Journal of Theoretical Biology 421 (2017) 153–167

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

# Optimal point process filtering and estimation of the coalescent process

Kris V. Parag*, Oliver G. Pybus

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

ARTICLE INFO

*Article history:*
Received 4 November 2016
Revised 18 March 2017
Accepted 2 April 2017
Available online 3 April 2017

*Keywords:*
Coalescent theory
Non-linear filters
Bayesian inference
Parametric estimation
Snyder filters

ABSTRACT

The coalescent process is a widely used approach for inferring the demographic history of a population, from samples of its genetic diversity. Several parametric and non-parametric coalescent inference methods, involving Markov chain Monte Carlo, Gaussian processes, and other algorithms, already exist. However, these techniques are not always easy to adapt and apply, thus creating a need for alternative methodologies. We introduce the Bayesian Snyder filter as an 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 datasets. We analytically solve the filter equations for the constant population size Kingman coalescent, derive expressions for its mean squared estimation error, and estimate its robustness to prior distribution specification. For populations with deterministically time-varying size we numerically solve the Snyder equations, and test this solution on common 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, dataset of hepatitis C virus sequences and show that the filter compares well to a popular phylodynamic inference method. The Snyder filter is an exact (given discretised priors, it does not approximate the posterior) and direct Bayesian estimation method that has the potential to become a useful alternative tool for coalescent inference.

## 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 $N(t) \gg n$ (Kingman, 1982). 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) (Kingman, 1982). Since then, the coalescent model has been generalised to incorporate deterministically varying population sizes (Griffiths and Tavare, 1994), stochastic population fluctuations (Kaj and Krone, 2003), geographically structured populations (Notohara, 1990), and data sets containing sequences sampled at different time points (heterochronous sampling) (Rodrigo et al., 1990). As a result, the coalescent model has been applied to a range of problems in many biological disciplines including conservation biology, anthropology and epidemiology (Strimmer and Pybus, 2001). 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 (Pybus et al., 2003), the oscillating behaviour of dengue virus in Vietnam (Rasmussen et al., 2014) and to estimate the generation time of HIV-1 within individual infected patients (Rodrigo et al., 1990). 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 (Kingman, 2000).

* Corresponding author.
*E-mail address:* kris.parag@zoo.ox.ac.uk (K.V. Parag).

http://dx.doi.org/10.1016/j.jtbi.2017.04.001
0022-5193/© 2017 Elsevier Ltd. All rights reserved.

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 (Nordberg, 2001). 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 (Diggle et al., 2000). 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 $n$ (Palacios and Minin, 2013; Yang, 2014). 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 (Diggle et al., 2000). 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 rigorous 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 (De Maio et al., 2015; Kuhner, 2008; Kuhner et al., 1995). These approaches, while readily able to account for genealogical uncertainty, can be complex or difficult to implement (Kim et al., 2015).

The motivation for our work is most similar to that of Palacios and Minin (2012). 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 (Drummond et al., 2002), and related approaches.

In this paper, we introduce and analyse the Snyder filter (Snyder, 1972), 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. (2008) and to invertebrate visual phototransduction by Parag (2014).

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 subsequently 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, examine its computational performance and, for the Kingman coalescent, reiterate the link between sequential data from a single tree and parallel data from many trees. We also present an informative relation between the Snyder filter approach and a popular non-parametric coalescent inference technique called the classic skyline plot (Pybus et al., 2000). We prove that the Snyder filter naturally generalises the skyline in a non-linear parametric manner. Furthermore, we show that the likelihood being implicitly optimised by the Snyder filter is equivalent to the standard coalescent likelihood from the literature for any deterministically time varying demographic model.

## 2. Methods

### 2.1. The Snyder filter

Consider a Poisson process, $u(t)$, at time $t \geq 0$ with instantaneous intensity $\lambda(t, \vec{x}(t))$ on the space of non-negative real numbers. The function $u(t)$ is an integer valued process that counts the number of points at $t$. The vector $\vec{x}(t)$ is called the information process and 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 $u(t)$ is a doubly stochastic Poisson process (DSPP). Let the counting process stream from time 0 to time $t$ be denoted $u_t = u(s), \forall s : 0 \leq s \leq t$. 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 $u_t$ and the basic statistics of $\vec{x}(t)$ (Snyder, 1972). 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 $u_t$, lead to the causal posterior $\mathrm{P}(\vec{x}(t) \,|\, u_t)$. We assume that we observe (without ambiguity) a known function of $u(t)$, which we write as $\mathcal{F}(t)$ and refer to as the observation process. Consequently $\mathrm{P}(\vec{x}(t) \,|\, u_t) = \mathrm{P}(\vec{x}(t) \,|\, \mathcal{F}_t)$.