Bayesian Nonparametric Analysis of Kingman's Coalescent

Shui Feng
McMaster University

(Joint Work With Stefano Favaro of University of Torino and Paul A. Jenkins of University of Warwick)

The 15th Workshop on Markov Processes and Related Topics at Jilin University.

July 11-15, 2019

- Kingman Coalescent
- Bayesian Nonparametrics
- Fleming-Viot Process With Parent Independent Mutation
- New Results

Kingman Coalescent

For each $n \geq 1$, let $E_n = \{1, 2, ..., n\}$ and $\mathcal{E}_n$ denote the collection of equivalence relations of E_n. Each element of $\mathcal{E}_n$ is thus a subset of $E_n \times E_n$. For example, in the case of $n = 3$, the set

$$\{(1,1), (2,2), (3,3), (1,3), (3,1)\}$$

defines an equivalence relation that results in two equivalence classes $\{1, 3\}$ and $\{2\}$. The set $\mathcal{E}_n$ is clearly finite and its elements will be denoted by $\eta, \xi,$ etc.

In genetic applications, the equivalence relations are defined through the ancestral structures. Two individuals are equivalent if they have the same ancestor at some time t in the past. For ξ, η in $\mathcal{E}_n$, we write $\xi \prec \eta$ if η is obtained from ξ by combining exactly two equivalence classes of ξ into one. For distinct ξ, η in $\mathcal{E}_n$, set

$$q_{\xi\eta} = \begin{cases} 1, & \xi \prec \eta \\ 0, & \text{else.} \end{cases}$$

Let $|\xi|$ be the number of equivalence classes induced by ξ. Define

$$q_\xi := -q_{\xi,\xi} = \binom{|\xi|}{2}.$$

Definition: Kingman's n-coalescent is a $\mathcal{E}_n$-valued, continuous-time, Markov chain X_t with infinitesimal matrix $(q_{\xi\eta})$ starting at $X_0 = \{(i,i) : i = 1, ..., n\}$. Kingman's coalescent corresponds to the limit process as n tends to infinity.

Let $D(t) = |X_t|$.

Then D_t is a pure-death process with death rate

$$\lim_{h \to 0} h^{-1} P\{D(t+h) = k-1 \mid D(t) = k\} = \binom{k}{2}, \quad k \geq 2.$$

Replace $\binom{k}{2}$ with $n(n+\theta-1)/2$ leads to Kingman's coalescent with mutation.

Tavaré's Formula

Let $t > 0$, $\lambda_m = \frac{m(m-1+\theta)}{2}$ and $(a)_k = a(a+1)\cdots(a+k-1)$. Set

$$d_n^\theta(t) = P\{D(t) = n\}, \rho_m(t) = (-1)^m(2m-1+\theta)\mathsf{e}^{-\lambda_m t}.$$

It is obtained in Tavare [5] that for $n \geq 1$

$$d_n^\theta(t) = (-1)^n \sum_{m \geq n} \rho_m(t) \frac{\binom{m}{n}(n+\theta)_{(m-1)}}{m!}.$$

For $n = 0$,

$$d_0^\theta(t) = 1 + \sum_{m=1}^{\infty} \rho_m(t)(\theta)_{(m-1)} m!.$$

Bayesian Nonparametrics

Consider a random variable X with distribution depending on parameter θ. In Bayesian statistics, the parameter is modelled as a random variable Θ. This reflects the basic principle of Bayesian statistics: all forms of uncertainty should be expressed as randomness.

The distribution Q of Θ is called the *prior distribution*.

Under a Bayesian model, data is generated in two stages, as

$$\begin{aligned} &\Theta \sim Q \\ &X_1, X_2, \ldots | \Theta \sim \text{iid with common distribution } P_\Theta. \end{aligned}$$

Here the sequence $X_1, X_2, \ldots$ is conditionally i.i.d. or exchangeable.

The main objective is then to determine the *posterior distribution*, the conditional distribution of Θ given the data,

$$Q\{\Theta \in \cdot | X_i = x_i, i = 1, \ldots, n\}.$$

This corresponds to parameter estimation in the classical approach.

A *Bayesian nonparametric model* is a Bayesian model with infinite dimensional parametric space.

Examples include the space of probability measures and the space of probability density functions.

An Example

For any $\theta > 0$, let $U_1, U_2, \ldots$ be a sequence of iid random variables with common distribution $Beta(1, \theta)$. Set

$$V_1 = U_1, V_n = (1 - U_1) \cdots (1 - U_{n-1}) U_n, \quad n \geq 2.$$

Let S be Polish space and ν_0 a probability on S. The **Dirichlet process** (first appeared in Ferguson [4]) with parameters θ, ν_0 is the random measure

$$\Xi_{\theta,\nu_0} = \sum_{i=1}^{\infty} V_i \delta_{\xi_i}.$$

where $\xi_1, \xi_2, \ldots$ are i.i.d. with common distribution ν_0 and is independent of $\{V_i\}_{i \geq 1}$. Denote the law of Ξ_{θ,ν_0} by Π_{θ,ν_0}.

Clearly Ξ_{θ,ν_0} belongs to an infinite dimensional space. Thus Π_{θ,ν_0} can serve as a nonparametric prior.

Fleming-Viot Process With Parent Independent Mutation

Let S be a compact metric space, $C(S)$ be the set of continuous functions on S, $M_1(S)$ the space of probability measures on S equipped with the usual weak topology, and ν_0 a diffuse probability in $M_1(S)$. Consider operator A of the form

$$Af(x) = \frac{\theta}{2} \int (f(y) - f(x))\nu_0(dy), \ f \in C(S).$$

Define

$$\mathcal{D} = \{u : u(\mu) = f(\langle \phi, \mu \rangle), \ f \in C_b^\infty(\mathbf{R}), \phi \in C(S), \mu \in M_1(S)\},$$

where $\langle \phi, \mu \rangle$ is the integration of ϕ with respect to μ and $C_b^\infty(\mathbf{R})$ denotes the set of all bounded, infinitely differentiable functions on $\mathbf{R}$. Then the **Fleming-Viot process with parent independent mutation** (FV process) is a pure atomic measure-valued Markov process with generator

$$\mathcal{A}u(\mu) = \langle A\delta u(\mu)/\delta\mu(\cdot), \mu\rangle + \frac{f''(\langle\phi,\mu\rangle)}{2}\langle\phi,\phi\rangle_\mu,\ \ u \in \mathcal{D},$$

where

$$\begin{aligned}\delta u(\mu)/\delta\mu(x) &= \lim_{\varepsilon\to 0+}\varepsilon^{-1}\{u((1-\varepsilon)\mu+\varepsilon\delta_x)-u(\mu)\},\\ \langle\phi,\psi\rangle_\mu &= \langle\phi\psi,\mu\rangle - \langle\phi,\mu\rangle\langle\psi,\mu\rangle,\end{aligned}$$

and δ_x stands for the Dirac measure at $x \in S$.

Theorem 1. (Ethier [1]) *The Fleming-Viot process with parent independent mutation is reversible with the Dirichlet process Π_{θ,ν_0} as the reversible measure.*

Transition Function:

The probability transition function of the Fleming-Viot process has the form

$$P(t,\mu,d\,\nu) = d_0^\theta(t)\Pi_{\theta,\nu_0}(d\,\nu) \\ +\sum_{n=1}^{\infty} d_n^\theta(t)\int_{S^n}\mu^n(d\,x_1\times\cdots\times d\,x_n)\Pi_{n+\theta,\frac{n}{\theta+n}\eta_n+\frac{\theta}{\theta+n}\nu_0}(d\,\nu).$$

where

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}.$$

A natural connection to Bayesian nonparametrics is the fact that $\Pi_{n+\theta,\frac{n}{\theta+n}\eta_n+\frac{\theta}{\theta+n}\nu_0}$ turns out to be the posterior distribution of Dirichlet process Π_{θ,ν_0} given a sample of $X_i = x_i, i = 1,\ldots,n$.

New Results in Favaro, F and Jenkins [3]

Main Questions

1 How to makes use of the observed sample of individuals to infer quantities related to the genealogy of an additional unobservable sample?

2 How many non-mutant lineages would one expect a time t ago if the initial observable sample is enlarged by certain number of unobservable samples?

3 How many of these non-mutant lineages have small or large frequencies?

Distributional Identity for $D_m(t)$

For any $m \geq 1$, let $D_m(t)$ denote the pure death process in the Kingman m-coalescent.

For any $n \geq 1$ let $(Z_1^*, \dots, Z_n^*)$ be independent random variables identically distributed according to a non-atomic probability measure. For any $m \geq 1$, let $\mathbf{X}_m = (X_1, \dots, X_m)$ be a random sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1 \leq i \leq n} \delta_{Z_i^*}$.

The random variables $(Z_1^*, \dots, Z_n^*)$ denote the genetic types of the ancestors.

The composition of the sample $\mathbf{X}_m$ can be described as follows. We denote by $\{X_1^*, \dots, X_{K_m}^*\}$ the labels identifying the K_m distinct types in $\mathbf{X}_m$ which do not coincide with any of the atoms Z_i^*'s.

For any $j = 1, \dots, n$, let

$$M_{j,m} = \sum_{1 \le i \le m} \mathbb{1}_{\{Z_j^*\}}(X_i)$$

denote the number of X_i's that coincide with the atom Z_j^*, and $\mathbf{M}_m = (M_{1,m}, \dots, M_{n,m})$. Introduce the random variable

$$R_{n,m} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m} > 0\}}. \qquad (1)$$

It is clear that $R_{n,m}$ denotes the number of distinct types in the sample $\mathbf{X}_m$ that coincide with the atoms Z_i^*'s.

Theorem 2. *For any* $m \geq 1$ *let* $\mathbf{X}_m$ *be a sample from a Dirichlet process with atomic base measure* $\theta\nu_0 + \sum_{1\leq i\leq n} \delta_{Z_i^*}$*, for* $n \geq 0$*. Then, for* $r = 0, \ldots, \min(n, m)$

$$\mathbb{P}[R_{n,m} = r] = r! \frac{\binom{n}{r}\binom{m}{r}(\theta + r)_{(m-r)}}{(\theta + n)_{(m)}}.$$

Furthermore,

$$D_m(t) \overset{d}{=} R_{D(t),m}$$

for each $t > 0$*, where* $\{D(t) : t \geq 0\}$ *is the death process of the cardinality of Kingman's coalescent starting from infinity.*

Kingman's Coalescent on the left hand side is represented through the posterior distribution on the right.

Refined Distributional Identity

Noting that the process $D_m(t)$ describes the non-mutant lineages surviving from time 0 to time t (Griffiths [2]). For any $l = 1, \dots, m$, let $D_{l,m}(t)$ denote the number of non-mutant lineages having frequency l. Define

$$R_{l,n,m} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m}=l\}}.$$

In orther words $R_{l,n,m}$ denotes the number of distinct types in $\mathbf{X}_m$ that coincide with the atoms Z_i^*'s and have frequency l. From the discussion above it is clear that, given $D(t) = n$, $D_{l,m}(t)$ has the same distribution as $R_{l,n,m}$. Thus we obtain that

$$D_{l,m}(t) \overset{\mathsf{d}}{=} R_{l,D(t),m},$$

Note that the random variable $D_{l,m}(t)$ represents a natural refinement of $D_m(t)$ in the sense that

$$D_m(t) = \sum_{l=1}^{m} D_{l,m}(t).$$

Theorem 3. *For any* $m \geq 1$ *let* $\mathbf{X}_m$ *be a sample from a Dirichlet process with atomic base measure* $\theta\nu_0 + \sum_{1\leq i\leq n} \delta_{Z_i^*}$*, for* $n \geq 0$*. Then, for* $r = 0, \ldots, \min(n, \lfloor m/l \rfloor)$

$$\mathbb{P}[R_{l,n,m} = r] = \frac{m!}{(\theta+n)_{(m)}} \sum_{i=r}^{\min(n,\lfloor m/l \rfloor)} (-1)^{i-r} \frac{\binom{i}{r}\binom{n}{i}(\theta+n-i)_{(m-il)}}{(m-il)!},$$

where $\min(n, \lfloor m/l \rfloor)$ *denotes the minimum between* n *and the integer part of* m/l*.*

Conditional Distributional Identities

Let $\mathbf{X}_m$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$ and, for any $m' \ge 0$, let $\mathbf{X}_{m'} = (X_{m+1}, \ldots, X_{m+m'})$ be an additional sample. More precisely $\mathbf{X}_{m'}$ may be viewed as a sample from the conditional distribution of the Dirichlet process with base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$, given the initial sample $\mathbf{X}_m$. We denote by $M_{j,m'} = \sum_{1\le i\le m'} \mathbb{1}_{\{Z_j^*\}}(X_{m+i})$ the number of X_{m+i}'s that coincide with the atom Z_j^*, and we introduce the random variable

$$R_{n,m+m'} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m}+M_{i,m'}>0\}},$$

which denotes the number of distinct types in the enlarged sample $\mathbf{X}_{m+m'} = \{\mathbf{X}_m, \mathbf{X}_{m'}\}$ that coincide with the atoms Z_i^*'s. Similarly we introduce the following random variable

$$\tilde{R}_{l,n,m'} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m'}>0\}} \mathbb{1}_{\{M_{i,m}=l\}}.$$

Theorem 4. *For any $m \geq 1$ and $m' \geq 0$ let $\mathbf{X}_{m+m'}$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1 \leq i \leq n} \delta_{Z_i^*}$, for $n \geq 0$. Then one has*

(i) *for* $x = y, \ldots, \min(n, y + m')$

$$\mathbb{P}[R_{n,m+m'} = x \,|\, R_{n,m} = y] = (x-y)! \frac{\binom{n-y}{x-y}\binom{m'}{x-y}(\theta + m + x)_{(m'-x+y)}}{(\theta + n + m)_{(m')}};$$

(ii) *for* $x = 0, \ldots, \min(y, m')$

$$\mathbb{P}[\tilde{R}_{l,n,m'} = x \,|\, R_{l,n,m} = y] = \frac{\binom{y}{x}}{(\theta + n + m)_{(m')}}$$

$$\times \sum_{i=y-x}^{y} (-1)^{i-(y-x)} \binom{x}{y-i} (\theta + n + m - i(1+l))_{(m')}.$$

Therefore, $R_{n,m}$ and $R_{l,n,m}$ are sufficient to predict $R_{n,m+m'}$ and $\tilde{R}_{l,n,m'}$, respectively.

A direct application of this result leads to

$$\mathbb{P}[D_{m+m'}(t) = x \mid D_m(t) = y]$$
$$= \frac{\binom{m}{y}\binom{m'}{x-y}(\theta+y)_{(x-y)}(m+m'+\theta)_{(y)}\mathbb{P}[D_{m+m'}(t)=x]}{\binom{m'+m}{x}(\theta+m)_{(x)}\mathbb{P}[D_m(t)=y]}$$

for any $x = y, \ldots, m + m'$ and each $t > 0$.

Generalizations to refined distributions involving frequencies can also be obtained in similar manner.

References

[1] Ethier, S.N.: The infinitely-many-neutral-alleles diffusion model with ages. *Adv. Appl. Probab.* **22**, (1990),1–24.

[2] Griffiths, R.C.: Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Popul. Biol.* **17**, (1980), 37–50.

[3] Favaro, S., Feng, S. and Jenkins, P. A.: Bayesian nonparametric analysis of Kingman's coalescent. *Ann. Inst. H. Poincaré Probab. Statist.*, Vol. **55**, No. 2, 1087–1115 .

[4] Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, (1973), 209–230.

[5] Tavaré, S.: Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26**, (1984),119–164.