Random graph representation of the Chinese restaurant process

Annalisa Cerquetti

Istituto di Metodi Quantitativi, Bocconi University Milano, Italy

annalisa.cerquetti@unibocconi.it

Abstract

The Chinese restaurant process (Pitman, 1996) is a well-known sequential random construction which generates observations from the exchangeable partition of the positive integers induced by sampling from a Dirichlet process. A generalization is provided by Ishwaran and James (2003) for sampling from Pitman's *species sampling* models. Here we derive a random graph representation of the exchangeable partition induced by sampling from the species sampling models. The growing random graph model is characterized in terms of attachment rules deduced by a variation of the generalized Chinese restaurant process, based on the associated sampling distribution.

1. Exchangeable random partitions

According to Kingman's theory (1978) if $(X_1,\dots,X_n)$ is a sample from a random distribution F_0, the random partition $\Pi_n = \{A_1,\dots,A_k\}$ of $[n]=\{1,2,\dots,n\}$, induced by the exchangeable equivalence relation

$$i \approx j \Leftrightarrow X_i(\omega) = X_j(\omega),$$

is an *exchangeable* partition of $[n]$. This means that for each n, the distribution of Π_n is such that for each particular partition $\{A_1,\dots,A_k\}$ of $[n]$, with $|A_j| = n_j$, for $1 \le j \le k$, where $n_j \ge 1$ and $\sum_{j=1}^k n_j = n$,

$$P(\Pi_n = \{A_1,\dots,A_k\}) = p(n_1,\dots,n_k), \tag{1}$$

for some symmetric function p of k-tuples of non-negative integers with sum n (compositions of n), called the *exchangeable partition probability function (EPPF)* of Π_n. An infinite random partition $\Pi_\infty := (\Pi_n)$ of the set of the positive integers $\mathbb{N}$ is exchangeable if its restriction Π_n to $[n]$ is exchangeable for every n.

Different ways to encode the random sizes $N_j = |A_j|$ of the blocks of an exchangeable partition as random compositions of n may be considered. One way is to consider the *block counts vector*, i.e. the random vector of non-negative integers

$$M_i^{(n)} := \sum_{j=1}^k 1(N_j = i),$$

for $i = 1,\dots,n$, subject to $\sum_{i=1}^n i M_i^{(n)} = n$ and $\sum_{i=1}^n M_i^{(n)} = k$, which counts how many blocks of size i there are in a given partition of $[n]$, for $i=1,\dots,n$.
Due to the bijection between random compositions of n and possible vectors of counts, for each partition of $[n]$ the probability that $M_i^{(n)} = m_i, 1 \le i \le n$ depends on the EPPF and is given by the following formula, (see e.g. Pitman, 1996):

$$p^*(m_1,\dots,m_n) = \frac{n!}{\prod_{i=1}^n (i!)^{m_i} m_i!} p(n_1,\dots,n_k). \tag{2}$$

The best known case of EPPF is the one associated with the Dirichlet process of parameter $\mu = \theta\nu$, and is given by the following formula:

$$p_{(\theta)}(n_1,\dots,n_k) = \frac{\theta^{k-1}\prod_{j=1}^k (n_j-1)!}{[1+\theta]_{n-1}}, \tag{3}$$

where $n = \sum_j n_j$ and $[x]_m = \prod_{j=1}^m (x+j-1)$.

The Chinese restaurant sequential description of sampling from (3) is well known to be as follows.
Assume that an *unlimited* number of customers arrives sequentially in a restaurant with an *unlimited* number of circular tables, each capable of sitting an *unlimited* number of customers. Let the first customer to arrive be seated at the first table.
For $n \ge 1$, given $n_1,\dots,n_k$ the placement of the first n customers at k tables, the $n+1$th customer is:

- seated at the table j, with probability $p_{j,n} = \frac{n_j}{n+\theta}$, for $1 \le j \le k$,
- seated at a *new* table with probability $p_{0,n} = \frac{\theta}{n+\theta}$.

2. Prediction rules for the species sampling models

By Theorem 2. in Hansen and Pitman (2000), the class of *species sampling sequences*, i.e. of all exchangeable sequences (X_n) admitting a prediction rule of the form:

$$P(X_{n+1} \in \cdot | X_1,\dots,X_n) = \sum_{j=1}^{k_n} p_{j,n}\delta_{X_j^*}(\cdot) + p_{0,n}\nu(\cdot), \tag{4}$$

[where $(X_1^*,\dots,X_{k_n}^*)$, are the k_n distinct values in $X_1,\dots,X_n$, i.i.d with non-atomic probability measure ν, and n_j is the multiplicity of X_j^*], is characterized by constraints on $p_{j,n}$, and $p_{0,n}$. It is shown that these quantities can be expressed in terms of EPPF associated with the random partition generated by $(X_1,\dots,X_n)$ as follows, provided $p(\mathbf{n}) > 0$:

$$p_{j,n} = p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})}, \tag{5}$$

for $1 \le j \le k_n$, where $p(\mathbf{n}^{j+}) = p(n_1,\dots,n_j+1,\dots,n_k)$,

$$p_{0,n} = p_{0,n}(\mathbf{n}) = \frac{p(\mathbf{n}^{l+})}{p(\mathbf{n})} \tag{6}$$

for $l = k_n + 1$.

In Ishwaran and James (2003) the previous constraints are exploited to give a generalized version of the Chinese restaurant sequential construction, which provides samples from the partition structure induced by a species sampling model.

It may be shown (Cerquetti, 2005) that an alternative characterization of the class of species sampling sequences may be obtained resorting to the following alternative prediction rules, expressed in terms of the sampling formula (2):

$$p_{i,n}^* = p_{i,n}^*(\mathbf{m}) = \frac{(i+1)(m_{i+1}+1)}{n+1}\frac{p^*(\mathbf{m}_{i-1}^{(i+1)+})}{p^*(\mathbf{m})} \tag{7}$$

$$p_{0,n}^* = p_{0,n}^*(\mathbf{m}) = \frac{(m_1+1)}{n+1}\frac{p^*(\mathbf{m}^{1+})}{p^*(\mathbf{m})} \tag{8}$$

where, for $i = 1,\dots,n$,

$$p^*(\mathbf{m}_{i-1}^{(i+1)+}) = p^*(m_1,\dots,m_i-1,m_{i+1}+1,\dots,m_n). \tag{9}$$

The corresponding expression for the prediction rule (4) results:

$$P(X_{n+1} \in \cdot | X_1,\dots,X_n) = \sum_{i=1}^{n} p_{i,n}^*\delta_{\tilde{X}_i}(\cdot) + p_{0,n}^*\nu(\cdot), \tag{10}$$

where for each i, $\tilde{X}_i$ is one of the distinct values in $(X_1,\dots,X_n)$ with the same multiplicity i.

3. Growing random graphs derived from the CRP

A graph $G = (V,E)$ is a mathematical tool for abstract representation and modeling network structures, in which the vertices set V represents units and the edges set E represents the interactions between pairs of units. The classical Poisson random graph model of Erdős and Rény, which is inadequate to describe some important properties of real-world networks, has been generalized in a variety of ways.

Here we consider a growing random graph model defined as follows (recall that given a graph G, a *clique* is a maximal complete subgraph of G):

A graph starts with a single isolated vertex.
For $n \ge 1$, given the observed *clique counts vector*, $(m_1,\dots,m_n)$, the $n+1$th adding vertex:

- joins one of the existing cliques of order i (i.e. establishes a connection with each vertex of the clique) with probability $p_{i,n}^*(\mathbf{m})$, for $i = 1,\dots,n$,
- stands alone, i.e. starts a *new* clique, with probability $p_{0,n}^*(\mathbf{m})$.

It easy to see that at each step the observed *clique counts vector* is a sample from (2).

Example 1. An explicit form for the sampling distribution (2) is known to be as follows for the *two-parameter* species sampling model discussed in Pitman (1996):

$$p_{\alpha,\theta}^*(m_1,\dots,m_n) =$$

$$= n!\frac{\prod_{l=1}^{k-1}(\theta+l\alpha)}{[1+\theta]_{n-1}}\prod_{i=1}^{n}\left(\frac{[1-\alpha]_{i-1}}{i!}\right)^{m_i}\frac{1}{m_i!}. \tag{11}$$

for $0 \le \alpha < 1$ and $\theta > -\alpha$. For $\alpha = 0$, $\theta > 0$ this is the Ewens sampling formula.
A growing random graph model governed by the two-parameter model may be easily obtained by applying rules (7) and (8) to (11).

The attachment rules defining the random evolution of the graph turn out to be as follows.
Let a graph starts with a single vertex.
For $n \ge 1$, given $(m_1,\dots,m_n)$ the observed clique counts vector, the $(n+1)$th adding vertex:

- joins one of the existing cliques of order i, with probability:

$$p_{i,n}^* = \frac{m_{i,n}(i-\alpha)}{n+\theta}, \quad \text{for } i = 1,\dots,n$$

- starts a *new* clique, (i.e. remains isolated), with probability:

$$p_{0,n}^* = \frac{\theta + \alpha\sum_{i=1}^n m_{i,n}}{n+\theta}.$$

At each time step, the observed clique counts vector is a sample from (11).

Remark 1. Consider a sequence of random graphs $G_n = (V_n, E_n)$ whose vertices are labelled with an exchangeable sequence (X_n), and edges arise between nodes i and j such that $X_i = X_j$ as in Cerquetti and Fortini (2003). If the sequence of the vertices labels is a sample from a species sampling model, then, for each n, rules (7) and (8) define a sequential construction of samples from G_n. Moreover the random graph G_n decomposes almost surely into random cliques, and the clique counts vector distribution is given by (2).

Remark 2. An alternative way to construct generalizations of the Erdős-Rényi model is to define a random graph by specifying its *degree distribution*. It easy to show that the model defined in section 3. is characterized by a *random* degree distribution governed by a variation of the sampling formula (2).

Acknowledgements

This poster was prepared while the author was visiting the research group in Probability and its Applications at the Boole Centre for Research in Informatics, University College Cork, Ireland. Thanks to Neil O'Connell for the kind hospitality and to SFI for supporting me.

References

CERQUETTI, A. (2005) Random graph representation of species sampling models. *Tech. Rep., IMQ, Bocconi University, Milano.* (In preparation)

CERQUETTI, A. AND FORTINI, S. (2003) A Poisson approximation for colored graphs under exchangeability. *Tech. Rep. n.79, IMQ, Bocconi University, Milano.*

HANSEN, B. AND PITMAN, J. (2000) Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters* **46**, 251-256.

ISHWARAN, H. AND JAMES, L.F. (2003) GWCR processes for species sampling mixture models. *Statistica Sinica* **13**, 1211-1235.

PITMAN, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson, Shapley L.S. and MacQueen J.B., editors, *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes*, pages 245-267. IMS, Hayward, CA.