Optimal Point Process Filtering and Estimation of the Coalescent Process

Kris V Parag[1], Oliver G Pybus

Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK

Abstract

The coalescent process is an important and widely used model for inferring the dynamics of biological populations from samples of genetic diversity. Coalescent analysis typically involves applying statistical models to either samples of genetic sequences or an estimated genealogy in order to infer the demographic history of the population from which the samples originated. Several parametric and non-parametric estimation techniques involving a range of Markov chain Monte Carlo, Gaussian process and other algorithms, already exist. However, these techniques sometimes trade estimation accuracy and sophistication for methodological flexibility and ease of use. Thus, there is a need for coalescent estimation techniques that can be easily implemented for a range of inference problems while still maintaining statistical optimality. We introduce the Bayesian Snyder filter as a natural, easily implementable and flexible minimum mean square error estimator for parametric demographic functions on fixed genealogies.

By reinterpreting the coalescent as a self-exciting Markov process, we show that the Snyder filter can be applied to both isochronously and heterochronously sampled coalescent datasets. We analytically solve the filter equations for the constant population size Kingman coalescent and derive expressions for the mean squared estimation error and the estimate robustness to prior distribution specification. For populations with deterministically time-varying size we numerically solve the Snyder equations, and test this solution on common epidemiological demographic models. We find that the Snyder filter accurately recovers the (true) demographic history for these models. We also apply the filter to a well-studied, empirical hepatitis C virus sequence dataset and show that its output compares well with that of a relevant benchmark phylodynamic inference method. The Snyder filter is an exact, direct Bayesian estimation method (given discretised priors, it does not approximate the posterior) that has the potential to become as a useful, alternative technique for coalescent inference.

Keywords: Coalescent theory, non-linear filters, Bayesian inference, parametric estimation, Snyder filters

1. Introduction

Genetic sequences contain information about the dynamics of the population from which they were sampled. The coalescent process provides a framework for extracting this information by describing the shared ancestry among n individuals randomly sampled from a population of effective size $N(t) \gg n$ [1]. The shared ancestry of the sampled individuals can be modelled as a random, ultrametric, bifurcating genealogy with n tips and $n-1$ branches. The branch lengths give the times at which sampled lineages coalesce. These coalescence times depend on $N(t)$ which is sometimes also called the demographic function. $N(t)$ essentially describes the dynamics of the population. A key problem in coalescent inference is the estimation of $N(t)$ or its embedded parameters either directly from a genealogy, or indirectly from a set of sampled genetic sequences.

The original, standard coalescent was developed by Kingman for a constant $N(t)$ and for sets of genetic sequences that are sampled at one time point (isochronous sampling) [1]. Since then, the coalescent model has been generalised to incorporate deterministically varying population sizes [2], stochastic population fluctuations [3], geographically structured populations [4], and data sets containing sequences sampled at different time points (heterochronous sampling) [5]. As a result, the coalescent model has been applied to a range of problems in many biological disciplines

[1]Email: kris.parag@zoo.ox.ac.uk

including conservation biology, anthropology and epidemiology [6]. Our work is geared towards infectious disease epidemiology in which pathogen populations, due to their large size and very rapid molecular evolution, are often treated as deterministically varying in size and heterochronously sampled. In this setting, the coalescent process has been successfully used to infer the growth and history of the hepatitis C epidemic in Egypt [7], the oscillating behaviour of dengue virus in Vietnam [8] and to estimate the generation time of HIV-1 within individual infected patients [5]. The accuracy and efficiency of such inferences are linked to the statistical techniques used. Consequently, the design of good coalescent demographic inference methods is important [9].

We focus here on the coalescent inference problem for a haploid population with deterministically varying population size, under both isochronous and heterochronous sampling. We follow the typical coalescent assumptions of a panmictic (well mixed) and neutrally evolving population that is sparsely and randomly sampled [10]. Several methods for inferring the demographic function, $N(t)$, have been developed and can be broadly categorised into parametric (model based) and non-parametric (design based) approaches [11]. The parametric approach characterises $N(t)$ using a biologically-inspired function (model) with a fixed number of explicit demographic parameters. These parameters interact in a preset manner and the model dimensionality is independent of n. In contrast, non-parametric methods use more generalised descriptions for $N(t)$ or rely on summary statistics derived from the data. Non-parametric methods therefore make weaker assumptions about demographic dynamics. This may allow a more robust description of population size but comes at the expense of less statistical power, and with the possibility that model dimensionality increases with n [12] [13]. Consequently, the choice of parametric or non-parametric methods depends on how much one knows about the sampled population. If the nature of the dynamics can be reliably encoded in a predefined function then parametric methods should lead to more efficient estimation [11]. However, if little is known about the study population, or the possibility of model misspecification is high then non-parametric methods should be used.

Here we assume that a suitable parametric demographic model $N(t, \vec{x})$ has already been chosen and that its parameters $\vec{x}$, or a function of them, are to be estimated from the data in an optimal way. We limit our current work to parametric demographic inference for two reasons. First, our interest is in developing new inference techniques that minimise approximations and that are theoretically rigourous enough to allow analytic results when possible. To do this explicit models are useful and so we apply parametric descriptions. Our metric for defining inference performance will be the classical mean squared estimation error (defined later). Secondly, we want to use techniques that avoid numerical issues such as optimisation to local minima or poor algorithm convergence. These could hamper the flexibility of an inference method and reduce reproducibility among analyses. Such issues can sometimes be encountered in (but are not limited to) advanced non-parametric coalescent inference methods that approximate the posterior distribution using Markov Chain Monte Carlo (MCMC) or importance sampling [14] [15] [16]. These approaches, while readily able to account for genealogical uncertainty, can be complex or difficult to implement [17].

The motivation for our work is most similar to that of Palacios and Minin [18]. They presented a non-parametric technique for fixed genealogies aimed at replacing MCMC approaches. Their method traded a little accuracy to achieve large computational accelerations relative to existing MCMC techniques. We also assume a fixed genealogy in our work but instead focus on analytical tractability and statistical efficiency (minimising mean squared error). The method we will introduce avoids the need to specify and modify MCMC operators, as found in the phylodynamic inference software BEAST [19], and related approaches.

In this paper, we introduce and analyse the Snyder filter [20], a technique from electrical and systems engineering, as a means of achieving the aforementioned inference goals. The Snyder filter is an explicit, parametric, Bayesian inference technique that solves dynamical equations for the joint posterior distribution of $\vec{x}$. These equations can then be used to obtain a conditional mean estimator that minimises the mean square error between the true parameter (or function) and its estimate. The Snyder filter is unlike other existing Bayesian methods for coalescent inference because it directly computes the posterior distribution, given a model and priors. We show how the Snyder filter, which treats coalescent data as a point process stream, can be used as an alternative and useful Bayesian estimator. The Snyder filter has remained largely unknown to the biological sciences and, to our knowledge, has only been applied to neuronal spiking by Bobrowski *et al* [21] and to invertebrate visual phototransduction by Parag [22].

We start by defining the Snyder filter and provide its equations for the estimation of random variables embedded within the self-exciting rate of a point process. We then demonstrate how the deterministically time-varying coalescent process can be reinterpreted so that it is amenable to Snyder based inference and describe how to incorporate heterochronous sampling. Next we show that when population size is constant (the Kingman coalescent) then the filter can be solved analytically. From these equations, we recover the known maximum likelihood estimator of the

Kingman coalescent and we derive an approximate, explicit minimum mean square error (MMSE) function. We also provide a measure of robustness of the Snyder filter to prior specification. This completes our theoretical treatment of the Snyder filter approach. We subseqeuntly explore and quantify the performance of the Snyder filter by applying it to (i) data simulated under several canonical, deterministically time-varying, epidemiologically relevant demographic models and (ii) a well-studied empirical dataset comprising hepatitis C virus (HCV) gene sequences from Egypt. This HCV dataset has been widely used in previous studies and thus allows us to compare our method with previous approaches. In the appendices we give other formulations of the general Snyder filter and reiterate the link between sequential data from a single tree and parallel data from many trees, for the Kingman coalescent. We also present an informative relation between the Snyder filter approach and a popular non-parametric coalescent inference technique called the classic skyline plot [23]. We show that the Snyder filter naturally generalises the skyline in a non-linear parametric manner.

2. Methods

2.1. The Snyder Filter

Consider a Poisson process, $\mathcal{F}(t)$, at time $t \geq 0$ with instantaneous intensity $\lambda(t, \vec{x}(t))$ on the space of non-negative real numbers. $\mathcal{F}(t)$ is an integer valued process that counts the number of points at t and the vector $\vec{x}(t)$ is called the information process. The information process is what we want to infer (section 2.2 will show that this process encodes the parameters of a coalescent demographic function). If $\vec{x}(t)$ is stochastic then $\mathcal{F}(t)$ is a doubly stochastic Poisson process (DSPP). Let the counting process stream from time 0 to time t be denoted $\mathcal{F}_t = \mathcal{F}(s)$, $\forall s : 0 \leq s \leq t$. In 1972 Snyder introduced an exact Bayesian filter for the optimal, causal estimation of this stochastic hidden information process $\vec{x}(t)$ given only observations of $\mathcal{F}_t$ and the basic statistics of $\vec{x}(t)$ [20]. We call this the Snyder filter in this work. The Snyder filter is a set of non-linear differential equations on the probability distribution of $\vec{x}(t)$ which, when solved across the observation process $\mathcal{F}_t$, lead to the causal posterior $\vec{q}(t) = \mathrm{P}(\vec{x}(t) \,|\, \mathcal{F}_t)$. The information process $\vec{x}(t)$ can follow quite general models including Markov diffusions and continuous time Markov jump processes (see [24] for details). The filter also works if the intensity is generalised to depend on the counting process as $\lambda(t, \vec{x}(t), \mathcal{F}_t)$. This makes the DSPP self-exciting since its rate of producing points depends on the points themselves [24].

The Snyder filter is 'causal' because the estimate at t only uses observations or data up to time t. It is 'exact' because it directly solves for the joint posterior $\mathrm{P}(\vec{x}(t) \,|\, \mathcal{F}_t)$ without approximating either the observation process $\mathcal{F}_t$ or the dynamics of the information process $\vec{x}(t)$. The only approximations come from standard implementation issues such as the numerical integration of differential equations and the fact that distributions must be represented discretely. In some instances it is possible to solve the equations analytically, in which case there are no approximations.

If $\vec{x}(t)$ is modelled as a continuous time Markov jump process, then the generally non-linear filter on the probabilities $\vec{q}(t)$ can be described with a set of linear differential equations on an un-normalised distribution $q^*(t)$. This is then normalised afterwards to $\vec{q}(t)$ [25] by integrating the joint distribution over its domain. We focus on a simplification of this case in which $\vec{x}(t)$ is reduced to a vector of random variables $\vec{x}$ and allow for the intensity to depend on $\mathcal{F}_t$. However, we restrict this dependence so that only the current count matters (the process is then Markovian [24]). The intensity is then $\lambda(t, \vec{x}, \mathcal{F}(t))$ instead of $\lambda(t, \vec{x}(t), \mathcal{F}_t)$ and the Snyder filter [24] [25] is defined by equations 1-3 below. $\Lambda_{\mathcal{F}(t)}$ is a diagonal matrix called the rate matrix. It has entries for each possible value of $\lambda(t, \vec{x}, \mathcal{F}(t))$ at any given t due to the possible values of $\vec{x}$ and the current event count $\mathcal{F}(t)$. Let an arbitrary value of $\vec{x}$ be μ and denote the corresponding normalised and unnormalised probabilities as $\vec{q}(t, \mu)$ and $q^*(t, \mu)$. Then $\vec{q}(t) = \{\vec{q}(t, \mu)\}$; that is $\vec{q}(t)$ is the complete probability distribution and $\vec{q}(t, \mu)$ is the single value corresponding to μ. Assume we have observed an event stream from time 0 until time T so that $\max(t) = T$. If the first event time is $t = \tau_1$ then τ_1^- and τ_1^+ are infinitesimally before and after that event time. The initial condition for the differential equations is the prior:

$\vec{q}(0) = \mathrm{P}(\vec{x})$. The following equations then describe the dynamics of $\vec{q}(t)$ with time until τ_1^+.

$$\frac{\mathrm{d}q^*(t)}{\mathrm{d}t} = -q^*(t)\Lambda_{\mathcal{F}(t)}, \text{ for } 0 \leq t \leq \tau_1^- \tag{1}$$

$$\vec{q}(t) = \frac{q^*(t)}{\int q^*(t,\, \mu)\, \mathrm{d}\mu}, \text{ for } 0 \leq t \leq \tau_1^- \tag{2}$$

$$\vec{q}(\tau_1^+) = \vec{q}(\tau_1^-)\Lambda_{\mathcal{F}(\tau_1^-)} \left(\int \vec{q}(\tau_1^-,\, \mu)\lambda(t,\, \mu,\, \mathcal{F}(\tau_1^-))\, \mathrm{d}\mu \right)^{-1} \tag{3}$$

Equation 1 describes a continuous exponential decaying trajectory on the state probabilities until the observed event at τ_1^-. At this point a discontinuous update of the posterior occurs according to equation 3. The updated posterior $\vec{q}(\tau_1^+)$ is then used as a new initial condition and the equations are solved again until the next event (over $\tau_1^+ \leq t \leq \tau_2^-$). This procedure repeats until the completely observed event stream ends and we obtain $\vec{q}(T)$. The integrals in these equations mean that every possible value of $\vec{x}$ is enumerated. The solution $\vec{q}(t)$ across time follows a piecewise deterministic Markov process since it involves a continuous function that is punctuated by random jumps [26]. Conceptually, the filter is placing a particle on each probability mass and then evolving the interconnected dynamics of all the particles along the observed process until a posterior results. The original non-linear form of the filter together with some more analytical insight into its mechanics are given in the Appendix A.

We will show that coalescent process inference falls within the Snyder filtering framework and derive the appropriate rate matrix $\Lambda_{\mathcal{F}(t)}$. Since we focus on parametric estimation in this work, $\vec{q}(T) = \mathrm{P}(\vec{x}(t)\,|\,\mathcal{F}_T)$ will be our posterior of interest, as it uses all of the observed data (numbers of extant lineages) $\mathcal{F}_T$. This posterior allows calculation of the minimum mean squared error (MMSE) estimate of any given function of the parameters, $f(\vec{x})$. This is defined as: $\mathrm{MMSE} = \mathbb{E}\left[\left(f(\vec{x}) - \hat{f}(\vec{x})\right)^2\right]$. The estimate $\hat{f}(\vec{x}) = \mathbb{E}\left[f(\vec{x})\,|\,\mathcal{F}_T\right]$ is called the conditional mean [24]. The Snyder filter is optimal because it directly calculates the posterior, $\vec{q}(T)$, that minimises the mean squared error.

2.2. Snyder Inference for the Coalescent Process

We focus on coalescent processes with deterministically time-varying demographic functions, denoted $N(t)$ at time t. We assume this process is heterochronously sampled, with $n \ll N(t)$ samples (lineages) taken across K distinct times so that $\sum_{i=1}^{K} n_i = n$. The pair $(s_i,\, n_i)$ represents the i^{th} sample time and the corresponding number of tree tips (sampled lineages) introduced at that time. Here $s_1 = 0$ is always taken as the present (the most recent time of sampling). Time therefore increases into the past. If $K = 1$ and $(s_1,\, n_1) = (0,\, n)$, then sampling is termed isochronous. We define the k^{th} coalescent time, c_k, as the time point at which k coalescent events have occurred. For a total of n sampled lineages, a maximum of $n - 1$ coalescent events are possible. In this notation both the sample time index, i, and the coalescent time index, k increase with time. All the information for estimation is encoded in the coalescent event times, $(c_k,\, k)$ for $1 \leq k \leq n - 1$ and in the knowledge of the sampling scheme, $(s_i,\, n_i)$ for $1 \leq i \leq K$. We focus on estimation given the values of c_k, s_i and n_i and we do not consider the bifurcating genealogy from which they arise. Our observed process is therefore the trajectory of the number of existing lineages with time along the tree from the tips to the root. This can be thought of as a lineages through time plot from the present to the past [27].

The coalescent process for n lineages is usually described as an inhomogeneous Poisson process with a maximum count of $n - 1$ (the process ends with a single lineage). In this interpretation, the stream of $n - 1$ coalescent events is separated into intervals that end with either a coalescent or sampling event. The coalescent rate is described for each interval by conditioning on the number of lineages in that interval. If an interval starts with i lineages then the rate is $\binom{i}{2} N(t)^{-1}$ [2]. Inference on the coalescent process then usually involves writing down a product likelihood in terms of these intervals (or groups of them). MCMC or some other technique is then used to either maximise the likelihood function or sample the posterior of the process [28] [29].

We take a different approach. We do not explicitly condition on the number of lineages in an interval or derive a product form distribution. Instead we embed the trajectory of the number of lineages causally (sequentially with time) within the coalescent rate. Both the fall in lineages due to coalescent events and the rise due to sample introductions are explicitly within the rate function and contribute a source of randomness (the coalescent rate is a function of a random counting process). Since the rate of producing stochastic coalescent events is stochastic, and depends on the events themselves, the overall process is self-exciting. This will be explained further in the following paragraphs.

We treat the observed sampled coalescent trajectory as an instance of this self-exciting point process. The inference problem is then to estimate the parameters of the driving process given the observed sampled coalescent. Note that while the coalescent rate is stochastic, the demographic function that we estimate is deterministically time-varying (it is the non-lineage dependent component of the coalescent rate).

We start by assuming the demographic function can be parametrised with a vector of random variables $\vec{x}$ so that it is written $N(t, \vec{x})$. Let $u(t) = \sum_{k=1}^{n-1} 1(t \geq c_k)$ be the count of the number of coalescent events from the present (time 0) to time t in the past, and $h(t) = \sum_{i=1}^{K} n_i 1(t \geq s_i)$ count the total number of samples introduced into the process by time t. The function 1(:) is an indicator. The observed data at any given time t is then $\mathcal{F}(t) = h(t) - u(t)$, the number of extant lineages at t. The coalescent rate is equivalent to the count dependent birth rate of a Markov birth process (births are new coalescent events) [24], between sample times. This birth rate is defined in equations 4 and 5 and uses the fact that between sample times only coalescent events change the lineage count. The lineages introduced at sample times are jump discontinuities on the rate of the process [24] and merely serve as initial conditions for the Markov birth process.

$$\lambda(t, \vec{x}, \mathcal{F}(t))|_{\{s_i \leq t < s_{i+1}\}} := \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}\left(u(t + \Delta t) - u(t) \geq 1 \,|\, \mathcal{F}(t)\right) \tag{4}$$

$$\lambda(t, \vec{x}, \mathcal{F}(t)) = \binom{\mathcal{F}(t)}{2} \frac{1}{N(t, \vec{x})} \tag{5}$$

In this form the coalescent rate is amenable to Snyder filtering techniques and the coalescent process can also be described as a Markov self-exciting point process. This is because the counting statistics of a self-exciting point process with 0-memory are identical to those of pure birth process with a population dependent birth rate [24]. For comparison, note that $\mathcal{F}(t) \,|\, \lambda(t, \vec{x}, \mathcal{F}(t))$ is an inhomogeneous Poisson process that corresponds to the usual (aformentioned) approach to coalescent inference. The difference in treating the coalescent process as a self-exciting Markov process versus a conditioned inhomogeneous Poisson process, while subtle and in some contexts trivial, is what allows us to apply Snyder inference techniques.

We now explain the numerical implementation of the filter for coalescent inference. Let the vector of random variables (parameters) to be estimated, $\vec{x}$, be described with l elements so that $\vec{x} = [x_1, x_2, \ldots x_l]$. The function $N(t, \vec{x})$ is then an l parameter multivariate demographic function. Assume the distribution of the i^{th} random variable can be described on a domain of m_i points, so that the demographic function is on a joint Cartesian grid of $m = \prod_{i=1}^{l} m_i$ possible values. As a result there are m possible sets of values for the l element vector $\vec{x}$. We will denote some arbitrary value of $\vec{x}$ as μ. The Snyder filter solves a differential equation for the joint probability mass across all μ. Consequently, the filter has dimension m. The discretisation of the parameter space together with any errors from standard numerical integration form the only approximations of the Snyder inference method.

Due to this discretisation, the prior $\mathrm{P}(\vec{x})$ and posterior $\vec{q}(t) = \mathrm{P}(\vec{x} \,|\, \mathcal{F}_t)$ have m elements. $\Lambda_{\mathcal{F}(t)}$ is then an $m \times m$ diagonal matrix with entries at each time t given by enumerating $\lambda(t, \vec{x}, \mathcal{F}(t))$ across the possible μ. Therefore the diagonal entry in $\Lambda_{\mathcal{F}(t)}$ corresponding to $\vec{x} = \mu$ is $\lambda(t, \mu, \mathcal{F}(t))$. The integrals in equations 1-3 are then across the m elements of the relevant vectors or matrices and solved across the observed stream $\mathcal{F}(t)$ for $0 \leq t \leq T$. Note that at the coalescent event times (at which equation 3 is solved): $\mathcal{F}(c_k^+) = \mathcal{F}(c_k^-) - 1$. At sample times: $\mathcal{F}(s_i^+) = \mathcal{F}(s_i^-) + n_i$.

When the coalescent is heterochronously sampled ($K > 1$) an extra condition must be included to account for $\mathcal{F}(t)$ falling to 1 after some time τ, where $\tau < s_K$ (the last sample time). This condition can be written as: $\{\vec{q}(\tau < t < s_i) : [\mathcal{F}(t) = 1] \wedge [\exists i \leq K : s_i > t]\} = \vec{q}(\tau)$. This means that the posterior, at τ, is maintained until the next sampling time, at which point $\mathcal{F}(t)$ rises above 1 again. This follows because over this time period the rate $\Lambda_{\mathcal{F}(t)} = 0_{m \times m}$ so that equation 1 gives $\frac{\mathrm{d}q^*(t)}{\mathrm{d}t} = 0_{1 \times m}$. The subscripts reflect the dimensions on the 0s.

We are interested in parametric estimates of the random variables $\vec{x}$ and the demographic function $N(t, \vec{x})$. We take the joint posterior generated by the Snyder filter at the very end of the observed data $\mathcal{F}_T$ which is $\vec{q}(T) = \mathrm{P}(\vec{x} \,|\, \mathcal{F}_T) = \{\vec{q}(T, \mu) = \mathrm{P}(\vec{x} = \mu \,|\, \mathcal{F}_T)\}$. Denote the marginalisation of this joint posterior for the i^{th} parameter, x_i as $q_i(T, \mu_i)$ with μ_i an arbitrary value that x_i can take. The MMSE estimates of interest at any time t given all the data are then:

$$\hat{x}_i = \mathbb{E}\left[x_i \,|\, \mathcal{F}_T\right] = \int q_i(T, \mu_i) \mu_i \, \mathrm{d}\mu_i \tag{6}$$

$$\hat{N}(t) = \mathbb{E}\left[N(t) \,|\, \mathcal{F}_T\right] = \int \binom{\mathcal{F}(t)}{2} \frac{\vec{q}(T, \mu)}{\lambda(t, \mu, \mathcal{F}(t))} \, \mathrm{d}\mu \tag{7}$$

Note that $\hat{N}(t, \vec{x}) \neq N(t, \hat{\vec{x}})$ by the properties of expectations. The integrals are either over the m-point Cartesian grid of the parameters or the m_i-point line of x_i and the algebraic operations involve vectors of appropriate size. These MMSE estimates and the joint posterior $\vec{q}(T)$ form the core of the results that we present in sections 3.1 - 3.4. We take estimates at T because this makes use of all the information that exists in the event time series. T is also known as the time to the most recent common ancestor (TMRCA) of the coalescent process.

3. Results

We apply Snyder filtering theory to several coalescent inference problems. We start by showing that the constant population size Kingman coalescent inference problem can be solved analytically with Snyder filtering. This enables us to recover the standard maximum likelihood estimator and explicitly derive an approximate MMSE. We also describe the Bayesian sensitivity of this solution by constructing a differential equation that characterises how the posterior estimate changes with choice of prior. We subsequently consider more general deterministically time-varying coalescent processes which cannot be solved analytically. Hence we solve the Snyder filter equations numerically. We simulate genealogies under different multivariate demographic functions and use these models to explore the performance and flexibility of Snyder based inference under both isochronous and heterochronous sampling. Lastly, we apply the Snyder filter approach to a well studied dataset of Egyptian hepatitis C viral sequences and compare our results with those of previous analyses. To keep the comparison fair we focus on a standard parametric method [7] and match our priors and model to that technique.

3.1. Analytic Snyder Filter Solution to the Standard Kingman Coalescent

The standard Kingman coalescent has a constant demographic function $N(t) = x_1$ and is isochronously sampled so that $K = 1$. The Kingman coalescent inference problem is to find the conditional mean $\hat{x}_1 = \mathbb{E}[x_1 \,|\, \mathcal{F}_T]$ where $\mathcal{F}_T$ is the coalescent counting process from 0 to T (the only sample time is at $t = 0$). We show that, for the Kingman coalescent, we can reformulate the data structure of this problem so that the Snyder equations 1 and 3 admit an analytical solution. This reformulation (and analytic treatment) does not work for more complex coalescent models because it rests on manipulating the structure of the rate matrix $\Lambda_{\mathcal{F}(t)}$ such that the matrix loses its dependence on $\mathcal{F}(t)$.

For isochronous sampling of a single n tip tree of $n-1$ coalescent events, the time of the k^{th} coalescent event, c_k corresponds to the end of a coalescent interval that starts with $n-k+1$ lineages (it is the k^{th} branch of the tree). The coalescent rate over this interval is therefore: $\lambda(t, x_1, \mathcal{F}(t)) = \binom{n-k+1}{2} x_1^{-1}$ as $\mathcal{F}(t) = n-k+1$ for $c_{k-1} \leq t < c_k$. In Appendix B we show that instead of using this $n-1$ event data stream, we can alternatively use a single event from $n-1$ independent trees (unlinked loci) and lose no information for inference (for a fixed number of observations).

Let us take the k^{th} coalescent event interval and scale it by the number of lineages existing in that interval so that $\delta_k = \binom{n-k+1}{2}(c_k - c_{k-1})$. Calculate this scaled interval for $n-1$ independent trees and randomly order them as $\{\delta_k(1), \delta_k(2) \ldots \delta_k(n-1)\}$. Then sum them cumulatively to create an observed process $\mathcal{G}(\bar{t}) = \max_{1 \leq i \leq n-1} \sum_i \delta_k(i) \leq \bar{t}$. We use the notation $\bar{t}$ because the time frame is now different and goes from 0 to $\bar{T} \neq T$. In fact T can never be infinite by the properties of the Kingman coalescent while $\bar{T} \to \infty$ as $n \to \infty$. However, note that $\mathcal{F}(t) = \mathcal{G}(\bar{t})$ (t and $\bar{t}$ correspond to the same event count). Each $\delta_k(i)$ is exponentially distributed with the same rate, $\eta = x_1^{-1}$, by the properties of exponential scaling. Consequently, the rate to be inferred is only a function of x_1, $\lambda(x_1)$. It has no dependence on the number of coalescent events or even time. As a result, $\mathcal{G}(\bar{t})$ is no longer self-exciting. Instead it is a simple homogeneous Poisson process. This results in the rate matrix $\Lambda_{\mathcal{G}(\bar{t})}$ being replaced by the constant matrix $\Lambda = \text{diag}\,(\lambda\,(\mu_1))$ where μ_1 is an arbitrary value that x_1 can take. Snyder showed that for a homogeneous Poisson process, equations 1-3 can be analytically solved [20]. The data transformation from $\mathcal{F}_t$ to $\mathcal{G}_{\bar{t}}$ allows this solution, given in equation 8 to be applied to the Kingman coalescent.

$$\mathrm{P}(\eta = z \,|\, \mathcal{G}_{\bar{t}}) = \frac{z^{\mathcal{G}(\bar{t})} e^{-z\bar{t}} \mathrm{P}(\eta = z)}{\int_0^\infty z^{\mathcal{G}(\bar{t})} e^{-z\bar{t}} \mathrm{P}(\eta = z)\, \mathrm{d}z} \tag{8}$$

$$\hat{x}_1 = \int_0^\infty z^{-1} \mathbb{P}(\eta = z \,|\, \mathcal{G}_{\bar{T}})\, \mathrm{d}z \geq \hat{\eta}^{-1} \tag{9}$$

The MMSE estimate of x_1 is given in equation 9. It is the same as the previously defined $\mathbb{E}[x_1 \,|\, \mathcal{F}_T]$, due to the aforementioned information equivalence. The inequality involving $\hat{\eta} = \mathbb{E}[\eta \,|\, \mathcal{G}_{\bar{T}}] = \mathbb{E}[\eta \,|\, \mathcal{F}_T] = \mathbb{E}[x_1^{-1} \,|\, \mathcal{F}_T]$ in the above expression comes from Jensen's inequality. We apply Taylor's expansion to $\hat{\eta}$ and achieve the relation 10.

$$\mathbb{E}\left[\eta \,|\, \mathcal{F}_T\right] \approx \mathbb{E}\left[x_1 \,|\, \mathcal{F}_T\right]^{-1} + \mathrm{var}\left(x_1 \,|\, \mathcal{F}_T\right) \mathbb{E}\left[x_1 \,|\, \mathcal{F}_T\right]^{-3} \tag{10}$$

$$\mathrm{var}\left(x_1 \,|\, \mathcal{F}_T\right) \approx \hat{x}_1^2 \left(\hat{\eta}\hat{x}_1 - 1\right) \tag{11}$$

For a given observed trajectory $\mathcal{F}_T$ with associated scaled process $\mathcal{G}_{\bar{T}}$, the conditional variance $\mathrm{var}\left(x_1 \,|\, \mathcal{F}_T\right) = \mathrm{var}\left(x_1 \,|\, \mathcal{G}_{\bar{T}}\right)$ is equivalent to our calculated MMSE. This allows an approximate expression for the MMSE to be derived. This is given in equation 11. To our knowledge, these are the first explicit MMSE solutions for Bayesian inference on the Kingman coalescent. In the limit of an infinitely long observed coalescent information stream, $\mathcal{G}_{\bar{T}} \to \infty$. At this limit, since equation 8 has $z^{\mathcal{G}(\bar{T})}$ terms then the conditional mean $\hat{\eta}$ would have a numerator term of $z^{\mathcal{G}(\bar{T})+1}$ and $\hat{x}_1$ would have $z^{\mathcal{G}(\bar{T})-1}$, and their difference would be negligible. Consequently, $x_1 \to \eta^{-1}$ so that MMSE $\to 0$, as expected. Similar analytic expressions are not available when the demographic function varies with time because it is not possible to perform exponential scaling so as to remove the count-dependent component of the process.

We now solve equation 8 exactly for a uniform prior. In this case the prior term cancels from the numerator and denominator of this expression and we recognise the Laplace transform definition: $\mathcal{L}_{\bar{t}}[z] := \int_0^\infty z e^{-z\bar{t}}\, \mathrm{d}z$. Using this with equation 9 leads to a straightforward solution (when all the data is used):

$$\hat{x}_1 = \mathcal{L}_{\bar{T}}\left[z^{\mathcal{G}(\bar{T})-1}\right] \mathcal{L}_{\bar{T}}\left[z^{\mathcal{G}(\bar{T})}\right]^{-1} = \bar{T}\mathcal{G}(\bar{T})^{-1} = \bar{T}\mathcal{F}(T)^{-1} \tag{12}$$

Since there are $n-1$ events $\mathcal{F}(T) = n-1$. Further $\bar{T} = \sum_{i=1}^{n-1} \binom{n-k+1}{2} (c_k(i) - c_{k-1}(i))$. This is the same as the sufficient statistic for the Kingman coalescent (see equation 18 of Appendix B). As a result we recover the well known maximum likelihood Kingman coalescent solution: $\hat{x}_1 = \frac{\bar{T}}{n-1} = \hat{x}_1^{(1)}$ [29].

Lastly we provide some Bayesian sensitivity analysis of the Snyder posterior for the Kingman coalescent. Consider the maximum a posteriori (MAP) estimate of the parameters $\vec{x}$ underlying a Poisson process whose rate is only a function of these parameters and time (i.e. with no self excitation). The MAP estimate is the mode of the posterior $\vec{q}(T, \vec{x}) = \mathrm{P}(\vec{x} \,|\, \mathcal{H}_T)$ with $\mathcal{H}_T$ as the observed event stream from a Poisson process that can, in its most general form, be inhomogeneous. Snyder showed that the MAP estimate results by solving [20]:

$$\vec{0} = \frac{\partial}{\partial \vec{x}} \left[\log \vec{q}(0, \vec{x}) - \int_0^t \lambda(\tau, \vec{x})\, \mathrm{d}\tau + \sum_{k=1}^{\mathcal{F}(t)} \log \lambda(c_k, \vec{x}) \right] \tag{13}$$

Here $\frac{\partial}{\partial \vec{x}}$ is the gradient operator, c_k are the event times, and $\vec{q}(0, \vec{x})$ the prior on the parameters of the demographic function. Due to the data transformation previously mentioned, this equation is valid for inference of the Kingman coalescent because the rate is simply $\lambda(x_1) = x_1^{-1}$ when the scaled time stream $\mathcal{G}_{\bar{T}}$ is used as the observed data (observations are until $\bar{T}$). We apply this equation with the notation of $q_0(x_1) = \vec{q}(0, x_1)$ and $q_0(x_1)' = \frac{\partial \vec{q}(0, x_1)}{\partial x_1}$.

$$\bar{T} - (n-1)x_1 + \frac{q_0(x_1)'}{q_0(x_1)} x_1^2 = 0 \tag{14}$$

The solution to the expression above is the MAP estimate x_1^{map}. It clearly depends on the logarithmic derivative of the prior, $L = \frac{q_0(x_1)'}{q_0(x_1)}$. This derivative represents the sensitivity of the Bayesian MAP estimate to the prior. If the prior is uniform, $L = 0$ and $x_1^{\mathrm{map}} = \frac{\bar{T}}{n-1}$ which is the original maximum likelihood estimate of x_1 for the Kingman coalescent. If an inverse prior: $q_0(\mu_1) = \mu_1^{-1}$ is used then $L = \mu_1^{-1}$ with μ_1 as some arbitrary value of x_1 and hence $x_1^{\mathrm{map}} = \frac{\bar{T}}{n}$. Most importantly, as n (and hence $\bar{T}$) gets large then for any prior the logarithmic derivative term becomes relatively negligible and all MAP estimates converge to the maximum likelihood estimate.

3.2. Snyder MMSE Estimation of Simulated Demographic Models

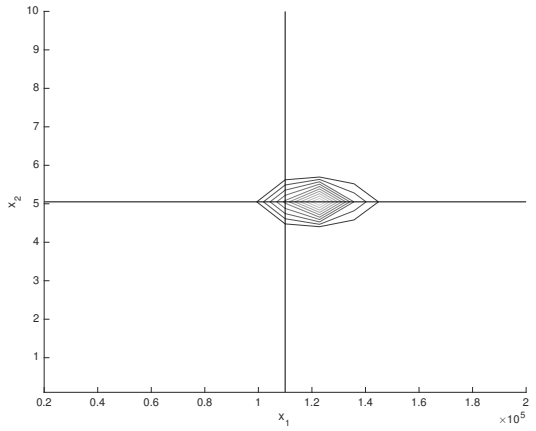
We now apply the Snyder filter to coalescent data simulated under several deterministically time-varying demographic models that have been previously used in the epidemiological literature. Our aim is to estimate the underlying

parameters of these models. As noted above (sections 2.2 and 3.1), no analytic solution exists for these models so instead we discretise our parameter space and then numerically solve equations 1-3. As developed in section 2.2, a demographic function with $l \geq 1$ parameters is denoted $N(t, \vec{x})$ (or just $N(t)$ for short). The i^{th} parameter, x_i, is inferred over a discrete space of m_i values. This search space forms the domain on which the resulting marginal posterior $\vec{q}(T, x_i) = \mathrm{P}(x_i \,|\, \mathcal{F}_T)$ is defined. The observed stream $\mathcal{F}_T$ is a set of $n-1$ simulated event times obtained from an appropriate n tip coalescent tree. The Snyder filter iteratively processes the coalescent events and directly yields the joint posterior $\vec{q}(T, \vec{x}) = \mathrm{P}(\vec{x} \,|\, \mathcal{F}_T)$. The MMSE conditional estimates of the parameters $\hat{x}_i$ and of the overall demographic history $\hat{N}(t)$ are obtained by either marginalising $\vec{q}(T, \vec{x})$ or an appropriate function of it.

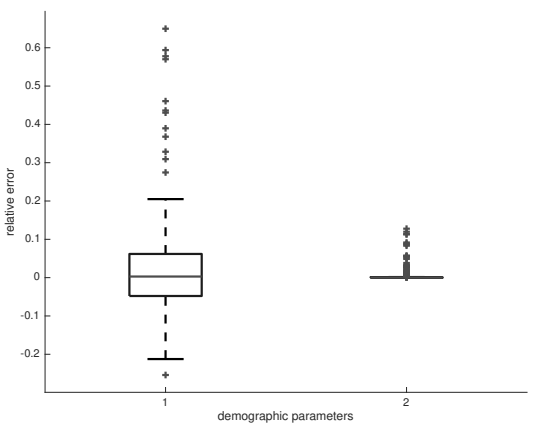
We simulate and analyse three demographic functions. In each case isochronous sampling is assumed so there are $v = n - k + 1$ lineages existing between the $(k-1)^{\text{th}}$ and k^{th} coalescent event. The first model is exponential growth $N(t) = x_1 e^{-x_2 t}$ [30], which describes a rapidly growing population (in forward time). Here $x_1 = N(0)$ is the population size at the present and x_2 is the exponential growth rate. We used a time rescaling algorithm to iteratively generate the k^{th} coalescent time, c_k, from the $(k-1)^{\text{th}}$ coalescent time: $c_k = r^{-1} \log\left(e^{r c_{k+1}} + r z x_1 \binom{v}{2}^{-1}\right)$, $z \sim \exp(1)$. The second model is logistic growth, which represents density dependent population growth and is defined as $N(t) = x_1 \frac{1+e^{-x_2 x_3}}{1+e^{-x_2(x_3 - t)}}$ [23]. An offset parameter, x_4, is added so that $N(t) \geq 0$ for all $t > 0$. Parameters x_1 and x_2 are defined as in the exponential growth model and x_3 is a half life parameter. The logistic model was simulated using a rejection sampling algorithm that generates points from a homogeneous process with rate $\lambda_{\max} = \binom{n}{2} {x_4}^{-1}$ and then chooses one as the next coalescent event with acceptance probability $p_{(v,t)} = \binom{v}{2}(N(t)\lambda_{\max})^{-1} \geq v(v-1)x_4 \left(n(n-1)(x_1 + x_4)\right)^{-1}$. The third model represents cyclical population dynamics using a sinusoid defined by the function $N(t) = x_1 \sin(x_2 t + x_3) + x_4$. Here x_1 is the cycle amplitude and x_2 its frequency of occurrence, with phase and magnitude offsets given by x_3 and x_4 respectively. Data from this model was also generated using a rejection sampling algorithm with $\lambda_{\max} = \binom{n}{2}(x_4 - x_1)^{-1}$ and $p_{(v,t)} = \binom{v}{2}(N(t)\lambda_{\max})^{-1} \geq v(v-1)(x_4 - x_1)\left(n(n-1)(x_4 + x_1)\right)^{-1}$.

For every demographic model, we simulated a total of $M = 500$ trees, each with $n = 200$ tips and then applied the Snyder filter to the coalescent times of each tree to obtain informed joint posterior distributions. We set the true values of each parameter at approximately the midpoint of its parameter space so that it would be well described by the prior. We chose an uninformative uniform prior over the grid space with probability mass of $\frac{1}{m}$ at each point. We define T_j (the TMRCA) as the final observation time for the j^{th} tree. Rather than calculating the MMSE of the i^{th} parameter, we instead calculate the relative error $1 - \frac{\hat{x}_i}{x_i}$ in order to evaluate both the bias and variance of the estimator. We summarise the results from the Snyder filter in figures below. If needed, the MMSE can be calculated with the frequentist approximation $M^{-1} \sum_{j=1}^{M} \left(x_i(T_j) - \hat{x}_i(T_j)\right)^2$.

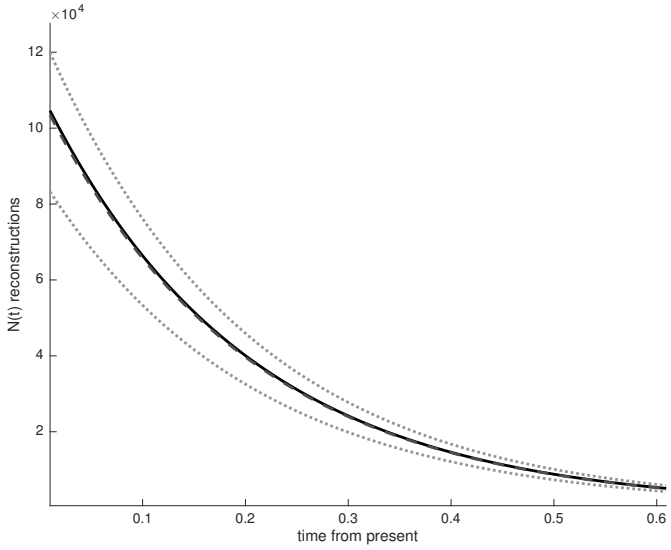
Figures 1a, 2a and 3a give representative smoothed posteriors from a single Snyder run for the exponential, logistic and sinusoidal models respectively. For the exponential model, the full joint posterior is given. They provide an idea of what can be inferred from a a single tree by the Snyder filter. The posteriors often seem well positioned relative to the true value of the parameter and provide a good cover. Figures 1c, 2c and 3c summarise Snyder demographic estimates from the full M runs. In this case a demographic function estimate is generated from each tree and the resulting trajectories combined and summarised by its mean and 95% credible interval. The Snyder filter estimates each $N(t)$ well with the true trajectory well tracked by the mean and easily covered by the credible interval. Figures 1b, 2b and 3b also focus on all M runs and present an idea of the bias and uncertainty in the relative error on each estimated parameter. For all models the boxplots of these errors seem centred on 0 and so the method appears unbiased. The uncertainty on each parameter is at most about the order of 10% away from the true value (relative errors of ± 0.1).



(a) Joint posteriors from single run

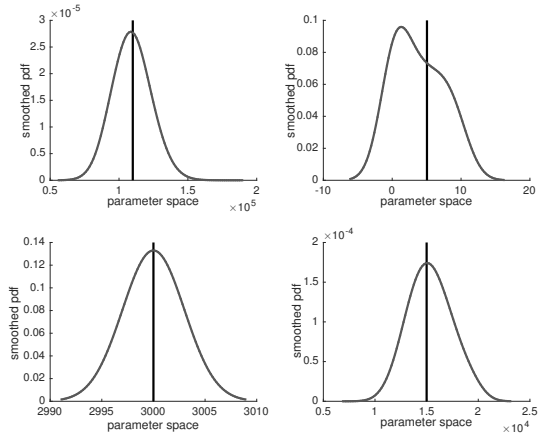


(b) Relative estimate error boxplot

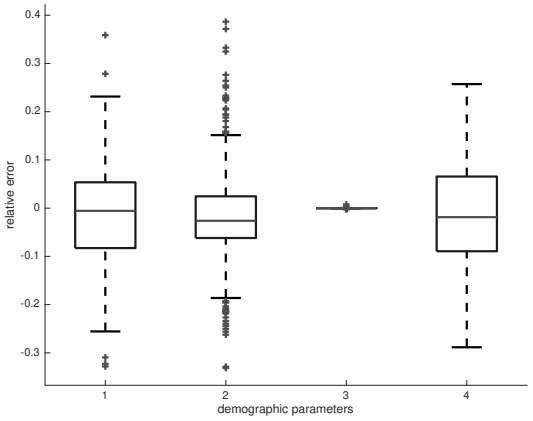


(c) Combined demographic estimate from M runs

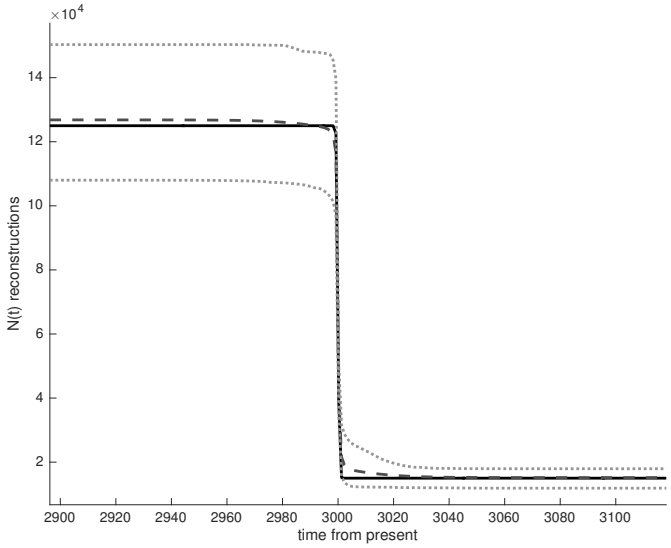
Figure 1: **Snyder estimates under the exponential growth coalescent model:** $N(t) = x_1 e^{-x_2 t}$**.** a) Contour plot of the joint posterior for the two model parameters, x_1 and x_2 from an individual run. Contour lines show values of the posterior P(x_1, x_2 | data). Thick grey lines show the true values of the two parameters. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^2, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.



(a) Smoothed posteriors from single run

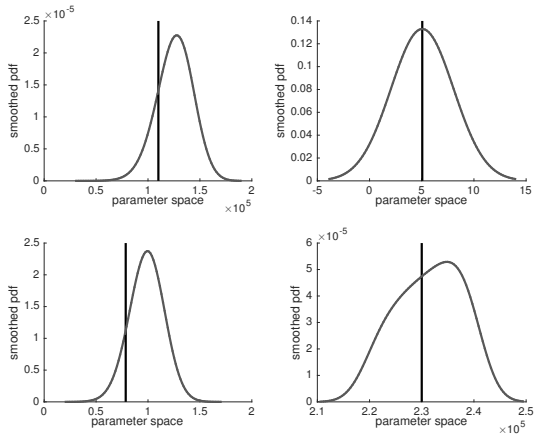


(b) Relative estimate error boxplot

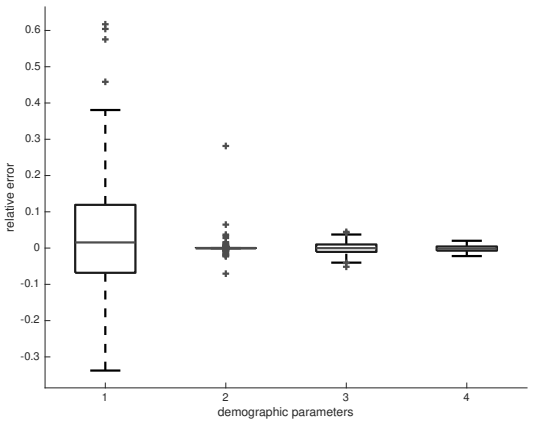


(c) Combined demographic estimate from M runs

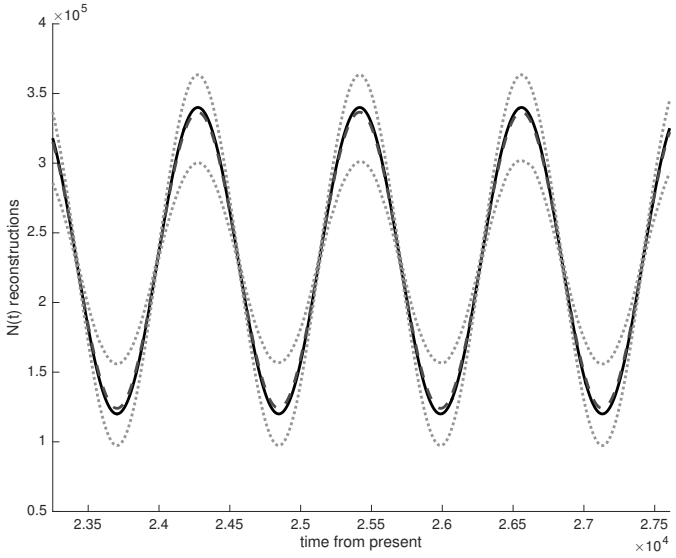
Figure 2: **Snyder estimates under the logistic growth coalescent model:** $N(t) = x_1 \frac{1+e^{-x_2 x_3}}{1+e^{-x_2(x_3-t)}} + x_4$**.** a) Marginal posteriors for the model parameters, x_1 and x_2 from an individual run. The grey outline shows values of the posterior $\mathrm{P}(x_i \,|\, \mathrm{data})$ after kernel smoothing with a normal distribution. The black vertical line is the true value of x_i. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^4, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.



(a) Smoothed posteriors from single run



(b) Relative estimate error boxplot



(c) Combined demographic estimate from M runs

Figure 3: **Snyder estimates under a sinusoidal coalescent model:** $N(t) = x_1 \sin(x_2 t + x_3) + x_4$**.** a) Marginal posteriors for the model parameters, x_1 and x_2 from an individual run. The grey outline shows values of the posterior $\mathrm{P}(x_i \,|\, \mathrm{data})$ after kernel smoothing with a normal distribution. The black vertical line is the true value of x_i. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^4, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.

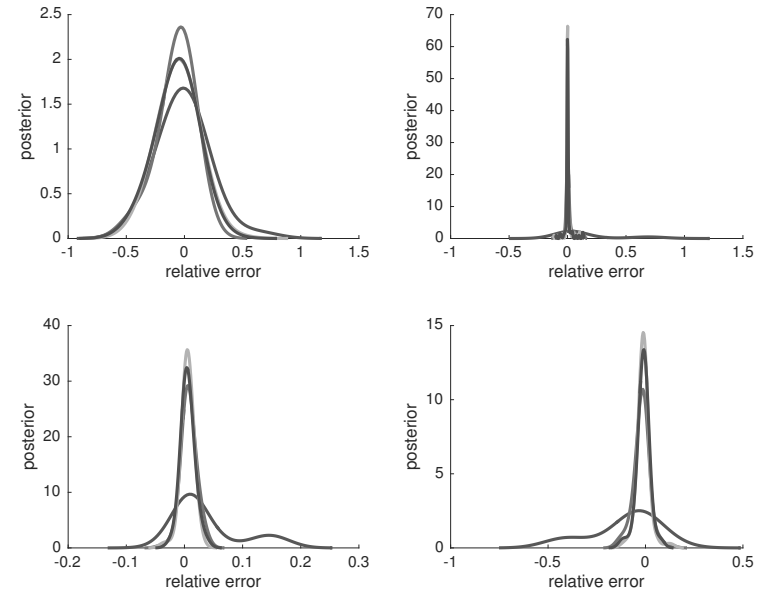
3.3. Extension to Demographic Models with Heterochronous Sampling

The simulations in section 3.2 considered only isochronous sampling in which all the sequences were sampled only at the present ($t = 0$). Here we show how Snyder filtering can be applied to the heterochronous coalescent process. We examine a classical demographic function that models delayed rapid growth forward in time. The constant-exponential-constant (con-exp-con) model [7]: $N(t) = x_1 1(t \leq x_3) + x_1 e^{-x_2(t-x_3)} 1(t \in (x_3, x_4)) + x_1 e^{-x_2(x_4-x_3)} 1(t \geq x_3 + x_4)$. $N_2(t)$ is simply a translated and time limited version of the exponential growth model [30]: $N_{\text{exp}}(t) = x_1 e^{-x_2 t}$ from section 3.2. In fact $N_{\text{exp}}(t) = N_2(t)|_{x_3 \to 0,\, x_4 \to \infty}$. The con-exp-con function can also be considered as a piecewise equivalent of the logistic function from section 3.2 with parameters x_3 and x_4 controlling the times at which the population switches between phases of constant size and exponential growth. We chose the con-exp-con function because its multiple phases mean that sample timing could be an important factor in its inference. For example if no coalescent events occurred during a phase then there would be a dearth of information about this phase. This is one of the reasons heterochronous sampling is useful. It spreads the coalescent observations across time in a more uniform manner. This especially helps since coalescent trees often have regions with a few long branches (long times between informative events).

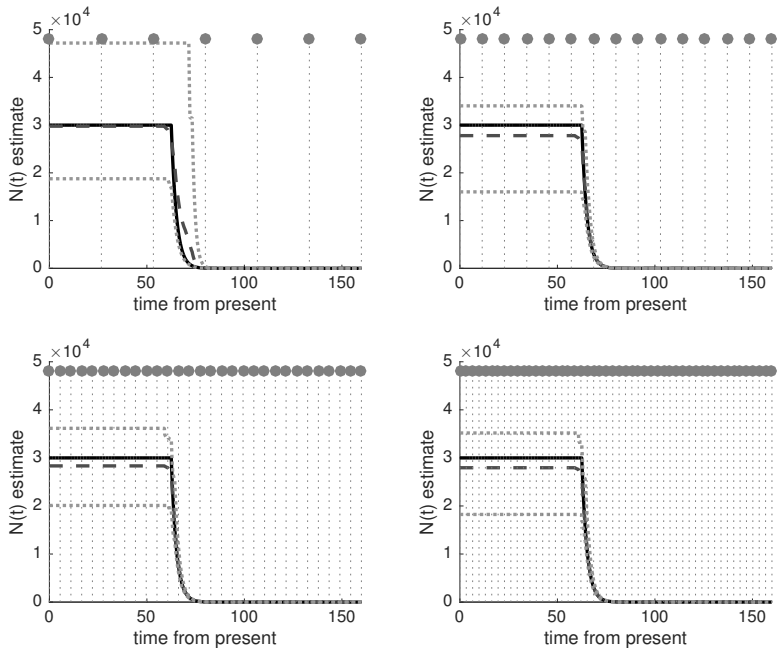
We modify the time rescaling technique in [12] to simulate coalescent trees under heterochronous sampling. The key step of this algorithm involves solving for c_{next}, the time of the next coalescent event. If c_{next} is less than the next sample time then the lineage count is reduced by 1 since the next event is a coalescent one. If c_{next} is larger than the next sampling time, then the time is set to that sampling time and the number of lineages is updated with the new sample count. The process of calculating c_{next} is then restarted. If the number of lineages falls to 1 before the last sampling time then the coalescent rate is set to 0 until the next sampling time and the posterior distribution stays unchanged over this interval (see section 2.2 for the mathematical expression of this condition). The specific time rescaling solution for the con-exp-con model is given below with $z \sim \exp(1)$.

$$c_{\text{next}} = \begin{cases} t + x_1 z \binom{\mathcal{F}(t)}{2}^{-1}, & \text{if } t \leq x_3 \\ t + x_1 e^{-x_2 x_4} z \binom{\mathcal{F}(t)}{2}^{-1}, & \text{if } t \geq x_3 + x_4 \\ x_3 + x_2^{-1} \log\left(e^{x_2(t-x_3)} + x_1 x_2 z \binom{\mathcal{F}(t)}{2}^{-1} \right), & \text{otherwise} \end{cases}$$

We simulated the con-exp-con coalescent using n sampled lineages that are introduced at times $s_i = (i-1) T_{\text{samp}} K^{-1}$, $1 \leq i \leq K$, until some maximum sampling time T_{samp}. At each sampling time $n_i = nK^{-1} = n^*$ lineages were added (to the nearest integer). The figures below illustrate the performance of Snyder filter estimation when applied to 100 heterochronous trees, sampled at K uniform times. We use a maximum sampling time of $T_{\text{samp}} = x_3 + (x_3 + x_4)$, so that the sampling times are symmetric with respect to the period of exponential growth. Figure 4a gives the smoothed density of the relative conditional estimate error across the runs for each parameter. For example the relative error contribution of the i^{th} parameter from the j^{th} run is $x_i^{-1}(x_i - \mathbb{E}[x_i \,|\, \mathbb{F}_T^j])$. For every parameter it is clear that the densities are approximately centred around 0 indicating unbiased estimation. The lowest K case has the widest densities. This relates to the fact that there are fewer observable events across some of the con-exp-con phases. Figure 4b summarises the estimates of $N(t)$ across the runs with the mean reconstruction and its 2.5% and 97.5% quantiles. The con-exp-con function is well estimated across each sampling scheme as the true trajectory is always within the 95% credible interval and accurately tracked by the mean combined estimate. In the simulations below we used a joint uniform prior with probability mass of m^{-1} on each of the m grid points.



(a) Smoothed density of parameter estimates



(b) Demographic reconstruction summary

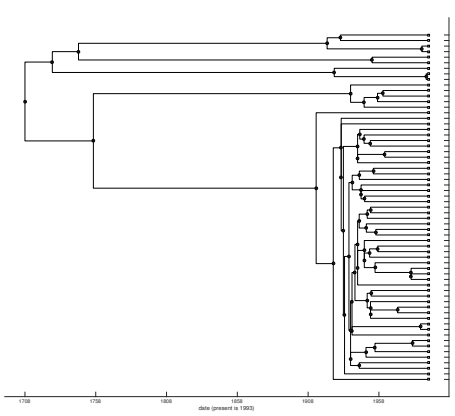
Figure 4: **Snyder estimates under heterochronously sampled con-exp-con models.** A con-exp-con model was simulated with uniform heterochronous sampling at $K = [7, 15, 30, 50]$ distinct times for trees with $n \approx 200$ tips. Each sampling introduces $n^* = \frac{n}{K}$ samples to the nearest integer. The simulation was repeated 200 times and the conditional mean estimate taken from each run. Subfigure a) gives the smoothed density (normal kernel) of the relative error, $1 - \frac{\hat{x}_i}{x_i}$ on these estimates, for each parameter. Panel b) gives a summary of reconstructed demographic functions from these runs. The true demographic function is in solid grey and the estimated function in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. The dotted vertical stems are the sample times. Simulations were done at: $[m_i, m, M] = [15, 15^2, 100]$ with a joint prior of $\frac{1}{m}$ on each grid point.

3.4. Estimation of the Hepatitis C (HCV) Epidemic in Egypt

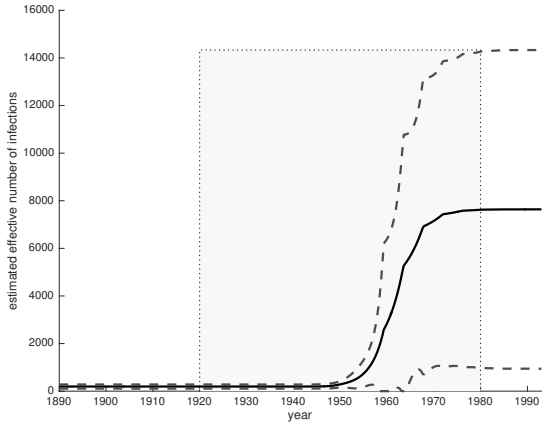
We now illustrate the application of the Snyder filter to empirical data. One particular dataset, comprising hepatitis C virus (HCV) gene sequences sampled in Egypt, has been repeatedly used as a benchmark for the performance of parametric and non-parametric coalescent estimators in molecular epidemiology (see [28] [31] [32] [33] [34]). In addition to allowing different inference approaches to be compared, a further benefit of this dataset is that its demographic history is partially known, through independent epidemiological information. Specifically, the high current prevalence of HCV in Egypt is very likely due to the rapid transmission of the virus during a widespread public health campaign of parenteral antischistosomal therapy (PAT). Poor sterilisation of needles used to deliver this drug treatment led to the inadvertent transmission of HCV [35]. Since these injections were primarily given between 1920 and 1980, we expect coalescent inference methods to reconstruct a rapid rise in HCV effective population size during this period. The Egyptian HCV dataset comprises 63 HCV genotype 4 gene sequences, 411 nucleotides in length, sampled isochronously in 1993 [36]. In previous work [7] a con-exp-con model was fitted to this dataset using a Bayesian MCMC approach. To enable direct comparison, we apply the Snyder filter to the same demographic function. The con-exp-con demographic model is the same as that introduced in section 3.3, but with the indicator functions expanded. Setting $x_1 = N_C$, $x_2 = r$, $x_3 = x$ and $x_4 = y - x$, this demographic model can be written as in [7] with $t > 0$ describing time in the past from 1993 (the time of sampling).

$$N(t) = \begin{cases} N_C, & \text{if } t \leq x \\ N_C e^{-r(t-x)}, & \text{if } x < t < y \\ N_A = N_C e^{-r(y-x)}, & \text{if } t \geq y \end{cases}$$

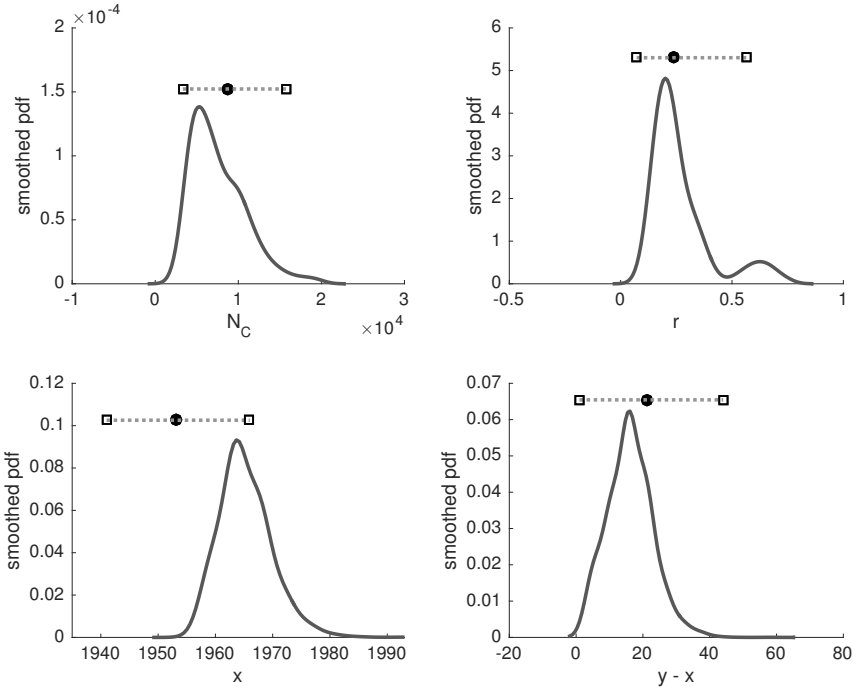
To obtain coalescent times we estimated a molecular clock phylogeny from the Egypt HCV dataset as follows. First, we used Garli [37] to estimate a maximum likelihood (ML) phylogeny of the sequences under the general time reversible nucleotide substitution model with gamma distributed site rate heterogeneity. This was the nucleotide substitution model previously used in [7]. The heuristic search implemented by Garli was run 50 times and the tree with the highest likelihood retained. Next, we used the program R8s [38] to convert the ML tree into an ultrametric phylogeny whose branches are scaled in units of years. This conversion was performed using a Langley-Fitch 3-rate molecular clock mode. Divergence times were estimated by constraining the TMRCA of the tree to the range of values reported in [7]. The resulting time-scaled tree is shown in Figure 5a. The coalescent times from this tree formed the input data to the Snyder filter with $[m_i, m] = [20, 20^4]$. Prior distributions were set to match those of [7] as closely as possible. The resulting marginalised posterior distributions are given in Figures 5c-f and are compared to the posteriors obtained by Bayesian MCMC sampling in [7]. The Snyder MMSE estimate of the demographic function is shown in Figure 5b. It is clear that the Snyder filter gives results that are consistent with the known epidemic history of HCV in Egypt. It also achieves parameter estimates that compare well with those obtained by MCMC sampling in [7] (see the legend of Figures 5c-f).



(a) Maximum likelihood HCV tree



(b) Demographic reconstruction of epidemic



(c) Smoothed density of parameter estimates

Figure 5: **Snyder estimates of the HCV epidemic showing exponential growth in the early 20$^{\text{th}}$ century.** (a) The ultrametric time-scaled tree derived from the 63 sequence HCV dataset using Garli and R8s. The branching times of this tree provide the coalescent event times for Snyder filtering. (b) The Snyder estimate of HCV population size using a con-exp-con demographic model. The continuous black line is the conditional mean estimate while the dotted lines are delimited by twice the standard deviation of its posterior. Rapid growth during the PAT period (shaded) is clear. (c) Smoothed marginal posteriors for each parameter of the con-exp-con demographic model (solid black line) (normal kernel smoothing). The Snyder filter used: $[m_i, n] = [20, 63]$ with uniform priors set to $\frac{1}{m_i}$ for each parameter. The parameter estimates from [7], which were obtained by Bayesian MCMC sampling, are shown as box-type plots (median estimate, and 95% highest posterior density credible intervals) above each marginal posterior.

4. Discussion and Conclusion

In this paper we introduced the Bayesian Snyder filter as a viable technique for coalescent inference and we evaluated its performance on simulated and empirical data. We showed that by reinterpreting the variable population size coalescent process as a self-exciting Markov process, it is possible to apply the Snyder filter to coalescent estimation problems. Further, for the standard Kingman coalescent, we derived an analytic solution for the Synder MMSE estimator, which was equivalent to the known maximum likelihood estimate. We also explored the Bayesian sensitivity of the filter for this case and found that sensitivity decreased as the observable data stream length, n, increased.

We applied the Snyder filter to isochronous phylogenies simulated under three different parametric demographic models. Two of these models (exponential and logistic) are commonly used to represent emerging epidemics whilst the third model (sinusoidal) represents the cyclical epidemic behaviour exhibited seasonal viruses like influenza. In all cases, the Snyder filter correctly inferred the underlying dynamics and achieved good MMSE estimates of the model parameters which were unbiased. We also showed that Snyder filtering could be readily extended to heterochronous sampling and applied this extension to two nested epidemic growth models. Estimates were once again found accurate both at the parameter and complete demographic function scales. We also note in Appendix E that the Snyder filter is implicitly calculating exactly the well known deterministically time-varying coalescent likelihood from the literature [2] [12]. Consequently, the Snyder filter solves, in a sequential MMSE manner, the correct inference problem and is hence verified as a useful coalescent estimation tool.

We then assessed performance of Snyder filter inference on empirical data from the Egyptian HCV epidemic. In keeping with previous analyses on this dataset, we used a parametric con-exp-con demographic model. The Snyder filter estimates compared well with those previously obtained using Bayesian MCMC sampling [7]. It is worth noting that the Bayesian MCMC sampling approach incorporated uncertainty in the phylogeny and the molecular clock model, whereas we assumed that the coalescent event times were known without error and used the best ML tree over 50 runs. However, previous authors have concluded that phylogenetic uncertainty is not significant in the estimation of Egyptian HCV transmission history. Consequently, our comparison seems valid and fair. In fact Minin *et al* [32] found that estimates from a single fixed genealogy bore little difference to those obtained directly from the molecular sequences of this HCV dataset. We also apply our single tree method to this dataset and compare it to known MCMC results in order to be consistent with the literature [18]. For other sequence alignments that contain comparatively less phylogenetic information, further work is necessary to understand how genealogical sources of error can be best incorporated into the Snyder filter framework.

The Snyder filter has the potential to be a capable inference method for coalescent models. Since it only involves the solution of linear differential equations, it is easy to implement and provides stable estimates (i.e. the filter is generally not susceptible to numerical explosion or lack of convergence). Further, because it allows straightforward calculation of conditional means, it easily leads to optimal (MMSE) estimates. Hence the Snyder filter may serve as a useful benchmark for testing the limits of other estimation schemes, and for providing bounds on achievable estimation accuracy [22]. A key benefit of the Snyder filter as an inference framework is its flexibility. For example, in our simulation study, the Snyder filter algorithm remained unchanged across all the different demographic models and data types (isochronous and heterochronous) that we explored. The only changes to implementation that were required were the specification of the demographic function and its priors. Given its flexibility we propose that the Snyder filter could be useful for model selection problems.

The Snyder filter, from a theoretical perspective, is also the right type of filter to solve optimal coalescent inference problems. If our system was linear, discrete-time and corrupted by Gaussian noise, the optimal (MMSE) filter would be the famous Kalman filter [39]. The Snyder filter is the Kalman analogue for continuous time processes with point process noise. This paper will show that coalescent inference is within this class of problems. One may wonder about the relation of the Snyder technique to popular algorithms such as the expectation-maximisation (EM) [40] or sequential Monte-Carlo (SMC) [41]. The EM algorithm is usually used to solve estimation problems with incomplete information. This is a more complex set of problems than those considered here and so this technique is not necessary for this work. More importantly when the EM algorithm is adapted for filtering problems it usually employs a Kalman filter within its mechanism [42]. Consequently, it may be possible to integrate the Snyder filter within an EM framework to solve more complex problems. SMC is maybe a more relevant method for the problems investigated in this work. It applies simulation based methods (just as in MCMC) to solve online (data is sequentially available) estimation problems for systems that do not admit finite dimensional filters. However, the Snyder filter is finite

dimensional for coalescent inference problems and treats the tree as a sequential data source. As a result, we propose the Snyder filter as an alternative to this technique as well.

While its implementation is linear, the Snyder filter is a non-linear MMSE filter (its linear equations must be normalised; see Appendix A). The importance of this is apparent when the parametric Snyder filter is compared to the non-parametric classic skyline plot [23]. The classic skyline plot is a common non-parametric coalescent inference technique that uses coalescent waiting times to estimate the harmonic mean of the demographic function. In Appendix C, we show that the classic skyline plot is a minimal information linear estimator of the coalescent process. This can be improved to a linear MMSE estimator if more information is introduced via first and second order process statistics (mean and covariance functions). If this linearity constraint is maintained, then no further improvements (defined as reductions in estimation MSE) are possible, even if more coalescent process information is available [24]. At this point the MSE can be further minimised only by allowing for non-linear filter formulations that make use of extra information, usually embedded in the form of some demographic model structure (parameters). The parametric Snyder filter is exactly the non-linear MMSE estimator that makes maximum use of this additional coalescent process information. These connections hint at why we think the Snyder filter could serve as a useful benchmark estimator.

The filter was originally developed for doubly stochastic Poisson processes [24]. As a result, it should be capable of handling stochastic demographic functions, or input streams with multiple types of stochastic events (for example coalescent events coloured by geographical area of origin). The Snyder filter could also be extended to use alternate descriptors of the relation between population dynamics and phylogenetic structure, such as birth-death models [27] or even diffusion processes [20]. Extending the filter to these cases will form the basis of future research. The flexibility, exactness and robustness of the Snyder filter suggest it has promise, but its benefits will become most apparent when it is applied to more complex models that are computationally challenging to implement in other frameworks, such as Bayesian MCMC sampling.

We have provided some example code on GitHub at `https://github.com/kpzoo/snyder-coalescent-code`. These are a set of Matlab m files that were used to generate all the results for the isochronous and heterochronous simulated models used in this work. The code generates a coalescent process for a given demographic functions and then runs a Snyder filter to infer its underlying parameters. We provide this code so users can get a feel for how simple it is to implement the Snyder filter along with an idea of its complexity and ease of computation.

Acknowledgements

This work was supported by the European Research Council under the European Commission Seventh Framework Programme (FP7/2007-2013)/European Research Council grant agreement 614725-PATHPHYLODYN.

5. Bibliography

[1] J. Kingman, On the Genealogy of Large Populations, Journal of Applied Probability 19 (1982) 27–43.
[2] R. Griffiths, S. Tavare, Sampling Theory for Neutral Alleles in a Varying Environment, Phil Trans R Soc B 344 (1994) 403–10.
[3] I. Kaj, S. Krone, The Coalescent Process in a Population with Stochastically Varying Size, J. Appl. Prob 40 (2003) 33–48.
[4] M. Notohara, The Coalescent and the Genealogical Process in Geographically Structured Population, J Math Biol 29 (1990) 59–75.
[5] A. Rodrigo, U. Shpaer, E. Delwart, et al., Coalescent Estimates to HIV-1 Generation Time in vivo, PNAS 96 (5) (1990) 2187–91.
[6] K. Strimmer, O. Pybus, Exploring the Demographic History of DNA Sequences using the Generalized Skyline Plot, Mol. Biol. Evol 18 (12) (2001) 2298–305.
[7] O. Pybus, A. Drummond, T. Nakano, et al., The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach, Mol. Biol. Evol 20 (3) (2003) 381–7.
[8] D. Rasmussen, M. Boni, K. Koelle, Reconciling Phylodynamics with Epidemiology: the case of Dengue Virus in Southern Vietnam, Mol. Biol. Evol 2 (31) 258–71.
[9] J. Kingman, Origins of the Coalescent: 1974–1982, Genetics 156 (2000) 1461–3.
[10] M. Nordberg, Handbook of Statistical Genetics: Coalescent Theory, John Wiley and Sons, 2001.
[11] P. Diggle, J. Mateu, H. Clough, A Comparison between Parametric and Non-parametric Approaches to the Analysis of Replicated Spatial Point Patterns, Advances in Applied Probability 32 (2000) 331–43.
[12] J. Palacios, V. Minin, Gaussian Process-Based Bayesian Nonparametric Inference of Population Trajectories from Gene Genealogies, Biometrics 69 (2013) 8–18.
[13] Z. Yang, Molecular Evolution: A Statistical Approach., Oxford University Press, 2014, Ch. Coalescent Theory and Species Trees.
[14] M. Kuhner, J. Yamato, J. Felsenstein, Estimating Effective Population Size and Mutation Rate from Sequence Data using Metropolis-Hastings Sampling, Genetics 140 (1995) 1421–30.
[15] M. Kuhner, Coalescent Genealogy Samplers: Windows in Population History, Trends in Ecology and Evolution 24 (2) (2008) 86–93.

[16] N. De Maio, C. Wu, K. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, PLoS Genetics 11 (8) (2015) e1005421.
[17] J. Kim, M. E, M. Racz, N. Ross, Can one Hear the Shape of a Population History?, Theoretical Population Biology 100 (2015) 26–38.
[18] J. Palacios, V. Minin, Integrated Nested Laplace Approximation for Bayesian Nonparametric Phylodynamics, Proceedings of the Twenty-Eighth International Conference on Uncertainty in Artificial Intelligence, 2012, pp. 726–35.
[19] A. Drummond, G. Nicholls, A. Rodrigo, W. Solomon, Estimating Mutation Parameters, Population History and Genealogy Simultaneously from Temporally Spaced Sequence Data, Genetics 161 (2002) 1307–20.
[20] D. Snyder, Filtering and Detection for Doubly Stochastic Poisson Processes, IEEE Transactions on Information Theory 18 (1972) 91–102.
[21] O. Bobrowski, R. Meir, Y. Eldar, Bayesian Filtering in Spiking Neural Networks; Noise, Adaptation and Multisensory Integration, Neural Computation 21 (2008) 1277–1320.
[22] K. Parag, Point Process Noise in Fundamental Molecular Reactions and Invertebrate Vision, Ph.D. thesis, University of Cambridge (2014).
[23] O. Pybus, A. Rambaut, P. Harvey, An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies, Genetics 155 (2000) 1429–37.
[24] D. Snyder, M. Miller, Random Point Procresses in Time and Space, 2nd Edition, Springer-Verlag, 1991.
[25] M. Rudemo, Doubly-Stochastic Poisson Processes and Process Control, Advances in Applied Probability 2 (1972) 318–338.
[26] M. Davis, Piecewise-Deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models, J. R. Statist. Soc. B 46 (3) (1984) 353–88.
[27] S. Nee, R. May, P. Harvey, The Reconstructed Evolutionary Process, Phil Trans R Soc B 344 (1994) 305–11.
[28] A. Drummond, A. Rambaut, B. Shapiro, O. Pybus, Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences, Mol. Biol. Evol 22 (5) (2005) 1185–92.
[29] J. Felsenstein, Estimating Effective Population Size from Samples of Sequences: Inefficiency of Pairwise and Segregating Sites as compared to Phylogenetic Estimates, Genet Res 59 (1992) 139–47.
[30] M. Slatkin, R. Hudson, Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations, Genetics 129 (1991) 555–62.
[31] J. Heled, A. Drummond, Bayesian Inference of Population Size History from Multiple Loci, BMC Evolutionary Biology 8 (209).
[32] V. Minin, E. Bloomquist, M. Suchard, Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics, Mol. Biol. Evol 25 (7) (2008) 1459–71.
[33] R. Opgen-Rhein, L. Fahrmeir, K. Strimmer, Inference of Demographic History from Genealogical Trees using Reversible Jump Markov Chain Monte Carlo, BMC Evolutionary Biology 5 (6).
[34] T. Stadler, D. Kuhnert, S. Bonhoeffer, A. Drummond, Birth-death Skyline Plot reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (hcv), PNAS 110 (1) (2013) 228–33.
[35] C. Frank, M. Mohamed, T. Strickland, et al., The Role of Parenteral Antischistosomal Therapy in the Spread of Hepatitis C Virus in Egypt, The Lancet 355.
[36] S. Ray, R. Arthur, A. Carella, et al., Genetic Epidemiology of Hepatitis C Virus throughout Egypt, J. Infect. Dis 182 (2000) 698–707.
[37] D. Zwickl, Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion, Ph.D. thesis, University of Texas at Austin (2006).
[38] M. Sanderson, R8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock, Bioinformatics 19 (2) (2003) 301–2.
[39] R. Kalman, A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engineering 82 (1960) 35–45.
[40] A. Dempster, N. Laird, D. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society 39 (1) (1977) 1–38.
[41] A. Doucet, J. de Freitas, N. Gordon, An Introduction to Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York, 2001.
[42] Z. Ghahramani, G. Hinton, Parameter Estimation for Linear Dynamical Systems, Tech. rep., Dept. Comp. Sci. Univ. Toronto (1996).
[43] P. Taylor, L. Jonker, Evolutionarily Stable Strategies and Game Dynamics, Mathematical Biosciences 40 (1978) 145–56.
[44] M. Harper, The Replicator Equation as an Inference Dynamic, arXiv (2010) 0911.1763.
[45] C. Shalizi, Dynamics of Bayesian Updating with Dependent Data and Misspecified Models, Electron. J. Stat 3 (2009) 1039–74.
[46] D. Snyder, Information Processing for Observed Jump Processes, Information and Control 22 (1973) 69–78.
[47] M. Cowles, B. Carlin, Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, Journal of the American Statistical Association 91 (1996) 883–904.
[48] S. Lan, J. Palacios, M. Karcher, et al., An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics, Bioinformatics 31 (20) (2015) 3282–89.

Figure Legends

Figure 1: **Snyder estimates under the exponential growth coalescent model:** $N(t) = x_1 e^{-x_2 t}$**.** a) Contour plot of the joint posterior for the two model parameters, x_1 and x_2 from an individual run. Contour lines show values of the posterior $P(x_1, x_2 \,|\, \text{data})$. Thick grey lines show the true values of the two parameters. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories,

with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^2, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.

Figure 2: **Snyder estimates under the logistic growth coalescent model:** $N(t) = x_1 \frac{1+e^{-x_2 x_3}}{1+e^{-x_2(x_3-t)}} + x_4$**.** a) Marginal posteriors for the model parameters, x_1 and x_2 from an individual run. The grey outline shows values of the posterior $P(x_i \,|\, \text{data})$ after kernel smoothing with a normal distribution. The black vertical line is the true value of x_i. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^4, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.

Figure 3: **Snyder estimates under a sinusoidal coalescent model:** $N(t) = x_1 \sin(x_2 t + x_3) + x_4$**.** a) Marginal posteriors for the model parameters, x_1 and x_2 from an individual run. The grey outline shows values of the posterior $P(x_i \,|\, \text{data})$ after kernel smoothing with a normal distribution. The black vertical line is the true value of x_i. This plot shows typical estimates from a single simulated tree. (b) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$, measured across the 500 replicate trees with 200 tips simulated under exponential growth. (c) The averaged reconstruction of the estimated demographic function, obtained from the trees in b). The true demographic function is in solid grey and the estimated one in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. Simulations were done at: $[m_i, m, n, M] = [15, 15^4, 200, 500]$ with a joint prior of $\frac{1}{m}$ on each grid point.

Figure 4: **Snyder estimates under heterochronously sampled con-exp-con models.** A con-exp-con model was simulated with uniform heterochronous sampling at $K = [7, 15, 30, 50]$ distinct times for trees with $n \approx 200$ tips. Each sampling introduces $n^* = \frac{n}{K}$ samples to the nearest integer. The simulation was repeated 200 times and the conditional mean estimate taken from each run. Subfigure a) gives the smoothed density (normal kernel) of the relative error, $1 - \frac{\hat{x}_i}{x_i}$ on these estimates, for each parameter. Panel b) gives a summary of reconstructed demographic functions from these runs. The true demographic function is in solid grey and the estimated function in dashed black. Dotted lines are uncertainty bounds capturing 95% of the trajectories, with each tree yielding a single estimated trajectory. The dotted vertical stems are the sample times. Simulations were done at: $[m_i, m, M] = [15, 15^2, 100]$ with a joint prior of $\frac{1}{m}$ on each grid point.

Figure 5: **Snyder estimates of the HCV epidemic showing exponential growth in the early** 20^{th} **century.** (a) The ultrametric time-scaled tree derived from the 63 sequence HCV dataset using Garli and R8s. The branching times of this tree provide the coalescent event times for Snyder filtering. (b) The Snyder estimate of HCV population size using a con-exp-con demographic model. The continuous black line is the conditional mean estimate while the dotted lines are delimited by twice the standard deviation of its posterior. Rapid growth during the PAT period (shaded) is clear. (c) Smoothed marginal posteriors for each parameter of the con-exp-con demographic model (solid black line) (normal kernel smoothing). The Snyder filter used: $[m_i, n] = [20, 63]$ with uniform priors set to $\frac{1}{m_i}$ for each parameter. The parameter estimates from [7], which were obtained by Bayesian MCMC sampling, are shown as box-type plots (median estimate, and 95% highest posterior density credible intervals) above each marginal posterior.

Appendix

A. The Non-linear Snyder Filter

In this appendix we provide some more detail and points of reference for the Snyder filter. Using the same notation as in section 2.1, with $\vec{x}$ as a vector of parameters to be inferred, μ as a variable describing an arbitrary value of $\vec{x}$ and $\mathcal{F}_t$ as all the observable information up until time t the Snyder filter is written as below [24]. The Markov rate matrix $\Lambda_{\mathcal{F}(t)}$ is integrated with respect to the parameter space, and this introduces the non-linearity. Note that the differential equations are now directly in normalised probability. As noted in section 2.1 the differential equation is solved until the left limit of the first event time, τ_1^-, then a discontinuous update applied leading to $\vec{q}(\tau_1^+)$. This is then used as the new initial condition and the differential equations solved again until τ_2^-. These steps are repeated until the entire

observed process is traversed, resulting in the final posterior $\vec{q}(T)$.

$$\frac{\mathrm{d}\vec{q}(t)}{\mathrm{d}t} = -\vec{q}(t)\Lambda_{\mathcal{F}(t)} + \int \vec{q}(t,\,\mu)\lambda(t,\,\mu,\,\mathcal{F}(t))\,\mathrm{d}\mu \;\text{ for } 0 \leq t \leq \tau_1^- \tag{15}$$

$$\vec{q}(\tau_1^+) = \vec{q}(\tau_1^-)\Lambda_{\mathcal{F}(\tau_1^-)}\left(\int \vec{q}(\tau_1^-,\,\mu)\lambda(t,\,\mu,\,\mathcal{F}(\tau_1^-))\,\mathrm{d}\mu\right)^{-1} \tag{16}$$

If we define $\phi = -\Lambda_{\mathcal{F}(t)}$, then the non-linearity is an average $\langle\phi\rangle$ taken across the current posterior $\vec{q}(t)$. We can then write $\frac{\mathrm{d}\vec{q}(t)}{\mathrm{d}t} = \vec{q}(t)(\phi - \langle\phi\rangle)$. The Snyder differential equation therefore describes movement away from a local mean with time. This form contains an interesting conceptual detail. If ϕ can be considered a fitness vector then this is a continuous time replicator equation from evolutionary game theory [43]. Further, the discontinuous update equation obeys the discrete time version of the same replicator equation as step size, away from an event time, becomes infinitesimally small. The correspondence between Bayesian statistics and the replicator equation has been identified in discrete time [44] [45]. Here we note, for the first time (to our knowledge), that the Snyder filter presents a continuous time analogue. The Snyder solution for parametric inference may therefore be seen as the evolution of a series of competing strategies with fitness given by the negative of the rate matrix (which encodes a measure of stability). The strategies pertain to the probability posterior describing $\vec{x}$ sequentially across time.

Whereas the linear equations given in the main text only hold for continuous time Markov processes (of which random variable parameter estimation is a simplification [24]), the non-linear equations apply to several types of processes. If $\vec{x}$ becomes either a Markov diffusion, Markov jump or Poisson driven Markov process $\vec{x}(t)$ then the differential equations of 15 simply include the extra term $\mathbb{L}[\vec{q}(t)]$. Here $\mathbb{L}[\vec{q}(t)]$ is an operator describing the dynamics of the stochastic process. For example, it is $\vec{q}(t)Q$ for a Markov jump process with Q as the infinitesimal generator of the Markov process. $\mathbb{L}[\vec{q}(t)]$ is the Kolmogorov operator for a diffusion process [24]. Extensions such as these make the Snyder filter quite flexible and powerful. More details on possible inference problems solved with Snyder filtering can be found in [20] [24] [46], [21] and [22].

B. Information Equivalence for the Kingman Coalescent

Consider the standard isochronously sampled Kingman coalescent with rate $\binom{i}{2}{x_1}^{-1}$ over an interval with i lineages [1]. Let $\eta = {x_1}^{-1}$ and consider a single coalescent tree with n tips and $n-1$ waiting times of duration τ_k, $1 \leq k \leq n-1$. The interval τ_k ends at time c_k and has $n-k+1$ lineages until that point. We start by outlining the known likelihood calculation for the Kingman coalescent based on the series of coalescent events along one tree [29]. We then compare it to a likelihood derived from data composed of a single coalescent event from several independent trees. The likelihood function for the first case, $L_1(\eta)$, follows from the independence of the coalescent intervals and standard Poisson process theory. We can use the sufficient statistic $\mathcal{T}_1$ to Fisher-Neyman factorise the likelihood (into a product of functions h_1 and g_1) and to then derive the maximum likelihood (ML) estimator, $\hat{x}_1^{(1)}$. Note that $\binom{n-k+1}{2}\tau_k \sim \exp(\eta)$ (exponential distribution properties) [29]. While in general $\hat{x}_1 \neq \hat{\eta}^{-1}$, it is valid here because the same value of $\hat{x}_1$ maximises the partial derivatives of the likelihood with respect to η or x_1.

$$L_1(\eta) = \left[\prod_{k=1}^{n-1}\binom{n-k+1}{2}\right]\eta^{n-1}e^{-\eta\sum_{k=1}^{n-1}\binom{n-k+1}{2}\tau_k} = [h_1(n)]g_1(\eta,\,\mathcal{T}_1) \tag{17}$$

$$\mathcal{T}_1 = \sum_{k=1}^{n-1}\binom{n-k+1}{2}\tau_k \implies \hat{x}_1^{(1)} = \hat{\eta}_{1\mathrm{mle}}^{-1} = \frac{\mathcal{T}_1}{n-1} \tag{18}$$

Now, let us assume a single j^{th} coalescent event is observed from each of $n-1$ independent trees (unlinked loci). The data are now $\tau_j(k)$ for $1 \leq k \leq n-1$. By the properties of exponential scaling, the product $\binom{n-j+1}{2}\tau_j(k) \sim \exp(\eta)$. The

likelihood, $L_2(\eta)$, sufficient statistic, $\mathcal{T}_2$ and ML estimator $\hat{x}_1^{(2)}$ can then be derived as below.

$$L_2(\eta) = \binom{n-j+1}{2}^{n-1} \eta^{n-1} e^{-\eta \sum_{k=1}^{n-1} \binom{n-j+1}{2} \tau_j(k)} = [h_2(n)] g_2(\eta, \mathcal{T}_2) \tag{19}$$

$$\mathcal{T}_2 = \sum_{k=1}^{n-1} \binom{n-j+1}{2} \tau_j(k) \implies \hat{x}_1^{(2)} = \hat{\eta}_{2\text{mle}}^{-1} = \frac{\mathcal{T}_2}{n-1} \tag{20}$$

Since both sufficient statistics, $\mathcal{T}_1$ and $\mathcal{T}_2$, comprise a sum of $n-1$ independent exponential variables with rate η, and the corresponding ML estimators only depend on the sufficient statistic and the sample size, n, then it is equally efficient to sample coalescent waiting times from either a single tree or from multiple trees. Further, $\text{var}\left(\hat{x}_1^{(1)}\right) = \text{var}\left(\hat{x}_1^{(2)}\right) = (n-1)^{-1} x_1^2$ [29]. Hence, for a constant sized population, the $n-1$ coalescent times of one tree contain the same information as the j^{th} coalescent times from $n-1$ independent trees (unlinked loci).

C. The Relation between the Snyder Filter and the Classical Skyline

Consider the coalescent process with deterministically time varying population size. In this appendix, we focus on characterising the relationship between the Snyder filter and non-parametric coalescent estimation schemes for this coalescent process. The classic skyline plot, developed in [23] is a well known non-parametric method for estimating coalescent demographic functions. As in section 3.1 the time c_k denotes the start of an interval with $n-k+1$ lineages (sampling is isochronous). For a given coalescent waiting time $\tau_k = c_k - c_{k-1}$, starting with $n-k+1$ lineages, the skyline estimate, $\hat{N}_{\text{sky}}$ is given below, with $U \sim \mathcal{U}(0, 1)$ and $W = -\log(U) \sim \exp(1) \equiv \text{Gam}(1, 1)$. It is an estimate of the harmonic mean of the population, $\mathcal{H}[N(t)]$, over each event interval, corrupted by multiplicative noise, W.

$$\hat{N}_{\text{sky}} = \binom{n-k+1}{2} \tau_k = W \mathcal{H}[N(t)], \; \mathcal{H}[N(t)] = \left(\int_{c_{k-1}}^{c_k} N(t)^{-1} \tau_k^{-1} \, \mathrm{d}t \right)^{-1} \tag{21}$$

Define $\eta(t) = N(t)^{-1}$ as the intensity to be estimated and let $\nu(t) = \binom{n-k+1}{2} \eta(t)$. Note that $\mathcal{H}[N(t)] = \mathcal{A}[N(t)^{-1}]^{-1}$ where $\mathcal{A}$ is a functional that calculates the arithmetic mean. The skyline then implies an estimator $\hat{\eta}_{\text{sky}} = \left(\binom{n-k+1}{2} \tau_k \right)^{-1} = W^{-1} \mathcal{A}[\eta(t)]$ with the multiplicative noise now being $W^{-1} \sim$ Inverse Gam(1, 1). In [24] an expression is given for linear filters of doubly stochastic Poisson processes (DSPPs). When scaled with $\binom{n-k+1}{2}$, all linear filters of $\eta(t)$ over τ_k satisfy equation 22 for arbitrary functions $g(t)$ and $h(c_{k-1}, s)$. Here $u(s)$ is the usual counting process for coalescent events up to time $s \leq t$ and $\hat{\eta}_{\text{lin}}$ is the intensity estimate resulting from the linear filter.

$$\hat{\eta}_{\text{lin}} = \binom{n-k+1}{2}^{-1} \left[g(t) + \int_{c_{k-1}}^{c_k} h(c_{k-1}, s) \, \mathrm{d}u(s) \right] \tag{22}$$

Choosing $g(t) = 0$ and $h(c_{k-1}, s) = \tau_k^{-1}$ results in $\hat{\eta}_{\text{lin}} = \left(\binom{n-k+1}{2} \tau_k \right)^{-1} [u(c_k) - u(c_{k-1})] = \hat{\eta}_{\text{sky}}$. The classic skyline plot is therefore a type of linear filter for DSPPs. More specifically, it is a moving average filter that only requires knowledge of the interval time between coalescent events, which serves as the averaging time. Generalisations of the skyline method essentially alter the construction of the integral in equation 22 [6]. While we speak of DSPPs in this appendix, the results directly hold for the time-varying coalescent. As mentioned in section 2.2, this is because the deterministically time-varying coalescent estimation problem can be described as a simplification of the DSPP inference problem.

The benefit of moving average filters is that they require minimal process information. If the first and second order statistics of the intensity process are also known then a linear MMSE estimator can be obtained which outperforms all possible other linear filters. Provided the estimator is constrained to be linear this MMSE filter cannot be improved upon, even with extra knowledge of higher order statistics or structural information (such as a parametric model) [24]. The linear MMSE estimator is equivalent to the optimal estimate of an intensity, given that it was corrupted by additive Gaussian noise [24]. The Snyder filter generalises the linear MMSE filter for DSPPs by removing the linearity constraint. This allows extra information from a parametric model to be used and leads to superior performance (smaller MSE). In summary, the skyline is non-parametric, linear and makes use of a minimum of

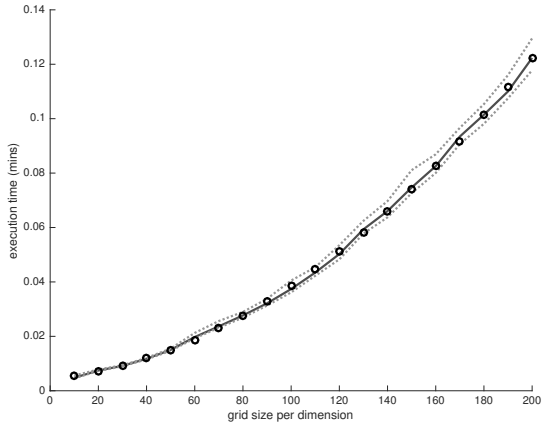
coalescent information. Given mean and covariance information of the coalescent intensity, the skyline can be adapted to a linearly constrained MMSE filter. If a coalescent rate model is known then this can be further improved with the (unconstrained) Snyder non-linear MMSE estimator. Beyond this, no performance improvements are possible without introducing additional observable information.

D. Computational Performance, Implementation and Limits of the Snyder Filter

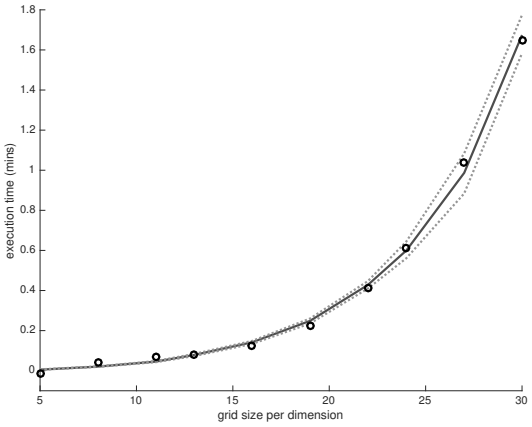
As mentioned in the main text, the Snyder filter involves solving ordinary differential equations for each point on a parametric space grid. Consequently, its complexity is directly related to the dimension and granularity of the grid. These ideas were encapsulated by the filter dimensionality, previously defined as $m = \prod_{i=1}^{l} m_i$ (for estimation of an l parameter $N(t)$). Obviously as either l or m_i increase the Snyder filter becomes more computationally complex. Consequently, in its current implementation, the curse of dimensionality is a possible limitation to the filter functionality. To gain an idea of the its complexity we simulated and estimated the exponential and logistic growth coalescents for $M = 100$ trees with $n = 100$ isochronously sampled tips across several grid sizes. Our results are compiled below and done using an iMac with 8GB RAM and a 4 core 2.7 GHz Intel i5 processor (though parallelisation was not used). Figures 6a and 6b give the median and 95% credible intervals of the execution time across the runs. The filter runs in polynomial time with respect to m for a given l. The order of the polynomial increases with l and of course the maximum grid resolution computable decreases with dimension. The execution times are tightly bounded. This seems sensible since the only stochastic component of the computation relates to the different trees generated on each run.

However, in spite of this, the Snyder filter is able to compute its estimates relatively fast (a few minutes), over the values investigated. Figures 6c and 6d show how relative parametric estimation error varies with grid size for both models. Note the lack of any significant trend in accuracy with grid size from about the midpoint of the scale investigated. This suggests that good accuracy can be obtained with reasonably sized grids and underlies the seemingly coarse grid sizes chosen for simulations in the main text. These estimates will improve in precision and bias as the size of the tree increases (more observable information on the underlying demographic function). The (non-optimised) code we used to generate these performance measures can be found at

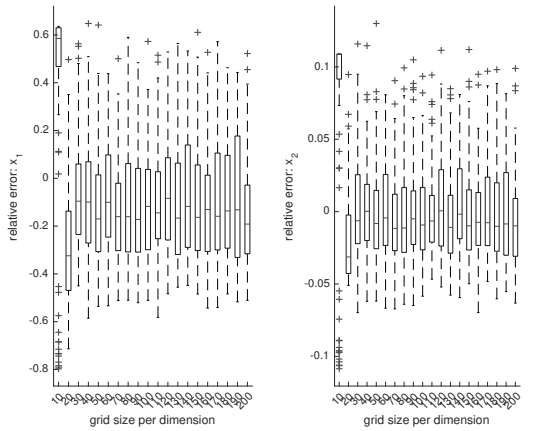
Generally, one could argue that a major benefit of MCMC based methods is their better robustness to the curse of dimensionality, especially when compared to grid based filters. However, we claim that the Snyder filter trades this for the advantage of deterministic inference (the method will produce the same results for a given observed data stream). This negates issues like the convergence problems and local minima that can affect MCMC inference [47]. Furthermore, if many parameters are needed to describe a demographic function then it is probably wiser to use existing non-parametric methods (such as the skyline techniques [23] [28]). Thus, we envisioned the Snyder filter as a useful tool for inference when one has a parametric model that is not over-parametrised and one wants to learn the physics of the problem. Additionally, it may even be possible to integrate the Snyder filter within existing MCMC and other inference techniques. Usually these involve a forward algorithm or a calculation of the coalescent likelihood. The Snyder filter could provide a different way of fulfilling these computations.



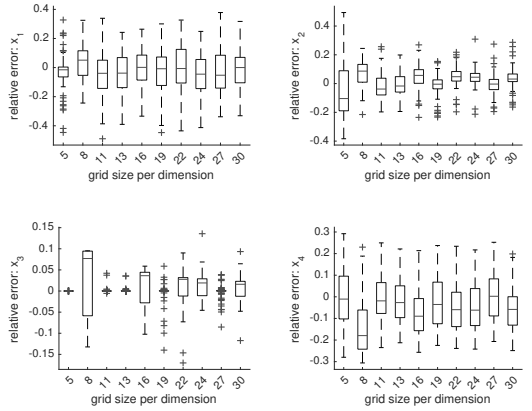
(a) Execution time for exponential



(b) Execution time for logistic



(c) Error boxplot for exponential



(d) Error boxplot for logistic

Figure 6: **Snyder computational complexity and performance.** (a)-(b) Execution time of the filter in minutes across varying grid sizes for the exponential and logistic demographic functions respectively. This includes all overheads such as the time to simulate the tree data. The solid black line is the median time across 100 runs at each grid size. The dotted lines enclose the 95% credibility interval. The black circles are fitted points from a least squares error polynomial. (c)-(d) Box plot of the relative estimation errors of each model parameter $1 - \frac{\hat{x}_i}{x_i}$ against grid size. Each box plot uses the Snyder estimates obtained from the 100 replicate trees which were used in (a) and (b). Simulations were done at: $[n, M] = [100, 100]$, and with the usual uniform priors of $\frac{1}{m}$ on each grid point.

E. The Coalescent Likelihood and the Snyder Filter

Here we explain how the Snyder filter relates to the standard single tree coalescent likelihood for deterministically time-varying demographic functions. We assume an isochronously sampled coalescent with effective population size $N(t, \vec{x})$ and observed coalescent times c_k for $1 \leq k \leq n-1$. Here c_k denotes the first time when exactly k coalescent events have occurred (therefore the interval ending at c_k had $n - k + 1$ lineages). We define c_0 as 0 to indicate the present (n lineages exist). With this notation the known coalescent likelihood, $L_{\text{coal}} = \mathrm{P}\left(\{c_k\}_{1\leq k\leq n-1} \,|\, N(t, \vec{x})\right)$ [2] [48] can be written as follows with $A_k = \binom{n-k+1}{2}$. It is a product of densities across the $n-2$ intervals containing the $n-1$ coalescent events that form a tree of n tips.

$$L_{\text{coal}} = \prod_{k=1}^{n-1} \frac{A_k}{N(c_k, \vec{x})} \exp\left[- \int_{c_{k-1}}^{c_k} \frac{A_k}{N(t, \vec{x})} \, \mathrm{d}t \right] \tag{23}$$

Now we can directly incorporate the coalescent rate $\lambda(t, \vec{x}, \mathcal{F}(t)) = \frac{\mathcal{F}(t)}{N(t, \vec{x})}$ in the above equation, take natural logs and sum corresponding terms to get the expression below. Note that $\mathcal{F}(t)$ gives the appropriate k for A_k at any time t.

$$\log L_{\text{coal}} = \sum_{k=1}^{n-1} \log \lambda(c_k,\, \vec{x},\, k) - \sum_{k=1}^{n-1} \left[\int_{c_{k-1}}^{c_k} \lambda(t,\, \vec{x},\, k)\, \mathrm{d}t \right] \tag{24}$$

However, as noted by Snyder and Miller [24], for an inhomogeneous Poisson process with $n-1$ observed events and rate $\lambda(u,\, \vec{x})$, the filter differential equations can be solved to get the following analytic expression.

$$\vec{q}(t,\, \vec{x}) = e^{H(t,\, \vec{x})} \vec{q}(0,\, \vec{x}) \mathbb{E}\left[e^{H(t,\, \vec{x})} \right]^{-1} \tag{25}$$

$$H(t,\, \vec{x}) = -\int_0^t \lambda(\tau,\, \vec{x})\, \mathrm{d}\tau + \int_0^t \log \lambda(\tau,\, \vec{x})\, \mathrm{d}\mathcal{F}(\tau) \tag{26}$$

This is a Bayes law expression with $e^{H(t,\, \vec{x})}$ serving as the likelihood function. $\mathcal{F}(\tau)$ is the observation process at time τ and $\mathrm{d}\mathcal{F}(\tau) = 1$ at every observed event time and 0 otherwise (it is called a counting integral). Consequently, the Snyder inhomogeneous log likelihood is $\log L_{\text{sny}} = H(t,\, \vec{x})$. However, the coalescent inference problem is not strictly inhomogeneous due to the lineage dependent term. If we condition on the number of lineages in each inter-event interval then we can describe each of these as a single interval from a different inhomogeneous process. Since these intervals are independent we can collect them together and treat the updated posterior from an interval as the prior to the next interval. As a result we can decompose $H(t,\, \vec{x})$ as $\sum_{k=1}^{n-1} H_k$ where $H_k = -\int_{c_{k-1}}^{c_k} \lambda(t,\, \vec{x},\, k)\, \mathrm{d}t + \log \lambda(c_k,\, \vec{x},\, k)$. This means that $L_{\text{sny}} \equiv L_{\text{coal}}$. This confirms that the Snyder filter (implicitly) solves the correct likelihood and is comparable to all coalescent inference methods on fixed trees [18] [23].

Our preference for using the differential equation Snyder description versus this likelihood formulation are twofold. Firstly, equation 25 only works here because of the lineage conditioning. If our coalescent rate changed due to some other stochastic process, this could not be easily integrated into this solution, unlike in the differential equation case. Thus this solution is inflexible and specialised. Secondly, the integrals in this formulation (together with additional ones required to calculate conditional means) become difficult for multidimensional problems and may need to be solved with Monte Carlo techniques [24]. We are only interested in fully deterministic solutions for coalescent inference problems that are flexible enough to accommodate potentially complex demographic functions.