The talk is based on joint results with Jim Pitman [2]. We consider a sample \(X_1, \ldots, X_n\) from a random discrete distribution \(P = (P_1, P_2, \ldots)\) where

\[
P[X_i = j | P] = P_j, \quad (j = 1, 2, \ldots; i = 1, \ldots, n).
\]

In the context of population genetics the sample generates a partition \(\{C_1, \ldots, C_k\}\) of the index set \([n] = \{1, \ldots, n\}\) on clusters of equal values, thought as of individuals labelled by \([n]\) sampled from an idealized infinite population and partitioned on species. Values \(X_i\) carry some additional species characteristics, say a relative age of the species in the process of evolution. Clusters can be ordered in different ways. We deal with order by appearance, with \(C_1^* = \{i : X_i = X_1\}\), \(C_2^* = \{i : X_i = X_{\min[j \in [n]: C_j]}\}\), etc., and order by age (sample value), with \(C_1^\alpha = \{i : X_i = \min[j \in [n]: X_j]\}\), \(C_2^\alpha = \{i : X_i = \min[j \in [n]: C_j X_j]\}\), etc.

We suppose that the distribution \(P\) is constructed by a residual allocation model:

\[
P_j = H_j (1 - H_1) \cdots (1 - H_{j-1}), \quad j = 1, 2, \ldots, \quad H_j \in (0, 1) \text{ almost surely},
\]

with independent (but not necessarily identically distributed) factors \(H_1, H_2, \ldots\). The most known model of this type is the GEM(\(\theta\)) distribution arising when all \(H_j\) have beta(1, \(\theta\)) distribution, \(\theta > 0\), with sampling described by the Ewens sampling formula. It has been generalized to the two parameter GEM(\(\alpha, \theta\)) model when \(H_j\) has beta(1 - \(\alpha\), \(\theta + j\alpha\)) distribution, for some \(\alpha \in (0, 1)\), and further generalized to the so-called Gibbs partitions.

Let \(N_i^\alpha = \#C_{\alpha i}\), \(N_i^\theta = \#C_i\), \(i = 1, \ldots, k\), be the cluster sizes in appearance and age order. It is well-known that given the list of cluster sizes \(\{n_1, \ldots, n_k\}\) in any order, the distribution of \(\{N_1^\alpha, \ldots, N_k^\alpha\}\) can be constructed by a sequence of size-biased picks, with cluster of size \(n_i\) chosen with probability \(n_i/(n_1 + \cdots + n_k)\) from all clusters. This is true for any \(P\) and for GEM(\(\theta\)) model the same was shown to be true for the distribution of \(\{N_1^\alpha, \ldots, N_k^\alpha\}\) by Donnelly and Tavaré [1]. We show that for GEM(\(\alpha, \theta\)) and more generally for Gibbs(\(\alpha\)) models the size-biased pick should be replaced by \((\text{size} - \alpha)\)-biased pick with probability \((n_i - \alpha)/(n_1 + \cdots + n_k - k\alpha)\) to choose a part of size \(n_i\).

References
