Random permutations and partition models

Peter McCullagh [1]
University of Chicago
January 2010

Set partitions

For $n \geq 1$, a partition B of the finite set $[n] = \{1, \ldots, n\}$ is

- a collection $B = \{b_1, \ldots\}$ of disjoint non-empty subsets, called blocks, whose union is $[n]$;
- an equivalence relation or Boolean function $B\colon [n] \times [n] \to \{0, 1\}$ that is reflexive, symmetric and transitive;
- a symmetric Boolean matrix such that $B_{ij} = 1$ if i, j belong to the same block.

These equivalent representations are not distinguished in the notation, so B is a set of subsets, a matrix, a Boolean function, or a subset of $[n] \times [n]$, as the context demands. In practice, a partition is sometimes written in an abbreviated form, such as $B = 2|13$ for a partition of [3]. In this notation, the five partitions of [3] are

$$123, \quad 12|3, \quad 13|2, \quad 23|1, \quad 1|2|3.$$

The blocks are unordered, so 2|13 is the same partition as 13|2 and 2|31.

A partition B is a sub-partition of B^* if each block of B is a subset of some block of B^* or, equivalently, if $B_{ij} = 1$ implies $B^*_{ij} = 1$. This relationship is a partial order denoted by $B \leq B^*$, which can be interpreted as $B \subset B^*$ if each partition is regarded as a subset of $[n]^2$. The partition lattice $\mathcal{E}_n$ is the set of partitions of $[n]$ with this partial order. To each pair of partitions B, B' there corresponds a greatest lower bound $B \wedge B'$, which is the set intersection or Hadamard component-wise matrix product. The least upper bound $B \vee B'$ is the least element that is greater than both, the transitive completion of $B \cup B'$. The least element of $\mathcal{E}_n$ is the partition $\mathbf{0}_n$ with n singleton blocks, and the greatest element is the single-block partition denoted by $\mathbf{1}_n$.

A permutation $\sigma\colon [n] \to [n]$ induces an action $B \mapsto B^\sigma$ by composition such that $B^\sigma(i, j) = B(\sigma(i), \sigma(j))$. In matrix notation, $B^\sigma = \sigma B \sigma^{-1}$, so the action by conjugation permutes both the rows and columns of B in the same way. The block sizes are preserved and are maximally invariant under conjugation. In this way, the 15 partitions of [4] may be grouped into five orbits or equivalence classes as follows:

$$1234, \quad 123|4\,[4], \quad 12|34\,[3], \quad 12|3|4\,[6], \quad 1|2|3|4.$$

Thus, for example, 12|34 is the representative element for one orbit, which also includes 13|24 and 14|23.

The symbol $\#B$ applied to a set denotes the number of its elements, so $\#B$ is the number of blocks, and $\#b$ is the size of block $b \in B$. If $\mathcal{E}_n$ is the set of equivalence relations on $[n]$, or the set of partitions of $[n]$, the first few values of $\#\mathcal{E}_n$ are 1, 2, 5, 15, 52, called Bell numbers. More generally, $\#\mathcal{E}_n$ is the nth moment of the unit Poisson distribution whose exponential generating function is $\exp(e^t - 1)$. In the discussion of explicit probability models on $\mathcal{E}_n$, it is helpful to use the ascending and descending factorial symbols

$$\alpha^{\uparrow r} = \alpha(\alpha + 1) \cdots (\alpha + r - 1) = \Gamma(r + \alpha)/\Gamma(\alpha)$$
$$k^{\downarrow r} = k(k - 1) \cdots (k - r + 1)$$

for integer $r \geq 0$. Note that $k^{\downarrow r} = 0$ for positive integers $r > k$. By convention $\alpha^{\uparrow 0} = 1$.

[1] Support for this research was provided in part by NSF Grant DMS-0906592.

Dirichlet partition model

The term *partition model* refers to a probability distribution, or family of probability distributions, on the set $\mathcal{E}_n$ of partitions of $[n]$. In some cases, the probability is concentrated on the the subset $\mathcal{E}_n^k \subset \mathcal{E}_n$ of partitions having k or fewer blocks. A distribution on $\mathcal{E}_n$ such that $p_n(B) = p_n(\sigma B \sigma^{-1})$ for every permutation $\sigma\colon [n] \to [n]$ is said to be finitely exchangeable. Equivalently, p_n is exchangeable if $p_n(B)$ depends only on the block sizes of B.

Historically, the most important examples are Dirichlet-multinomial random partitions generated for fixed k in three steps as follows.

- First generate the random probability vector $\pi = (\pi_1, \ldots, \pi_k)$ from the Dirichlet distribution with parameter $(\theta_1, \ldots, \theta_k)$.
- Given π, the sequence $Y_1, \ldots, Y_n, \ldots$ is independent and identically distributed, each component taking values in $\{1, \ldots, k\}$ with probability π. Each sequence of length n in which the value r occurs $n_r \geq 0$ times has probability

$$\frac{\Gamma(\theta_\bullet) \prod_{j=1}^k \theta_j^{\uparrow n_j}}{\Gamma(n + \theta_\bullet)},$$

  where $\theta_\bullet = \sum \theta_j$.
- Now forget the labels $1, \ldots, k$ and consider only the partition B generated by the sequence Y, i.e. $B_{ij} = 1$ if $Y_i = Y_j$. The distribution is exchangeable, but an explicit simple formula is available only for the uniform case $\theta_j = \lambda/k$, which is now assumed. The number of sequences generating the same partition B is $k^{\downarrow \#B}$, and these have equal probability in the uniform case. Consequently, the induced partition has probability

$$p_{nk}(B, \lambda) = k^{\downarrow \#B} \frac{\Gamma(\lambda) \prod_{b \in B} (\lambda/k)^{\uparrow \#b}}{\Gamma(n + \lambda)}, \tag{1}$$

  called the uniform Dirichlet-multinomial partition distribution. The factor $k^{\downarrow \#B}$ ensures that partitions having more than k blocks have zero probability.

In the limit as $k \to \infty$, the uniform Dirichlet-multinomial partition becomes

$$p_n(B, \lambda) = \frac{\lambda^{\#B} \prod_{b \in B} \Gamma(\#b)}{\lambda^{\uparrow n}}. \tag{2}$$

This is the celebrated Ewens distribution, or Ewens sampling formula, which arises in population genetics as the partition generated by allele type in a population evolving according to the Fisher-Wright model by random mutation with no selective advantage of allele types (Ewens, 1972). The preceding derivation, a version of which can be found in chapter 3 of Kingman (1980), goes back to Watterson (1974). The Ewens partition is the same as the partition generated by a sequence drawn according to the Blackwell-McQueen urn scheme (Blackwell and McQueen, 1973).

Although the derivation makes sense only if k is a positve integer, the distribution (1) is well defined for negative values $-\lambda < k < 0$. For a discussion of this and the connection with GEM distributions and Poisson-Dirichlet distributions, see Pitman (2006, section 3.2).

Partition processes and partition structures

Deletion of element n from the set $[n]$, or deletion of the last row and column from $B \in \mathcal{E}_n$, determines a map $D_n \colon \mathcal{E}_n \to \mathcal{E}_{n-1}$, a projection from the larger to the smaller lattice. These deletion maps make the sets $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ into a projective system

$$\cdots \mathcal{E}_{n+1} \xrightarrow{D_{n+1}} \mathcal{E}_n \xrightarrow{D_n} \mathcal{E}_{n-1} \cdots$$

A family $p = (p_1, p_2, \ldots)$ in which p_n is a probability distribution on $\mathcal{E}_n$ is said to be mutually consistent, or Kolmogorov-consistent, if each p_{n-1} is the marginal distribution obtained from p_n under deletion of element n from the set $[n]$. In other words, $p_{n-1}(A) = p_n(D_n^{-1}A)$ for $A \subset \mathcal{E}_{n-1}$. Kolmogorov consistency guarantees the existence of a random partition of the natural numbers whose finite restrictions are distributed as p_n. The partition is infinitely exchangeable if each p_n is finitely exchangeable. Some authors, for example Kingman (1980), refer to p as *a partition structure*.

An exchangeable partition process may be generated from an exchangeable sequence $Y_1, Y_2, \ldots$ by the transformation $B_{ij} = 1$ if $Y_i = Y_j$ and zero otherwise. The Dirichlet-multinomial and the Ewens processes are generated in this way. Kingman's (1978) paintbox construction shows that every exchangeable partition process may be generated from an exchangeable sequence in this manner.

Let B be an infinitely exchangeable partition, $B[n] \sim p_n$, let B^* be a fixed partition in $\mathcal{E}_n$, and suppose that the event $B[n] \leq B^*$ occurs. Then $B[n]$ lies in the lattice interval $[\mathbf{0}_n, B^*]$, which means that $B[n] = B[b_1]|B[b_2]|\ldots$ is the concatenation of partitions of the blocks $b \in B^*$. For each block $b \in B^*$, the restriction $B[b]$ is distributed as $p_{\#b}$, so it is natural to ask whether, and under what conditions, the blocks of B^* are partitioned independently given $B[n] \leq B^*$. Conditional independence implies that

$$p_n(B \,|\, B[n] \leq B^*) = \prod_{b \in B^*} p_{\#b}(B[b]), \tag{3}$$

which is a type of non-interference or lack-of-memory property not dissimilar to that of the exponential distribution on the real line. It is straightforward to check that the condition is satisfied by (2) but not by (1). Aldous (1996) shows that conditional independence uniquely characterizes the Ewens family. Mixtures of Ewens processes do not have this property.

Further exchangeable partition models

Although Dirichlet partition processes are the most common in applied work, it is useful to know that many alternative partition models exist. Although some of these are easy to simulate, most do not have simple expressions for the distributions, but there are exceptions of the form

$$p_n(B; \lambda) = \frac{\Gamma(B)\, Q_n(B; \lambda)}{\lambda^{\uparrow n}}, \tag{4}$$

for certain polynomials $Q_n(B; \lambda)$ of degree $\#B$ in λ. One such polynomial is

$$Q_n(B, \lambda) = \sum_{B \leq B' \leq \mathbf{1}_n} \lambda^{\#B'} / B',$$

which depends on B only through the block sizes. The functions $\Gamma(B) = \prod_{b \in B} \Gamma(\#b)$ and $B^\alpha = \prod_{b \in B} (\#b)^\alpha$ are multiplicative $\mathcal{E}_n \to \mathcal{R}$, and $1/B = B^{-1}$ is the inverse of the product of block sizes.

For each $\lambda > 0$, $p_n(B; \lambda)$ depends on B only through the block sizes, so the distribution is exchangeable. Moreover, it can be shown that the family is mutually consistent in the Kolmogorov sense. However, the conditional independence property (3) is not satisfied unless $Q_n(B; \lambda) = \lambda^{\#B}$.

The expected number of blocks grows slowly with n, approximately $\lambda \log(n)$ for the Ewens process, and $\lambda \log^2(n)/\log\log(n)$ for the process shown above.

Chinese restaurant process

A partition process is a random partition $B \sim p$ of a countably infinite set $\{u_1, u_2, \ldots\}$, and the restriction $B[n]$ of B to $\{u_1, \ldots, u_n\}$ is distributed as p_n. The conditional distribution of $B[n+1]$ given $B[n]$ is determined by the probabilities assigned to those events in $\mathcal{E}_{n+1}$ that are consistent with $B[n]$, i.e. the events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$. For the uniform Dirichlet-multinomial model (1), these are

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} (\#b + \lambda/k)/(n+\lambda) & b \in B \\ \lambda(1 - \#B/k)/(n+\lambda) & b = \emptyset. \end{cases} \tag{5}$$

In the limit as $k \to \infty$, we obtain

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} \#b/(n+\lambda) & b \in B \\ \lambda/(n+\lambda) & b = \emptyset, \end{cases} \tag{6}$$

which is the conditional probability for the Ewens process.

To each partition process p there corresponds a sequential description called the Chinese restaurant process, in which $B[n]$ is the arrangement of the first n customers at $\#B$ tables. The placement of the next customer is determined by the conditional distribution $p_{n+1}(B[n+1] \,|\, B[n])$. For the Ewens process, the customer chooses a new table with probability $\lambda/(n+\lambda)$ or one of the occupied tables with probability proportional to the number of occupants. The term, due to Pitman, Dubins and Aldous, is used primarily in connection with the Ewens and Dirichlet-multinomial models.

Exchangeable random permutations

Beginning with the uniform distribution on permutations of $[n]$, the exponential family with canonical parameter $\theta = \log(\lambda)$ and canonical statistic $\#\sigma$ equal to the number of cycles is

$$p_n(\sigma) = \lambda^{\#\sigma}/\lambda^{\uparrow n}.$$

The Stirling number of the first kind, $S_{n,k}$, is the number of permutations of $[n]$ having exactly k cycles, for which $\lambda^{\uparrow n} = \sum_{k=1}^{n} S_{n,k}\lambda^k$ is the generating function. The cycles of the permutation determine a partition of $[n]$ whose distribution is (2), and a partition of the integer n whose distribution is (7). From the cumulant function

$$\log(\lambda^{\uparrow n}) = \sum_{j=0}^{n-1} \log(j+\lambda)$$

it follows that $\#\sigma = X_0 + \cdots + X_{n-1}$ is the sum of independent Bernoulli variables with parameter $E(X_j) = \lambda/(\lambda+j)$, which is evident also from the Chinese restaurant representation. For large n, the number of cycles is roughly Poisson with parameter $\lambda \log(n)$, implying that $\hat{\lambda} \simeq \#\sigma/\log(n)$ is a consistent estimate as $n \to \infty$, but practically inconsistent.

A minor modification of the Chinese restaurant process also generates a random permutation by keeping track of the cyclic arrangement of customers at tables. After n customers are seated, the next customer chooses a table with probability (5) or (6), as determined by the partition process. If the table is occupied, the new arrival sits to the left of one customer selected uniformly at random from the table occupants. The random permutation thus generated is $j \mapsto \sigma(j)$ from j to the left neighbour $\sigma(j)$.

Provided that the partition process is consistent and exchangeable, the distributions p_n on permutations of $[n]$ are exchangeable and mutually consistent under the projection $\Pi_n \to \Pi_{n-1}$ on permutations in which element n is deleted from the cyclic representation (Pitman, 2006, section 3.1). In this way, every infinitely exchangeable random partition also determines an infinitely exchangeable random permutation $\sigma \colon \mathbb{N} \to \mathbb{N}$ of the natural numbers. Distributional exchangeability in this context is not to be confused with uniformity on Π_n.

On the number of unseen species

A partition of the set $[n]$ is a set of blocks, and the block sizes determine a partition of the integer n. For example, the partition $15|23|4$ of the set $[5]$ is associated with the integer partition $2+2+1$, one singleton and two doubletons. An integer partition $m = (m_1, \ldots, m_n)$ is a list of multiplicities, also written as $m = 1^{m_1} 2^{m_2} \cdots n^{m_n}$, such that $\sum jm_j = n$. The number of blocks, usually called the number of parts of the integer partition, is the sum of the multiplicities $m_\bullet = \sum m_j$.

Under the natural action $B \mapsto \pi B \pi^{-1}$ of permutations π on set partitions, each orbit is associated with a partition of the integer n. The multiplicity vector m contains all the information about block sizes, but there is a subtle transfer of emphasis from block sizes to the multiplicities of the parts.

By definition, an exchangeable distribution on set partitions is a function only of the block sizes, so $p_n(B) = q_n(m)$, where m is the integer partition corresponding to B. Since there are

$$\frac{n!}{\prod_{j=1}^n (j!)^{m_j} m_j!}$$

set partitions B corresponding to a given integer partition m, to each exchangeable distribution p_n on set partitions there corresponds a marginal distribution

$$q_n(m) = p_n(B) \times \frac{n!}{\prod_{j=1}^n (j!)^{m_j} m_j!}$$

on integer partitions. For example, the Ewens distribution on integer partitions is

$$\frac{\lambda^{m_\bullet} \Gamma(\lambda) \prod \Gamma(j)^{m_j}}{\Gamma(n+\lambda)} \times \frac{n!}{\prod_{j=1}^n (j!)^{m_j} m_j!} = \frac{\lambda^{m_\bullet}\, n!\, \Gamma(\lambda)}{\Gamma(n+\lambda) \prod_j j^{m_j} m_j!}. \tag{7}$$

This version leads naturally to an alternative description of the Ewens distribution as follows. Let $M = M_1, \ldots, M_n$ be independent Poisson random variables with mean $E(M_j) = \lambda\theta^j/j$ for some positive number θ. Then $\sum jM_j$ is sufficient for θ, and the conditional distribution $\mathrm{pr}(M = m \mid \sum_{j=1}^n jM_j = n)$ is the Ewens integer-partition distribution with parameter λ. This representation leads naturally to a simple method of estimation and testing, using Poisson log-linear models with model formula $1 + j$ and offset $-\log(j)$ for response vectors that are integer partitions.

The problem of estimating the number of unseen species was first tackled in a paper by Fisher (1943), using an approach that appears to be entirely unrelated to partition processes. Specimens from species i occur as a Poisson process with rate ρ_i, the rates for distinct species being independent and identically distributed gamma random variables. The number $N_i \geq 0$ of occurrences of species i in an interval of length t is a negative binomial random variable

$$\mathrm{pr}(N_i = x) = (1-\theta)^\nu \theta^x \frac{\Gamma(\nu + x)}{x!\, \Gamma(\nu)}. \tag{8}$$

In this setting, $\theta = t/(1+t)$ is a monotone function of the sampling time, whereas $\nu > 0$ is a fixed number independent of t. Specimen counts for distinct species are independent and identically distributed random variables with parameters $\nu > 0$ and $0 < \theta < 1$.

The probability that no specimens from species i occur in the sample is $(1-\theta)^\nu$, the same for every species. Most species are unlikely to be observed if either θ is small, i.e. the time interval is short, or ν is small.

Let M_x be the number of species occurring $x \geq 0$ times, so that $M_\bullet$ is the unknown total number of species of which $M_\bullet - M_0$ are observed. The approach followed by Fisher is to estimate

the parameters θ, ν by conditioning on the number of species observed and regarding the observed multiplicities M_x for $x \geq 1$ as multinomial with parameter vector proportional to the negative binomial frequencies (8). For Fisher's entomological examples, this approach pointed to $\nu = 0$, consistent with the Ewens distribution (7), and indicating that the data are consistent with the number of species being infinite. Fisher's approach using a model indexed by species is less direct for ecological purposes than a process indexed by specimens. Nonetheless, subsequent analyses by Good and Toulmin (1956), Holgate (1969) and Efron and Thisted (1976) showed how Fisher's model can be used to make predictions about the likely number of new species in a subsequent temporal extension of the original sample. This amounts to a version of the Chinese restaurant process.

At this point, it is worth clarifying the connection between Fisher's negative binomial formulation and the Ewens partition formulation. The relation between them is the same as the relation between binomial and negative binomial sampling schemes for a Bernoulli process: they are not equivalent, but they are complementary. The partition formulation is an exchangeable process indexed by *specimens*: it gives the distribution of species numbers in a sample consisting of a fixed number of *specimens*. Fisher's version is also an exchangeable process, in fact an iid process, but this process is indexed by *species*: it gives the distribution of the sample composition for a fixed set of *species* observed over a finite period. In either case, the conditional distribution given a sample containing k species and n specimens is the distribution induced from the uniform distribution on the set of $S_{n,k}$ permutations having k cycles. For the sorts of ecological or literary applications considered by Good and Toulmin (1956) or Efron and Thisted (1976), the partition process indexed by specimens is much more direct than one indexed by species.

Fisher's finding that the multiplicities decay as $E(M_j) \propto \theta^j/j$, proportional to the frequencies in the log-series distribution, is a property of many processes describing population structure, either social structure or genetic structure. It occurs in Kendall's (1975) model for family sizes as measured by surname frequencies. One explanation for universality lies in the nature of the transition rates for Kendall's process, a discussion of which can be found in section 2.4 of Kelly (1978).

Equivariant partition models

A family $p_n(\sigma;\theta)$ of distributions on permutations indexed by a parameter matrix θ, is said to be equivariant under the induced action of the symmetric group if $p_n(\sigma;\theta) = p_n(g\sigma g^{-1}; g\theta g^{-1})$ for all σ, θ, and for each group element $g\colon [n] \to [n]$. By definition, the parameter space is closed under conjugation: $\theta \in \Theta$ implies $g\theta g^{-1} \in \Theta$. The same definition applies to partition models. Unlike exchangeability, equivariance is not a property of a distribution, but a property of the family. In this setting, the family associated with $[n]$ is not necessarily the same as the family of marginal distributions induced by deletion from $[n+1]$.

Exponential family models play a major role in both theoretical and applied work, so it is natural to begin with such a family of distributions on permutations of the matrix-exponential type

$$p_n(\sigma;\theta) = \alpha^{\#\sigma} \exp(\mathrm{tr}(\sigma\theta))/M_\alpha(\theta),$$

where $\alpha > 0$ and $\mathrm{tr}(\sigma\theta) = \sum_{j=1}^n \theta_{\sigma(j),j}$ is the trace of the ordinary matrix product. The normalizing constant is the α-permanent

$$M_\alpha(\theta) = \mathrm{per}_\alpha(K) = \sum_\sigma \alpha^{\#\sigma} \prod_{j=1}^n K_{\sigma(j),j}$$

where $K_{ij} = \exp(\theta_{ij})$ is the component-wise exponential matrix. This family of distributions on permutations is equivariant.

The limit of the α-permanent as $\alpha \to 0$ gives the sum of cyclic permutations

$$\mathrm{cyp}(K) = \lim_{\alpha \to 0} \alpha^{-1} \mathrm{per}_\alpha(K) = \sum_{\sigma : \#\sigma = 1} \prod_{j=1}^{n} K_{\sigma(j),j},$$

giving an alternative expression for the α-permanent

$$\mathrm{per}_\alpha(K) = \sum_{B \in \mathcal{E}_n} \alpha^{\#B} \prod_{b \in B} \mathrm{cyp}(K[b])$$

as a sum over partitions. The induced marginal distribution (11) on partitions is of the product-partition type recommended by Hartigan (1990), and is also equivariant. Note that the matrix θ and its transpose determine the same distribution on partitions, but they do not usually determine the same distribution on permutations.

The α-permanent has a less obvious convolution property that helps to explain why this function might be expected to occur in partition models:

$$\sum_{b \subset [n]} \mathrm{per}_\alpha(K[b]) \, \mathrm{per}_{\alpha'}(K[\bar{b}]) = \mathrm{per}_{\alpha+\alpha'}(K). \tag{9}$$

The sum extends over all 2^n subsets of $[n]$, and $\bar{b}$ is the complement of b in $[n]$. A derivation can be found in section 2.4 of McCullagh and Møller (2006). If B is a partition of $[n]$, the symbol $K \cdot B = B \cdot K$ denotes the Hadamard component-wise matrix product for which

$$\mathrm{per}_\alpha(K \cdot B) = \prod_{b \in B} \mathrm{per}_\alpha(K[b])$$

is the product over the blocks of B of α-permanents restricted to the blocks. Thus the function $B \mapsto \mathrm{per}_\alpha(K \cdot B)$ is of the product-partition type.

With α, K as parameters, we may define a family of probability distributions on $\mathcal{E}_n^k$, i.e. partitions of $[n]$ having k or fewer blocks, as follows:

$$p_{nk}(B) = k^{\downarrow \#B} \, \mathrm{per}_{\alpha/k}(K \cdot B) / \, \mathrm{per}_\alpha(K). \tag{10}$$

The fact that (10) is a probability distribution on $\mathcal{E}_n$ follows from the convolution property of permanents. The limit as $k \to \infty$

$$p_n(B) = \alpha^{\#B} \prod_{b \in B} \mathrm{cyp}(K[b]) / \, \mathrm{per}_\alpha(K), \tag{11}$$

is a product-partition model satisfying the conditional independence property (3). For $K = \mathbf{1}_n$, the $n \times n$ matrix whose elements are all one, $\mathrm{per}_\alpha(\mathbf{1}_n) = \alpha^{\uparrow n}$ is the ascending factorial function. Thus the uniform Dirichlet-multinomial model (1) and the Ewens model (2) are both obtained by setting $\theta = 0$.

Leaf-labelled trees

Kingman's $[n]$-coalescent is a non-decreasing, $\mathcal{E}_n$-valued Markov process (B_t) in continuous-time starting from the partition $B_0 = \mathbf{0}_n$ with n singleton blocks at time zero. The coalescence intensity is one for each pair of blocks regardless of size, so each coalescence event unites two blocks chosen uniformly at random from the set of pairs. Consequently, the first coalescence occurs after a random time T_n exponentially distributed with rate $\rho(n) = n(n-1)/2$ and mean $1/\rho(n)$. After k

coalescences, the partition consists of $n-k$ blocks, and the waiting time T_k for the next subsequent coalescence is exponential with rate $\rho(n-k)$. The time to complete coalescence is the sum of independent exponentials $T = T_n + T_{n-1} + \cdots + T_2$, which is a random variable with mean $2 - 2/n$ and variance increasing from 1 at $n = 2$ to a little less than 1.16 as $n \to \infty$. In the context of the Fisher-Wright model, the coalescent describes the genealogical relationships among a sample of individuals, and T is the time to the most recent common ancestor of the sample.

The $[n]$-coalescent is exchangeable for each n, but the property that makes it interesting mathematically, statistically and genetically is its consistency under selection or sub-sampling (Kingman, 1982). If we denote by p_n the distribution on $[n]$-trees implied by the specific Markovian model described above, it can be shown that the embedded tree obtained by deleting element n from the sample $[n]$ is not only Markovian but also distributed as p_{n-1}, i.e. the same coalescent rule applied to the subset $[n-1]$. This property is mathematically essential for genealogical trees because the occurrence or non-occurrence of individual n in the sample does not affect the genealogical relationships among the remainder.

A fragmentation $[n]$-tree is a non-increasing $\mathcal{E}_n$-valued Markov process starting from the trivial partition $B_0 = \mathbf{1}_n$ with one block of size n at time $t = 0$. The simplest of these are the consistent binary Gibbs fragmentation trees studied by Aldous (1996), Bertoin (2001, 2006) and McCullagh, Pitman and Winkel (2008). The first split into two branches occurs at a random time T_n exponentially distributed with parameter $\rho(n)$. Subsequently, each branch fragments independently according to the same family of distributions with parameter $\rho(\#b)$ for branch b, which is a Markovian conditional independence property analogous to (3). Consistency and conditional independence put severe limitations on both the splitting distribution and the rate function $\rho(n)$, so the entire class is essentially one-dimensional.

A rooted leaf-labelled tree T is also a non-negative symmetric matrix. The interpretation of T_{ij} as the distance from the root to the junction at which leaves i, j occur on disjoint branches implies the inequality $T_{ij} \geq \min(T_{ik}, T_{jk})$ for all $i, j, k \in [n]$. The set of $[n]$-trees is a subset of the positive definite symmetric matrices, not a manifold, but a finite union of manifolds of dimension $2n-1$, or n if the diagonal elements are constrained to be equal. Like partitions, rooted trees form a projective system within the positive definite matrices. A fragmentation tree is an infinitely exchangeable random tree, which is also a special type of infinitely exchangeable random matrix.

Cluster processes and classification models

A $\mathcal{R}^d$-valued cluster process is a pair (Y, B) in which $Y = (Y_1, \ldots)$ is an $\mathcal{R}^d$-valued random sequence and B is a random partition of $\mathbb{N}$. The process is said to be exchangeable if, for each finite sample $[n] \subset \mathbb{N}$, the restricted process $(Y[n], B[n])$ is invariant under permutation $\sigma\colon [n] \to [n]$ of sample elements.

The Gauss-Ewens process is the simplest non-trivial example for which the distribution for a sample $[n]$ is as follows. First fix the parameter values $\lambda > 0$, and Σ^0, Σ^1 both positive definite of order d. In the first step B has the Ewens distribution on $\mathcal{E}_n$ with parameter λ. Conditionally on B, Y is a zero-mean Gaussian matrix of order $n \times d$ with covariance matrix

$$\operatorname{cov}(Y_{ir}, Y_{js} \,|\, B) = \delta_{ij}\Sigma^0_{rs} + B_{ij}\Sigma^1_{rs},$$

where δ_{ij} is the Kronecker symbol. A scatterplot colour-coded by blocks of the Y values in $\mathcal{R}^2$ shows that the points tend to be clustered, the degree of clustering being governed by the ratio of between to within-cluster variances.

For an equivalent construction we may proceed using a version of the Chinese restaurant process in which tables are numbered in order of occupancy, and $t(i)$ is the number of the table at which customer i is seated. In addition, $\epsilon_1, \ldots$ and $\eta_1, \ldots$ are independent Gaussian sequences with

independent components $\epsilon_i \sim N_d(0, \Sigma^0)$, and $\eta_i \sim N_d(0, \Sigma^1)$. The sequence t determines B, and the value for individual i is a vector $Y_i = \eta_{t(i)} + \epsilon_i$ in $\mathcal{R}^d$, or $Y_i = \mu + \eta_{t(i)} + \epsilon_i$ if a constant non-zero mean vector is included.

Despite the lack of class labels, cluster processes lend themselves naturally to prediction and classification, also called supervised learning. The description that follows is taken from McCullagh and Yang (2006) but, with minor modifications, the same description applies equally to more complicated non-linear versions associated with generalized linear mixed models (Blei, Ng and Jordan 2003). Given the observation $(Y[n], B[n])$ for the 'training sample' $[n]$, together with the feature vector Y_{n+1} for specimen u_{n+1}, the conditional distribution of $B[n+1]$ is determined by those events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$ that are compatible with the observation. The assignment of a positive probability to the event that the new specimen belongs to a previously unobserved class seems highly desirable, even logically necessary, in many applications.

If the classes are tree-structured with two levels, we may generate a sub-partition $B' \leq B$ whose conditional distribution given B is Ewens restricted to the interval $[\mathbf{0}_n, B]$, with parameter λ'. This sub-partition has the effect of splitting each main clusters randomly into sub-clusters. For the sample $[n]$, let $t'(i)$ be the number of the sub-cluster in which individual i occurs. Given B, B', the Gauss-Ewens two-level tree process is a sum of three independent Gaussian processes $Y_i = \eta_{t(i)} + \eta'_{t'(i)} + \epsilon_i$ for which the conditional distributions may be computed as before. In this situation, however, events that are compatible with the observation $B[n], B'[n]$ are of three types as follows:

$$u_{n+1} \mapsto b' \in B'[n], \quad u_{n+1} \mapsto \emptyset \subset b \in B[n], \quad u_{n+1} \mapsto \emptyset.$$

In all, there are $\#B'+\#B+1$ disjoint events for which the conditional distribution given $B[n], B'[n], Y[n+1]$ must be computed. An event of the second type is one in which the new specimen belongs to the major class $b \in B$, but not to any of the sub-types previously observed for this class.

Further applications of partition models

Exchangeable partition models are used to construct non-trivial, exchangeable processes suitable for cluster analysis and density estimation. See Frayley and Raftery (2002) or Booth, Casella and Hobert (2008) for a discussion of computational techniques. Cluster analysis means a partitioning of the sample units into dissimilar non-overlapping blocks that are internally homogeneous. Densiy estimation refers to the conditional distribution of Y_{n+1} given the sample values. Usually, this is to be done for an exchangeable process in the absence of external covariate or relational information about the units. In the computer-science literature, cluster detection is also called unsupervised learning. The simplest of these models is the marginal Gauss-Ewens process in which only the sequence $Y[n]$ is observed, and $B[n]$ is to be inferred. The conditional distribution $p_n(B \,|\, Y[n])$ on $\mathcal{E}_n$ is the posterior distribution on clusterings or partitions of $[n]$, and $E(B \,|\, Y)$ is the one-dimensional marginal distribution on pairs of units. In estimating the number of clusters, it is important to distinguish between the sample number $\#B[n]$, which is necessarily finite, and the population number $\#B[\mathbb{N}]$, which could be infinite (McCullagh and Yang, 2008).

Exchangeable partition models are also used to provide a Bayesian solution to the multiple comparisons problem. The key idea is to associate with each partition B of $[k]$ a subspace $V_B \subset \mathcal{R}^k$ equal to the span of the columns of B. Thus, V_B consists of vectors x such that $x_r = x_s$ if $B_{rs} = 1$. For a treatment factor having k levels $\tau_1, \ldots, \tau_k$, the Gauss-Ewens prior distribution on R^k puts positive mass on the subspaces V_B for each $B \in \mathcal{E}_k$. Likewise, the posterior distribution also puts positive probability on these subspaces, which enables us to compute in a coherent way the posterior probability $\mathrm{pr}(\tau \in V_B \,|\, y)$ or the marginal posterior probability $\mathrm{pr}(\tau_r = \tau_s \,|\, y)$. For details, see Gopalan and Berry (1998).

References

[1] Aldous, D. (1996) Probability distributions on cladograms. In *Random Discrete Structures*. IMA Vol. Appl. Math **76**. Springer, New York, 1–18.

[2] Booth, J.G., Casella, G. and Hobert, J.P. (2008) Clustering using objective functions and stochastic search. *J. Roy. Statist. Soc. B* **70**, 119–139.

[3] Bertoin, J. (2001) Homogeneous fragmentation processes. *Probab. Theory and Related Fields* **121** 301–318.

[4] Bertoin, J. (2006) *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Math, **102**.

[5] Blackwell, D. and MacQueen, J. (1973) Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

[6] Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet allocation. *J. Machine learning Research* **3**, 993–1022.

[7] Efron, B. and Thisted, R.A. (1976) Estimating the number of unknown species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

[8] Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

[9] Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* **12**, 42–58.

[10] Fraley, C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611–631.

[11] Good, I.J. and Toulmin, G.H. (1956) The number of new species, and the increase in population coverage when a sample is increased. *Biometrika* **43**, 45–63.

[12] Gopalan, R. and Berry, D.A. (1998) Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93**, 1130–1139.

[13] Hartigan, J.A. (1990) Partition models. *Communications in Statistics: Theory and Methods* **19**, 2745–2756.

[14] Holgate, P. (1969) Species frequency distributions. *Biometrika* **65**, 651–660.

[15] Kelly, F.P. (1978) *Reversibility and Stochastic Networks*. Wiley, Chichester.

[16] Kendall, D.G. (1975) Some problems in mathematical genealogy. In *Perspectives in Probability and statistics: Papers in Honour of M.S. Bartlett.* Academic Press, London, 325–345.

[17] Kingman, J.F.C. (1975) Random discrete distributions (with discussion). *J. Roy. Statist. Soc. B* **37**, 1–22.

[18] Kingman, J.F.C. (1977) The population structure associated with the Ewens sampling formula. *Theoretical Population Biology* **11**, 274–283.

[19] Kingman, J.F.C. (1978) The representation of partition structures. *J. Lond. Math. Soc.* **18**, 374–380.

[20] Kingman, J.F.C. (1980) *Mathematics of Genetic Diversity*. CBMS-NSF conference series in applied math, **34** SIAM, Philadelphia.

[21] Kingman, J.F.C. (1982) The coalescent. *Stochastic Processes Appl.* **13**, 235–248.

[22] McCullagh, P. and Møller, J. (2006) The permanental process. *Adv. Appl. Prob.* **38**, 873–888.

[23] McCullagh, P. and Yang, J. (2006) Stochastic classification models. Proc. International Congress of Mathematicians, 2006, vol. III, 669–686.

[24] McCullagh, P. and Yang, J. (2008). How many clusters? *Bayesian Analysis* **3**, 1–19.

[25] McCullagh, P., Pitman, J. and Winkel, M. (2008) Gibbs fragmentation trees. *Bernoulli* **14**, 988–1002.

[26] Pitman, J. (2006) *Combinatorial Stochastic Processes*. Springer-Verlag, Berlin.

[27] Watterson, G.A. (1974) The sampling theory of selectively neutral alleles. *Adv. Appl. Prob.* **6**, 217–250.