Lossless Bayesian inference in infinite dimension without discretisation or truncation: a case study on Λ-coalescents

Jere Koskela
Joint work with Paul A. Jenkins and Dario Spanò

International conference on Monte Carlo techniques
5-8 July 2016, Paris

Mathematics and Statistics Centre for Doctoral Training

THE UNIVERSITY OF WARWICK

EPSRC
Engineering and Physical Sciences Research Council

Outline

Exposition: Likelihood-informed subspaces

The finite alleles Λ-coalescent

Projection onto moments

Consistency

Sampling posterior moments

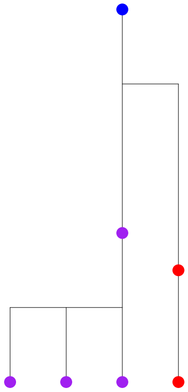
Likelihood-informed subsaces

- Consider inferring an unknown function $f \in \mathcal{C}$ from samples $\mathbf{n} := \{x_i, f(x_i)\}_{i=1}^{n}$.
- Choose a Gaussian prior $\mu \in \mathcal{M}_1(\mathcal{C})$ and sample $\mu(df|\mathbf{n})$ using MCMC.
- Speed up mixing (but lose some signal) by choosing a finite-dimensional subspace $\mathcal{C}_d$, computing the push-forward μ_d and sampling $\mu_d(df|\mathbf{n})$.
- Also yields an easily implementable algorithm.

In this talk:

- An example inference problem (the Λ-coalescent) for which the mapping $\mathcal{C} \mapsto \mathcal{C}_d$ is *lossless*, μ_d can be computed explicitly and (some of) the "residual" uncertainty between $\mu_d(df|\mathbf{n})$ and $\mu(df|\mathbf{n})$ can be controlled.

The finite alleles Λ-coalescent

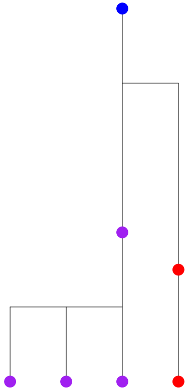


- In reverse time, each $k \leq n$ lineages merges at rate

$$\lambda_{n,k} := \int_{[0,1]} r^{k-2}(1-r)^{n-k}\Lambda(dr).$$

- Each lineage mutates with rate θ.
- Sample type of most recent common ancestor.
- Mutations resolved forwards in time through stochastic matrix M.

The inference problem



- Data: a vector of observed type frequencies $\mathbf{n} \in \mathbb{N}^d$.
- Missing data: the ancestral tree and mutation events.
- The likelihood

$$\mathbb{P}_{\Lambda,\theta,M}(\mathbf{n}) = \int_{\mathcal{A}} \mathbb{1}_{\{\mathbf{n}\}}(A_0)\mathbb{P}_{\Lambda,\theta,M}(dA)$$

  has no known closed form expression.
- (Relatively) efficient importance sampling algorithms are available for pointwise evaluation.
- Standing assumption: M and θ are known.

Proposition 1

Let genetic labels be identified with $\{1, \ldots, d\}$ and let $\mathbf{n} = (n_1, \ldots, n_d)$ denote the observed type frequencies. The likelihood $\mathbb{P}_\Lambda(\mathbf{n})$ is constant across any measures Λ which share the first $n-2$ moments.

Proof. The likelihood solves

$$\mathbb{P}_\Lambda(\mathbf{n}) = \frac{\theta}{n\theta - q_{nn}} \sum_{i,j=1}^{d} (n_j - 1 + \delta_{ij}) M_{ji} \mathbb{P}_\Lambda(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j)$$

$$+ \frac{1}{n\theta - q_{nn}} \sum_{i:n_i \geq 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i - k + 1}{n - k + 1} \mathbb{P}_\Lambda(\mathbf{n} - (k-1)\mathbf{e}_i).$$

with boundary condition $\mathbb{P}_\Lambda(\mathbf{e}_i) = m(i)$, where m is the unique M-invariant distribution on $\{1, \ldots, d\}$.

Parametrisation

- Let $\sim_n$ denote the equivalence relation on Λ's of agreement of first $n-2$ moments.
- Let $\mu \in \mathcal{M}_1(\mathcal{M}_1([0,1]))$ denote a prior. Proposition 1 implies $\mu(d\Lambda | \sim_n) = \mu(d\Lambda)|_{\sim_n}$.
- This suggests parametrising an inference problem with n observations with $n-2$ moments.
- Procedure can be interpreted as analytically integrating "$\infty - (n-2)$" dimensions, and leaving $n-2$ to sample (Rao-Blackwellisation)...
- ...provided a suitable prior can be found.

The Dirichlet process mixture model

- $\{z_i\}_{i=1}^{\infty} \overset{\text{i.i.d}}{\sim} H$.
- $\{\beta'_i\}_{i=1}^{\infty} \overset{\text{i.i.d}}{\sim} \mathrm{Beta}(1, \alpha)$.
- $\beta_i := \prod_{j=1}^{i-1}(1 - \beta'_j)\beta'_j$.
- $\{\sigma_i\}_{i=1}^{\infty} \overset{\text{i.i.d}}{\sim} F$.
- $\Lambda(r) = \sum_{i=1}^{\infty} \beta_i \phi(\sigma_i^{-1}(r - z_i))$, where ϕ is the standard Gaussian density conditioned on $[\eta, 1]$ for any $\eta > 0$.
- Easy (and exponentially accurate) to truncate, or...

Moments of the Dirichlet process mixture model

Let $C_0, \ldots, C_n \in \mathbb{R}^{n+1}$ solve

$$C_n = -1,$$
$$\sum_{k=0}^{n-r-1} \binom{n-r}{k} C_{r+k} = 1 \text{ for } r \in \{0, \ldots, n-1\}.$$

Then

$$(-1)^{n+1} 2^n F_n(\boldsymbol{\sigma}, \mathbf{g}_n, \alpha) = C_0 + \sum_{k=1}^{n} \frac{C_k}{(\pi i)^k} \times$$
$$\times \sum_{1 \le j_1 < \ldots < j_k \le n} \int_0^\infty \cdots \int_0^\infty \frac{h_k(\mathbf{s}_k; g_{j_1} - \sigma_{j_1}, \ldots, g_{j_k} - \sigma_{j_k}; \alpha)}{s_1 \times \cdots \times s_k} d\mathbf{s}_k,$$

where h_k is the characteristic function of a γ_α-random measure and F_n is the joint distribution of n moments $\mu(g_1), \ldots, \mu(g_n)$.

Proposition 2

If the observed allele frequencies come from a bounded number of time points, then the posterior is always inconsistent.

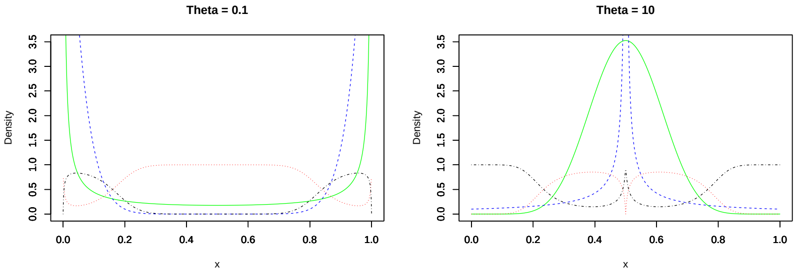


Figure 1 : $\mu = \frac{1}{2}(\delta_{\delta_0} + \delta_{\delta_1})$. Two types. Single-time sampling distributions of the $\lim_{n\to\infty}$ type fractions in blue and green, corresponding posterior probabilities in black and red. At $\theta = 1$ everything is uniform.

Proposition 3

Let $\Delta > 0$ be a fixed sampling interval, and let $\mathbf{n} := (\mathbf{n}_1, \ldots, \mathbf{n}_k)$ denote samples of size n sampled at times $\{\Delta j\}_{j=0}^{k-1}$. Suppose the prior μ places full mass on a $\mathcal{D}_\eta$, set of strictly positive, bounded densities on $[\eta, 1]$ for some $\eta > 0$, and for any $\varepsilon > 0$ and $\phi_0 \in \mathcal{D}_\eta$ suppose that

$$\mu\left(\phi \in \mathcal{D}_\eta : \int_\eta^1 \left\{\left|\log\left(\frac{\phi_0(r)}{\phi(r)}\right)\right| + \left|\frac{\phi_0(r)}{\phi(r)} - 1\right|\right\} r^{-2}\phi_0(r)dr < \varepsilon\right) > 0.$$

Then the posterior is consistent as both n and $k \to \infty$.

Consistency of a finite number of moments follows immediately since $\phi \mapsto \int_\eta^1 r^j \phi(r)dr$ is continuous and bounded.

Pseudo-marginal MCMC

Algorithm 1 The pseudo-marginal algorithm

Require: Prior $P(x)$, unbiased likelihood estimator $L(x)$, transition kernel $q(x, y)$, and run length n.

```
Initialise X_0 = x and L_0 = L(x).
for i = 1, ..., n do
    Sample y ~ q(x, ·) and L = L(y).
    Set a = 1 ∧ q(y,x) L P(y) / (q(x,y) L_{i-1} P(x)) and sample u ~ U(0,1).
    if u < a then
        Set X_i = y and L_i = L.
    else
        Set X_i = X_{i-1} and L_i = L_{i-1}.
    end if
end for
return X
```

Initialise $X_0 = x$ and $L_0 = L(x)$.

for i = 1, . . . , n **do**

  Sample $y \sim q(x, \cdot)$ and $L = L(y)$.

  Set $a = 1 \wedge \frac{q(y,x)LP(y)}{q(x,y)L_{i-1}P(x)}$ and sample $u \sim U(0, 1)$.

  if $u < a$ **then**

    Set $X_i = y$ and $L_i = L$.

  else

    Set $X_i = X_{i-1}$ and $L_i = L_{i-1}$.

  end if

end for

return X

Algorithm 2 The noisy pseudo-marginal algorithm

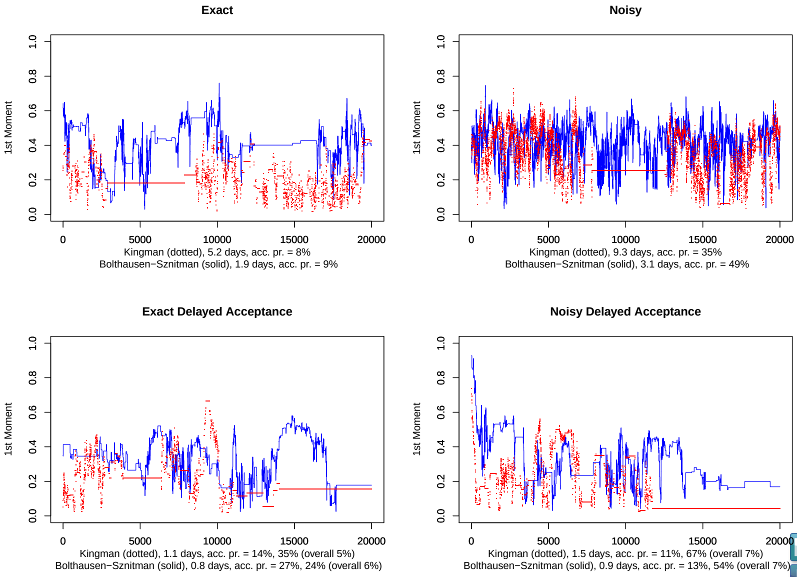
Require: Prior $P(x)$, unbiased likelihood estimator $L(x)$, transition kernel $q(x, y)$, and run length n.

Initialise $X_0 = x$ and $L_0 = L(x)$.
for i = 1, . . . , n **do**
  Sample $y \sim q(x, \cdot)$ and $L = L(y)$.
  Sample $L' = L(x)$.
  Set $a = 1 \wedge \frac{q(y,x)LP(y)}{q(x,y)L'P(x)}$ and sample $u \sim U(0, 1)$.
  if $u < a$ **then**
    Set $X_i = y$ and $L_i = L$.
  else
    Set $X_i = X_{i-1}$ and $L_i = L'$.
  end if
end for
return X

Simulation study: set up

- Prior on Λ: truncated Dirichlet process mixture with 4 components and $\eta = 10^{-6}$.
- Quantity of interest: $\lambda_{3,3}$, the first moment of Λ.
- Two simulated data sets of 5×20 individuals each, with $d = 2^{15}$:
  - Kingman coalescent: $\Lambda = \delta_0$, $\lambda_{3,3} = 0$.
  - Bolthausen-Sznitman coalescent: $\Lambda = U(0,1)$, $\lambda_{3,3} = 0.5$.
- Gaussian random walk Metropolis-Hastings proposal (with conditioning for boundaries).
- Likelihood estimator uses 180 and 75 particles, respectively.

Simulation study: short runs



Simulation study: long runs