Robust nonparametric Bayesian inference of Λ-measures from molecular data

Jere Koskela, Paul A. Jenkins and Dario Spanò
Mathematics Institute and Department of Statistics, University of Warwick

Introduction

The Λ-coalescent family [DK99, Pit99, Sag99] models the ancestry of a sample of haplotypes from a large population with an infinite variance family size distribution. Recent studies have indicated that these coalescents sometimes provide better modeling fits to high-fecundity populations, such as Atlantic cod or Pacific oysters, than the seminal Kingman's coalescent [BBB94, Á04, EW06, BB08, BBS11, SBB13]. The Λ-coalescent likelihood can be estimated pointwise using importance sampling [BB08, BBS11, KJS15b], also facilitating Bayesian inference. Natural parameter spaces for Λ-coalescents are infinite dimensional, motivating a nonparametric approach. In this poster we focus on inferring the measure $\Lambda \in \mathcal{M}_1([0,1])$ which drives the merger events of Λ-coalescent trees. We assume the simple scenario of recurrent mutation and negligible recombination, selection or any other evolutionary dynamics. In brief, suppose $P \in \mathcal{M}_1(\mathcal{M}_1([0,1]))$ is a prior probability measure on the space of possible Λ-measures, which captures the prior information about plausibility of different Λ-measures as explanations for population dynamics. Such a prior can be elicited from a wide variety of biological and ecological information, such as the family size distribution as outlined in the box below. The Bayesian procedure is a means of refining prior beliefs by using observed data, and is outlined in the box to the right.

Example: the Beta$(2-\alpha,\alpha)$-coalescent

The Beta$(2-\alpha,\alpha)$-coalescent [Sch03] can be obtained as the high density limit of a finite population which evolves as follows. Suppose there are N individuals evolving in discrete generations, each with a haplotype drawn from a finite set $\mathcal{H}$ (e.g. $\{T,C,A,G\}$ if we are modeling a single locus of DNA). Each individual produces a random number of potential offspring distributed according to a power law tail $r^{-\alpha}, \alpha \in [1,2)$. Offspring inherit the type of their parent. The next generation is formed by sampling N of these offspring without replacement. Those offspring not sampled are assumed dead, so that the population is of constant size. Measuring time in units of N generations and letting $N \to \infty$ yields a population whose type-frequencies are described by the $|\mathcal{H}|$-dimensional Beta$(2-\alpha,\alpha)$-Fleming-Viot jump-diffusion, and the ancestries of samples are given by Beta$(2-\alpha,\alpha)$-coalescent trees.

General Λ-coalescents with recurrent mutation

More generally, the Λ-coalescent family is parametrised by probability measures on the unit interval: $\Lambda \in \mathcal{M}_1([0,1])$. This measure determines the coalescence rates. When there are $n \in \mathbb{N}$ lineages, any $k \in \{2,\dots,n\}$ of them will merge at rate

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx).$$

In addition to Beta measures, other famous examples are $\Lambda = \delta_0$, i.e. Kingman's coalescent, and $\Lambda = \frac{2}{2+\psi^2}\delta_0 + \frac{\psi^2}{2+\psi^2}\delta_\psi$, where $\psi \in (0,1]$ [EW06]. Mutation can be incorporated into Λ-coalescents similarly to Kingman's coalescent. Conditional on an ancestral tree, mutations occur along branches as points of a Poisson process with rate $\theta > 0$. In the finite alleles context mutation probabilities for each parental haplotype can be specified by a stochastic matrix M, which is assumed to have a unique stationary distribution m. Figure 1 depicts a realisation of a Λ-coalescent tree with four leaves.

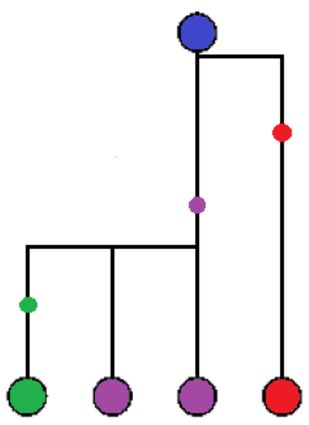

Figure 1 : A Λ-coalescent tree annotated with mutations. The most recent common ancestor (MRCA) is sampled from the stationary law of the mutation stochastic matrix M, and happens to be blue. Three mutations take place along the leaves of the tree, resulting in the haplotype configuration green, purple, purple, red at the leaves of the tree.

Λ-Fleming-Viot processes and duality

The evolution of population allele frequencies of populations with Λ-coalescent ancestries are modelled by the Λ-Fleming-Viot processes [BLG03]: jump diffusions on the $|\mathcal{H}|$-dimensional probability simplex with generator

$$Gf(\mathbf{x}) = \sum_{i,j\in\mathcal{H}} \left\{ \frac{\Lambda(\{0\})}{2}\mathbf{x}_i(\delta_{ij}-\mathbf{x}_j)f_{ij}(\mathbf{x}) + \theta\mathbf{x}_j(M_{ji}-\delta_{ij})f_i(\mathbf{x}) \right\} + \sum_{i\in\mathcal{H}} \int_0^1 \mathbf{x}_i\{f((1-r)\mathbf{x}+r\mathbf{e}_i) - f(\mathbf{x})\}r^{-2}\Lambda(dr),$$

where $\mathbf{e}_i$ is the unit vector with 1 in the i^{th} place and zeros elsewhere, and subscripts on functions denote partial derivatives. We denote the Λ-Fleming-Viot process by $(\mathbf{X}_t)_{t\geq 0}$ and its stationary law by π^Λ. The law of n leaves generated by a Λ-coalescent as outlined in the previous box can be expressed as an i.i.d. sample from a random measure sampled from the corresponding Λ-Fleming-Viot process. Denote the Λ-coalescent death process of n untyped lineages by $(\Pi_t)_{t\in[0,T]}$ where $T := \inf\{t > 0 : |\Pi_t| = 1\}$ is the hitting time of the MRCA. Let $f : \mathcal{H}^n \mapsto \mathbb{R}$ be a bounded function, and for a partition $\phi = \{C_1,\dots,C_m\}$ of $\{1,\dots,n\}$ let $f_\phi(a_1,\dots,a_m) = f(h_1,\dots,h_n)$ with $h_i = a_k$ when $i \in C_k$. Duality can then be stated as

$$\mathbb{E}_n\left[\sum_{a\in\mathcal{H}} f_{\Pi_T}(a)m(a)\right] = \mathbb{E}_{\pi^\Lambda}\left[\sum_{h_1,\dots,h_n\in\mathcal{H}} f(h_1,\dots,h_n)\mathbf{X}_0(h_1)\dots\mathbf{X}_0(h_n)\right],$$

where the expectation on the left is with respect to Λ-coalescent trees with n leaves, and on the right with respect to the stationary Λ-Fleming-Viot process.

Bayesian inference of Λ-measures

We aim to infer the jump measure Λ assuming θ and M are known. Fix $\eta > 0$ and assume that $\Lambda \in \mathcal{D}_\eta^b$, the space of strictly positive, bounded probability densities on $[\eta,1]$. Let $P \in \mathcal{M}_1(\mathcal{D}_\eta^b)$ be the prior. These assumptions on Λ rule out all of the examples mentioned on this poster but they are needed for technical reasons. Moreover, we are able to ensure P supports measures which lie arbitrarily close to any desired $\Lambda \in \mathcal{M}_1([0,1])$ in a sense we make precise below. The posterior given an observed configuration of type frequencies $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ is given by

$$P(B|\mathbf{n}) = \frac{\int_B \mathbb{P}^\Lambda(\mathbf{n})P(d\Lambda)}{\int_{\mathcal{D}_\eta^b}\mathbb{P}^\Lambda(\mathbf{n})P(d\Lambda)}$$

for measurable $B \in \mathcal{B}(\mathcal{D}_\eta^b)$. The posterior is said to be *consistent* if $P(U_{\Lambda_0}^c|\mathbf{n}) \to 0$ as $n \to \infty$, where U_{Λ_0} is any open neighbourhood of the data-generating $\Lambda_0 \in \mathcal{D}_\eta^b$. Intuitively this corresponds to it being possible to learn the true Λ_0 from a sufficiently large data set.

Likelihood evaluation: Ancestry as missing data

Let A denote an ancestral tree of the sample $\mathbf{n}$ annotated with mutations. The likelihood can be written as

$$\mathbb{P}^\Lambda(\mathbf{n}) = \sum_{A_0,\dots,A_N} \mathbb{P}^\Lambda(\mathbf{n}|A)\prod_{i=0}^{-N+1}\mathbb{P}^\Lambda(A_i|A_{i-1})\mathbb{P}^\Lambda(A_{-N}), \qquad (1)$$

where $\mathbb{P}^\Lambda(\mathbf{n}|A) = 1$ if the leaves of A are compatible with $\mathbf{n}$ and 0 otherwise, N is the number of transitions required to reach the MRCA, A_0 denotes the leaves of the tree, A_{-N} is the root and the other A_i's are intermediate states separated by mutations or coalescences. For example, in Figure 1 we have $N = 5$, $A_0 = (G,P,P,R)$, $A_{-1} = (P,P,P,R)$, $A_{-2} = (P,R)$, $A_{-3} = (B,R)$, $A_{-4} = (B,B)$ and $A_{-5} = (B)$, where B, G, P and R stand for blue, green, purple and red. While the summations in (1) are over too many terms to evaluate, the conditional distributions $\mathbb{P}^\Lambda(A_i|A_{i-1})$ can be obtained explicitly and the resulting system of equations can be shown to depend only on the first $n-2$ moments of Λ, when n lineages have been observed [KJS15a]. Thus the likelihood is constant within equivalence classes of measures which agree on their first $n-2$ moments, and we can use these moments to parametrise the inference problem with no loss of signal. It is also possible to make the $\|\cdot\|_\infty$-distance between the moment sequence of any $\Lambda \in \mathcal{M}_1([0,1])$ and some $\Lambda' \in \mathcal{D}_\eta^b$ arbitrarily small by choosing η sufficiently small, so that restricting attention to measure $\Lambda \in \mathcal{D}_\eta^b$ is not as restrictive as it first appears. Recursion (1) can also be used to design efficient importance sampling algorithms to approximate the intractable likelihood in practice [BBS11, KJS15b].

Simultaneous observations: inconsistency

As before let $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ denote a discrete set of observations, and let $\mathbf{x} := \lim_{n\to\infty}\frac{\mathbf{n}}{n}$ denote the limiting observed allele frequencies. If all observations are simultaneous then the posterior is inconsistent, and in fact

$$\lim_{n\to\infty} P(B|\mathbf{n}) = \frac{\int_B \pi^\Lambda(\mathbf{x})P(d\Lambda)}{\int_{\mathcal{D}_\eta^b}\pi^\Lambda(\mathbf{x})P(d\Lambda)}$$

so that the support of the limiting posterior coincides with that of the prior [KJS15a]. It is easy to see why this is: by duality the sample $\mathbf{n}$ can be viewed as a discrete sample from a stationary Λ-Fleming-Viot process $(\mathbf{X}_t)_{t\geq 0}$. As $n \to \infty$ the normalised allele frequencies converge to a single draw from the law of $\mathbf{X}_0$, which is π^Λ by stationarity.

Temporally structured observations: consistency

Suppose now that the data sets of size n are available from $m+1$ distinct time points with separation $\Delta > 0$. Let $(\mathbf{n}_0,\dots,\mathbf{n}_m) \in \mathbb{N}^{|\mathcal{H}|\times(m+1)}$ denote the observed allele counts, and let $(\mathbf{x}_0,\dots,\mathbf{x}_m)$ denote the corresponding, limiting allele frequencies. Fixing m and letting $n \to \infty$ yields the limiting posterior

$$\lim_{n\to\infty} P(B|\mathbf{n}_0,\dots,\mathbf{n}_m) = \frac{\int_B \pi^\Lambda(\mathbf{x}_0)\prod_{i=1}^m p_\Delta^\Lambda(\mathbf{x}_{i-1},\mathbf{x}_i)P(d\Lambda)}{\int_{\mathcal{D}_\eta^b}\pi^\Lambda(\mathbf{x}_0)\prod_{i=1}^m p_\Delta^\Lambda(\mathbf{x}_{i-1},\mathbf{x}_i)P(d\Lambda)}, \qquad (2)$$

where $p_\Delta^\Lambda(\mathbf{x},\mathbf{y})$ is the transition density of the Λ-Fleming-Viot process with time step Δ. It is clear from (2) that posterior consistency of the discretely observed Λ-Fleming-Viot process implies posterior consistency of P on $\mathcal{D}_\eta^b$ as $n, m \to \infty$. This can be verified [KJS15a], and so posterior consistency holds for time series data. Note that sequences of finitely many moments can be written as bounded functionals of Λ, and hence posterior consistency of moments is inherited.

A pseudo-marginal Metropolis-Hastings algorithm

As outlined above, the signal a data set of size n carries about the data generating measure $\Lambda_0 \in \mathcal{D}_\eta^b$ can be captured fully by computing the posterior distribution of the first $n-2$ moments. This posterior can be sampled by using the pseudo-marginal Metropolis-Hastings algorithm [AR09], which inherits the efficient exploration of high-dimensional parameter space from the standard Metropolis-Hastings algorithm and replaces likelihood evaluations with unbiased estimators. Such estimators are readily available for the Λ-coalescent [BBS11, KJS15b], so that the algorithm is implementable. A pseudo-code specification is provided below, with $\boldsymbol{\lambda}$ denoting a moment sequence of length n, $\hat{L}(\boldsymbol{\lambda};\mathbf{n})$ denoting a generic likelihood estimator and K an irreducible transition kernel between moment sequences.

1: Initialise $Y_0 \leftarrow (\boldsymbol{\lambda}, \hat{L}(\boldsymbol{\lambda};\mathbf{n}))$.
2: **for** $j = 1,\dots,N$ **do**
3: Sample $\boldsymbol{\lambda}' \sim K(\boldsymbol{\lambda},\cdot)$, $U \sim U(0,1)$.
4: Compute $\hat{L}(\boldsymbol{\lambda}';\mathbf{n})$.
5: Set $\alpha \leftarrow 1 \wedge \frac{K(\boldsymbol{\lambda}',\boldsymbol{\lambda})\hat{L}(\boldsymbol{\lambda}';\mathbf{n})P(\boldsymbol{\lambda}')}{K(\boldsymbol{\lambda},\boldsymbol{\lambda}')\hat{L}(\boldsymbol{\lambda};\mathbf{n})P(\boldsymbol{\lambda})}$.
6: **if** $U < \alpha$ **then**
7: Set $Y_j \leftarrow (\boldsymbol{\lambda}', \hat{L}(\boldsymbol{\lambda}';\mathbf{n}))$.
8: **else**
9: Set $Y_j \leftarrow Y_{j-1}$.
10: **end if**
11: **end for**

Robust bounds for functionals of Λ

Consider a credible region C specified by k functions $\{f_j\}_{j=1}^k$ and constants $\{c_j\}_{j=1}^k$ obtained from a sample of moment sequences from the posterior as $f_j(\lambda_{3,3},\dots,\lambda_{n,n}) \leq c_j$ for $j \in \{1,\dots,k\}$. Then $\int_\eta^1 q(r)\Lambda(dr)$ can be bounded on $\{\Lambda : \{\lambda_{3,3},\dots,\lambda_{n,n}\} \in C\}$ by setting C_k as the subspace of C consisting of discrete measures with at most k atoms:

$$C_k = \left\{\Lambda \in C : \Lambda = \sum_{i=1}^p w_i\delta_{r_i} \text{ where } 1 \leq p \leq k, w_i \geq 0 \text{ and } r_i \in [\eta,1]\right\}.$$

Then

$$\inf_{\Lambda\in C}\int_{\eta,1} q(r)\Lambda(dr) = \inf_{\Lambda\in C_k}\int_{\eta,1} q(r)\Lambda(dr)$$
$$\sup_{\Lambda\in C}\int_{\eta,1} q(r)\Lambda(dr) = \sup_{\Lambda\in C_k}\int_{\eta,1} q(r)\Lambda(dr)$$

by a result of [Win88], and the optimisation problems on the R.H.S. involve a finite number of locations and weights. Hence they can be solved numerically, yielding robust bounds.

Example: The Dirichlet process mixture model prior

Let $g_\tau(r) := \frac{\phi_\tau(r)\mathbb{1}_{[\eta,1]}(r)}{\phi_\tau([\eta,1])}$, where ϕ_τ is the Gaussian density with mean 0 and precision τ. Let DP(α) denote the law of the Dirichlet process with mean $\alpha \in \mathcal{M}_f([\eta,1])$ and let $F \in \mathcal{M}_1((0,\infty))$ assign positive probability to all open sets. The Dirichlet process mixture model is a prior on strictly positive, bounded densities on $[\eta,1]$ sampled as follows. Let $Q \sim \text{DP}(\alpha)$ with size-ordered atoms $\{z_i\}_{i\in\mathbb{N}}$, and $\{\tau_i\}_{i\in\mathbb{N}} \sim F$ be i.i.d. samples. Then $\sum_{i=1}^\infty Q(z_i)g_{\tau_i}(z-z_i)$ is a draw from the Dirichlet process mixture model prior, whose support is dense [BD12].

Acknowledgements and References

Jere Koskela is supported by EPSRC as part of the MASDOC DTC at the University of Warwick. Grant No. EP/HO23364/1. Paul Jenkins is supported in part by EPSRC grant EP/L018497/1.

[Á04] E. Árnason. Mitochondrial cytochrome b DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885, 2004.

[AR09] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.*, 37(2):697–725, 2009.

[BB08] M. Birkner and J. Blath. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.*, 57(3):435–463, 2008.

[BBB94] J. D. G. Boom, E. G. Boulding, and A. T. Beckenback. Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. *Can. J. Fish. Aquat. Sci.*, 51:1608–1614, 1994.

[BBS11] M. Birkner, J. Blath, and M. Steinrücken. Importance sampling for Lambda–coalescents in the infinitely many sites model. *Theor. Popln Biol.*, 79(4):155–173, 2011.

[BD12] A. Bhattacharya and D. B. Dunson. Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds. *Ann. Inst. Stat. Math.*, 64:687–714, 2012.

[BLG03] J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probab. Theory Related Fields*, 126:261–288, 2003.

[DK99] P. Donnelly and T. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205, 1999.

[EW06] B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633, 2006.

[KJS15a] J. Koskela, P. A. Jenkins, and D. Spanò. Bayesian nonparametric inference for Λ-coalescents: posterior consistency and a parametric method. *In preparation*, 2015.

[KJS15b] J. Koskela, P. A. Jenkins, and D. Spanò. Computational inference beyond Kingman's coalescent. *J. Appl. Probab.*, 52(2), 2015.

[Pit99] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902, 1999.

[Sag99] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lineages. *J. Appl. Probab.*, 36(4):1116–1125, 1999.

[SBB13] M. Steinrücken, M. Birkner, and J. Blath. Analysis of DNA sequence variation within marine species using Beta–coalescents. *Theor. Popln Biol.*, 87:15–24, 2013.

[Sch03] J. Schweinsberg. Coalescent processes obtained from super-critical Galton-Watson processes. *Stoch. Proc. Appl.*, 106:107–139, 2003.

[Win88] G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 30(4):581–587, 1988.