An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics

Shiwei Lan [1*], Julia A. Palacios [2,3], Michael Karcher [4], Vladimir N. Minin [4,5] and Babak Shahbaba [6*]

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL.
[2] Department of Ecology and Evolutionary Biology, Brown University.
[3] Department of Organismic and Evolutionary Biology, Harvard University.
[4] Department of Statistics, University of Washington.
[5] Department of Biology, University of Washington.
[6] Department of Statistics, University of California, Irvine.

ABSTRACT

Phylodynamics focuses on the problem of reconstructing past population size dynamics from current genetic samples taken from the population of interest. This technique has been extensively used in many areas of biology, but is particularly useful for studying the spread of quickly evolving infectious diseases agents, e.g., influenza virus. Phylodynamics inference uses a coalescent model that defines a probability density for the genealogy of randomly sampled individuals from the population. When we assume that such a genealogy is known, the coalescent model, equipped with a Gaussian process prior on population size trajectory, allows for nonparametric Bayesian estimation of population size dynamics. While this approach is quite powerful, large data sets collected during infectious disease surveillance challenge the state-of-the-art of Bayesian phylodynamics and demand computationally more efficient inference framework. To satisfy this demand, we provide a computationally efficient Bayesian inference framework based on Hamiltonian Monte Carlo for coalescent process models. Moreover, we show that by splitting the Hamiltonian function we can further improve the efficiency of this approach. Using several simulated and real datasets, we show that our method provides accurate estimates of population size dynamics and is substantially faster than alternative methods based on elliptical slice sampler and Metropolis-adjusted Langevin algorithm.

1 INTRODUCTION

Population genetics theory states that changes in population size affect genetic diversity, leaving a trace of these changes in individuals' genomes. The field of *phylodynamics* relies on this theory to reconstruct past population size dynamics from current genetic data. In recent years, phylodynamic inference has become an essential tool in areas like ecology and epidemiology. For example, a study of human influenza A virus from sequences sampled in both hemispheres pointed to a source-sink dynamics of the influenza evolution [Rambaut et al., 2008].

Phylodynamics connects population dynamics and genetic data using coalescent-based methods [Griffiths and Tavare, 1994, Kuhner et al., 1998, Drummond et al., 2002, Strimmer and Pybus, 2001, Drummond et al., 2005, Opgen-Rhein et al., 2005, Heled and Drummond, 2008, Minin et al., 2008, Palacios and Minin, 2013]. Typically, phylodynamics relies on Kingman's coalescent model, which is a probability model that describes formation of genealogical relationships of a random sample of molecular sequences. The coalescent model is parameterized in terms of the *effective population size*, an indicator of genetic diversity [Kingman, 1982].

While recent studies have shown promising results in alleviating computational difficulties of phylodynamic inference [Palacios and Minin, 2012, 2013], existing methods still lack the level of computational efficiency required to realize the full potential of phylodynamics: *developing surveillance programs that can operate similarly to weather monitoring stations allowing public health workers to predict disease dynamics in order to optimally allocate limited resources in time and space*. To achieve this goal, we present an accurate and computationally efficient inference method for modeling population dynamics given a genealogy. More specifically, we concentrate on a class of Bayesian nonparametric methods based on Gaussian processes [Minin et al., 2008, Gill et al., 2013, Palacios and Minin, 2013]. Following Palacios and Minin [2012] and Gill et al. [2013], we assume a log-Gaussian process prior on the effective population size. As a result, the estimation of effective population size trajectory becomes similar to the estimation of intensity of a log-Gaussian Cox process [LGCP; Møller et al., 1998], which is extremely challenging since the likelihood evaluation becomes intractable: it involves integration over an infinite-dimensional random function. We resolve the intractability in likelihood evaluation by discretizing the integration interval with a regular grid to approximate the likelihood and the corresponding score function.

For phylodynamic inference, we propose a computationally efficient Markov chain Monte Carlo (MCMC) algorithm using Hamiltonian Monte Carlo [HMC; Duane et al., 1987, Neal, 2010] and one of its variation, called Split HMC [Leimkuhler and Reich, 2004, Neal, 2010, Shahbaba et al., 2013], which speeds

---

*to whom correspondence should be addressed

up standard HMC's convergence. Our proposed algorithm has several advantages. First, it updates all model parameters jointly to avoid poor MCMC convergence and slow mixing rates when there are strong dependencies among model parameters [Knorr-Held and Rue, 2002]. Second, unlike a recently proposed Integrated Nested Laplace Approximation [INLA Rue et al., 2009, Palacios and Minin, 2012] method, our approach can be extended to a more realistic setting where instead of a genealogy of the sampled individuals, we observe their genetic data that indirectly inform us about genealogical relationships. Third, we show that our method is up to an order of magnitude more efficient than MCMC algorithms, such as Metropolis-Adjusted Langevin algorithm [MALA; Roberts and Tweedie, 1996], adaptive MALA [aMALA; Knorr-Held and Rue, 2002], and Elliptical Slice Sampler [ES^2; Murray et al., 2010], frequently used in phylodynamics. Finally, although in this paper we focus on phylodynamic studies, our proposed methodology can be easily applied to more general point process models.

The remainder of the paper is organized as follows. In Section 2, we provide a brief overview of the coalescent model and HMC algorithms. Section 3 presents the details of our proposed sampling methods. Experimental results based on simulated and real data are provided in Section 4. Section 5 is devoted to discussion and future directions.

2 PRELIMINARIES

2.1 Coalescent

Assume that a genealogy with time measured in units of generations is available. The coalescent model allows us to trace the ancestry of a random sample of n genomic sequences as tree: two sequences or lineages merge into a common ancestor as we go back in time. Those "merging" times are called *coalescent times*. The coalescent with variable population size can be viewed as an inhomogeneous Markov death process that starts with n lineages at present time, $t_n = 0$, and decreases by one at each of the consequent coalescent times, $t_{n-1} < \cdots < t_1$, until reaching their most recent common ancestor [Griffiths and Tavare, 1994].

Suppose we observe a genealogy of n individuals sampled at time 0. Under the standard (*isochronous*) coalescent model, the coalescent times $t_n = 0 < t_{n-1} < \cdots < t_1$, conditioned on the *effective population size trajectory*, $N_e(t)$, have the density

$$P[t_1, \ldots, t_n \mid N_e(t)] = \prod_{k=2}^{n} P[t_{k-1} \mid t_k, N_e(t)]$$
$$= \prod_{k=2}^{n} \frac{C_k}{N_e(t_{k-1})} \exp\left\{-\int_{I_k} \frac{C_k}{N_e(t)} dt\right\}, \tag{1}$$

where $C_k = \binom{k}{2}$ and $I_k = (t_k, t_{k-1}]$. Note that the larger the population size, the longer it takes for two lineages to coalesce. Further, the larger the number of lineages, the faster two of them meet their common ancestor [Palacios and Minin, 2012, 2013].

For rapidly evolving organisms, we may have different sampling times. When this is the case, the standard coalescent model can be generalized to account for such *heterochronous* sampling [Rodrigo and Felsenstein, 1999]. Under the heterochronous coalescent the number of lineages change at both coalescent times and sampling times. Let $\{t_k\}_{k=1}^{n}$ denote the coalescent times as before, but now
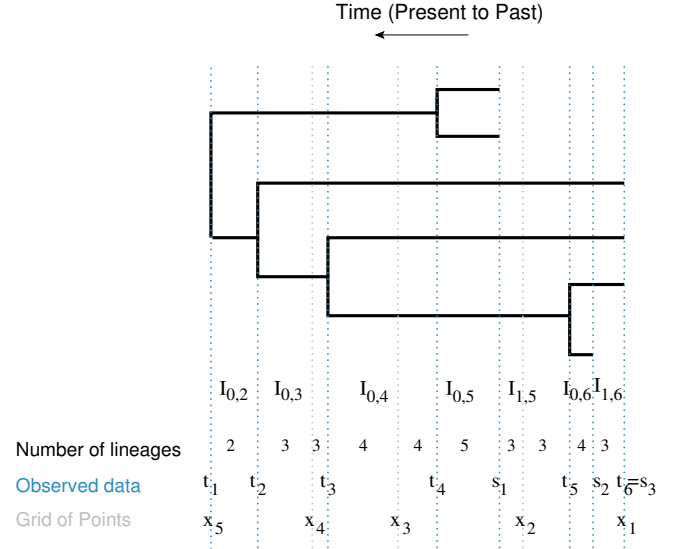
Fig. 1. A genealogy with coalescent times and sampling times. Blue dashed lines indicate the observed times: coalescent times $\{t_1, \cdots, t_6\}$ and sampling times $\{s_1, s_2, s_3\}$. The intervals where the number of lineages change are denoted by $I_{i,k}$. The superimposed grid $\{x_1, \cdots, x_5\}$ is marked by gray dashed lines. We count the number of lineages in each interval defined by grid points, coalescent times and sampling times.

let $s_m = 0 < s_{m-1} < \cdots < s_1$ denote sampling times of $n_m, \ldots, n_1$ sequences respectively, where $\sum_{j=1}^{m} n_j = n$. Further, let $\mathbf{s}$ and $\mathbf{n}$ denote the vectors of sampling times $\{s_j\}_{j=1}^{m}$ and numbers of sequences $\{n_j\}_{j=1}^{m}$ sampled at these times, respectively. Then the coalescent likelihood of a single genealogy becomes

$$P[t_1, \ldots, t_n \mid \mathbf{s}, \mathbf{n}, N_e(t)] =$$
$$\prod_{k=2}^{n} \frac{C_{0,k} \exp\left\{-\int_{I_{0,k}} \frac{C_{0,k}}{N_e(t)} dt - \sum_{i \geq 1} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt\right\}}{N_e(t_{k-1})}, \tag{2}$$

where the coalescent factor $C_{i,k} = \binom{l_{i,k}}{2}$ depends on the number of lineages $l_{i,k}$ in the interval $I_{i,k}$ defined by coalescent times and sampling times. For $k = 2, \ldots, n$, we denote half-open intervals that end with a coalescent event by

$$I_{0,k} = (\max\{t_k, s_j\}, t_{k-1}], \tag{3}$$

for $s_j < t_{k-1}$, and half-open intervals that end with a sampling event by ($i > 0$)

$$I_{i,k} = (\max\{t_k, s_{j+i}\}, s_{j+i-1}], \tag{4}$$

for $t_k < s_{j+i-1} \leq s_j < t_{k-1}$. In density (2), there are $n - 1$ intervals $\{I_{i,k}\}_{i=0}$ and $m - 1$ intervals $\{I_{i,k}\}_{i>0}$ for all i, k. Note that only those intervals satisfying $I_{i,k} \subset (t_k, t_{k-1}]$ are non-empty. See Figure 1 for more details.

We can think of isochronous coalescence as a special case of heterochronous coalescence when $m = 1, C_{0,k} = C_k, I_{0,k} = I_k, I_{i,k} = \emptyset$ for $i > 0$. Therefore, in what follows, we refer to Equation (2) as the general case.

We assume the following log-Gaussian Process prior on the effective population size, $N_e(t)$:

$$N_e(t) = \exp[f(t)], \quad f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})), \quad (5)$$

where $\mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$ denotes a Gaussian process with mean function $\mathbf{0}$ and covariance function $\mathbf{C}(\boldsymbol{\theta})$. *A priori*, $N_e(t)$ is a log-Gaussian Process.

For computational convenience, we use Brownian motion, which is a special case of GP, as our prior for $f(t)$. We define the covariance function as $\mathbf{C}(\kappa) = \frac{1}{\kappa} \mathbf{C}_{BM}$, where the precision parameter κ has a $\mathrm{Gamma}(\alpha, \beta)$ prior. For ascending times $0 = x_0 < x_1 < x_2 < \cdots < x_{D-1}$, the (i, j)-th element of $\mathbf{C}_{BM}$ is set to $\min\{x_i, x_j\}$. This way, we reduce the computational complexity of inverting the covariance matrix from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$ since the inverse covariance matrix is tri-diagonal [Rue and Held, 2005, Palacios and Minin, 2013] with elements

$$\mathbf{C}_{BM}^{-1}(i, j) = \begin{cases} \dfrac{1_{\{i+1 \leq D-1\}}}{x_{i+1} - x_i} + \dfrac{1}{x_i - x_{i-1}}, & \text{if } i = j, \\[2mm] -\dfrac{1}{|x_i - x_j|}, & \text{if } |i - j| = 1, \\[2mm] 0, & \text{otherwise.} \end{cases}$$

In practice we modify the $(1, 1)$ element, $1/(x_2 - x_1) + 1/x_1$, to be $1/(x_2 - x_1)$ and denote the resulting precision matrix as $\mathbf{C}_{in}^{-1}$ to indicate that it comes from an intrinsic autoregression [Besag and Kooperberg, 1995, Knorr-Held and Rue, 2002]. Note that $\mathbf{C}_{in}^{-1}$ is degenerate in 1 eigen direction so we add a small number, e, to the diagonal elements of $\mathbf{C}_{in}^{-1}$ to make it invertible when $\mathbf{C}_{in}$ is needed.

2.2 HMC

Hamiltonian Monte Carlo [Duane et al., 1987, Neal, 2010] is a Metropolis sampling algorithm that suppresses the random walk behavior by proposing states that are distant from the current state, but nevertheless accepts new proposals with high probability. These distant proposals are found by numerically simulating Hamilton dynamics, whose state space consists of *position*, denoted by the vector $\boldsymbol{\theta}$, and *momentum*, denoted by the vector $\mathbf{p}$. The objective is to sample from the continuous probability distribution with the density function $\pi(\boldsymbol{\theta})$. It is common to assume $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is a symmetric, positive-definite matrix known as the *mass matrix*, often set to the identity matrix $\mathbf{I}$ for convenience.

In this simulation of Hamiltonian dynamics, the *potential energy*, $U(\boldsymbol{\theta})$, is defined as the negative log density of $\boldsymbol{\theta}$ (plus any constant); the *kinetic energy*, $K(\mathbf{p})$ for momentum variable $\mathbf{p}$, is set to be the negative log density of $\mathbf{p}$ (plus any constant). Then the total energy of the system, the *Hamiltonian* function, is defined as their sum: $H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p})$. Then the system of $(\boldsymbol{\theta}, \mathbf{p})$ evolves according to the following set of *Hamilton's equations*:

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) = \mathbf{M}^{-1} \mathbf{p}, \\ \dot{\mathbf{p}} &= -\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}). \end{aligned} \quad (6)$$

In practice, we use a numerical method called *leapfrog* to approximate the Hamilton's equations [Neal, 2010] when the analytical solution is not available. We numerically solve the system for L steps, with some step size, ε, to propose a new state in

the Metropolis algorithm, and accept or reject it according to the Metropolis acceptance probability. [See Neal, 2010, for more discussions].

3 METHOD

3.1 Discretization

As discussed above, the likelihood function (2) is intractable in general. We can, however, approximate the likelihood using discretization. To this end, we use a fine regular grid, $\mathbf{x} = \{x_d\}_{d=1}^{D}$, over the observation window and approximate $N_e(t)$ by a piece-wise constant function as follows:

$$N_e(t) \approx \sum_{d=1}^{D-1} \exp[f(x_d^*)] 1_{t \in (x_d, x_{d+1}]}, \quad x_d^* = \frac{x_d + x_{d+1}}{2}. \quad (7)$$

Note that the regular grid $\mathbf{x}$ does not coincide with the sampling coalescent times, except for the first sampling time $s_m = x_1$ and the last coalescent time $t_1 = x_D$. To rewrite (2) using the approximation (7), we sort all the time points $\{\mathbf{t}, \mathbf{s}, \mathbf{x}\}$ to create new $D + m + n - 4$ half-open intervals $\{I_\alpha^*\}$ with either coalescent time points, sampling time points, or grid time points as the end points (See Figure 1).

For each $\alpha \in \{1, \cdots, D + m + n - 4\}$, there exists some i, k and d such that $I_\alpha^* = I_{i,k} \cap (x_d, x_{d+1})$. Each integral in density (2) can be approximated as a sum:

$$\int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \approx \sum_{I_\alpha^* \subset I_{i,k}} C_{i,k} \exp[-f(x_d^*)] \Delta_d,$$

where $\Delta_d := x_{d+1} - x_d$. This way, the likelihood of coalescent times (2) can be rewritten as a product of the following terms:

$$\left\{ \frac{C_{i,k}}{\exp[f(x_d^*)]} \right\}^{y_\alpha} \exp \left\{ -\frac{C_{i,k} \Delta_d}{\exp[f(x_d^*)]} \right\}, \quad (8)$$

where y_α is set to 1 if I_α^* ends with a coalescent time, and to 0 otherwise. This happens to be proportional to the probability mass of a Poisson random variable y_α with intensity $\lambda_\alpha := C_{i,k} \Delta_d \exp[-f(x_d^*)]$. Therefore, the likelihood of coalescent times (2) can be approximated as follows:

$$P[t_1, \ldots, t_n \mid N_e(t)] \approx \prod_{\alpha=1}^{D+m+n-4} P[y_\alpha \mid N_e(t)]$$

$$= \prod_{d=1}^{D-1} \prod_{I_\alpha^* \subset (x_d, x_{d+1})} \left\{ \frac{C_{i,k}}{\exp[f(x_d^*)]} \right\}^{y_\alpha} \exp \left\{ -\frac{C_{i,k} \Delta_d}{\exp[f(x_d^*)]} \right\}, \quad (9)$$

where the coalescent factor $C_{i,k}$ on each interval I_α^* is determined by the number of lineages $l_{i,k}$ in I_α^*.

3.2 Sampling methods

Denote $\mathbf{f} := \{f(x_d^*)\}_{d=1}^{D-1}$. Our model can be summarized as

$$y_\alpha | \mathbf{f} \sim \mathrm{Poisson}[\lambda_\alpha(\mathbf{f})],$$

$$\mathbf{f} | \kappa \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\kappa} \mathbf{C}_{in}\right), \quad (10)$$

$$\kappa \sim \mathrm{Gamma}(\alpha, \beta).$$

Transforming the coalescent times, sampling times and grid points into $\{y_\alpha, C_{i,k}, \Delta_d\}$, we condition on these data to generate posterior samples for $\mathbf{f} = \log N_e(\mathbf{x}^*)$ and κ, where $\mathbf{x}^* = \{x_d^*\}$ is the set of the middle points in (7). We use these posterior samples to make inference about $N_e(t)$.

For sampling $\mathbf{f}$ using HMC, we use (9) to compute the discretized log-likelihood

$$l = -\sum_{d=1}^{D-1} \sum_{I_\alpha^* \subset (x_d, x_{d+1}]} \{y_\alpha f(x_d^*) + C_{i,k}\Delta_d \exp[-f(x_d^*)]\}$$

and the corresponding gradient (score function)

$$s_d = -\sum_{I_\alpha^* \subset (x_d, x_{d+1}]} \{y_\alpha - C_{i,k}\Delta_d \exp[-f(x_d^*)]\}.$$

Because the prior on κ is conditionally conjugate, we could directly sample from its full conditional posterior distribution,

$$\kappa|\cdot \sim \text{Gamma}(\alpha + (D-1)/2, \beta + \mathbf{f}^\top \mathbf{C}_{in}^{-1}\mathbf{f}/2). \quad (11)$$

However, updating $\mathbf{f}$ and κ separately is not recommended in general because of their strong interdependency [Knorr-Held and Rue, 2002]: large value of precision κ strictly confines the variation of $\mathbf{f}$, rendering slow movement in the space occupied by $\mathbf{f}$. Therefore, we update $(\mathbf{f}, \kappa)$ jointly in MCMC sampling algorithms. In practice, it is better to sample $\boldsymbol{\theta} := (\mathbf{f}, \tau)$, where $\tau = \log(\kappa)$ is in the same scale as $\mathbf{f} = \log N_e(\mathbf{x}^*)$. Note that the log-likelihood of $\boldsymbol{\theta}$ is the same as that of $\mathbf{f}$ because equation (2) does not involve τ. The log-density prior on $\boldsymbol{\theta}$ is defined as follows:

$$\log P(\boldsymbol{\theta}) \propto ((D-1)/2 + \alpha - 1)\tau - (\mathbf{f}^\top \mathbf{C}_{in}^{-1}\mathbf{f}/2 + \beta)e^\tau. \quad (12)$$

3.3 Speed up by splitting Hamiltonian

Splitting the Hamiltonian is a technique used to speed up HMC [Leimkuhler and Reich, 2004, Neal, 2010, Shahbaba et al., 2013]. The underlying idea is to divide the total Hamiltonian into several terms, such that the dynamics associated with some of these terms can be solved analytically. For these parts, typically quadratic forms, the simulation of the dynamics does not introduce a discretization error, allowing for faster movements in the parameter space.

We split the Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p})$ as follows:

$$H(\boldsymbol{\theta}, \mathbf{p}) = \frac{-l - [(D-1)/2 + \alpha - 1]\tau + \beta e^\tau}{2} + \frac{\mathbf{f}^\top \mathbf{C}_{in}^{-1}\mathbf{f}e^\tau + \mathbf{p}^\top \mathbf{p}}{2} + \frac{-l - [(D-1)/2 + \alpha - 1]\tau + \beta e^\tau}{2}. \quad (13)$$

We further split the middle part (which is the dominant part) into two dynamics involving $\mathbf{f}|\tau$ and $\tau|\mathbf{f}$ respectively,

$$\begin{cases} \dot{\mathbf{f}}|\tau = \mathbf{p}_{-D}, \\ \dot{\mathbf{p}}_{-D} = -\mathbf{C}_{in}^{-1}\mathbf{f}e^\tau. \end{cases} \quad (14a) \quad \begin{cases} \dot{\tau}|\mathbf{f} = p_D, \\ \dot{p}_D = -\mathbf{f}^\top \mathbf{C}_{in}^{-1}\mathbf{f}e^\tau/2. \end{cases} \quad (14b)$$

Using the spectral decomposition $\mathbf{C}_{in}^{-1} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$ and denoting $\mathbf{f}^* := \sqrt{\boldsymbol{\Lambda}}e^{\tau/2}\mathbf{Q}^{-1}\mathbf{f}$ and $\mathbf{p}_{-D}^* := \mathbf{Q}^{-1}\mathbf{p}_{-D}$, we can analytically solve the dynamics (14a) as follows [Lan, 2013] (more details are in the appendix):

$$\begin{bmatrix} \mathbf{f}^*(t) \\ \mathbf{p}_{-D}^*(t) \end{bmatrix} = \begin{bmatrix} \cos(\sqrt{\boldsymbol{\Lambda}}e^{\tau/2}t) & \sin(\sqrt{\boldsymbol{\Lambda}}e^{\tau/2}t) \\ -\sin(\sqrt{\boldsymbol{\Lambda}}e^{\tau/2}t) & \cos(\sqrt{\boldsymbol{\Lambda}}e^{\tau/2}t) \end{bmatrix} \begin{bmatrix} \mathbf{f}^*(0) \\ \mathbf{p}_{-D}^*(0) \end{bmatrix}.$$

We then use the standard leapfrog method to solve the dynamics (14b) and the residual dynamics in (13) in a symmetric way. Note

that we only need to diagonalize $\mathbf{C}_{in}^{-1}$ once before the sampling, and calculate $\mathbf{f}^\top \mathbf{C}_{in}^{-1}\mathbf{f}e^\tau = \mathbf{f}^{*\top}\mathbf{f}^*$; therefore, the overall computational complexity of the integrator is $\mathcal{O}(D^2)$.

4 EXPERIMENTS

We illustrate the advantages of our HMC-based methods using three simulation studies and apply these methods to analysis of a real dataset. We evaluate our methods by comparing them to INLA in terms of accuracy and to several sampling algorithms, MALA, aMALA, and ES^2, in terms of sampling efficiency. We define sampling efficiency as time-normalized effective sample size (ESS). Given B MCMC samples for each parameter, we calculate the corresponding ESS $= B[1 + 2\Sigma_{k=1}^K \gamma(k)]^{-1}$, where $\Sigma_{k=1}^K \gamma(k)$ is the sum of K monotone sample autocorrelations [Geyer, 1992]. We use the minimum ESS normalized by the CPU time, s (in seconds), as the overall measure of efficiency: $\min(\text{ESS})/\text{s}$.

We tune the stepsize and number of leapfrog steps for our HMC-based algorithm such that their overall acceptance probabilities are in a reasonable range (close to 0.70). Since MALA [Roberts and Tweedie, 1996] and aMALA [Knorr-Held and Rue, 2002] can be viewed as HMC with one leap frog step for numerically solving Hamiltonian dynamics, we implement MALA and aMALA proposals using our HMC framework. MALA, aMALA, and HMC-based methods update $\mathbf{f}$ and τ jointly. aMALA uses a joint block-update method designed for GMRF models: it first generates a proposal $\kappa^*|\kappa$ from some symmetric distribution independently of $\mathbf{f}$, and then updates $\mathbf{f}^*|\mathbf{f}, \kappa^*$ based on a local Laplace approximation. Finally, $(\mathbf{f}^*, \kappa^*)$ is either accepted or rejected. It can be shown that aMALA is equivalent to a form of Riemannian MALA [Roberts and Stramer, 2002, Girolami and Calderhead, 2011, also see Appendix B]. In addition, aMALA closely resembles the most frequently used MCMC algorithm in Gaussian process-based phylodynamics [Minin et al., 2008, Gill et al., 2013].

ES^2 [Murray et al., 2010] is a commonly used sampling algorithm designed for models with Gaussian process priors. It was also adopted by Palacios and Minin [2013] for phylodynamic inference. ES^2 implementation relies on the assumption that the target distribution is approximately normal. This, of course, is not a suitable assumption for the joint distribution of $(\mathbf{f}, \tau)$. Therefore, we alternate the updates $\mathbf{f}|\kappa$ and $\kappa|\mathbf{f}$ when using ES^2.
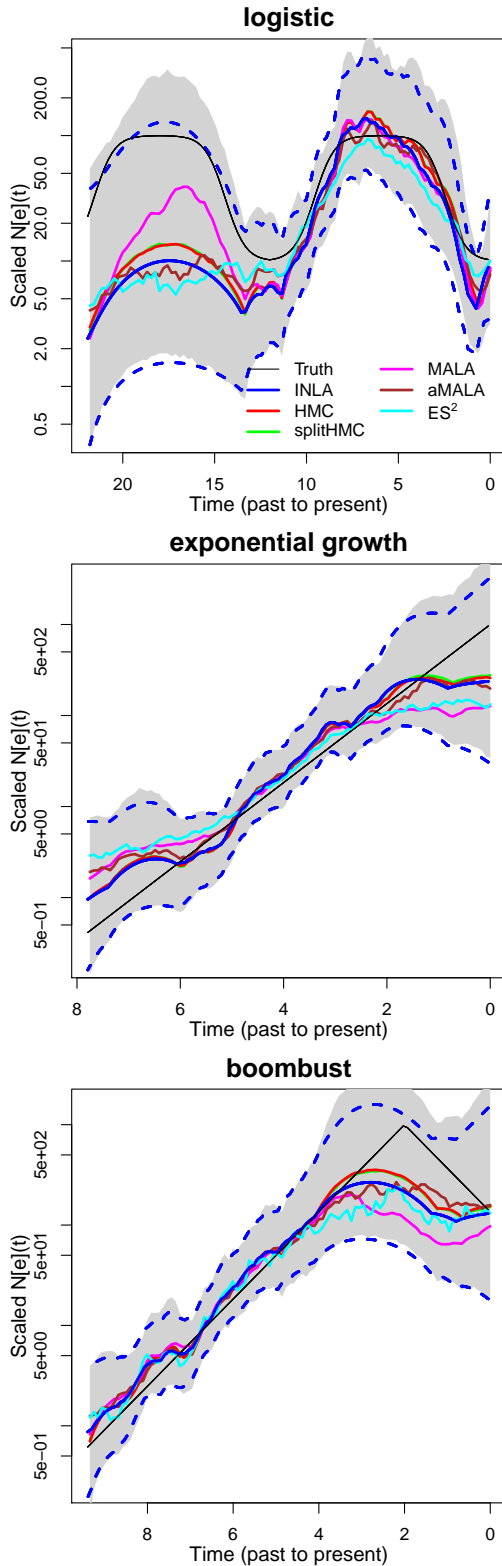
4.1 Simulations

We simulate three genealogies relating $n = 50$ individuals with the following true trajectories:

1. logistic trajectory:

$$N_e(t) = \begin{cases} 10 + \frac{90}{1 + \exp(2(3 - (t \mod 12)))}, & t \mod 12 \le 6, \\ 10 + \frac{90}{1 + \exp(2(-9 + (t \mod 12)))}, & t \mod 12 > 6; \end{cases}$$

2. exponential growth: $N_e(t) = 1000 \exp(-t)$;

3. boombust:

$$N_e(t) = \begin{cases} 1000 \exp(t - 2), & t \le 2, \\ 1000 \exp(-t + 2), & t > 2. \end{cases}$$

Fig. 2. INLA vs MCMC: simulated data under logistic (top), exponential growth (middle) and boombust (bottom) population size trajectories. Dotted blue lines show 95% credible intervals given by INLA and shaded regions show 95% credible interval estimated with MCMC samples given by splitHMC.

	Method	AP	s/iter	minESS($\mathbf{f}$)/s	spdup($\mathbf{f}$)	ESS(τ)/s	spdup(τ)
	ES^2	1.00	1.56E-03	0.22	1.00	0.25	1.00
	MALA	0.84	1.81E-03	0.41	1.90	1.47	5.89
I	aMALA	0.53	4.60E-03	0.08	0.38	0.13	0.53
	HMC	0.80	9.51E-03	1.78	8.23	1.81	7.25
	splitHMC	0.75	7.30E-03	2.19	**10.13**	2.51	**10.04**
	ES^2	1.00	1.57E-03	0.21	1.00	0.23	1.00
	MALA	0.78	1.82E-03	0.32	1.53	1.14	5.03
II	aMALA	0.53	4.61E-03	0.09	0.41	0.18	0.79
	HMC	0.76	1.19E-02	2.73	12.99	1.34	5.91
	splitHMC	0.76	7.84E-03	4.31	**20.50**	2.35	**10.40**
	ES^2	1.00	1.56E-03	0.20	1.00	0.20	1.00
	MALA	0.83	1.82E-03	0.37	1.87	1.05	5.18
III	aMALA	0.53	4.61E-03	0.08	0.40	0.14	0.67
	HMC	0.81	1.18E-02	2.22	11.24	1.17	5.75
	splitHMC	0.72	7.41E-03	2.87	**14.53**	1.90	**9.33**

Table 1. Sampling efficiency in modeling simulated population trajectories. The true population trajectories are: I) logistic, II) exponential growth, and III) boombust respectively. AP is the acceptance probability. s/iter is the seconds per sampling iteration. "spdup" is the speedup of efficiency measurement minESS/s using ES^2 as baseline.

We use $D = 100$ equally spaced grid points in the approximation of likelihood when applying INLA and MCMC algorithms (HMC, splitHMC, MALA, aMALA and ES^2).

Figure 2 compares the estimates of $N_e(t)$ using INLA and MCMC algorithms in for the three simulations. In general, the results of MCMC algorithms match closely with those of INLA. It is worth noting that MALA and ES^2 are occasionally slow to converge. Also, INLA fails when the number of grid points is large, e.g. 10000, while MCMC algorithms can still perform reliably.

For each experiment we run 15000 iterations with the first 5000 samples discarded. We repeat each experiment 10 times. The results provided in Table 1 are averaged over 10 repetitions. As we can see, our methods substantially improve over MALA, aMALA and ES^2. Note that although aMALA has higher ESS compared to MALA, its time-normalized ESS is worse than that of MALA because of its high computational cost of calculating the Fisher information.

Figure 3 compares different sampling methods in terms of their convergence to the stationary distribution when we increase the size of grid points to $D = 1000$. As we can see in this more challenging setting, Split HMC has the fastest convergence rate. Neither MALA nor aMALA, on the other hand, has converged within the given time. aMALA is not even getting close to the stationarity, making it much worse than MALA.

4.2 Human Influenza A in New York

Next, we analyze real data based on a genealogy estimated from 288 H3N2 sequences sampled in New York state from January 2001 to March 2005 in order to estimate population size dynamics of human influenza A in New York [Palacios and Minin, 2012, 2013]. The key feature of the influenza A virus epidemic in temperate regions like New York are the epidemic peaks during winters followed by strong bottlenecks at the end of the winter season. We use 120 grid points in the likelihood approximation. Figure 4 shows that with $\mathbf{C}_{BM}^{-1}$, MCMC algorithms identify such peak-bottleneck pattern more clearly than INLA. However, their results based on intrinsic precision matrix, $\mathbf{C}_{in}^{-1}$, are quite comparable to that of INLA. In
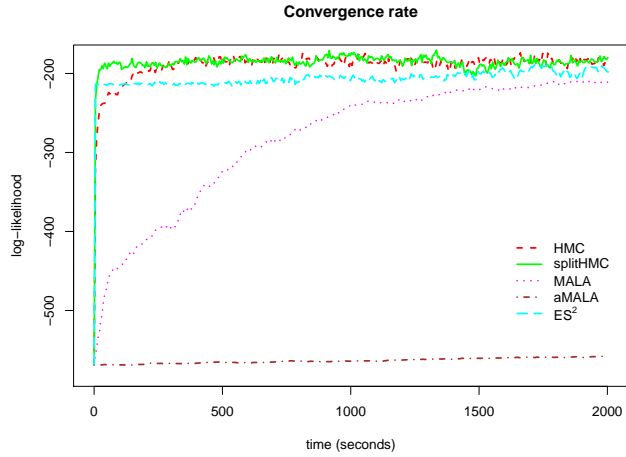
Fig. 3. Trace plots of log-likelihoods for different sampling algorithms based on a simulated coalescent model with logistic population trajectory. splitHMC converges the fastest.

Table 2, we can see that the speedup by HMC and splitHMC over other MCMC methods is substantial.

| Method	AP	s/Iter	minESS(**f**)/s	spdup(**f**)	ESS(τ)/s	spdup(τ)
ES^2	1.00	1.92E-03	0.34	1.00	0.27	1.00
MALA	0.78	2.12E-03	0.29	0.86	0.60	2.20
aMALA	0.77	8.48E-03	0.002	0.01	0.33	1.23
HMC	0.71	1.74E-02	1.69	4.98	0.57	2.12
splitHMC	0.75	1.11E-02	2.90	**8.53**	1.06	**3.92**

Table 2. Sampling efficiency of MCMC algorithms in influenza data.

5 DISCUSSION

In this paper, we have proposed new HMC-based sampling algorithms for phylodynamic inference, which are substantially more efficient than existing methods. Although we have focused on phylodynamics, the methodology presented here can be applied to general LGCP models.

There are several possible future directions. One possibility is to use ES^2 as a proposal generating mechanism in updating **f** as opposed to using it for sampling from the posterior distribution. Finding a good proposal for $\kappa(\tau)$, however, remains challenging.

An important extension of the methods presented here is to allow for genealogical uncertainty. The MCMC methods analyzed here can be incorporated into a hierarchical framework to infer population size trajectories from sequence data directly. In contrast, INLA cannot be adapted easily to perform inference from sequence data. This greatly limits its generalizability.
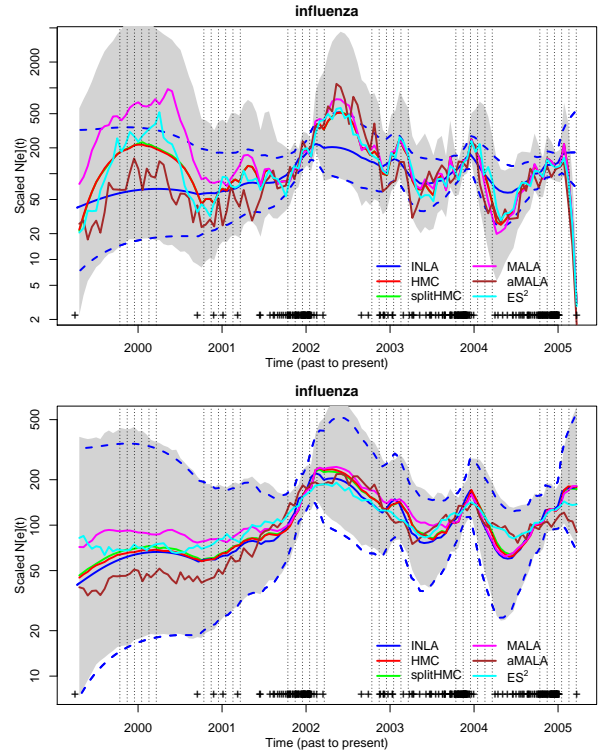
Fig. 4. Population dynamics of influenza A in New York (2001-2005): shaded region is the 95% credible interval calculated with samples given by splitHMC. Results using $\mathbf{C}_{BM}^{-1}$ (upper) vs results using $\mathbf{C}_{in}^{-1}$ (lower).

APPENDIX A: SPLITHMC

Here, we show how to solve Hamiltonian dynamics defined by (13) using the "splitting" strategy. Denote $\mathbf{z} := (\mathbf{f}, \mathbf{p})$, $\mathbf{M} := \mathbf{C}_{in}^{-1}\mathrm{e}^{\tau}$, and $\mathbf{A} := \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M} & \mathbf{0} \end{bmatrix}$. The dynamics (14a) can be written as

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}. \qquad (15)$$

We then have the analytical solution to (15) as

$$\mathbf{z}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{z}(0), \qquad (16)$$

where $\mathrm{e}^{\mathbf{A}t}$ is a matrix defined as $\mathrm{e}^{\mathbf{A}t} := \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbf{A}^i$, which in turn can be written as follows:

$$\mathrm{e}^{\mathbf{A}t} = \begin{bmatrix} \mathbf{I} - \frac{t^2}{2!}\mathbf{M} + \frac{t^4}{4!}\mathbf{M}^2 + \cdots & \mathbf{I}t - \frac{t^3}{3!}\mathbf{M} + \frac{t^5}{5!}\mathbf{M}^2 + \cdots \\ -\mathbf{M}t + \frac{t^3}{3!}\mathbf{M}^2 - \frac{t^5}{5!}\mathbf{M}^3 + \cdots & \mathbf{I} - \frac{t^2}{2!}\mathbf{M} + \frac{t^4}{4!}\mathbf{M}^2 + \cdots \end{bmatrix}$$
$$= \begin{bmatrix} \cos(\sqrt{\mathbf{M}}t) & \mathbf{M}^{-\frac{1}{2}}\sin(\sqrt{\mathbf{M}}t) \\ -\mathbf{M}^{\frac{1}{2}}\sin(\sqrt{\mathbf{M}}t) & \cos(\sqrt{\mathbf{M}}t) \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{M}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \cos(\sqrt{\mathbf{M}}t) & \sin(\sqrt{\mathbf{M}}t) \\ -\sin(\sqrt{\mathbf{M}}t) & \cos(\sqrt{\mathbf{M}}t) \end{bmatrix} \begin{bmatrix} \mathbf{M}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

For positive definite matrix $\mathbf{M}$, we can use the spectral decomposition $\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}$, where $\mathbf{Q}$ is orthogonal matrix, i.e.

$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathsf{T}}$. Therefore we have

$$
\begin{aligned}
\mathrm{e}^{\mathbf{A}t} = & \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{D}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \cos(\sqrt{\mathbf{D}}t) & \sin(\sqrt{\mathbf{D}}t) \\ -\sin(\sqrt{\mathbf{D}}t) & \cos(\sqrt{\mathbf{D}}t) \end{bmatrix} \\
& \cdot \begin{bmatrix} \mathbf{D}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Q}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^{-1} \end{bmatrix} .
\end{aligned}
\tag{17}
$$

In practice, we only need to diagonalize $\mathbf{C}_{in}^{-1}$ once: $\mathbf{C}_{in}^{-1} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$, then $\mathbf{D} = \boldsymbol{\Lambda}\mathrm{e}^{\tau}$. If we let $\mathbf{f}^* := \sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}\mathbf{Q}^{-1}\mathbf{f}$, $\mathbf{p}^*_{-D} := \mathbf{Q}^{-1}\mathbf{p}_{-D}$, we have the following solution (16):

$$
\begin{bmatrix} \mathbf{f}^*(t) \\ \mathbf{p}^*_{-D}(t) \end{bmatrix} = \begin{bmatrix} \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) & \sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) \\ -\sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) & \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) \end{bmatrix} \begin{bmatrix} \mathbf{f}^*(0) \\ \mathbf{p}^*_{-D}(0) \end{bmatrix} .
$$

We then apply leapfrog method to the remaining dynamics. Algorithm 1 summarizes these steps.

Algorithm 1 splitHMC for the coalescent model (splitHMC)

Initialize $\boldsymbol{\theta}^{(1)}$ at current $\boldsymbol{\theta} = (\mathbf{f}, \tau)$
Sample a new momentum value $\mathbf{p}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Calculate $H(\boldsymbol{\theta}^{(1)}, \mathbf{p}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\mathbf{p}^{(1)})$ according to (13)
for $\ell = 1$ to L **do**

$$
\mathbf{p}^{(\ell+1/2)} = \mathbf{p}^{(\ell)} + \varepsilon/2 \begin{bmatrix} \mathbf{s}^{(\ell)} \\ ((D-1)/2 + \alpha - 1) - \beta\exp(\tau^{(\ell)}) \end{bmatrix}
$$

$$
p_D^{(\ell+1/2)} = p_D^{(\ell)} - \varepsilon/2\mathbf{f}^{*(\ell)\mathsf{T}}\mathbf{f}^{*(\ell)}/2
$$

$$
\tau^{(\ell+1/2)} = \tau^{(\ell)} + \varepsilon/2 p_D^{(\ell+1/2)}
$$

$$
\begin{bmatrix} \mathbf{f}^{*(\ell+1)} \\ \mathbf{p}^{*(\ell+1/2)}_{-D} \end{bmatrix} \leftarrow \begin{bmatrix} \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\frac{1}{2}\tau^{(\ell+1/2)}}\varepsilon) & \sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\frac{1}{2}\tau^{(\ell+1/2)}}\varepsilon) \\ -\sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\frac{1}{2}\tau^{(\ell+1/2)}}\varepsilon) & \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\frac{1}{2}\tau^{(\ell+1/2)}}\varepsilon) \end{bmatrix}
$$
$$
\cdot \begin{bmatrix} \mathbf{f}^{*(\ell)} \\ \mathbf{p}^{*(\ell+1/2)}_{-D} \end{bmatrix}
$$

$$
\tau^{(\ell+1)} = \tau^{(\ell+1/2)} + \varepsilon/2 p_D^{(\ell+1/2)}
$$

$$
p_D^{(\ell+1)} = p_D^{(\ell+1/2)} - \varepsilon/2\mathbf{f}^{*(\ell+1)\mathsf{T}}\mathbf{f}^{*(\ell+1)}/2
$$

$$
\mathbf{p}^{(\ell+1)} = \mathbf{p}^{(\ell+1/2)} + \varepsilon/2 \begin{bmatrix} \mathbf{s}^{(\ell+1)} \\ ((D-1)/2 + \alpha - 1) - \beta\exp(\tau^{(\ell+1)}) \end{bmatrix}
$$

end for
Calculate $H(\boldsymbol{\theta}^{(+1)}, \mathbf{p}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\mathbf{p}^{(L+1)})$ according to (13)
Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\boldsymbol{\theta}^{(+1)}, \mathbf{p}^{(L+1)}) + H(\boldsymbol{\theta}^{(1)}, \mathbf{p}^{(1)})]\}$
Accept or reject the proposal according to α for the next state $\boldsymbol{\theta}'$

APPENDIX B: ADAPTIVE MALA

We now show that the joint block updating in Knorr-Held and Rue [2002] can be recognized as an adaptive MALA algorithm. First, we sample $\kappa^* | \kappa \sim p(\kappa^*|\kappa) \propto \frac{\kappa^*+\kappa}{\kappa^*\kappa}$ on $[\kappa/c, \kappa c]$ for some $c > 1$ controlling the step size of κ. Denote $\mathbf{w} := \{C_{i,k}\Delta_d\}_1^{D+m+n-4}$

Algorithm 2 Adaptive MALA (aMALA)

Given current state $\boldsymbol{\theta} = (\mathbf{f}, \kappa)$ calculate potential energy $U(\boldsymbol{\theta})$
repeat
  $z \sim \mathrm{Unif}[1/c, c], \; u \sim \mathrm{Unif}[0, 1]$
until $u < \frac{z+1/z}{c+1/c}$
update precision parameter $\kappa^* = \kappa z$
Sample momentum $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\mathbf{f}, \kappa^*)^{-1})$
Calculate log of proposal density $\log p(\mathbf{f}^*|\mathbf{f}, \kappa^*) = -\frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\mathbf{f}, \kappa^*)\mathbf{p} + \frac{1}{2}\log\det\mathbf{G}(\mathbf{f}, \kappa^*)$
update momentum $\mathbf{p} \leftarrow \mathbf{p} - \varepsilon/2\mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\nabla U(\mathbf{f}, \kappa^*)$
update latent variables $\mathbf{f}^* = \mathbf{f} + \varepsilon\mathbf{p}$
update momentum $\mathbf{p} \leftarrow \mathbf{p} - \varepsilon/2\mathbf{G}(\mathbf{f}^*, \kappa)^{-1}\nabla U(\mathbf{f}^*, \kappa)$
Calculate log of reverse proposal density $\log p(\mathbf{f}|\mathbf{f}^*, \kappa) = -\frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\mathbf{f}^*, \kappa)\mathbf{p} + \frac{1}{2}\log\det\mathbf{G}(\mathbf{f}^*, \kappa)$
Calculate new potential energy $U(\boldsymbol{\theta}^*)$
Accept/reject the proposal according to $\log\alpha = -U(\boldsymbol{\theta}^*) + U(\boldsymbol{\theta}) - \log p(\mathbf{f}^*|\mathbf{f}, \kappa) + \log p(\mathbf{f}|\mathbf{f}^*, \kappa)$ for the next state $\boldsymbol{\theta}'$

---

and use the following Taylor expansion for $\log p(\mathbf{f}|\kappa)$ about $\hat{\mathbf{f}}$:

$$
\begin{aligned}
\log p(\mathbf{f}|\kappa) &= -\mathbf{y}^{\mathsf{T}}\mathbf{f} - \mathbf{w}^{\mathsf{T}}\exp(-\mathbf{f}) - \frac{1}{2}\mathbf{f}^{\mathsf{T}}\kappa\mathbf{C}_{in}^{-1}\mathbf{f} \\
&\approx -\mathbf{y}^{\mathsf{T}}\mathbf{f} - (\mathbf{w}\exp(-\hat{\mathbf{f}}))^{\mathsf{T}}[\mathbf{1} - (\mathbf{f}-\hat{\mathbf{f}}) + (\mathbf{f}-\hat{\mathbf{f}})^2/2] - \frac{1}{2}\mathbf{f}^{\mathsf{T}}\kappa\mathbf{C}_{in}^{-1}\mathbf{f} \\
&= (-\mathbf{y} + \mathbf{w}\exp(-\hat{\mathbf{f}})(1+\hat{\mathbf{f}}))^{\mathsf{T}}\mathbf{f} + \frac{1}{2}\mathbf{f}^{\mathsf{T}}[\kappa\mathbf{C}_{in}^{-1} + \mathrm{diag}(\mathbf{w}\exp(-\hat{\mathbf{f}}))]\mathbf{f} \\
&=: \mathbf{b}^{\mathsf{T}}\mathbf{f} - \frac{1}{2}\mathbf{f}^{\mathsf{T}}\mathbf{G}\mathbf{f},
\end{aligned}
$$

where $\mathbf{b}(\hat{\mathbf{f}}) := -\mathbf{y} + \mathbf{w}\exp(-\hat{\mathbf{f}})(1+\hat{\mathbf{f}})$, $\mathbf{G}(\hat{\mathbf{f}}, \kappa) := \kappa\mathbf{C}_{in}^{-1} + \mathrm{diag}(\mathbf{w}\exp(-\hat{\mathbf{f}}))$. Setting $\hat{\mathbf{f}}$ to the current state, $\mathbf{f}$, and propose $\mathbf{f}^*|\mathbf{f}, \kappa^*$ from the following Gaussian distribution:

$$
\mathbf{f}^*|\mathbf{f}, \kappa^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),
$$

with

$$
\boldsymbol{\mu} = \mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\mathbf{b}(\mathbf{f}) = \mathbf{f} + \mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\nabla_{\mathbf{f}}\log p(\mathbf{f}|\kappa^*) \text{ and}
$$

$$
\boldsymbol{\Sigma} = \mathbf{G}(\mathbf{f}, \kappa^*)^{-1},
$$

which has the same form as Langevin dynamical proposals. Interestingly, $\mathbf{G}(\hat{\mathbf{f}}, \kappa)$ is exactly the (observed) Fisher information. That is, this approach is equivalent to Riemannian MALA [Girolami and Calderhead, 2011].

Finally, $\boldsymbol{\theta}^* = (\mathbf{f}^*, \kappa^*)$ is jointly accepted with the following probability:

$$
\begin{aligned}
\alpha &= \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})}{p(\boldsymbol{\theta}|\mathcal{D})}\frac{p(\kappa|\kappa^*)p(\mathbf{f}|\mathbf{f}^*, \kappa)}{p(\kappa^*|\kappa)p(\mathbf{f}^*|\mathbf{f}, \kappa^*)}\right\} \\
&= \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})}{p(\boldsymbol{\theta}|\mathcal{D})}\frac{p(\mathbf{f}|\mathbf{f}^*, \kappa)}{p(\mathbf{f}^*|\mathbf{f}, \kappa^*)}\right\},
\end{aligned}
$$

where $p(\kappa^*|\kappa)$ is a symmetric proposal. Algorithm 2 summarizes the steps for adaptive MALA.

REFERENCES

Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005. doi: 10.1093/molbev/msi103.

Alexei J. Drummond, Geoff K. Nicholls, Allen G. Rodrigo, and Wiremu Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

S. Duane, A. D. Kennedy, B J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

Mandev S Gill, Philippe Lemey, Nuno R Faria, Andrew Rambaut, Beth Shapiro, and Marc A Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724, 2013.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, (with discussion) 73(2):123–214, 2011.

R. C. Griffiths and Simon Tavare. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410, 1994. doi: 10.1098/rstb.1994.0079.

Joseph Heled and Alexei Drummond. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1): 289, 2008. ISSN 1471-2148. doi: 10.1186/1471-2148-8-289.

J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235 – 248, 1982. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/0304-4149(82)90011-4.

Leonhard Knorr-Held and Håvard Rue. On Block Updating in Markov Random Field Models for Disease Mapping. *Scandinavian Journal of Statistics*, 29(4):597–614, 2002.

Mary K. Kuhner, Jon Yamato, and Joseph Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149(1):429–434, 1998.

Shiwei Lan. Advanced bayesian computational methods through geometric techniques, 2013. Copyright - Copyright ProQuest, UMI Dissertations Publishing 2013; Last updated - 2014-02-13; First page - n/a; M3: Ph.D.

B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.

Vladimir N. Minin, Erik W. Bloomquist, and Marc A. Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008. doi: 10.1093/molbev/msn090.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):pp. 451–482, 1998. ISSN 03036898.

Iain Murray, Ryan Prescott Adams, and David J.C. MacKay. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X. L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2010.

Rainer Opgen-Rhein, Ludwig Fahrmeir, and Korbinian Strimmer. Inference of demographic history from genealogical trees using reversible jump markov chain monte carlo. *BMC Evolutionary Biology*, 5(1):6, 2005. ISSN 1471-2148. doi: 10.1186/1471-2148-5-6.

Julia A. Palacios and Vladimir N. Minin. Integrated nested laplace approximation for bayesian nonparametric phylodynamics. In Nando de Freitas and Kevin P. Murphy, editors, *UAI*, pages 726–735. AUAI Press, 2012.

Julia A. Palacios and Vladimir N. Minin. Gaussian process-based bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics*, 69(1):8–18, 2013. ISSN 1541-0420. doi: 10.1111/biom.12003.

Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, 2008.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):pp. 341–363, 1996. ISSN 13507265.

Allen G Rodrigo and Joseph Felsenstein. Coalescent approaches to hiv population genetics. *The evolution of HIV*, pages 233–272, 1999.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x.

Babak Shahbaba, Shiwei Lan, Wesley O. Johnson, and RadfordM. Neal. Split hamiltonian monte carlo. *Statistics and Computing*, pages 1–11, 2013. ISSN 0960-3174. doi: 10.1007/s11222-012-9373-1.

Korbinian Strimmer and Oliver G Pybus. Exploring the demographic history of dna sequences using the generalized skyline plot. *Molecular Biology & Evolution*, 18(12):2298–2305, January 2001 2001.