

Exact Bayesian Inference for Phylogenetic Birth-Death Models

KV Parag and OG Pybus

April 30, 2018

Abstract

Motivation: Inferring the rates of change of a population from a reconstructed phylogeny of genetic sequences is a central problem in macro-evolutionary biology, epidemiology, and many other disciplines. A popular solution involves estimating the parameters of a birth-death process (BDP), which links the shape of the phylogeny to its birth and death rates. Modern BDP estimators rely on random Markov chain Monte Carlo (MCMC) sampling to infer these rates. Such methods, while powerful and scalable, cannot be guaranteed to converge, leading to results that may be hard to replicate or difficult to validate.

Results: We present a conceptually and computationally different parametric BDP inference approach using flexible and easy to implement Snyder filter (SF) algorithms. This method is deterministic so its results are provable, guaranteed, and reproducible. We validate the SF on constant rate BDPs and find that it solves BDP likelihoods known to produce robust estimates. We then examine more complex BDPs with time-varying rates. Our estimates compare well with a recently developed parametric MCMC inference method. Lastly, we perform model selection on an empirical Agamid species phylogeny, obtaining results consistent with the literature. The SF makes no approximations, beyond those required for parameter quantisation and numerical integration, and directly computes the posterior distribution of model parameters. It is a promising alternative inference algorithm that may serve either as a standalone Bayesian estimator or as a useful diagnostic reference for validating more involved MCMC strategies.

1 Introduction

A common problem in biology involves inferring the rates of change of a population from an observed set of gene sequences, sampled from that population. Here the ‘population’ is composed of any set of members we are interested in, such as individual organisms or species, and the rates of change control its fluctuations. Usually a phylogeny (tree) that contains information about the population is constructed from the sampled

sequences. Statistical techniques are then applied to this tree in order to infer the underlying (and unobservable) rates of change.

The size of the biological population of interest can be described by the number, $l(t)$, of its constituent members at time $t \geq 0$. This count increments or decrements due to the random timings of birth (speciation) and death (extinction) events. We will refer to the members of our population as lineages (or taxa). If lineage births or deaths occur independently of one another, and only a single event is allowed at any time, then the population can be modelled as a continuous-time birth-death process (BDP) (Gernhard, 2008). The BDP leads to a rooted binary tree containing both extinct (dead) and extant (living) lineages. Extinct lineages cannot usually be observed, so an associated tree, called the reconstructed birth-death process (rBDP) is often defined (Nee *et al.*, 1994). The rBDP models the observable phylogeny (i.e. the tree reconstructed from the genetic sequences) at some observation time $T > 0$, as a pruning of the full BDP tree, such that only lineages with descendants at T remain. We use $\mathcal{F}(t)$ to count the number of lineages in the rBDP at time t . Harvey *et al.* (1994) showed that although the rBDP excludes extinct lineages, it still contains information about the lineage birth and death rates of the full BDP tree. Thus a link between the unobserved rates of population size change and the observed genealogy is derived.

Estimating the birth and death rates (which govern $l(t)$), from the rBDP ($\mathcal{F}(t)$) is an important problem in several disciplines, including macro-evolution, ecology and phylodynamics (Pyron and Burbink, 2013). Solutions to this problem for BDPs with time-varying or density dependent rates, incomplete sampling schemes, and multi-type behaviours have led to many biological insights, ranging from understanding the diversification behaviour in the animal kingdom to the space-time dynamics of viral epidemics (Stadler, 2009) (Stadler *et al.*, 2013) (Morlon *et al.*, 2011) (Hohna *et al.*, 2011) (Hohna, 2015) (Kuhnert *et al.*, 2016).

However, the information lost in going from a true BDP to the observed rBDP leads to several difficulties. Many combinations of birth and death rates can produce the same rBDP, which can obfuscate estimation of the true diversification process (Kubo and Iwasa, 1995) (Pyron and Burbink, 2013). Moreover, several standard inference methods struggle to infer non-zero death rates from certain empirical datasets despite the existence of known extinction events (Pyron and Burbink, 2013) (Purvis, 2008). This is often attributed to improper or overly assumptive modelling choices, which the data may then violate (Morlon, 2014). An ongoing and related issue is deciding whether a BDP or another phylogenetic model, such as the coalescent (Kingman, 1982), better describes an observed tree (Volz and Frost, 2014) (Stadler *et al.*, 2015). BDP inference therefore remains an important and active field of research.

We focus on inference for BDPs with time-varying per lineage birth and death rates, respectively denoted $\lambda(t)$ and $\mu(t)$. Biologically, such temporal variance may represent how external influences, such as changes in the abiotic or biotic environment, impact a population (Morlon, 2014). To this problem we apply a parametric Bayesian statistical algorithm from control and electrical engineering, which we term the Snyder filter (SF) (Snyder, 1972). We use the SF to infer $\lambda(t)$ and $\mu(t)$ given a reconstructed

phylogeny, which we assume to be observed without error. We have previously shown how the SF can be adapted to estimate coalescent processes (Parag and Pybus, 2017). Here we extend that work to the more complex time-varying BDP inference problem.

Several methods exist for inferring time-varying BDP rates from rBDPs. Nee *et al.* (1994) initiated this investigation by deriving (but not optimising) an appropriate likelihood function. Since then, several explicit likelihood approaches based on different joint probability density constructions, conditioning criteria and emphases on maximum likelihood versus Bayesian viewpoints, have dominated the field (Morlon, 2014) (Hohna, 2015). Morlon (2014) gives an overview of these methods. The most powerful among these techniques tend to use Markov chain Monte Carlo (MCMC) sampling. This allows one to accommodate complex BDP dynamics or include features such as incomplete sampling and genealogical uncertainty. However, due to the stochastic nature of MCMC, these benefits come at the expense of analytic tractability and methodological determinacy (Stadler *et al.*, 2013) (Morlon *et al.*, 2011). With no guarantee of convergence (Cowles and Carlin, 1996), it can sometimes be difficult, or time consuming, to assess or debug the performance of these MCMC samplers (Mossel and Vigoda, 2006).

Non-likelihood based methods typically utilise summary statistics, or analyse lineage through time plots (LTTs) (Pybus and Harvey, 2000) (Morlon, 2014) (Paradis, 2010). LTTs are plots of $\mathcal{F}(t)$ against t . Such methods do not make use of all the information in the rBDP. They, however, remain popular because they are easy to use, deterministic and usually interpretable. The SF algorithm that we introduce here melds some of the desirable properties of both MCMC and non-likelihood based approaches.

The SF directly computes the joint posterior distribution of a parametric BDP by exploiting the Poisson process nature of the inference problem. It achieves, over a defined parametric grid, provable minimum mean square error (MMSE) estimates by simply solving an appropriate set of coupled linear ordinary differential equations (Snyder, 1972). It is wholly deterministic and makes use of all the BDP information without having convergence issues. While we do not yet account for genealogical uncertainty, or non-uniform sampling, we do show how the SF has the potential to accommodate these features in the future.

We envision two potential applications for the SF. First, given its provable and MMSE nature, it can serve as a standalone Bayesian algorithm for learning about a phylogenetic dataset in a phenomenological manner. For example, it can be used to quickly estimate and evaluate competing parametric models. Second, its deterministic and reproducible posteriors make it a good diagnostic tool for validating MCMC methods, especially when their outputs appear implausible or vary among runs. This will become particularly valuable when complexities like genealogical uncertainty are included, as (i) convergence time increases exponentially and (ii) misleading or overconfident posteriors may result (Cowles and Carlin, 1996) (Mossel and Vigoda, 2006) (Yang and Zhu, 2018).

We define and adapt the SF for BDP inference problems in Methods. In Results, we apply the SF to constant rate BDPs and recover known

trends. We also compare the implicit SF likelihood with seven others from the literature. We then consider BDPs with time-varying birth and death rates, and validate the SF against a recent MCMC method (Hohna *et al.*, 2011), on data simulated from two illustrative models. We also perform model selection on empirical data (an Agamid phylogeny) and obtain results consistent with its original analysis (Rabosky and Lovette, 2008b).

2 Methods

2.1 Optimal Snyder Filtering

A doubly stochastic Poisson process (DSPP) is a Poisson process that has a stochastic rate of producing events. Let $\mathcal{F}(t)$ denote an observed DSPP at time $t \geq 0$ and let $\vec{x}(t)$ be a hidden vector state process that controls its stochastic rate. $\mathcal{F}(t)$ is then a non-decreasing integer valued process that counts the number of events at t . It has instantaneous intensity, $\beta(t, \vec{x}(t))$, on the space of non-negative real numbers. We want to infer the state process $\vec{x}(t)$ (Section 2.2 will show how $\vec{x}(t)$ encodes the parameters of a BDP) given past observations of the DSPP. We use $\mathcal{F}_t = \{\mathcal{F}(s) : 0 \leq s \leq t\}$ to denote all past observations up to t . Snyder (1972) derived a filter that optimally inferred $\vec{x}(t)$ (with respect to mean squared error), given \mathcal{F}_t and priors on $\vec{x}(t)$. We call this the SF. It is an exact, Bayesian inference method that generates the informed posterior, $\bar{q}(t) = P(\vec{x}(t) | \mathcal{F}_t)$, by solving a set of non-linear differential equations on the probability distribution of $\vec{x}(t)$, sequentially with time over \mathcal{F}_t . It is ‘exact’ because it computes the inferred joint posterior directly, without approximating either the observation process, \mathcal{F}_t , or the hidden process, $\vec{x}(t)$. The only approximations in the SF are inherited from the standard inaccuracies in numerically integrating differential equations, and in representing distributions discretely. For some problems the SF is analytically solvable, in which case there are no approximations.

The SF is general and applies to any hidden Markov state process that has dynamics describable by: $d\vec{x} = \vec{f}(\vec{x}) dt + \vec{g}(\vec{x}) d\vec{\chi}$. Here $\vec{\chi}$ is a martingale with independent increments and \vec{f} , \vec{g} are arbitrary vector functions of choice (see Snyder and Miller (1991) for details). The filter is also valid for DSPPs with intensities that additionally depend on the observed events, $\beta(t, \vec{x}(t), \mathcal{F}_t)$. Such DSPPs are called self-exciting (Snyder and Miller, 1991). We focus on inference problems with $\vec{f} = 0$ and $\vec{g} = 0$, which means our hidden state process is simply a vector of random variables \vec{x} . We also restrict the type of self-exciting rate dependence to be Markovian (0-memory). This DSPP intensity is then $\beta(t, \vec{x}, \mathcal{F}(t))$ instead of $\beta(t, \vec{x}(t), \mathcal{F}_t)$. These stipulations mean that the SF framework is applicable to the dynamics presented by BDPs with constant but unknown parameters. Under these conditions the SF can be transformed into a set of linear differential equations on an un-normalised distribution $q^*(t)$, which is then normalised to $\bar{q}(t)$ (Rudemo, 1972).

The resulting SF is described by Eq. (1)-Eq. (3) (Snyder and Miller, 1991) (Rudemo, 1972). The rate matrix, $\Lambda_{\mathcal{F}(t)}$, is diagonal with entries

for every value of $\beta(t, \vec{x}, \mathcal{F}(t))$ at any given t due to the possible values of \vec{x} and $\mathcal{F}(t)$. Let an arbitrary value of \vec{x} be ϵ and denote its normalised and un-normalised probabilities as $\tilde{q}(t, \epsilon)$ and $q^*(t, \epsilon)$. The complete posterior distribution is then $\tilde{q}(t)$ while $\tilde{q}(t, \epsilon)$ is the single value when $\vec{x} = \epsilon$. Assume that we have observed the set of events produced by $\mathcal{F}(t)$ over $0 \leq t \leq T$. If the first event is at $t = \tau_1$ then τ_1^- and τ_1^+ are infinitesimally before and after that event. The initial condition for the differential equations is the prior: $\tilde{q}(0) = P(\vec{x})$. Eq. (1)-Eq. (3) which describe the dynamics of $\tilde{q}(t)$ until τ_1^+ , form the core of the SF algorithm. Note that the integrals enumerate every possible value of \vec{x} .

$$\frac{dq^*(t)}{dt} = -q^*(t)\Lambda_{\mathcal{F}(t)}, \text{ for } 0 \leq t \leq \tau_1^- \quad (1)$$

$$\tilde{q}(t) = q^*(t) \left(\int q^*(t, \epsilon) d\epsilon \right)^{-1}, \text{ for } 0 \leq t \leq \tau_1^- \quad (2)$$

$$\tilde{q}(\tau_1^+) = \tilde{q}(\tau_1^-) \Lambda_{\mathcal{F}(\tau_1^-)} \left(\int \tilde{q}(\tau_1^-, \epsilon) \beta(t, \epsilon, \mathcal{F}(\tau_1^-)) d\epsilon \right)^{-1} \quad (3)$$

From 0 to τ_1^- , un-normalised state probabilities undergo a continuous exponential decay (Eq. (1)), before being normalised (Eq. (2)). At τ_1^- an event is observed and the posterior is discontinuously updated (Eq. (3)). The resulting $\tilde{q}(\tau_1^+)$ is then used as a new initial condition and the equations solved again until the next event (over the period $\tau_1^+ \leq t \leq \tau_2^-$). This repeats until we obtain $\tilde{q}(T) = P(\vec{x}(t) | \mathcal{F}_T)$. This joint posterior uses all of the observed data \mathcal{F}_T and yields the MMSE estimator (also known as the conditional mean) of any function of the parameters, $f(\vec{x})$. This is defined as $\hat{f}(\vec{x}) := \mathbb{E}[f(\vec{x}) | \mathcal{F}_T] = \int \tilde{q}(T, \epsilon) f(\epsilon) d\epsilon$ with the MMSE as $\mathbb{E}[(f - \hat{f})^2]$ (Snyder and Miller, 1991).

In the following section we will show how parameter estimation from an observed rBDP fits within this self-exciting SF framework. More information on the SF algorithm and some of its biological applications can be found in (Snyder, 1972), (Snyder and Miller, 1991), (Bobrowski *et al.*, 2008), (Parag and Vinnicombe, 2017) and (Parag and Pybus, 2017).

2.2 Birth-Death Process Inference

We consider a BDP with per lineage time-varying birth and death rates, $\lambda(t)$ and $\mu(t)$ and let n describe the number of extant lineages at observation time T . The full BDP and associated observed rBDP lineage counts satisfy the boundary conditions $(l(0), l(T)) = (\mathcal{F}(0), \mathcal{F}(T)) = (1, n)$. Note that $\mathcal{F}(t) \leq l(t)$ since $\mathcal{F}(t)$ is a pruned version of $l(t)$. If we denote the time of the k^{th} birth event in the rBDP as c_k for $1 \leq k \leq n-1$, then the observed process $\mathcal{F}(t) = 1 + \sum_{k=1}^{n-1} \mathbb{I}(t \geq c_k)$. \mathbb{I} is an indicator function that is 1 when its argument is true and 0 otherwise. We define the total diversification rate as $\alpha(t, \tau) := \int_t^\tau \lambda(u) - \mu(u) du$, and the probability that a single lineage at time t survives until T , $P(t, T)$, as in Eq. (4), from Kendall (1948).

$$P(t, T) := \left(1 + \int_t^T \mu(\tau) e^{-\alpha(t, \tau)} d\tau \right)^{-1} \quad (4)$$

The rBDP can be described as a generalised pure birth process (Kendall, 1948) with total birth or lineage growth rate, $\beta(t)$, defined below from Nee *et al.* (1994). This rate contains information about the unobservable death events since $P(t, T)$ depends on $\mu(t)$. Note that when $\mu(t) = 0$ over $0 \leq t \leq T$ then $P(t, T) = 1$ and $\beta(t) = \lambda(t)\mathcal{F}(t) = \lambda(t)l(t)$, thereby illustrating how deaths lead to information losses.

$$\beta(t) := \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{P}(\mathcal{F}(t + \Delta t) - \mathcal{F}(t) \geq 1 | \mathcal{F}(t)) \quad (5)$$

$$\beta(t) = \lambda(t)P(t, T)\mathcal{F}(t) \quad (6)$$

When considered from this perspective, the rBDP becomes amenable to SF inference. This follows because the counting statistics of a self-exciting DSPP with 0-memory are identical to those of a pure birth process with a population dependent birth rate (Snyder and Miller, 1991).

We assume that the BDP can be described with p parameters and define the parameter set as a vector $\vec{x} = (x_1, x_2, \dots, x_p)$. This set is partitioned so that the birth and death rates are parametrised as $\lambda(t, \vec{x}_\lambda)$ and $\mu(t, \vec{x}_\mu)$. Note that \vec{x}_λ and \vec{x}_μ may have common parameters but together they must span all of \vec{x} . We will usually just write $\lambda(t)$ and $\mu(t)$ as shorthand. This formulation means that $\lambda(t)P(t, T)$ from Eq. (6) is simply a function of \vec{x} and t . The rBDP rate can therefore be expressed as $\beta(t, \vec{x}, \mathcal{F}(t))$. As a result we can use the SF (Section 2.1) to solve the time-varying BDP inference problem and obtain $\mathbb{P}(\vec{x} | \mathcal{F}_T)$. Note that we have assumed isochronous and complete sampling, so that all n taxa or lineages observable in the rBDP are sampled at T with probability 1. Incomplete isochronous sampling would involve sampling each extant lineage with probability $\nu < 1$. Although we do not implement incomplete sampling here, in **supplementary material I** we show how it can be incorporated within the SF framework for any time-varying BDP model.

We now explain the numerical implementation of the SF algorithm. Let the vector, \vec{x} , of p random variables (parameters) to be estimated be such that the distribution of the i^{th} random variable can be described on a domain of m_i points. The parameter vector is then on a joint Cartesian grid of $m = \prod_{i=1}^p m_i$ possible values, so there are m possible vectors describing \vec{x} . We denote some arbitrary vector from this m -set as ϵ . The SF solves a differential equation for the joint probability mass across all ϵ . Consequently, the filter has dimension m and the prior $P(\vec{x})$ and posterior $\tilde{q}(t) = P(\vec{x} | \mathcal{F}_t)$ have m elements. The rate matrix, $\Lambda_{\mathcal{F}(t)}$ is then a diagonal matrix with m entries at each time t given by enumerating $\beta(t, \vec{x}, \mathcal{F}(t))$ over possible ϵ . The continuous differential equations of Eq. (1) are sequentially integrated along the inter-branch intervals of the rBDP and then normalised so that probabilities sum to 1. Discontinuous updates are applied every time we hit a branching time (at which $\mathcal{F}(c_k^+) = \mathcal{F}(c_k^-) + 1$) according to Eq. (3). All integrals are across the elements of the relevant vectors or matrices. This process is repeated until we reach the tip of the tree at T . Pseudo-code describing this implementation of the SF algorithm is given in **supplementary material II**. Further information on the numerical implementation, accuracy, and complexity of the SF can be found in Parag and Pybus (2017).

3 Results

3.1 Constant Rate Birth-Death Estimation

We apply the SF to the constant rate BDP, which is a special case of the time-varying model. The parameters to be estimated are $(x_1, x_2) = (\lambda, \mu)$ or equivalently $(\sigma = \lambda - \mu, \rho = \frac{\mu}{\lambda})$ with $\lambda \geq \mu > 0$ (Stadler, 2009). For this model, $P(t, T)$ can be explicitly written as $\frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda - \mu)(T - t)}}$. We can then parametrise $\beta(t)$ in terms of σ and ρ as in Eq. (7).

$$\beta(t) = \sigma \mathcal{F}(t) \left(1 - \rho e^{-\sigma(T-t)}\right)^{-1} \quad (7)$$

In **supplementary material I** we note that solutions to this problem are also applicable when sampling is incomplete and isochronous.

We simulate constant rate rBDPs with n extant lineages under known parameters, using the algorithms of Hartmann *et al.* (2010) and Stadler (2009). We reparametrise the inverse distributions from these algorithms to obtain the branching times c_k for $1 \leq k \leq n - 1$, with $c_{n-1} = T$. Working forwards through time, we numerically integrate the SF (Eq. (1)-Eq. (3)) over these rBDPs using $\beta(t)$ from Eq. (7). We use the (σ, ρ) parametrisation, since it allows easier prior definitions, and then convert our results to the (λ, μ) form. We obtain conditional mean estimates of these parameters over 10^4 replicate simulated trees. The results of two sets of simulations are shown in **Fig 1**, together with the percentage relative MMSE, defined as $100(1 - \frac{\hat{x}_i}{x_i})^2$ for parameter x_i . The SF estimates cover all the true values well. In general μ is more difficult to infer (it has a higher MMSE) and accuracy improves with reduced μ . We also observe a bias towards underestimating μ when it is high.

These findings confirm known trends (Nee, 2001) (Paradis, 2004). In **supplementary material III**, we further validate the SF performance by comparison with an alternative parametric least squares optimisation method from Paradis (2010). These results also suggest that the Yule model, which is a constant rate BDP with $\mu = 0$, should exhibit the lowest MMSE within this class of BDPs. In **supplementary material IV** we solve the SF equations analytically for the Yule model and characterise its optimal estimator. We also comment on the similarity of this solution to that of another popular phylogenetic model, the Kingman (1982) coalescent.

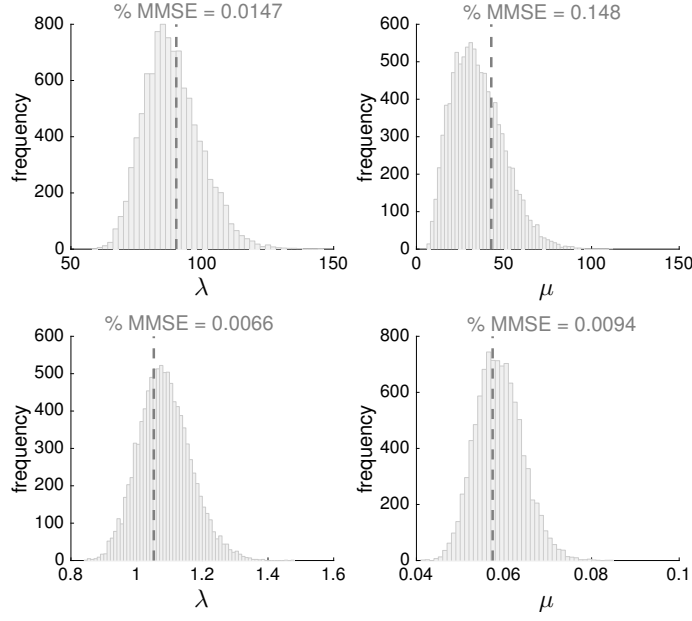


Figure 1: **Estimation of the constant rate BDP model.** Constant rate BDPs at a high and low (λ, μ) rate pair (vertical dashed lines) are estimated using a SF with grid dimension $m_i = 100$. A total of 10^4 independent, $n = 200$ tip, replicate trees were generated under each pair. The histograms show the conditional mean rate SF estimates across the replicate trees. The top subfigures are for a high death rate setting (priors between $[0.01, 0.99]$ for ρ and $[0.01, 100]$ for σ) and the bottom ones at a low setting (priors between $[0.01, 0.1]$ for ρ and $[0.1, 2]$ for σ). The % MMSE is given as a measure of accuracy.

3.2 The Snyder Filter Likelihood

Stadler (2013) noted that, in the literature, BDP inference problems have been solved under 7 distinct likelihood functions. These arise from different assumptions about the rBDP and correspond to conditioning on (1) a specific tree starting time, (2) survival of at least 1 starting lineage to T , (3) survival of exactly n lineages to T , (4) a specific most recent common ancestor (MRCA) time, (5) survival of both lineages subtending from the MRCA to T , (6) survival of exactly n lineages to T for a specific MRCA time and (7) survival of n lineages to T given a prior distribution on the tree starting time. Failing to properly account for the differences in these likelihoods can affect the accuracy, bias and comparability of BDP estimates, even for the simplest constant rate BDP. This is especially important when analysing empirical trees, as it may not always be clear if a chosen BDP inference scheme matches the conditions under which the data were obtained. We examine what conditioning assumptions the SF makes by comparing its implicit likelihood to the 7 in Stadler (2013).

If $\mathcal{F}(t)$ is a standard inhomogeneous Poisson process then the SF

can be analytically shown to solve the log-likelihood function: $H(\epsilon) = -\int_0^T \beta(s, \epsilon) ds + \sum_{k=1}^{n-1} \log \beta(c_k, \epsilon)$ (Snyder and Miller, 1991). Here ϵ is an arbitrary value of \vec{x} . Our BDP problem is self-exciting in addition to being inhomogeneous. We therefore modify $H(\epsilon)$ to account for the extra dependence $\beta(t)$ has on \mathcal{F}_T . If c_k is the k^{th} observed rBDP event time, then the rBDP is inhomogeneous between consecutive c_k values. We can therefore disaggregate the likelihood into interval sums with self-exciting birth rate $\beta(s, \epsilon, k)$ for times $s : k \leq \mathcal{F}(s) < k+1$. This delimits the k^{th} birth period and leads to Eq. (8) with $c_0 := 0$. This decomposition reflects the piecewise continuous nature of the SF equations from Section 2.1.

$$H(\epsilon) = \sum_{k=1}^{n-1} \left(-\int_{c_{k-1}}^{c_k} \beta(s, \epsilon, k) ds + \log \beta(c_k, \epsilon, k) \right) \quad (8)$$

The constant rate BDP admits a closed form for the integral in Eq. (8). See **supplementary material V** for a generalised form of this solution.

We examine a fine grid over the (σ, ρ) parametrisation of the constant rate BDP, with $m_1 = m_2 = g = 150$ and simulate a single rBDP tree with parameters at the grid median, $(50, 0.5)$. Given this observed tree, we evaluate log-likelihoods over this grid and marginalise for each parameter (we denote these generically as $H(\rho)$ and $H(\sigma)$). We plot the 7 marginal log-likelihoods from Stadler (2013) in **Fig 2** as labelled grey lines (likelihoods (2) and (5) are in darker grey). Their separation illustrates the notable impact of different conditioning assumptions on the BDP likelihoods. When the SF log-likelihood is computed from the start of the tree ($\mathcal{F}(0) = 1$), it exactly matches likelihood (2) (circle markers), provided that the correction $-g \log(n-1)!$ is added. Here g is a normalisation factor that depends on the Snyder grid size. This correction means that the SF is computing its likelihood on branching times, since for a given branching time vector, there are $(n-1)!$ different (equally likely) oriented trees (Stadler, 2013). If the SF log-likelihood is computed from the rBDP MRCA, where $\mathcal{F}(c_1) = 2$, then it matches likelihood (5) with the same correction constant (square markers in **Fig 2**).

These results makes sense, as the SF Markov birth rate, $\beta(t)$, depends on $P(t, T)$, a function that encodes survival time (see Eq. (4)). Such a survival condition is central to likelihoods (2) and (5). The equivalence shown here is important as it validates our SF implementation and clarifies how our algorithm relates to others in the literature. By ensuring that likelihood functions are consistent we can compare techniques directly. Moreover, Stadler (2013) noted that different likelihoods can lead to differing estimate biases, and recommended using likelihoods (2) and (5) due to their robustness and accuracy. The fact that these are the likelihoods solved by the SF is reassuring.

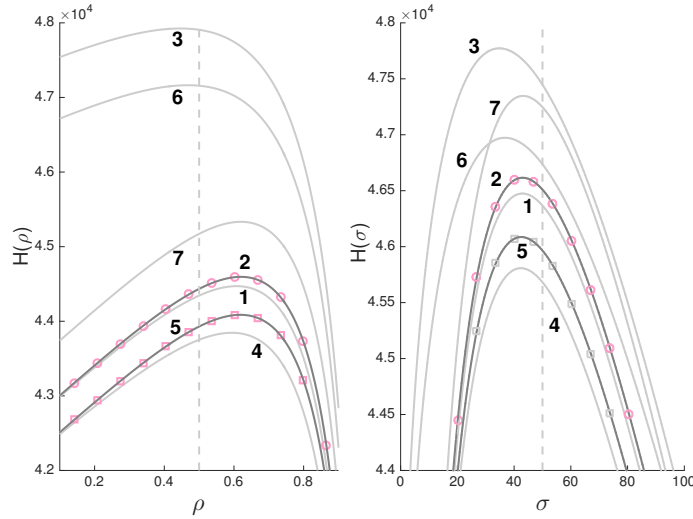


Figure 2: **Comparison of constant rate BDP marginal log-likelihood functions.** The 7 likelihoods given in Stadler (2013) and the SF likelihoods of Eq. (8) are examined over a grid with $m_1 = m_2 = 150$ points over $0.01 \leq \rho \leq 0.99$ and $0.01 \leq \sigma \leq 100$. The true parameter values used to simulate the tree are shown as vertical dashed lines. The 7 likelihoods are the labelled light grey lines with (2) and (5) in dark grey. All curves are based on the same tree with appropriate adjustments to ensure comparability (see text). The Snyder likelihoods conditioning on the survival of the tree from either its start (time 0) or MRCA (time c_1) are denoted by circle and square markers respectively.

3.3 Time Varying Rate Birth-Death Estimation

We now consider estimating BDPs with deterministically time-varying per lineage birth and death rates $\lambda(t, \vec{x}_\lambda)$ and $\mu(t, \vec{x}_\mu)$. As mentioned in Section 2.2, the vectors \vec{x}_λ and \vec{x}_μ are spanning subsets of the parameter set \vec{x} . Hence each parameter must appear at least once. Often we will use $\lambda(t)$ and $\mu(t)$ for convenience. These types of models, while not as tractable as the constant rate BDP, are important for describing complex diversification dynamics, such as adaptive radiations or mass extinctions (Paradis, 2010). Time varying BDPs admit no explicit inference solutions (Paradis, 2010), so we numerically integrate Eq. (1)-Eq. (3) across the branch times of a given phylogeny. In Section 2.2 we observed that this rBDP is equivalent to a Markov birth process, $\mathcal{F}(t)$, with rate $\beta(t, \vec{x}, \mathcal{F}(t)) \leq \lambda(t, \vec{x}_\lambda)\mathcal{F}(t)$, given by Eq. (6). Eq. (9) shows that the parametric form of this rate is a complex functional with nested integrals. The rate is Markov since it only depends on the current rBDP lineage count (Snyder and Miller, 1991).

$$\beta(t, \vec{x}, \mathcal{F}(t)) = \frac{\lambda(t, \vec{x}_\lambda)\mathcal{F}(t)}{1 + \int_t^T \mu(\tau, \vec{x}_\mu) e^{\int_t^\tau \mu(s, \vec{x}_\mu) - \lambda(s, \vec{x}_\lambda) ds} d\tau} \quad (9)$$

The inference problem is to find conditional mean estimates of the parameters, $\hat{x}_i = \mathbb{E}[x_i | \mathcal{F}_T]$. The SF will directly generate the joint posterior $\mathbb{P}(\vec{x} | \mathcal{F}_T)$ and \hat{x}_i can be obtained by marginalising and then integrating the marginalised posterior across its domain.

We investigate two BDP models with time-varying rates. The first has a constant death rate $\mu(t) = x_3$ and an exponentially decreasing birth rate $\lambda(t) = x_3 + x_1 e^{-x_2 t}$. We call this the speciation-decay model. It was introduced in Hohna (2014) to describe a speciation rate that initially starts above the extinction rate and then decays to $\lambda(t) = \mu(t)$. This model and various nested special cases of it, were used to solve a model selection problem on empirical ant and snake phylogenies. The second model uses a logistic function for both the birth and death rate so that: $\lambda(t) = (1 + e^{-x_1 t + x_2})^{-1}$ and $\mu(t) = (1 + e^{-x_3 t + x_4})^{-1}$. This logistic model was used by Paradis (2010) to capture the monotonically increasing or decreasing divergence rates commonly found in macroevolution.

We simulated rBDPs from each model using the algorithms of Hohna (2013), available in the R package TESS (Hohna *et al.*, 2016). This involved inversely sampling the k^{th} rBDP speciation time, c_k , by solving $r_k = \left(\int_0^{c_k} \lambda(t) P_1(t, T) dt \right) \left(\int_0^T \lambda(t) P_1(t, T) dt \right)^{-1}$ with $P_1(t, T) = P(t, T)^2 e^{-\alpha(t, T)}$ as the probability that a lineage at time t leaves exactly 1 surviving descendant at T . Note that $P(t, T)$ (see Eq. (4)) is the analogous probability for when at least 1 descendant survives to T . Each r_k is uniformly distributed in $[0, 1]$ and the tree generated is for n lineages at time T so that $\mathcal{F}(T) = n$. Where possible, we simulated under the parameter values reported in Hohna (2014) and Paradis (2010). We then applied the SF to the simulated trees and compared its estimates to the true parameter values. We benchmarked its performance by analysing the same simulated trees using a recent adaptive MCMC inference method (Hohna, 2013) (Hohna *et al.*, 2016) that is included in the TESS package.

We conditioned our rBDP trees to start from the MRCA of the observed lineages since this is a more practical scenario. We therefore start at c_1 instead of 0 and $\mathcal{F}(c_1) = 2$. For these inference problems, the MCMC method samples from the likelihood function given in Eq. (10). This is the time-varying form of likelihood (5) from Stadler (2013) (see Section 3.2).

$$L = \frac{P_1(0, T)^2}{P(0, T)^2} \prod_{k=2}^{n-1} k \lambda(c_k) P_1(c_k, T) \quad (10)$$

We estimated model parameters from 100 replicate simulated rBDP trees with $n = 100$ tips, using both the SF and MCMC methods. To keep comparisons fair we used the same uniform priors and parameter ranges for both methods. For the SF we set probability $\frac{1}{m_i}$ across the i^{th} parameter grid, with $m_i = [20, 15]$ respectively for the speciation-decay and logistic models. We checked MCMC convergence using the Geweke statistic (Cowles and Carlin, 1996) and by examining the autocorrelation function of the MCMC samples. **Fig 3** shows the resulting overall marginal posteriors from both methods, obtained by combining the estimated marginal posteriors across all replicate trees. The discrepancy between the estimates arises from the different parameter discretisations,

used implicitly in MCMC, and explicitly in the SF. We confirmed this by observing that, within numerical tolerances, the likelihoods from Eq. (8) and Eq. (10) are identical. Importantly, both methods give very similar marginal distributions and the true parameter values are within the coverage of the posteriors, thus confirming the performance of the SF.

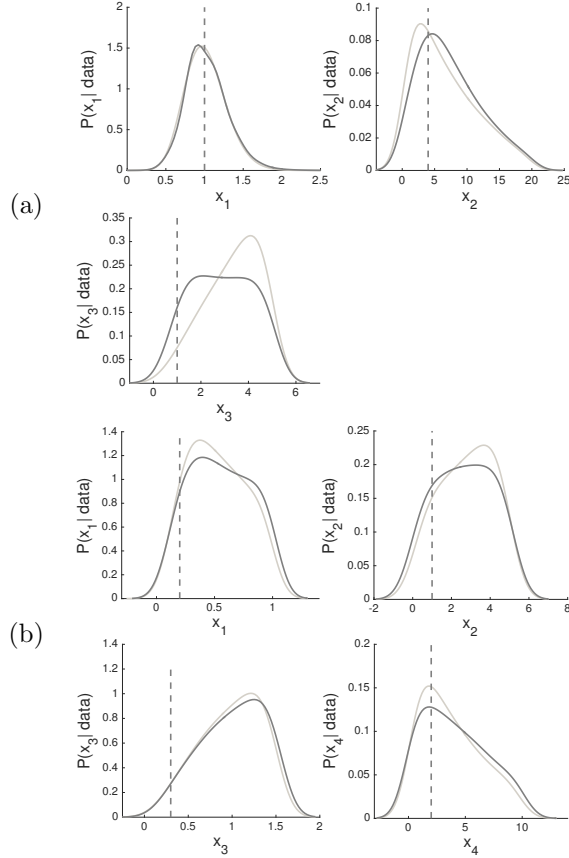


Figure 3: Estimated marginal posteriors for the parameters of time-varying BDP models. We simulated 100 trees with $n = 100$ tips and then estimated the underlying parameters of the model using the SF and MCMC methods. Results for the speciation-decay and logistic models are in (a) and (b) respectively. We used uniform priors over $m = 20^3$ points spanning $[0.1, 0.4, 0.1]$ to $[5, 20, 5]$ for parameters in (a) and over $m = 15^4$ points within $[0.02, 0.1, 0.03, 0.2]$ to $[1, 5, 1.5, 10]$ for those in (b). We applied a normal smoothing kernel to the resulting estimates. The MCMC posteriors are in light grey and the SF ones in darker grey. The true parameter values are vertical dashed lines.

3.4 Birth-death Estimation with Empirical Data

As the SF is flexible and easily implemented it should be useful for model selection problems. To show this, we analyse the Australian Agamid lizard dataset from Harmon *et al.* (2003), which is known to be almost completely sampled (93%) at the species level. Previous work by Rabosky and Lovette (2008b) tested 4 nested BDP models: (i) constant birth-death, *const*: $\lambda(t) = x_1$, $\mu(t) = x_2$; (ii) time-varying speciation, *spvar*: $\lambda(t) = x_1 e^{-x_2 t}$, $\mu(t) = x_3$; (iii) time-varying extinction, *exvar*: $\lambda(t) = x_1$, $\mu(t) = x_3(1 - e^{-x_2 t})$ and (iv) time-varying birth and death, *bothvar*: $\lambda(t) = x_1 e^{-x_2 t}$, $\mu(t) = x_4(1 - e^{-x_3 t})$. Rabosky and Lovette (2008b) found that the data was best described by the *spvar* model, thereby supporting a hypothesis of declining diversification. We apply the SF to these 4 parametric models. Our observable input data is the time scaled Agamid tree given in **supplementary material VI**.

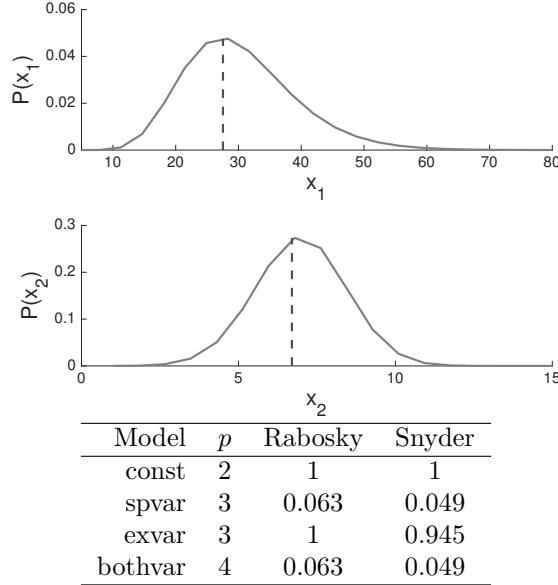


Figure 4: **Model selection for the Agamid phylogeny** ($n = 69$). The SF and the Rabosky and Lovette (2008b) methods were applied to the *const*, *spvar*, *exvar* and *bothvar* BDP models. Grids with $m_i = 30$ were used for all models except *bothvar* which had $m_i = 20$. SF priors were uniform over the ranges $[(0, 10), (0, 1)]$, $[(1, 100), (1, 25), (10^{-3}, 0.01)]$, $[(0.01, 10), (0.01, 10), (10^{-3}, 0.01)]$ and $[(1, 100), (1, 25), (0.01, 1), (10^{-3}, 0.01)]$. The table gives the relative fit based on a normalised Paradis (2010) metric (lower values imply better fits, p is the model dimension), and supports *spvar* as the best model. The sub-figures compare the SF marginal posteriors (solid) with the MLEs from Rabosky and Lovette (2008b) (dashed), for *spvar*. No comparison is provided for x_3 as it is not well informed by the data.

We performed model selection under both the SF and the Rabosky and Lovette (2008b) approaches by using the least squares technique developed by Paradis (2010). This method (i) converts the tree into an empirical cumulative distribution function (CDF), (effectively a scaled LTT plot), $F_e(t)$, (ii) computes the theoretically theoretical CDF, as $F(t, \epsilon) = \left(\int_0^t \lambda(s) e^{\alpha(0, s)} P(s, T)^2 ds \right) \left(\int_0^T \lambda(s) e^{\alpha(0, s)} P(s, T)^2 ds \right)^{-1}$, for any given value of the parameter vector, $\vec{x} = \epsilon$, and then (iii) assesses the accuracy of a model with parameters ϵ by the square error metric $\int_0^T (F_e(t) - F(t, \epsilon))^2 dt$. We calculated this metric for each model, with ϵ as its vector of parameter estimates, and then normalised by the maximum model square error. The results of this procedure are shown in the table in **Fig 4**. Here smaller values indicate better fits and p is the model dimension. The ‘Rabosky’ values are obtained by computing the maximum likelihood estimates (MLEs) from Rabosky and Lovette (2008b), using the R package LASER (Rabosky, 2006), and then applying the Paradis (2010) metric. The ‘Snyder’ values are similarly computed using a SF with uniform priors of $\frac{1}{m_i}$ on each parameter for every model.

The relative fit shown in this table is consistent between both inference methods, recommends *spvar* as the best model, and matches the results reported in the original Rabosky and Lovette (2008b) analysis. The expected CDF ($F(t, \epsilon)$) generated by the SF for each model, and the empirical $F_e(t)$ are provided in **supplementary material VI** for further visualisation of the relative fits. We used the Paradis (2010) metric, because it is easy to calculate and naturally links the estimates of unobserved parameters to the observable LTT. We also computed the Akaike information criteria for each model, from both methods. These are given in **supplementary material VI** and reaffirm our conclusions.

Fig 4 also presents the SF marginal posteriors (solid) and the MLEs from LASER (dashed) for the best fit *spvar* model. We find a close correspondence between these estimates. The posterior for the death parameter, x_3 , is not shown because it remained unchanged from the set prior. This suggests that either the death rate parameter is redundant or that the Agamid tree contains effectively no information about it. Rabosky and Lovette (2008b) also found x_3 to be a spurious parameter, although their MLE, which was just above 0, could be open to interpretation.

4 Discussion

We have introduced the Snyder filter as a new Bayesian algorithm for solving BDP inference problems, and demonstrated its efficacy on several BDP models and datasets. As it only depends on linear ordinary differential equations, the SF is simple, stable and deterministic. Provided that we can define parametric functions for the birth and death rates through time, the SF presents a direct and reproducible way of computing MMSE estimators.

We initially tested the SF on the constant rate BDP, since its behaviour is well understood. The SF not only confirmed known estimation trends, but also performed as well as a previous least squares optimisa-

tion method (see **supplementary material III**) (Paradis, 2010). Our method therefore maintains estimator accuracy whilst avoiding non-linear optimisation algorithms that may be susceptible to local extrema. Under the constant rate BDP, MMSEs fell with μ , implying that the smallest MMSE would be achieved as $\mu \rightarrow 0$. At this limit we obtain the Yule model. In **supplementary material IV** we analytically solved the SF for this model, deriving this MMSE estimator. We also found a sampling condition under which constant rate BDPs with $\mu \neq 0$ behaved like the Yule model. These results hint at the potential analytical usefulness of the SF.

At higher death rates, estimation, in addition to being less accurate, becomes more sensitive to rBDP conditioning (Stadler, 2013). This can create biases (not just for constant rate models) if the conditions under which empirical trees are obtained do not match those of the likelihood solved by a chosen inference scheme. Stadler (2013) examined 7 distinct rBDP likelihoods and found that those which condition only on the survival of the tree were the most robust to mismatches and hence the most useful for study. We found that the SF solves exactly these likelihoods. The $P(t, T)$ function in Eq. (7) appears to be the source of this implicit conditioning.

We then extended our analysis to time-varying BDPs. We benchmarked the SF against a modern MCMC method by Hohna *et al.* (2016), on trees simulated under the speciation-decay and logistic models given in Hohna (2014) and Paradis (2010) respectively. Both inference methods gave comparable marginal posteriors. This comparison highlighted a relative advantage of our algorithm. The MCMC method sometimes required multiple runs to avoid poor convergence and could give different results among runs on the same data. Result reproducibility is not guaranteed in MCMC techniques and often non-trivial indices need to be calculated to evaluate convergence (Cowles and Carlin, 1996). In contrast, the SF will always produce the same posteriors for the same observed phylogeny and parameter grid. The SF posteriors can therefore serve as useful references for debugging or validating MCMC and other randomised inference strategies. We present an example of this application for MCMC runs under the speciation-decay model, in **supplementary material VII**.

In terms of computational speed, we found, for the models we investigated, that our non-optimised Matlab implementation of the SF completes in a shorter time than the MCMC technique of Hohna *et al.* (2016). We provide a comparison of execution times for these BDP models in **supplementary material II**. MCMC methods could potentially be faster than the SF for higher dimensional parametric models, due to the grid based nature of the latter. However, in such scenarios there is always a question about whether non-parametric approaches are more suitable. A related discussion on SF computational and methodological complexity is given in Parag and Pybus (2017).

We also investigated a model selection problem on an empirical phylogeny of Agamids, first analysed by Rabosky and Lovette (2008b). The SF reproduced the relative model fit of Rabosky and Lovette (2008b), providing evidence for declining diversification, and matched the birth rate parameter MLEs for this dataset. The Rabosky and Lovette (2008b)

MLE for the death rate parameter was approximately 0, while the SF produced a posterior that matched its prior. This illustrates the transparency offered by working with complete distributions instead of point estimates. The SF clearly suggests that this parameter is redundant, or equivalently, that the data is not informative enough about this parameter. However, in using the Rabosky and Lovette (2008b) method, we would need to examine the complete likelihood function to distinguish between the competing hypotheses of an actual extinction rate of 0 (which is unlikely (Purvis, 2008)) and insufficient statistical power.

The SF presents a capable alternative BDP inference technique that, within numerical tolerances, provides exact MMSE estimates by directly computing the joint parameter posterior. It is simple and does not suffer from algorithmic stability issues like local minima or poor convergence. It exploits the Markov birth nature of rBDPs, which should allow easy extension to more complex BDPs. Our future work will generalise the SF to allow for genealogical uncertainty, incomplete sampling and non-linear dependence. To account for tree uncertainty we could run SFs on a covering set of trees, and then combine the results in a Bayesian manner. This is similar to superposing multiple Poisson streams and derives from taking the Markov birth process approach (Snyder and Miller, 1991).

We have shown in **supplementary material I** how to accommodate fixed incomplete sampling (also known as uniform taxon sampling) by replacing $P(t, T)$ with $P_\nu(t, T)$, the probability that a lineage has at least one descendant at T , and is also sampled. As long as a sampling process admits a description for $P_\nu(t, T)$ then the SF can be applied. The SF can also handle non-linear BDPs, in which birth and death rates become non-linear functions of $\mathcal{F}(t)$. Here the structure of β from Eq. (7) becomes more complex because the birth and death rates generalise to $\lambda(t, \vec{x}_\lambda, \mathcal{F}(t))$ and $\mu(t, \vec{x}_\mu, \mathcal{F}(t))$. The inference problem, however, is the same, as $\mathcal{F}(t)$ is known (observable), and the parameter space is unchanged. When the extinction rate is zero then density dependence, as defined in Rabosky and Lovette (2008a), falls within this class of models.

Funding

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.

References

- Bobrowski, O., Meir, R., and Eldar, Y. (2008). Bayesian Filtering in Spiking Neural Networks; Noise, Adaptation and Multisensory Integration. *Neural Computation*, **21**, 1277–1320.
- Cowles, M. and Carlin, B. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883–904.
- Gernhard, T. (2008). The Conditioned Reconstructed Process. *Journal of Theoretical Biology*, (253), 769–78.

- Harmon, L., Schulte II, J., Larson, A., and Losos, J. (2003). Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, **301**, 961–4.
- Hartmann, K., Wong, D., and Stadler, T. (2010). Sampling Trees from Evolutionary Models. *Syst. Biol.*, **59**(4), 465–76.
- Harvey, P., May, R., and Nee, S. (1994). Phylogenies without Fossils. *Evolution*, **48**(3), 523–9.
- Hohna, S. (2013). Fast Simulation of Reconstructed Phylogenies under Global Time-Dependent Birth–Death Processes. *Bioinformatics*, **29**(11), 1367–74.
- Hohna, S. (2014). Likelihood Inference of Non-Constant Diversification Rates with Incomplete Taxon Sampling. *PLoS ONE*, **9**(1).
- Hohna, S. (2015). The Time-dependent Reconstructed Evolutionary Process with a Key-role for Mass-extinction Events. *Journal of Theoretical Biology*, **380**, 321–31.
- Hohna, S., Stadler, T., Ronquist, F., and Britton, T. (2011). Inferring Speciation and Extinction Rates under Different Sampling Schemes. *Mol. Biol. Evol.*, **28**(9), 2577–89.
- Hohna, S., May, M., and Moore, B. (2016). Tess: an R Package for Efficiently Simulating Phylogenetic Trees and Performing Bayesian Inference of Lineage Diversification Rates. *Bioinformatics*, **32**(5), 789–91.
- Kendall, D. (1948). On the Generalized Birth and Death Process. *Ann. Math. Stat.*, **19**, 1–15.
- Kingman, J. (1982). On the Genealogy of Large Populations. *Journal of Applied Probability*, **19**, 27–43.
- Kubo, T. and Iwasa, Y. (1995). Inferring the Rates of Branching and Extinction from Molecular Phylogenies. *Evolution*, **49**(4), 694–704.
- Kuhnert, D., Stadler, T., Vaughan, T., and Drummond, A. (2016). Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. *Mol. Biol. Evol.*, **33**(8), 2102–16.
- Morlon, H. (2014). Phylogenetic Approaches for Studying Diversification. *Ecology Letters*, **17**, 508–25.
- Morlon, H., Parsons, T., and Plotkin, J. (2011). Reconciling Molecular Phylogenies with the Fossil Record. *PNAS*, **108**(39), 16327–32.
- Mossel, E. and Vigoda, E. (2006). Limitations of Markov Chain Monte Carlo Algorithms for Bayesian Inference of Phylogeny. *Ann. Appl. Prob.*, **16**(4), 2215–34.
- Nee, S. (2001). Inferring Speciation Rates from Phylogenies. *Evolution*, **55**(4), 661–8.
- Nee, S., May, R., and Harvey, P. (1994). The Reconstructed Evolutionary Process. *Phil Trans R Soc B*, **344**, 305–11.
- Paradis, E. (2004). Can extinction rates be estimated without fossils? *Journal of Theoretical Biology*, **229**, 19–30.
- Paradis, E. (2010). Time-Dependent Speciation and Extinction from Phylogenies: a Least Squares Approach. *Evolution*, **65**(3), 661–72.
- Parag, K. and Pybus, O. (2017). Optimal Point Process Filtering and Estimation of the Coalescent Process. *Journal of Theoretical Biology*, pages 153–67.
- Parag, K. and Vinnicombe, G. (2017). Point Process Analysis of Noise in Early Invertebrate Vision. *PLoS Computational Biology*, **13**(10), e1005687.

- Purvis, A. (2008). Phylogenetic Approaches to the Study of Extinction. *Ann. Rev. Ecol. Evol. Syst.*, **39**, 301–19.
- Pybus, O. and Harvey, P. (2000). Testing Macro-evolutionary Models using Incomplete Molecular Phylogenies. *Proc. R. Soc. Lond. B*, **267**, 2267–72.
- Pyron, R. and Burbink, F. (2013). Phylogenetic Estimates of Speciation and Extinction Rates for Testing Ecological and Evolutionary Hypotheses. *Trends in Ecology and Evolution*, **28**(12), 729–36.
- Rabosky, D. (2006). LASER: A Maximum Likelihood Toolkit for Detecting Temporal Shifts in Diversification Rates from Molecular Phylogenies. *Evolutionary Bioinformatics*, **2**, 247–50.
- Rabosky, D. and Lovette, I. (2008a). Density-Dependent Diversification in North American Wood Warblers. *Proc. R. Soc. B*, **275**, 2363–r71.
- Rabosky, D. and Lovette, I. (2008b). Explosive Evolutionary Radiations: Decreasing Speciation or Increasing Extinction through Time. *Evolution*, **62**(8), 1866–75.
- Rudemo, M. (1972). Doubly-Stochastic Poisson Processes and Process Control. *Advances in Applied Probability*, **2**, 318–338.
- Snyder, D. (1972). Filtering and Detection for Doubly Stochastic Poisson Processes. *IEEE Transactions on Information Theory*, **18**, 91–102.
- Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer-Verlag, 2 edition.
- Stadler, T. (2009). On Incomplete Sampling under Birth–Death Models and Connections to the Sampling-Based Coalescent. *Journal of Theoretical Biology*, **261**, 58–66.
- Stadler, T. (2013). How can we Improve Accuracy of Macroevolutionary Rate Estimates. *Syst. Biol.*, **62**(2), 321–9.
- Stadler, T., Kuhnert, D., Bonhoeffer, S., and Drummond, A. (2013). Birth-death Skyline Plot reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (hcv). *PNAS*, **110**(1), 228–33.
- Stadler, T., Vaughan, T., Gavryushkin, A., *et al.* (2015). How well can the Exponential-Growth Coalescent Approximate Constant-Rate Birth-Death Population Dynamics? *Proc. R. Soc. B*, **282**.
- Volz, E. and Frost, S. (2014). Sampling through Time and Phylodynamic Inference with Coalescent and Birth–death Models. *J. R. Soc. Interface*, **11**(20140945).
- Yang, Z. and Zhu, T. (2018). Bayesian Selection of Misspecified Models is Overconfident and may cause Spurious Posterior Probabilities for Phylogenetic Trees. *PNAS*, **115**(8), 1845–9.