Random Partitions and Bayesian Nonparametrics

Stefano Favaro (University of Torino and Collegio Carlo Alberto),
Shui Feng (McMaster University)

April 17, 2016–April 24, 2016

1 Overview of the Field

Random partitions arise naturally in many subjects including but not limited to Bayesian nonparametric, ecology, machine learning, number theory, physics, population genetics, and the theory of random matrices.

The basic framework for a Bayesian nonparametric model involves a two stage generation of data: a probability P is first chosen from the support of a prior distribution on the space of probability measures, followed by a sample of (conditionally) independent and identically distributed random variables from P. The main goal consists in determining and investigating the posterior distribution, that is the conditional distribution of P given the observable sample. The prior in a nonparametric Bayesian model corresponds to the law of the random probability measure P, and probability theory provides a large arsenal for studying distributional properties of P, especially under the assumption that P is discrete almost surely.

The natural link between random partitions and Bayesian nonparametric is through the celebrated de Finetti representation theorem. Exchangeable random partitions are the cornerstone of Bayesian nonparametric inference for a broad class of statistical problems, referred to as species sampling problems.

2 Recent Developments

Let $\mathbb{N}$ denote the set of natural numbers. A partition π of $\mathbb{N}$ is a collection of disjoint subsets $\{\pi_i : i \geq 1\}$ of $\mathbb{N}$ ordered by their least elements with $\cup_{i=1}^{\infty} \pi_i = \mathbb{N}$. Denote the collection of all partitions of $\mathbb{N}$ by Π. For any $n \geq 1$, a partition π^n of $\mathbb{N}^n = \{1, 2, \ldots, n\}$ is defined similarly. The set of all such partitions is denoted by Π^n. A random partition of $\mathbb{N}^n$ or $\mathbb{N}$ is a probability on Π^n or $\mathbb{N}$. Under certain consistency assumptions, one is able to construct a random partition on $\mathbb{N}$ from those on $\mathbb{N}^n$ by letting n tend to infinity. When the random partition depends only on the number of subsets and the size of each subsets of a partition π^n, it corresponds to a family of probability partition functions

$$\{p(n_1, n_2, \ldots, n_k) : 1 \leq i \leq k \leq n, 1 \leq n_i \leq n, \sum_{i=1}^{k} n_i = n\}$$

where k is the number of subsets and n_i is the size of the i-th set.

The most studied family of probability partition functions is the Ewens sampling formula ([4]) describing in the genetics context the sampling distribution of a neutral population. This is followed by the study of Kingman's partition structures and coalescent ([9],[10]). After the discovery of Pitman sampling formula ([12]) and the coalescent with multiple collisions ([13],[15]), there have been intensive studies on various generalizations of these models ([2],[14]).

In the Bayesian nonparametric settings, one starts with a discrete random probability measure $P = \sum_{i\geq 1} p_i \delta_{\xi_i}$ on an arbitrary space S, where $(p_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} p_i = 1$ almost surely, and $(\xi_i)_{i\geq 1}$ are S-valued random locations, or random labels, independent of $(p_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic distribution. By virtue of de Finetti representation theorem, there exists an exchangeable sequence of random variables $(X_i)_{i\geq 1}$ such that

$$\begin{aligned} X_i \,|\, P &\sim iid \quad P, \qquad i = 1, 2, \ldots, \\ P &\sim \quad \mathcal{P}, \end{aligned} \tag{1}$$

for any $n \geq 1$, with $\mathcal{P}$ being the distribution of $\lim_{n\to+\infty} n^{-1} \sum_{1\leq i\leq n} \delta_{X_i}$. Due to the discreteness of P, we expect ties in $(X_1, \ldots, X_n)$. Let $K_n = k \leq n$ denote the number of different types or species in the sample, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$ such that $\sum_{1\leq i\leq K_n} N_{i,n} = n$. The sample $(X_1, \ldots, X_n)$ induces a random partition Π_n of the set $\{1, \ldots, n\}$, in the sense that any index $1 \leq i \neq j \leq n$ belongs to the same partition set if and only if $X_i = X_j$. As shown by Kingman [9], for any $n \geq 1$ the distribution of the random partition is exchangeable.

Exchangeable random partitions are the cornerstone of Bayesian nonparametric inference for a broad class of statistical problems, referred to as species sampling problems, where samples are assumed to be drawn from a population of individuals belonging to an (ideally) infinite number of species $(X_i^*)_{i\geq 1}$ with unknown proportions $(p_i)_{i\geq 1}$. In such a species sampling framework, (1) takes on the natural interpretation of a Bayesian nonparametric model, where $\mathcal{P}$ is the prior distribution on the unknown species composition $(p_i)_{i\geq 1}$ of the population. Species sampling problems have originally appeared in ecology, and their importance has grown considerably in recent years, driven by challenging applications arising from bioinformatics, genetics, linguistics, design of experiments, machine learning, etc. Given an initial sample $(X_1, \ldots, X_n)$ featuring $K_n = k$ species with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$, interest lies in making inference on certain statistics of the random partition induced by an additional unobserved sample of size m. Statistics of interest are, among others, the $K_m^{(n)}$ new species and the $M_{l,m}^{(n)}$ species with frequency l to be observed in the additional sample. Given that, Bayesian nonparametric inference for species sampling problems relies on the study of the conditional distribution of the random partition induced by the additional sample given $(X_1, \ldots, X_n)$, i.e., the distribution of the random partition of a sample of size m from the posterior distribution of P. More details are found in [5] and [6] and references therein.

3 Objectives and Outcome of the Meeting

The random partitions discussed above are all associated with random measures describing the equilibrium behaviour of certain population. But in probabilistic literature, there are a plethora of random measures arising from stochastic processes. Random partitions constructed from these measures are thus associated with the non-equilibrium population structures. Models in this aspects include but not limited to the Fleming-Viot process ([8]), infinitely-many-neutral-alleles model ([3]), Petrov diffusion ([11]), coagulation and fragmentation processes ([2]), GEM process ([7]), and general coalescents ([1]). From a statistical perspective, these more general random structures suggest potential applications in modelling samples arising from populations with more complex compositional structures.

Our first objective is concerned with the equilibrium random partitions. Under the prior assumption that P is the two parameter Poisson Dirichlet process, we want to study the large m asymptotic behaviour of the posterior distribution of $K_m^{(m)}$ and $M_{l,m}^{(n)}$, given an initial observed sample of size n. From a Bayesian nonparametric perspective, this asymptotic analysis is mainly motivated by the need of deriving approximations of the posterior distributions $K_m^{(m)}$ and $M_{l,m}^{(n)}$. Indeed, while the two parameter Poisson Dirichlet prior leads to explicit expressions for these posterior distributions, such expressions involve combinatorial coefficients and special functions whose evaluation for large m is cumbersome, thus preventing their concrete implementation. In [5] and [6] we studied the large m asymptotic behaviour, in terms of fluctuations and large deviations, of the posterior distribution of $K_m^{(m)}$ and $M_{l,m}^{(n)}$. Recently we made progresses on the related problem of deriving central limit theorems and moderate deviation principles, for the posterior distribution of $K_m^{(m)}$. Of course such a result is of direct applicability for deriving large m asymptotic credible intervals for the Bayesian nonparametric estimator of $K_m^{(n)}$. We are able to complete this project during our stay at

BIRS. We also intend to discuss the problem of deriving, by means of tools from the theory of concentration inequalities, non-asymptotic credible intervals for the Bayesian nonparametric estimator of $K_m^{(n)}$. This part of research is currently an open project.

Our second objective is to study certain non-equilibrium random partitions. More specifically it is related to the genealogical structure of the Kingman's coalescent and, in particular, to the problem of making Bayesian nonparametric (predictive) inference on such a structure. Let $L_n(t)$ be the number of non mutant lineages at time t back in a Kingman's coalescent tree of a sample of n genes. While this distribution is well-known from Kingman [9], what seems unknown is the distribution of the number $L_{l,n}(t)$ of non mutant lineages with frequency l at time t back in a Kingman's coalescent tree of a sample of n genes. Recently we derived the distribution of $L_{l,n}(t)$, as well as related conditional distributions. These conditional distributions may be interpreted as genuine posterior distributions. During our stay at BIRS we completed this project by investigating some large n asymptotic properties of $L_{l,n}(t)$. A different project, still related to the Kingman's coalescent, concerns with the problem of making Bayesian nonparametric (predictive) inference on the genealogical structure of the Kingman's coalescent. Specifically, we aim at deriving the conditional distribution of the number of lineages in a Kingman's coalescent tree of a sample of $n+m$, given a Kingman's coalescent tree of a sample of n genes. From a Bayesian nonparametric perspective, such a conditional distribution takes on the interpretation of the posterior distribution of the number of lineages, and its expected value provides the corresponding Bayesian nonparametric estimator. Analogue of the celebrated Good-Turing and Good-Toulmin estimators has been introduced in the framework of lineages. This is also completed during our visit to BIRS.

References

[1] N. Berestycki, *Recent progress in coalescent theory*, Ensaios Matemticos, 16. Sociedade Brasileira de Matemtica, Rio de Janeiro, 2009.

[2] J. Bertoin, *Random fragmentation and coagulation processes*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2006.

[3] S.N. Ethier and T.G. Kurtz, The infinitely-many-neutral-alleles diffusion model, *Adv. Appl. Prob.* **13**(1981), 429–452.

[4] W. Ewens, The sampling theory of selectively neutral alleles, *Theor. Popul. Biol.*, **3**(1972), 87–112.

[5] S. Favaro and S. Feng, Asymptotics for the number of blocks in a conditional Ewens-Pitman sampling model, *Electron. J. Probab.*, **19**(2014), 1–15.

[6] S. Favaro and S. Feng, Large deviation principles for the Ewens-Pitman sampling model, *Electron. J. Probab.*, **20**(2015), 1–27.

[7] S. Feng and F.Y. Wang, A class of infinite-dimensional diffusion processes with connection to population genetics, *J. Appl. Prob.*, **44**(2007), 938–949.

[8] W.H. Fleming and M. Viot, Some measure-valued Markov processes in population genetics theory, *Indiana Univ. Math. J.*, **28**(1979), 817–843.

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

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

[11] L. Petrov, A two-parameter family of infinite-dimensional diffusions in the Kingman simplex, *Funktsional. Anal. i Prilozhen.*, **43** (2009), 45–66.

[12] J. Pitman, The two-parameter generalization of Ewens' random partition structure, *Technical report*, **345**(1992), Dept. Statistics. U.C. Berkeley.

[13] J. Pitman, Coalescents with multiple collisions, *Ann. Probab.*, **27**(1999), 1870–1902.

[14] J. Pitman, *Combinatorial Stochastic Processes*, Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer ,New York, 2006.

[15] S. Sagitov, The general coalescent with asynchronous mergers of ancestral lines, *J. Appl. Probab.*, **36**(1999), 1116–1125.