra smoothing step is necessary to employ a DSBP or a DDP. This can be written in a hierarchical structure with latent variables as

$$Y_i \,|\, X_i, Z_i \overset{\text{iid}}{\sim} \psi(\cdot, X_i, \varphi), \qquad X_i \,|\, Z_i \overset{\text{ind}}{\sim} P_{Z_i}, \qquad i = 1, 2, \ldots,$$

where ψ is a kernel with a (possibly additional) parameter φ, and $\{P_z, z \in \mathfrak{Z}\}$ may follow a DDP or DSBP.

14.9.1 Kernel Stick-Breaking Process

The *kernel stick-breaking process* (KSBP) is the combination of (14.27) with $G_j \overset{\text{iid}}{\sim} \text{DP}(MG)$ and stick-breaking weights (14.26) with relative stick lengths of the form $V_j(z) = U_j K(z, \Gamma_j)$, for random sequences (U_j) and (Γ_j), and a given "kernel function" $K : \mathfrak{Z} \times \mathfrak{Z} \to [0, 1]$. In the special case $M = 0$ we understand $\text{DP}(MG)$ to generate pointmasses $G_j = \delta_{\theta_j}$, for $\theta_j \overset{\text{iid}}{\sim} G$, and then (14.27) reduces to (14.25), with θ_j free of z.

It is sensible to choose the kernel K to have large values near the diagonal, so that $K(z, \gamma)$ and $K(z', \gamma)$ are similar if z and z' are close. For instance, K could be a decreasing function of the distance between its arguments. Typical choices are $U_j \overset{\text{ind}}{\sim} \text{Be}(a_j, b_j)$ and, independently, $\Gamma_j \overset{\text{iid}}{\sim} H$ for some distribution H, for $j = 1, 2, \ldots$. If $\text{E}\big[K(z, \Gamma)\big] > 0$ and $a_j/(a_j + b_j)$ remains bounded away from 0, then $\sum_{j=1}^{\infty} \log \text{E}[1 - V_j(z)] = -\infty$, and the construction marginally defines proper random probability distributions.

Since the G_j have expectation G and are independent of the weights, it follows that $\mathrm{E}(P_z | U_1, \Gamma_1, U_2, \Gamma_2, \ldots) = G$, for all $z \in \mathfrak{Z}$, as in Subsection 3.4.2. Similarly the (conditional) covariance between the measures at two values z, z' can be computed as

$$\operatorname{cov}\left(P_z(A), P_{z'}(A) | U_1, \Gamma_1, U_2, \Gamma_2, \ldots\right) = \sum_{j=1}^{\infty} W_j(z) W_j(z') \frac{G(A) G(A^c)}{M+1}.$$

Unconditional second moments can also be found easily, and particularly attractive formulas are found when the parameters a_j and b_j do not change with j.

Suppose that, given covariates (Z_i), we generated observations $X_i | Z_i \overset{\text{ind}}{\sim} P_{Z_i}$, for $i = 1, \ldots, n$. Since the realizations G_j from the Dirichlet distribution are discrete, with different supports for different j, two observations X_i and X_j can be equal only if they are sampled from the same component G_j. In that case they are as a sample of size 2 from a Dirichlet distribution, they are equal with probability $1/(M+1)$, by (4.13). Thus the probability of pairwise clustering $\mathrm{P}(X_i = X_j | Z_i, Z_j)$, in the KSBP model with i.i.d. vectors $(U_j, \Gamma_j) \sim (U, \Gamma)$ can be computed as

$$\begin{aligned}
\mathrm{P}(X_i = X_j | Z_i, Z_j) &= \sum_{k=1}^{\infty} \mathrm{E}\big[W_k(Z_i) W_k(Z_j)\big] \frac{1}{M+1} \\
&= \frac{\mathrm{E}\big[U^2 K(Z_i, \Gamma) K(Z_j, \Gamma)\big]}{M+1} \sum_{k=1}^{\infty} \Big(\mathrm{E}\big[(1 - UK(Z_i, \Gamma))(1 - UK(Z_j, \Gamma))\big]\Big)^{k-1} \\
&= \frac{\mathrm{E}\big[U^2 K(Z_i, \Gamma) K(Z_j, \Gamma)\big]}{(M+1)\big(\mathrm{E}\big[UK(Z_i, \Gamma)\big] + \mathrm{E}\big[UK(Z_j, \Gamma)\big] - \mathrm{E}\big[U^2 K(Z_i, \Gamma) K(Z_j, \Gamma)\big]\big)}.
\end{aligned}$$

For the important special case that $M = 0$, and $U \sim \mathrm{Be}(a, b)$ independent of Γ, this reduces to

$$\mathrm{P}(X_i = X_j | Z_i, Z_j) = \frac{\frac{a+1}{a+b+1} \mathrm{E}\big[K(Z_i, \Gamma) K(Z_j, \Gamma)\big]}{\mathrm{E}\big[K(Z_i, \Gamma)\big] + \mathrm{E}\big[K(Z_j, \Gamma)\big] - \frac{a+1}{a+b+1} \mathrm{E}\big[K(Z_i, \Gamma) K(Z_j, \Gamma)\big]}. \tag{14.28}$$

14.9.2 Local Dirichlet Process

Let $d: \mathfrak{Z} \times \mathfrak{Z}' \to \mathbb{R}^+$ be some function that "connects" the points of two measurable spaces $\mathfrak{Z}$ and $\mathfrak{Z}'$, and for given $\epsilon > 0$ let $N_\epsilon(z) = \{z' \in \mathfrak{Z}' : d(z, z') < \epsilon\}$ be the points in $\mathfrak{Z}'$ connected to a given $z \in \mathfrak{Z}$. These "neighborhoods" are used to couple the points z in a random fashion as follows. We generate a sample $\Gamma_1, \Gamma_2, \ldots \overset{\text{iid}}{\sim} H$, for a given probability measure H on $\mathfrak{Z}'$, and determine for every $z \in \mathfrak{Z}$ the set of variables Γ_j that fall into the neighborhood $N_\epsilon(z)$. Coupling arises as z with overlapping neighborhoods will share many Γ_j.

Assume that $H(N_\epsilon(z)) > 0$ for all $z \in \mathfrak{Z}$, so that each neighborhood $N_\epsilon(z)$ receives infinitely many Γ_j. For given z let $I_1(z, \epsilon) < I_2(z, \epsilon) < \cdots$ be the (random) indices such that $\Gamma_j \in N_\epsilon(z)$, in increasing order. Next for random samples $V_j \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$ and $\bar{\theta}_j \overset{\text{iid}}{\sim} G$ define stick lengths and support points by

$$W_j(z) = V_{I_j(z,\epsilon)} \prod_{l<j} (1 - V_{I_l(z,\epsilon)}), \qquad \theta_j(z) = \bar{\theta}_{I_j(z,\epsilon)}, \qquad j = 1, 2, \ldots$$

The resulting process (14.25) is called the *local Dirichlet process* (LDP).

Since for a given z, the variables $V_{I_l(z,\epsilon)}$ used for breaking the sticks are a random sample from the Be$(1, M)$-distribution, each process P_z is a DP(MG)-process, for any z. Thus a local Dirichlet process is a dependent Dirichlet process (an "LDP" is a "DDP"). The measures P_z are dependent through the auxiliary sequence Γ_j. If two points z and z' are close, then their sets of indices $(I_j(z, \epsilon))$ and $(I_j(z', \epsilon))$ will have a large intersection, and hence the stick breaks of P_z and $P_{z'}$ are more similar. On the other hand, measures P_z with disjoint sequences of indices are independent.

The parameter ϵ is a tuning parameter, and may be equipped with a prior.

The stick-breaking variables $(V_{I_j(z,\epsilon)})$ attached to z can be augmented to the sequence $V_j(z) = V_j \mathbb{1}\{\Gamma_j \in N_\epsilon(z)\}$, for $j = 1, 2, \ldots$, which has zero values at indices j between the values of the $I_j(z, \epsilon)$ and the values $V_{I_j(z,\epsilon)}$ inserted at their "correct" positions. The corresponding sequence of weights $\tilde{W}_j(z) = V_j(z) \prod_{l<j}(1 - V_l(z))$ has zeros at the same set of "in-between" indices and the values $W_j(z)$ at the indices $I_j(z, \epsilon)$. This shows that $P_z = \sum_{j=1}^{\infty} \tilde{W}_j(z)\delta_{\bar{\theta}_j}$ is an alternative representation of the local Dirichlet process P_z and hence the LDP is a special case of the KSBP with $U_j = V_j$ and the kernel $K(z, \Gamma) = \mathbb{1}\{\gamma \in N_\epsilon(z)\}$. Conequently, the clustering probability can be obtained from (14.28). It can be written in the form

$$\mathrm{P}(X_i = X_j \,|\, Z_i, Z_j) = \frac{2P_{Z_i,Z_j}}{(1 + P_{Z_i,Z_j})M + 2},$$

where $P_{z,z'} = H(N_\epsilon(z) \cap N_\epsilon(z'))/H(N_\epsilon(z) \cup N_\epsilon(z'))$.

14.9.3 Probit Stick-Breaking Process

A *probit stick-breaking process* is a dependent stick-breaking process (14.25) with stick-breaking weights (14.26)

$$V_j(z) = \Phi(\alpha_j + f_j(z)), \qquad \alpha_j \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, 1), \qquad f_j \overset{\text{iid}}{\sim} H, \qquad j = 1, 2, \ldots,$$

where Φ is the standard normal cumulative distribution function. The idea behind using the probit transformation is that it is easier to propose sensible models for (dependent) general real variables than for variables restricted to $[0, 1]$. For example, a Gaussian process prior may be used for the functions f_j. A potential disadvantage is that the $V_j(z)$ are not beta distributed, and hence the resulting process will not be a dependent Dirichlet process.

Conditions for the weights $W_j(z)$ to sum to one can be given in various ways. If the functions f_j are forced to be nonnegative, then $V_j(z) \geq \Phi(\alpha_j)$ by the monotonicity of Φ, so that $\sum_{j=1}^{\infty} \log \mathrm{E}[1 - V_j(z)] \leq \sum_{j=1}^{\infty} \log \mathrm{E}[1 - \Phi(\alpha_j)] = -\infty$ because the α_j are i.i.d. and the expectation is strictly smaller than 1.

14.9.4 Ordering Dependent Stick-Breaking Processes

A random discrete distribution $\sum_{j=1}^{\infty} W_j \delta_{\theta_j}$ is distributionally invariant under permutations of its terms. If for each given z the terms are permuted by a permutation of $\mathbb{N}$ that depends

on z, then the marginal distributions remain the same, and at the same time dependence across z arises. In particular, if the original discrete discrete distribution is a Dirichlet process, then the result is a dependent Dirichlet process. This process is called an *order based dependent Dirichlet process* (πDDP). In practice the permutation may be limited to a finite set $\{1, \ldots, N\}$, and this is clearly sufficient if the series is truncated at a finite number of terms.

Random permutations $\{\mathfrak{p}(z): z \in \mathfrak{Z}\}$ may be generated with the help of a point process $(T_1, T_2, \ldots)$ in $\mathfrak{Z}$. Assume that a given neighborhood $U(z)$ of z contains only finitely many points of this process, and given a metric on $\mathfrak{Z}$ define $\mathfrak{p}_1(z), \mathfrak{p}_2(z), \ldots$ by the relations

$$d(z, T_{\mathfrak{p}_z(1)}) < d(z, T_{\mathfrak{p}_z(2)}) < \cdots < d(z, T_{\mathfrak{p}_z(N_z)}).$$

Specifically, one can consider a Poisson process with absolutely continuous intensity measure. For an expression for the correlation between $P_z(A)$ and $P_{z'}(A)$, see Griffin and Steel (2006).

14.9.5 Nested Dirichlet Processes

For α a (atomless) finite measure on a (Polish) sample space $(\mathfrak{X}, \mathscr{X})$, the specification $Q \sim \mathrm{DP}(M\mathrm{DP}(\alpha))$ defines a random measure Q *on the space of measures* on $(\mathfrak{X}, \mathscr{X})$; equivalently Q is a random element in $\mathfrak{M}(\mathfrak{M}(\mathfrak{X}))$. The stick-breaking representation is helpful to understand this complicated structure: it takes the form $Q = \sum_{j=1}^{\infty} \pi_j \delta_{G_j}$, where $G_1, G_2, \ldots \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha)$ is a sample of *measures* on $(\mathfrak{X}, \mathscr{X})$. This structure is called a *nested Dirichlet process* (NDP).

A sample $F_1, F_2, \ldots \mid Q \overset{\text{iid}}{\sim} Q$ are again measures on $(\mathfrak{X}, \mathscr{X})$, which cluster by their dependence through Q, and thus might be a reasonable model for the distributions of multiple samples of observations in $\mathfrak{X}$ that share common features. The clustering of the F_k is the usual one in the Dirichlet process, and is clearest from the stick-breaking representation: every F_k will be one of the measures G_j in the realization of $Q = \sum_{j=1}^{\infty} \pi_j \delta_{G_j}$, and the F_k cluster in groups of equal distributions.

As $F_1, F_2, \ldots$ are a sample from a Dirichlet process, they are subject to the predictive formulas of Section 4.1.4, where $F_1, F_2, \ldots$ take the place of the observations $X_1, X_2, \ldots$ in that section. In particular, the marginal distribution of every F_k is equal to the center measure $\mathrm{DP}(\alpha)$, and F_2 is equal to F_1 or a new draw from $\mathrm{DP}(\alpha)$ with probabilities $1/(M+1)$ and $M/(M+1)$, respectively. By exchangeability it follows that, for every $k \neq k'$ and measurable set B,

$$\mathrm{P}(F_k = F_{k'}) = \frac{1}{M+1}, \quad \mathrm{cov}\,(F_k(B), F_{k'}(B)) = \frac{\mathrm{var}\,(F_k(B))}{M+1} = \frac{\bar{\alpha}(B)\bar{\alpha}(B^c)}{(M+1)(|\alpha|+1)}.$$

In particular, the correlation $\mathrm{corr}\,(F_k(B), F_{k'}(B)) = 1/(M+1)$ is free of B. The first equality in the display shows that the nested Dirichlet process is substantially different from the hierarchical Dirichlet process, introduced in Example 5.12, for which $\mathrm{P}(F_k = F_{k'}) = 0$, as only atoms are shared but not weights.

Samples $X_{k,i} \mid F_k, Q \overset{\text{iid}}{\sim} F_k$ from the distributions F_k cluster as well. As each F_k is marginally from a $\mathrm{DP}(\alpha)$-distribution, two different ($i \neq i'$) observations $X_{k,i}$ and $X_{k,i'}$ from the same sample are equal with probability $\mathrm{P}(X_{k,i} = X_{k,i'}) = 1/(|\alpha|+1)$, as for

an ordinary Dirichlet process. On the other hand, two different observations $X_{k,i}$ and $X_{k',i'}$ from different samples (thus $k \neq k'$ and $i \neq i'$) are equal with probability

$$\mathrm{P}(X_{k,i} = X_{k',i'}) = \frac{1}{M+1} \times \frac{1}{|\alpha|+1}.$$

This follows, because the distributions $G_k \sim \mathrm{DP}(\alpha)$ will have disjoint (countable) supports, and hence so will F_k and $F_{k'}$) unless they cluster on the same G_k.

Thus observations from the same sample are more likely to be equal than observations from different samples, but clustering across groups occurs with positive probability. This feature is shared with the hierarchical Dirichlet process.

14.10 The Indian Buffet Process

The Chinese restaurant process (see Section 14.1.2) induces clustering through the values of a single variable. In some situations individuals may be marked by multiple, potentially even infinitely many, features. Some features (or feature values) may be more popular than others, and be shared by more individuals. The Indian buffet process (IBP) is a stochastic process that models this phenomenon. Like the Chinese restaurant process, presence of a feature in more subjects enhances the probability that a future subject possesses the same feature, and thus builds up a clustering effect. However, feature sharing in an Indian buffet process occurs separately for each feature. In that sense, the process is a factorial analog of the Chinese restaurant process.

The Indian buffet process assumes that feature values are binary, with a 1 or 0 interpreted as possessing or not possessing the feature, and allows infinitely many features. Subjects are then characterized by an infinite row of 0s and 1s, and data can be represented in a binary matrix, with rows corresponding to subjects and columns corresponding to features. It will be assumed that every subject can possess only a finite number of features, so that the binary matrix has finite row sums. Thus the Indian buffet process specifies a probability distribution on sparse infinite binary matrices.

The Indian buffet process will be defined as a limit in distribution of $(n \times K)$-matrices $Z^K = (\!(Z_{ik})\!)$, corresponding to n subjects and K features, with n fixed and $K \to \infty$. For given K and parameter $M > 0$ the distribution of the matrix Z^K is specified as

$$Z_{ik} | \, \pi_1, \ldots, \pi_k \overset{\text{ind}}{\sim} \mathrm{Bin}(1, \pi_k), \qquad \pi_k \overset{\text{iid}}{\sim} \mathrm{Be}(M/K, 1).$$

Thus π_k is the probability that the kth feature is present in a subject, and given these probabilities all variables Z_{ik} are mutually independent. The probability of a particular $(n \times K)$ matrix (or "configuration") z with m_k values 1 and $n - m_k$ values 0 in particular locations in the kth column is then

$$\mathrm{P}(Z^K = z) = \prod_{k=1}^{K} \int \pi_k^{m_k} (1-\pi_k)^{n-m_k} \frac{M}{K} \pi_k^{M/K-1} \, d\pi_k = \prod_{k=1}^{K} \frac{M\Gamma(m_k + M/K)\Gamma(n - m_k + 1)}{K\Gamma(n+1+M/K)}.$$

The K columns of the matrix Z^K are i.i.d. and every column is an exchangeable n-vector, which makes the matrix "exchangeable in both rows and columns." The expected number of 1s per column is equal to $n(M/K)/(1+M/K)$, giving an expected total number of 1s equal to $nM/(1+M/K)$. For $K \to \infty$ this stabilizes to nM, so that sparsity is ensured. This is

01000110001011000010		11111110000000000000
10000100010001010001		11000001111000000000
01000100101000001001	left-ordering	10110001000100000000
01001001001001000010	→	01111000000011000000
00010100001001000100		11100000000000110000

Figure 14.6 Example of a matrix showing $K = 20$ features scored on $n = 5$ subjects (left) and the corresponding left-ordered matrix (right).

the rationale behind the choice M/K for the parameter of the beta distribution for the π_k — this parameter must decrease at the rate K^{-1} to balance the effect of the growing number of columns if the total number of 1s must remain finite.

The ordering of the features is considered irrelevant, and hence matrices that are identical up to a permutation of columns will be identified. It is convenient to choose a representative of every equivalence class of matrices thus created, defined by the *left-ordering procedure*. This consists of first permuting columns so that all 1s appear before all 0s in the first row. This forms two groups of columns, defined by their value in the first row, which are separately permuted by the same procedure, but now taking account of 1s and 0s in the second row. Next there are four groups of columns, defined by their values in the first two rows, and the procedure continues with four separate permutations based on the third row. The procedure continues sequentially until all rows have been operated on. (The final result can also be described as an ordering of the columns in terms of the value of the column viewed as a binary number.) Figure 14.6 shows an example of a feature matrix and the left-ordered representative of its equivalence class.

We define a probability distribution on the set of left-ordered matrices by assigning every such matrix the sum of all probabilities received by some member of its equivalence class. Since the columns of the matrix Z^K are exchangeable, all members of an equivalence class have the same probability, and this sum is simply equal to the cardinality of the equivalence class times $\mathrm{P}(Z^K = z)$. If the vector (K_h), for $h = 0, \ldots, 2^{n-1}$, gives the number of times that every of the 2^n possible columns (in $\{0, 1\}^n$) occurs in a matrix, then the cardinality of its equivalence class is equal to $K!/\prod_{h=0}^{2^n-1} K_h!$, being the number of permutations of the columns, divided by the number of permutations that yield the same left-ordered matrix. Thus with $[Z^K]$ denoting a left-ordered equivalence class, we assign to a configuration $[z]$ characterized by (K_h) and (m_k) the probability

$$\mathrm{P}([Z^K] = [z]) = \frac{K!}{\prod_{h=0}^{2^n-1} K_h!} \prod_{k=1}^{K} \frac{M\Gamma(m_k + M/K)\Gamma(n - m_k + 1)}{K\Gamma(n + 1 + M/K)}. \tag{14.29}$$

The numbers m_k enter this expression as an unordered set $\{m_1, \ldots, m_K\}$, but in their original interpretation they refer to the number of 1s in specific columns. In particular, after appropriate permutation every m_k refers to a column that belongs to one of the groups of columns whose cardinalities are counted by the K_h, for $h = 0, \ldots, 2^n - 1$. The values of the m_k as a group can for a given configuration be derived from the numbers K_h, which give not only the number of 1s but also their positions in the columns. Thus we can view the preceding probability as a probability distribution on the set of vectors (K_h). These have length 2^n, and nonnegative integer-valued coordinates with sum $\sum_h K_h = K$. Each vector corresponds to a left-ordered $(n \times K)$-matrix. For convenience of notation let $h = 0$ refer to the zero column, so that K_0 is the number of columns with only zero coordinates (i.e.

columns with $m_k = 0$). We shall take the limit as $K \to \infty$ with the numbers $(K_h)_{h>0}$ of nonzero columns fixed; hence $K_0 \to \infty$ and the number of columns k with $m_k > 0$ is also fixed.

Lemma 14.63 *In the preceding setting we have that* $K - \mathrm{E}K_0 = \sum_{i=1}^{n} M/i + O(1/K)$, *as* $K \to \infty$ *for fixed* n, *and hence* $\sum_{h>0} K_h$ *is bounded in probability. Furthermore, for fixed values* $(K_h)_{h>0}$, *expression* (14.29) *converges to, as* $K \to \infty$ *for fixed* n,

$$\mathrm{P}([Z] = [z]) := \frac{M^{\sum_{h>0} K_h}}{\prod_{h>0} K_h!} e^{-M \sum_{i=1}^{n} i^{-1}} \prod_{k:m_k>0} \frac{(n-m_k)!(m_k-1)!}{n!}.$$

This expression gives a probability density function of a variable $[Z]$ *with values in the set of left-ordered binary matrices with* n *rows and infinitely many columns. The corresponding distribution is exchangeable in the rows of the matrix.*

Proof The expected value $\mathrm{E}K_0$ is K times the probability that a given column of the $(n \times K)$ matrix Z^K has only zero coordinates. This latter probability is equal to $\mathrm{E}(1-\pi_k)^n$, for $\pi_k \sim \mathrm{Be}(M/K, 1)$. Since $\Gamma(1+\alpha) = \alpha\Gamma(\alpha)$, for $\alpha > 0$, this can be computed to be

$$\frac{\Gamma(n+1)\Gamma(1+M/K)}{\Gamma(n+1+M/K)} = \left[\left(1+\tfrac{M}{Kn}\right)\left(1+\tfrac{M}{K(n-1)}\right) \times \cdots \times \left(1+\tfrac{M}{K}\right)\right]^{-1}.$$

The right side can be seen to be $1-(M/K)\sum_{i=1}^{n} 1/i + O(1/K^2)$, from which the assertion on $K - \mathrm{E}K_0$ follows. Since $\sum_{h>0} K_h = K - K_0$, it follows that the sequence of nonnegative variables $\sum_{h>0} K_h$ has a bounded mean value and hence is bounded in probability.

To prove the convergence of (14.29), we first note that for any configuration (K_h) with $\sum_{h>0} K_h = K - K_0$ fixed, we have that $K!/(K_0! K^{K-K_0}) \to 1$, as $K \to \infty$. Furthermore, we have the identities, where $\Gamma(0)$ is interpreted as 1,

$$\begin{aligned}\prod_{k=1}^{K} \frac{\Gamma(m_k + M/K)}{\Gamma(m_k)} &= \Gamma\left(\tfrac{M}{K}\right)^{K_0} \prod_{k:m_k>1} \left[\left(1+\tfrac{M}{K(m_k-1)}\right)\right. \\ &\qquad \left. \times \cdots \times \left(1+\tfrac{M}{K}\right)\right]\left[\tfrac{M}{K}\Gamma\left(\tfrac{M}{K}\right)\right]^{K-K_0}, \\ \prod_{k=1}^{K} \frac{\Gamma(n+1+M/K)}{\Gamma(n+1)} &= \left[\left(1+\tfrac{M}{Kn}\right)\left(1+\tfrac{M}{K(n-1)}\right) \times \cdots \times \left(1+\tfrac{M}{K}\right)\right]^{K}\left[\tfrac{M}{K}\Gamma\left(\tfrac{M}{K}\right)\right]^{K}.\end{aligned}$$

When taking the quotient of these two expressions, the factors $\Gamma(M/K)$ cancel. Furthermore, for fixed $(K_h)_{h>0}$ the product on the right of the first equation has a bounded number of terms, each of which tends to 1, so that the product tends to 1. The first term in the second equation can be seen to converge to $e^{M\sum_{i=1}^{n} 1/i}$. The convergence of (14.29) follows by substituting these findings in the quotient of the latter expression and its claimed limit, and simplify the result.

By construction, the function $z \mapsto p_K(z)$ on the right side of (14.29) is a probability density function on the set of left-ordered $(n \times K)$-matrices z. We can view p_k also as a probability density function on the set $\mathfrak{K}$ of all vectors $(K_h)_{h>0}$ of length $2^n - 1$ with coordinates in the nonnegative integers with $\sum_{h>0} K_h < \infty$. The first part of the proof

shows that it gives probability at least $1 - \epsilon$ to the finite subset of all such vectors with $\sum_{h>0} K_h \leq L$, for a sufficiently large large constant L. Thus the sequence p_K is uniformly tight for the discrete topology on $\mathfrak{K}$. Then its pointwise limit must be a probability density function on $\mathfrak{K}$. □

Definition 14.64 (Indian buffet process) The (one-parameter) *Indian buffet process* (IBP) with parameter M is the probability distribution on the set of left-ordered binary matrices of dimension $(n \times \infty)$ given in Lemma 14.63, and defined for $n \in \mathbb{N}$.

Like the Chinese restaurant process, the Indian buffet process can be described by a culinary metaphor. Imagine that customers enter an Indian restaurant that serves a lunch buffet with an unlimited number of dishes. The first customer arriving in the restaurant chooses a Poisson(M) number of dishes from the buffet. The second customer chooses every dish tasted by the first customer independently and randomly with probability $1/2$ and also independently chooses an additional Poisson($M/2$) number of new dishes. The probability $1/2$ and parameter $M/2$ are chosen so that marginally the number of dishes selected by the second customer is also Poisson(M). In a similar manner, the ith customer chooses every previously tasted dish k with probability $m_{k,i-1}/i$, where $m_{k,i-1}$ is the number of previous customers (among $1, \ldots, i-1$) who tasted dish k, and also chooses an additional Poisson(M/i) number of so-far untasted dishes. After continuing up to n customers, we can form a binary matrix with the customers as rows and the columns as dishes, with 1s at coordinates (i, k) such that customer i tasted dish k.

Lemma 14.65 *The left-ordering of the binary matrix constructed in the preceding paragraph is a realization of the Indian buffet process.*

Proof This may be proved by writing down the explicit probability of obtaining a matrix configuration with given $(K_h)_{h>0}$, and verifying that this agrees with the expression given in Lemma 14.63. □

Conditional distributions in an Indian buffet process are easy to compute. These distributions can be used to implement a Gibbs sampling procedure when an Indian buffet prior is applied on latent features, similar as for a Dirichlet mixture process. If $Z_{-i,k}$ stands for the kth column of Z with the ith coordinate deleted and $m_{-i,k}$ is the corresponding column sum, then $\mathrm{P}(Z_{ik} = 1 | Z_{-i,k}) = m_{-i,k}/n$. This can be easily derived by taking the limit as $K \to \infty$ of the corresponding probability in the finite version of the model with K features. Then $Z_{-i,k}$ is a vector of Bernoulli variables Z_{jk}, for $j \neq i$, with success probability $\pi_k \sim \mathrm{Be}(M/K, 1)$, and the corresponding probability is $\mathrm{E}(\pi_k | Z_{-i,k}) = (m_{-i,k} + M/K)/(n + M/K)$, since the posterior distribution of π_k given $Z_{-i,k}$ is $\mathrm{Be}(M/K + m_{-i,k}, n - m_{-i,k})$.

Problem 14.19 describes a type of stick-breaking representation of the Indian buffet process.

A different way to construct the Indian buffet process is through the de Finetti representation of the distribution of the rows of the binary matrix, in the same way as the Chinese restaurant process arises from a random sample of observations from the Dirichlet process.

In the context of the Indian buffet process the observations (the rows of the matrix) are random variables with values in $\{0, 1\}^\infty$, and the mixing distribution is a measure on the set of distributions of such infinite vectors.

It is convenient to describe the construction in a continuous time framework, where the 1s are placed at random locations in $(0, \infty)$, and the remaining times are interpreted as 0s. Since we are interested in the pattern of sharing of 1s, the additional zeros do not affect the interpretation. Then the two distributions involved in the de Finetti representation are the Bernoulli process for the observations and the beta process for the mixing distribution. The beta process is described in Section 13.3 and Definition 13.3. It is a family of stochastic processes with independent increments on the time set $[0, \infty)$, parameterized by a pair (c, Λ) of a positive function c and a cumulative hazard function Λ;[13] the function c will here be restricted to be constant; the function Λ is the mean function of the process and will be important here only through its total mass, which will be assumed finite. The Bernoulli process is another process with independent increments, parameterized by a cumulative hazard function H, and defined as follows.

Definition 14.66 (Bernoulli process) A *Bernoulli process* with parameter H is an independent increment process on $[0, \infty)$ with intensity measure given by $\nu(dx, ds) = H(dx)\delta_1(ds)$, where H is a cumulative hazard function and δ_1 is the Dirac measure at 1.

The definition uses the language of completely random measures, as described in Appendix J, but the Bernoulli process can also be described directly as follows. In the case that H is a discrete measure the Bernoulli process can be identified with an infinite sequence of independent Bernoulli variables h_k, a variable $h_k \sim \text{Bin}(1, H\{x_k\})$ for every atom x_k of H. The sample paths of the Bernoulli process are then zero at time zero and jump up by 1 at every x_k such that $h_k = 1$.[14] If $H = H^c + H^d$ consists of a continuous and a discrete part, then the Bernoulli process is the sum $Z = Z^c + Z^d$ of two independent processes, where Z^c is a Poisson process with mean measure $\text{E}[Z^c(t)] = H^c(t)$ and $Z^d(t) = \sum_k h_k \mathbb{1}\{x_k \le t\}$ is as described previously (with $H = H^d$). While as an independent increment process the Bernoulli process is understood to increase by 1 at all jump points, below we view the process as a measure and identify it with its jumps, making the sample paths take values 0 and 1 rather than increasing functions.

Given a constant $c > 0$ and an atomless cumulative hazard function Λ, consider the model

$$Z_1, Z_2, \ldots \mid H \overset{\text{iid}}{\sim} \text{Bernoulli process}(H), \qquad H \sim \text{Beta process}(c, \Lambda). \tag{14.30}$$

Because the beta process produces a discrete cumulative hazard function H with probability one, the realizations of the Bernoulli processes Z_i can be identified with infinite sequences of 0s and 1s (where the order does not matter as long it is consistent across i). Thus for every n we can form the binary $(n \times \infty)$-matrix with rows $Z_1, \ldots, Z_n$. We show below that for $c = 1$ and Λ of total mass M, the left-ordered version of this matrix is distributed

[13] The cumulative distribution function of a measure on $(0, \infty)$ that is finite on finite intervals and has atoms strictly smaller than one; it may be identified with such a measure.

[14] The total number of jumps on finite intervals is finite, as it has expectation $\sum_{x_k \le t} H\{x_k\} < \infty$, for every t.

as the Indian buffet process with parameter M. For other values of c we obtain different distributions, each with exchangeable rows by construction.

Definition 14.67 (Two-parameter Indian buffet process) The *two-parameter Indian buffet process* with parameters c and M is the distribution of the left-ordering of the $(n \times \infty)$ matrix with rows the processes $Z_1, \ldots, Z_n$ defined in (14.30) (or rather the values of their jumps at a countable set containing all atoms of H). Here $c > 0$ and M is the total mass of Λ.

To connect this definition to the earlier one-parameter Indian buffet process, we derive the conditional distribution of a new row Z_{n+1} given the preceding ones. The following result is a process analog of the familiar binomial-beta conjugacy.

Lemma 14.68 *The posterior distribution of H given $Z_1, \ldots, Z_n$ in the model* (14.30) *is a beta process with parameters $c + n$ and $c\Lambda/(c + n) + \sum_{i=1}^{n} Z_i/(c + n)$. Furthermore, marginally every Z_i is a Bernoulli process with parameter Λ.*

Proof It suffices to give the proof for $n = 1$, because the general result can next be obtained by posterior updating. A proof can be based on a discrete approximation to the beta process, and calculations of the posterior distribution in the finite-dimensional case. Alternatively we can base a proof on the representations of the processes through point processes, as follows.

The beta process H can be identified with a point process $N = N^c + N^d$ on $(0, \infty) \times (0, 1)$, where N^c is a Poisson process with intensity measure $\nu^c(dx, ds) = cs^{-1}(1 - s)^{c-1}\, d\Lambda^c(x)\, ds$, and N^d consists of points (x, s), with x restricted to the atoms of Λ and $s \sim \mathrm{Be}(c\Lambda\{x\}, c(1 - \Lambda\{x\}))$, independently across atoms, and independently of N^c (see Section 13.3).

Given N (or H), the Bernoulli process Z_1 jumps (always by 1) with probability s at every point x with $(x, s) \in N$. In other words, the process Z_1 jumps at every x from the set of points (x, s) obtained from *thinning* N with probabilities s.

This thinning transforms the Poisson process N^c into a Poisson process M^c of intensity $s\nu^c(dx, ds) = c(1-s)^{c-1}\, d\Lambda^c(x)\, ds$. Given M^c, the process N^c is distributed as the union of M^c and a Poisson process K generated by thinning N^c with the complementary probabilities $1 - s$, hence with intensity measure $(1 - s)\nu^c(dx, ds) = cs^{-1}(1 - s)^c\, d\Lambda^c(x)\, ds$, which is the intensity measure of a beta process with parameters $(c + 1, c\Lambda^c/(c + 1))$. Observing Z_1 is less informative than observing M^c, as Z_1 only (but fully) reveals the horizontal locations x of the points (x, s) in M^c. Since the locations x and heights s of the points (x, s) in M^c can be realized by first spreading the locations x according to a Poisson process with intensity measure Λ^c and next generating the heights s independently from the $\mathrm{Be}(1, c)$-distribution, the conditional distribution of M^c given Z_1 is to reconstruct the heights independently from this same beta distribution. Since $\mathrm{Be}(1, c) = \mathrm{Be}((c+1)\Delta, (c+1)(1-\Delta))$, for $\Delta = 1/(c+1)$, it follows that the points in $M^c \,|\, Z_1$ can be described as the fixed atoms of the beta process with parameters $(c+1, Z_1^c/(c+1))$, for Z_1^c the jumps of Z_1 that do not belong to the support of Λ^d.

A point (x, s) of N^d leads to a jump in Z_1 with probability s, and this will sit at an atom of Λ^d. If x is such an atom and also a jump point of Z_1, then the conditional density of s is proportional to s times the $\mathrm{Be}(c\Lambda\{x\}, c(1 - \Lambda\{x\}))$-density, while if x is an atom of Λ^d and

not a jump point of Z_1, then the conditional density of s is proportional to $1-s$ times the latter density. This leads to $\mathrm{Be}(c\Lambda\{x\}+1, c(1-\Lambda\{x\}))$- and $\mathrm{Be}(c\Lambda\{x\}, c(1-\Lambda\{x\})+1)$-distributions for s in the two cases, which are consistent with fixed atoms of a beta process with parameters $c+1$ and discrete cumulative hazard component $c\Lambda^d/(c+1)+Z_1^d/(c+1)$, for Z_1^d the jumps of Z_1 that are also atoms of Λ^d.

The jumps of the process Z_1 occur at locations x obtained from the points (x,s) in N by thinning these points with probability s. The points (x,s) left from the Poisson process N^c form a Poisson process with intensity $s\nu^c(dx,ds) = c(1-s)^{c-1}\,d\Lambda^c(x)\,ds$, and project to a Poisson process with intensity $\Lambda^c(dx)$ on the first coordinate. The points (x,s) left from N^d are located at atoms of Λ^d, and an atom remains with probability s, where $s \sim \mathrm{Be}(c\Lambda\{x\}, c(1-\Lambda\{x\}))$, whence with marginal probability $\mathrm{E}[s] = \Lambda\{x\}$. □

Now specialize to the situation where $c > 0$ and Λ is an atomless measure of total mass M. Then the Bernoulli processes Z_i are marginally Poisson processes with intensity measure Λ, by the second assertion of Lemma 14.68. In particular, their total numbers of jumps are $\mathrm{Poi}(M)$-distributed.

Because given H the Z_i are i.i.d. Bernoulli processes with intensity H, the predictive distribution of Z_{n+1} given $Z_1,\ldots,Z_n$ is a Bernoulli process given an intensity process generated from the conditional distribution of H given $Z_1,\ldots,Z_n$. By the first assertion Lemma 14.68, this is a beta process with updated parameters, whence by the last assertion of the lemma Z_{n+1} given $Z_1,\ldots,Z_n$ is a Bernoulli process with the updated cumulative hazard function. The latter function can alternatively be written as

$$\frac{c}{c+n}\Lambda + \sum_{k=1}^{\infty} \frac{m_{k,n}}{c+n}\delta_{x_k},$$

where $m_{k,n}$ is the number of processes $Z_1,\ldots,Z_n$ that have a jump at atom x_k (and the sequence $x_1, x_2, \ldots$ must contain all jump points of all Z_i). Thus Z_{n+1} given $Z_1,\ldots,Z_n$ is a Bernoulli process with both a continuous component and a discrete component. By the definition of a Bernoulli process the continuous component is the Poisson process with intensity measure $c\Lambda/(c+n)$, while the discrete component is the process that jumps with probability equal to $m_{k,n}/(c+n)$ at every location x_k, with jump size 1. Because Λ is atomless by assumption, the events of the continuous component will with probability 1 not occur at any of the points x_k, and their number will be $\mathrm{Poi}(cM/(c+n))$. For $c=1$ the probability $m_{k,n}/(1+n)$ of a jump at an "old location" and the distribution $\mathrm{Poi}(M/(1+n))$ of the number of jumps at a "new location" agree exactly with the dish-sharing probability and the distribution of the number of new dishes in the Indian buffet process with parameter M. Since Z_1 starts with the same number of dishes and all conditionals are as in the Indian buffet process, the two-parameter Indian buffet process with parameter $c=1$ is equal to the one-parameter process of Definition 14.64.

The new number of dishes tasted by the ith customer arises from the Poisson part of $Z_{i+1}|\,Z_1,\ldots,Z_i$, and hence is Poisson distributed with mean $cM/(c+i)$. Therefore the expected total number of dishes tasted by at least one of the first n customers in a two-parameter Indian buffet process can be calculated as $Mc\sum_{i=0}^{n-1}(c+i)^{-1}$. The logarithmic growth as $n\to\infty$ for fixed c is reminiscent of the growth of the number of clusters in the Chinese restaurant process (see Proposition 4.8). The parameter c can be interpreted as a measure of concentration. If $c\to 0$, then the expected total number of dishes tasted

tends to M, and hence (on the average) all customers share the same dishes (maximum concentration), whereas if $c \to \infty$, the expected number of dishes tasted tends to Mn, meaning that (on the average) customers are not sharing at all (minimum concentration).

14.11 Historical Notes

Exchangeable random partitions were first studied by Kingman (1978, 1982) and later by Aldous (1985), Perman et al. (1992) and Pitman (1995, 2006). Kingman's representation for exchangeable partitions, Theorem 14.7, appeared in Kingman (1978); the short proof presented here was taken from Aldous (1985). Lemma 14.8 appeared explicitly in Lee et al. (2013). Partial exchangeability and Theorem 14.13 and Corollary 14.14 are due to Pitman (1995). Theorem 14.18 was obtained by Pitman (1996b). The important role of size-biased permutations is pointed out in Perman et al. (1992) and Pitman (1995, 1996a). The name *Chinese restaurant process* was apparently coined by Dubins and Pitman (see Aldous (1985) and Pitman (2006)). The sampling scheme was already known in Bayesian Nonparametrics as the *Blackwell-MacQueen urn scheme* or *generalized Pólya urn scheme*, and appeared in Blackwell and MacQueen (1973). In population genetics Hoppe (1984) introduced a different sampling scheme, which also leads to the EPPF induced by sampling from a Dirichlet process. The Chinese restaurant franchise process and the hierarchical Dirichlet process were studied by Teh et al. (2006). Proposition 14.9 characterizing the Dirichlet process as a special species sampling process is due to Gnedin and Pitman (2006); the proof given here is taken from Lee et al. (2013). The abstract posterior consistency result for species sampling process given by Theorem 14.19 was obtained for the case $\alpha_n = 1$ and $\beta_n = 0$ by Jang et al. (2010), by a longer proof. The present generalization covers the case of Gibbs processes, which were treated separately by De Blasi et al. (2013). Gibbs processes were introduced and systematically studied by Gnedin and Pitman (2006), although some of the results were anticipated in Pitman (2003). The main example, the two-parameter model, goes back to at least Perman et al. (1992) and Pitman (1995). They were used in Bayesian nonparametrics by Lijoi et al. (2007b,a). Theorems 14.26 and 14.27 and Lemma 14.28 are due to De Blasi et al. (2013). De Blasi et al. (2015) gave a review of the topic. The Pitman-Yor process appears explicitly in Perman et al. (1992), and Pitman (1995) who refer to Perman's 1990 thesis for the stick-breaking representation. This extended the characterization by McCloskey (1965) that the Dirichlet process is the only discrete random probability measure for which size-biased random permutation of random atoms admits a stick-breaking representation in terms of a collection of *i.i.d.* random variables. Pitman (1995) studied its partition structure, and Pitman and Yor (1997) explored connections with stable processes and gave an alternative construction via the Radon-Nikodym derivative of a σ-stable CRM. Proposition 14.34 is due to Pitman and Yor (1997), Theorem 14.37 is due to Pitman (1996b), Theorem 14.33 is due to Perman et al. (1992). Proposition 14.35 is due to Vershik et al. (2001). The distribution of linear functionals of a Pitman-Yor process was characterized in terms of generalized Cauchy-Stieltjes transform by James et al. (2008). Theorem 14.38 and Lemma 14.39 were obtained by James (2008). Poisson-Kingman processes were introduced and studied in Pitman (2003), following Perman et al. (1992). The latter reference obtains various distributional properties, including the characterization through residual allocation (or stick breaking) given in Theorem 14.49. Various applications of the gamma Poisson-Kingman process can be found in the monograph Feng (2010) and the references therein.

The EPPF and PPF of a special Poisson-Kingman process was also obtained by Regazzini et al. (2003) by using the normalized CRM approach. Gnedin and Pitman (2006) characterized all Gibbs type processes with type parameter $\sigma \in (0, 1)$ precisely as Poisson-Kingman processes based on σ-stable CRM. The *σ-diversity* for a Poisson-Kingman process based on a σ-stable or generalized gamma CRM was derived by Pitman (2003), based on a characterization of the diversity in terms of the weight sequence by Karlin (1967). Posterior inference about the σ-diversity with applications in Bayesian nonprametrics are discussed in Favaro et al. (2009) and Favaro et al. (2012a). Normalized completely random measures were introduced in the context of prior distributions on probability measures on the real line in Regazzini et al. (2003). The posterior updating rule was obtained in James et al. (2009). Results on the partition structure were also obtained in this paper, but follow from the case of Poisson-Kingman processes obtained in earlier work by Pitman (2003). The general stick-breaking representation Theorem 14.49 of homogeneous NCRM (or Poisson-Kingman processes) was given in Perman et al. (1992), and repeated in Pitman (2003), who also discussed the partitioning structure. It was specialized to the normalized inverse-Gaussian process as in Proposition 14.54 in Lijoi et al. (2005a). The generalized gamma NCRM was constructed as a Poisson-Kingman process in Pitman (2003) (apparently following a 1995 preprint), and as an NCRM in Lijoi and Prünster (2003); also see Cerquetti (2007). This class was characterized as the intersection of NCRM processes and Gibbs processes in Lijoi et al. (2008) and Cerquetti (2008). Bayesian analysis with such a prior distribution appeared in Lijoi et al. (2007b), extending the special case of the normalized inverse-Gaussian in Lijoi et al. (2005a). To date the normalized inverse-Gaussian and the Dirichlet process are the only processes with explicit descriptions of marginal distributions of sets in a partition. The generalized Dirichlet process discussed in Lijoi et al. (2005b) is another example for which many closed form expressions can be obtained, notwithstanding the fact that it falls outside the class of generalized gamma NCRM, and is not Gibbs; see Problem 14.12. Regazzini et al. (2003) and James et al. (2010) studied distributions of linear functionals of an NCRM process. Nieto-Barajas et al. (2004) obtained distribution of the mean functional of both prior and posterior process of NCRM processes. A review of many results on distribution of linear functionals of various type of processes is given by Lijoi and Prünster (2009). The first contribution on dependent discrete random probability distributions was made by Cifarelli and Regazzini (1978), who proposed a nonparametric prior for partially exchangeable arrays, defined as a mixture of Dirichlet processes. MacEachern (1999) developed the idea in a modern context and introduced the idea of dependent stick breaking. The concept was quickly followed up and led to various constructions of dependent processes, and applications in different contexts such as density regression. The order based dependent Dirichlet process appeared in Griffin and Steel (2006). The kernel stick-breaking process appeared in Dunson and Park (2008), the probit stick-breaking process in Rodríguez and Dunson (2011) and the local Dirichlet process in Chung and Dunson (2011). The nested Dirichlet process, which is similar in spirit to the hierarchical Dirichlet process, was introduced by Rodríguez et al. (2008). The one-parameter Indian buffet process was introduced by Griffiths and Ghahramani (2006). Thibaux and Jordan (2007) generalized it to the two-parameter situation and obtained the de Finetti representation through the beta process. The stick-breaking representation of the the Indian buffet process was obtained by Teh et al. (2007).

Problems

14.1 (Kingman's formula for EPPF) Show that the EPPF of a proper species sampling model can be written in the form

$$p(n_1, \dots, n_k) = \sum_{1 \le i_1 \ne \cdots \ne i_k < \infty} \mathrm{E} \prod_{j=1}^{k} W_{i_j}^{n_j}.$$

14.2 (De Blasi et al. 2013, Gnedin 2010) Suppose that P follows a Gibbs prior distribution with type-parameter $\sigma = -1$ and mixing distribution in (14.19) given by $\pi(j) = \gamma(1-\gamma)^{[j-1]}/j!$, where $0 < \gamma < 1$. Show that $V_{n+1,K_n+1}/V_{n,K_n} \to 1$ a.s. if K_n is the number of distinct values in a random sample of size n from a true distribution P_0, and conclude that the posterior distribution of P concentrates near the center measure G a.s. [Thus the posterior is *totally inconsistent*: the data have asymptotically no influence and the posterior concentrates near the prior mean.]

14.3 (De Blasi et al. 2013) Suppose that P follows a Gibbs distribution with type-parameter $\sigma = -1$ with mixing distribution in (14.19) the geometric distribution with parameter $\gamma \in (0, 1)$. Show that $V_{n+1,n+1}/V_{n,n} \to (2 - \gamma - 2\sqrt{1-\gamma})/\gamma$ a.s., and conclude that the posterior distribution based on a random sample from an atomless true distribution P_0 is inconsistent.

14.4 (Kingman 1993) Let $W_{1:n} \le \cdots \le W_{n:n}$ be the order statistics of a $\mathrm{Dir}(n; \alpha_{1,n}, \dots, \alpha_{n,n})$-distributed random vector. Assume that $\max\{\alpha_{n,j} : 1 \le j \le n\} \to 0$ and $\sum_{j=1}^{n} \alpha_{n,j} \to M \in (0, \infty)$ as $n \to \infty$. Show that $(W_{n:n}, W_{n-1:n}, \dots, W_{n-k:n}) \rightsquigarrow (P_1, P_2, \dots, P_k)$ for any fixed k, where $(P_1, P_2, \dots)$ follows a one-parameter Poisson-Dirichlet distribution with parameter M.

14.5 (Kingman 1993) If $(P_1, P_2, \dots)$ follows a one-parameter Poisson-Dirichlet distribution with parameter M, show that $-\log P_k \sim k/M$ a.s. as $k \to \infty$.

14.6 Show that Proposition 14.5 reduces to Ewens's sampling formula when applied to a Dirichlet process partition.

14.7 (Perman et al. 1992) If $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$, then show that the reordering $\tilde{w}_1, \tilde{w}_2, \dots$ of a probability vector $w_1, w_2, \dots$ such that the corresponding variables ε_j/w_j are in increasing order, is size-biased.

14.8 Reinterpret equation (14.23) in Theorem 14.49 to show that the sequence of variables $T_0 := Y$, $T_1 := \sum_{j \ge 2} \tilde{Y}_j$, $T_3 := \sum_{j \ge 3} \tilde{Y}_j, \dots$ is a Markov chain with stationary transition density.

14.9 Show that the Pitman-Yor process can be obtained as a $\mathrm{Ga}(M/\sigma, 1)$-mixture of generalized gamma processes $\rho_{\sigma,\xi}$, as claimed in Example 14.48(iv). [Hint: By Example 14.47 the Pitman-Yor distribution is a mixture of the $\mathrm{PK}(\rho_\sigma \,|\, t)$-distributions with respect to the density $h(t) = \sigma\Gamma(M)/\Gamma(M/\sigma) f_\sigma(t)$. Because conditional Poisson-Kingman distributions are invariant under exponential tilting of the intensity measure, $\mathrm{PK}(\rho_\sigma \,|\, t) = \mathrm{PK}(\rho_{\sigma,\xi} \,|\, t)$, and hence it suffices to show that the mixture $\int f_{\sigma,\xi}(t) g(\xi)\, d\mathrm{Ga}(M/\sigma, 1)(\xi)$. of the densities $f_{\sigma,\xi}$ associated with $\rho_{\sigma,\xi}$ is equal to h. Consider Laplace transforms and write $t^{-M} = \int \mu^{M-1} e^{-\mu t}\, d\mu/\Gamma(M)$.]

14.10 (Perman et al. 1992) Show that for a Poisson-Kingman process, with the random variables $Y, V_1, V_2, \ldots$ appearing in the stick-breaking representation given by Theorem 14.49, the following assertions are equivalent:

(a) P follows a Dirichlet process;
(b) Y and V_1 are independent;
(c) $Y \sim \text{Ga}(M, \lambda)$ and independently $V_1 \sim \text{Be}(1, M)$, for some $M, \lambda > 0$;
(d) $Y, V_1, V_2, \ldots$ are independent;
(e) $Y \sim \text{Ga}(M, \lambda)$ and $V_1, V_2, \ldots \overset{\text{iid}}{\sim} \text{Be}(1, M)$, for some $M, \lambda > 0$.

14.11 (Perman et al. 1992) Show that for a Poisson-Kingman process, with the random variables $Y, V_1, V_2, \ldots$ appearing in the stick-breaking representation given by Theorem 14.49, the following assertions are equivalent:

(a) $Y - \tilde{Y}_1$ and V_1 are independent;
(b) Y is a positive σ-stable random variable.

If either and hence both conditions are satisfied, then $Y_n, V_1, \ldots, V_n$ are independent for all n and $V_i \overset{\text{ind}}{\sim} \text{Be}(1-\sigma, M + i\sigma)$, $i = 1, 2, \ldots$

14.12 (Generalized Dirichlet distribution) A random probability vector $p \in \mathbb{S}_k$ has a *generalized Dirichlet distribution* with parameters $(a, b) \in \mathbb{R}^{k-1} \times \mathbb{R}^{k-1}$ if the density is

$$\prod_{i=1}^{k-1} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \Big(\prod_{i=1}^{k-1} p_i^{a_i - 1} \Big) p_k^{b_{k-1}-1} \Big[\prod_{i=1}^{k-2} \Big(1 - \sum_{j=1}^{i} p_j \Big)^{b_i - (a_{i+1} + b_{i+1})} \Big].$$

Show that if $V_i \overset{\text{ind}}{\sim} \text{Be}(a_i, b_i)$, for $i = 1, \ldots, k-1$, and $V_k = 1$, then the stick-breaking weights $(W_1, \ldots, W_k)$ obtained by $W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, $i = 1, \ldots, k$, follow the generalized Dirichlet distribution with parameters $((a_1, a_2, \ldots, a_{k-1}), (b_1, b_2, \ldots, b_{k-1}))$.

14.13 (Lijoi et al. 2005a) Let P follow the normalized inverse-Gaussian process with parameter α, where $\int_0^\infty \sqrt{x}\, d\alpha(x) < \infty$. Show that the cumulative distribution function of $L = \int x\, dP(x)$ satisfies

$$\begin{aligned} \text{P}(L \le s) = \frac{1}{2} - \frac{e^{|\alpha|}}{\pi} \int_0^\infty t^{-1} \exp\Big[- \int (1 + 4t^2(x-s)^2)^{1/4} \\ \cos\Big(\tfrac{1}{2} \arctan(2t(x-s)) \Big)\, d\alpha(x) \Big] \times \sin\Big[- \int \big[1 + 4t^2(x-s)^2\big]^{1/4} \\ \sin\Big(\frac{1}{2} \arctan(2t(x-s)) \Big)\, d\alpha(x) \Big]\, dt. \end{aligned}$$

14.14 (Lijoi et al. 2005a) Let P follow the normalized inverse-Gaussian process with parameter α, where α is atomless and has support in $[c, \infty)$, for $c \ge -\infty$. Show that the posterior density of $L = \int x\, dP(x)$ based on $(X_1, \ldots, X_n) | P \overset{\text{iid}}{\sim} P$ is given by $\pi^{-1} I_{c+}^{n-1} \text{Im}[\psi]$ if n is odd and $-\pi^{-1} I_{c+}^{n-1} \text{Re}[\psi]$ if n is even, where I_{c+}^m is the fractional integral (11.1) and

$$\psi(s) = \frac{1}{Q}(n-1)!\,2^{n-1}|\alpha|^{-(2n-(2+K))}\int_0^\infty t^{n-1}e^{-\int\sqrt{1-2it(x-s)}\,d\alpha(x)} \prod_{j=1}^{K}(1-2it(\tilde{X}_j - s))^{-N_{j,n}+1/2}\,dt,$$

for $Q = \sum_{r=0}^{n-1}\binom{n-1}{r}(-1)^r|\alpha|^{-2r}\Gamma(K+2+2r-2n;|\alpha|)$.

14.15 Let $(X, Y) \sim P$, where P follows a stick-breaking process with reference measure $\alpha = \alpha_1\alpha_{2|1}$. Let P_1 be the marginal of $X \sim P$ and $P_{2|1}(\cdot|\,X)$ be the regular conditional distribution of Y given X. Show that if α_1 is atomless, P_1 has a stick-breaking representation. If $\alpha_{2|1}(\cdot|\,x)$ is also atomless for all x, then show that $P_{2|1}(\cdot|\,X)$ is δ_Y with $Y \sim \alpha_{2|1}(\cdot|\,X)$.

14.16 Consider density estimation with a general kernel in the setting of Theorem 7.15, but the mixing distribution follows a stick-breaking process with independent stick-breaking variables $V_1, V_2, \ldots$ Formulate an analogous consistency theorem.

14.17 (Lijoi et al. 2008) Show that the intersection of the class of Gibbs processes and normalized (homogeneous) CRM is the class of normalized generalized gamma processes.

14.18 (James et al. 2006) Show that if the posterior distribution for an NCRM under i.i.d. sampling is again an NCRM, then the resulting prior must be a Dirichlet process.

14.19 (IBP stick breaking) Let $\pi_{(1),K} \geq \pi_{(2),K} \geq \cdots \geq \pi_{(K),K}$ be the ordered values of a sample $\pi_1, \pi_2, \ldots \pi_K \overset{\text{iid}}{\sim} \text{Be}(M/K, 1)$. Then $\pi_{(k),K} = \prod_{i=1}^{k}\tau_{i,K}$, for $\tau_{i,K} := \pi_{(i),K}/\pi_{(i-1),K}$ and $\pi_{(0),K}$ interpreted as 1. Show that $\tau_{i,K} \overset{\text{iid}}{\sim} \text{Be}(M, 1)$, for $i = 1, \ldots, K$.