Link between the birth-death and the Kingman coalescent - Applications to phylogenetic epidemiology

Josselin Cornuault, Fabio Pardi, Celine Scornavacca

Link between the birth-death and the Kingman coalescent

—

Applications to phylogenetic epidemiology

Josselin Cornuault[1,*], Fabio Pardi[2], Celine Scornavacca[1]

[1]ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France
[2]LIRMM, Université de Montpellier, CNRS, Montpellier, France
[*]joss.cornuault@gmail.com

Abstract

The two most popular tree models used in phylogenetics are the birth-death (BD) and the Kingman coalescent (KC). These two models differ in several respects, notably: (i) population size is random in the BD versus fixed in the KC, (ii) the BD makes assumptions about the way samples are collected, while the KC conditions on the number of samples and the collection times, thus bypassing the need to describe the sampling procedure. These two models have been applied to different contexts: the BD in macroevolutive studies of clades of species, and the KC for populations. The exception is the field of phylogenetic epidemiology which uses both models. It then asks the question of how such different models can be used in the same context. In this paper, we study large-population limits of the BD, in a search for a mathematical link between the BD and the KC. We show that the KC is the large-population limit of a BD conditioned on a given population trajectory, and we provide the formula for the parameter θ of the limiting KC. This formula appears in earlier studies, but the present article is the first to show formally how the correspondence arises as a large-population limit, and that the BD needs to be conditioned for the KC to arise. Besides these fundamentally mathematical results, we demonstrate how our findings can be used practically in phylogenetic inference. In particular, we propose a new method for phylogenetic epidemiology, ensuing from our results. We conjecture that this new method, used in conjunction with auxiliary data, should allow to estimate important epidemiological parameters (e.g. the prevalence and the effective reproduction number), in a way that is robust to the data-generating model and the sampling procedure. Future studies will be needed to put our claims to the test.

Keywords— birth-death model, Kingman coalescent, scaling limits, phylogenetic inference, phylogenetic epidemiology

Introduction

The field of phylogenetics makes ample use of tree models to obtain a probability distribution of a tree given an evolutionary model. Tree models may be used as tree priors in Bayesian phylogenetic inference (e.g. 1–3) or fitted to pre-estimated phylogenies in order to estimate evolutionary parameters (4). In either case, the choice of tree model is determined by the type of object that the phylogenetic tree represents.

The overwhelming majority of phylogenetic applications of tree models resorts to two canonical models: the birth-death (BD, 5) and the Kingman coalescent (KC, 6). The BD is typically used in macroevolution where the phylogenetic tree represents a species tree (7–9). It makes the simplification that speciation and extinction are instantaneous phenomena, whose frequency is governed by speciation (birth) and extinction (death) rates. In macroevolutionary studies, the BD is used to study the dynamics of species diversification across space (10, 11), time (12, 13), taxa (14), habitats (15, 16), etc... In contrast, the KC is generally used in population studies, where a phylogeny is taken to represent the genealogy of gene copies, borne by individuals in a population (17, 18). Its unique parameter is the instantaneous effective population size $\theta(t)$, which determines the rate of coalescence of any two gene copies. The KC has been most popular to study populations in part because of its robustness, as many population models converge to the KC in the limit of large population size (19). Its most popular application has been for demographic inference, in particular for estimating temporal variations of the effective population size (2, 20).

Thus, as a rule, the two tree models are used in different contexts: the BD for clades of species and the KC for populations of individuals (21, 22, but see 23 who applied the KC to a species tree). A notable exception to this schism is the phylogenetic study of epidemics, wherein the object of study is a population of infected hosts. Both the BD and the KC have been found suitable to model the phylogenetic relationships among pathogens borne by hosts (24, 25). On the one hand, it makes sense to assume that populations of pathogens too have an effective population size that governs the rate of coalescence of pathogens' gene copies, justifying the use of the KC (26–28). On the other hand, the transmission of a pathogen to an uninfected host (aka the "birth" of a new infected host) and the death (or recovery) of an infected host are seemly modeled with birth and death rates, justifying the use of the BD (29–31).

Given that the BD and the KC are both used in an epidemiological context, it makes sense to ponder why two models with such different mathematical foundations can be applied to the same object of study. The KC models the coalescence of the branches in the phylogeny of a sample from present to past. It is parametrized with a single function $\theta(t)$, and the coalescence of any two branches occurs at rate $1/\theta(t)$. Importantly, the KC is the large-population limit of population models whereby the census population size $N(t)$ is represented by a deterministic function of time (19). In con-

trast, the BD unravels from past to present and considers that $N(t)$ is random: $N(t)$ is increased (decreased) by births (deaths) occurring randomly at per-capita rate $\lambda(t)$ ($\mu(t)$). A phylogenetic tree of the sample is obtained by pruning unsampled lineages off of the entire population tree (32). The fact that population size is represented by a deterministic function in the KC and a random variable in the BD is certainly a serious obstacle to the identification of mathematical bridges between the two models.

Nonetheless, previous studies reported that, if the BD population size were known at some time t, any two branches in the tree should coalesce at that time at a rate given by $2\lambda(t)/N(t)$ (24, 33, 34). This suggests the following equivalence between the parameter of the KC and quantities of the BD:

$$\theta(t) \equiv \frac{N(t)}{2\lambda(t)}$$

To remedy the inconsistency of having a deterministic left-hand side and a random right-hand side in this equivalence relation, a deterministic function may be substituted for $N(t)$ to try and have the KC match the BD as closely as possible. For instance, one may evaluate whether a KC with $\theta(t) = \frac{E[N(t)|\lambda,\mu]}{2\lambda(t)}$ matches the corresponding BD (e.g. 24). The mere inspection of the formulae for the probability density of a tree under the two models indicates that the two models remain different, no matter what parametrization of $\theta(t)$ is chosen.

In the present paper, we build on this previous work and demonstrate a mathematical link between the BD and the KC. To do so we note that, if the BD population size were (approximately) known at all times (e.g. with $y(t)$ a non-random function, $N(t) \approx y(t)$ is given), then the branches of a BD tree should coalesce indeed at a rate close to $2\lambda(t)/y(t)$ at time t. This is a hint that the BD must be conditioned on the population size following some trajectory for the KC to arise. We thus study a BD conditioned on its population size being close to some curve (hereafter a *conditional* BD, in opposition to an *unconditional* BD without a condition on population size), and show that the KC arises as the large-population limit of the conditional BD. Beyond this fundamentally mathematical result, we demonstrate how this finding can be applied in an epidemiological context to estimate parameters of interest such as a pathogen prevalence or reproduction number with the KC (seen as the limit of a conditional BD). Importantly, the advantages of the KC (agnostic to the sampling procedure) and of the BD (whose parameters can be transformed into epidemiologically meaningful quantities) are combined into this new modelling approach.

The paper is structured as follows. After outlining the BD model, we consider various limits of the (un)conditional BD. First, we study the limit of the unconditional BD with fixed $\lambda(t)$ and $\mu(t)$ and with an initial number of individuals n going to infinity. We argue that the limit of this model is the trivial KC with $\theta(t) = \infty$. Second, we consider the unconditional BD with fixed $\lambda_1(t) := \lambda_n(t)/n$ and $r(t) = \lambda_n(t) - \mu_n(t)$ and show that as $n \to \infty$ the BD converges to something else than a KC. Third, we show that the conditional BD with fixed $\lambda_1(t)$ and $r(t)$ converges as $n \to \infty$ to a non-trivial KC, which constitutes the main result of this paper. After this mathematical section, we demonstrate how our results can be applied to the inference of epidemiological parameters from a phylogenetic tree.

Outline of the birth-death parameters and random outcomes

The variable t denotes the time before present. We consider a BD process starting at time T with n individuals. Each individual gives birth at rate $\lambda(t)$ and dies at rate $\mu(t)$. The growth rate is $r(t) = \lambda(t) - \mu(t)$. We denote by $N(t)$ the integer random population size at time t, and by $N := \big(N(t)\big)_t$ the value of $N(t)$ at all times. Initially, $N(T) = n$. We call N the *population process*. We call a realization of the population process a *population trajectory*. At the present, $Z(0)$ individuals are sampled randomly, each with probability ρ (Bernoulli sampling), and $Z(t)$ denotes the random number of ancestors of the sampled individuals at time t. We denote $Z := \big(Z(t)\big)_t$ the value of $Z(t)$ at all times and call Z the *sample process*. A valid realization $z = \big(z(t)\big)_t$ of Z is a decreasing integer-valued function of time with unit decrements. A value z induces a set $\alpha(z)$ of *coalescent times*, which are the time points at which z decreases.

Our purpose is to study the limiting distributions as $n \to \infty$ of some random outcomes of the BD process, namely the relative population size N/n and the sample process Z, and in particular to determine in which case Z converges to a KC.

Various limits of the BD process

The unconditional BD with fixed $\lambda(t)$ and $\mu(t)$ as $n \to \infty$

The most obvious large-population limit of the BD is to keep $\lambda(t)$ and $\mu(t)$ fixed and have $n \to \infty$. We study here the limits of $N(t)/n$ and Z and argue that, if Z does converge to a KC, it converges to the trivial KC with $\theta(t) = \infty$. The BD process considered here is called *unconditional*, as no condition on N is assumed.

Limit of the relative population size $N(t)/n$

The first result is that, as shown in Lemma C.5 in the Appendix, $N(t)/n$ converges in probability to $\mathbb{E}[N(t)|N(T) = n]/n = \exp\int_t^T r(s)ds$ as $n \to \infty$, uniformly on $[0, T]$ (see Figure 1 for an illustration), that is

$$\inf_{t\in[0,T]} \mathbb{P}\left(\left| \frac{N(t)}{n} - \frac{\mathbb{E}[N(t)|N(T) = n]}{n} \right| < \epsilon \;\middle|\; N(T) = n \right) \xrightarrow[n\to\infty]{} 1$$

This immediately implies that $N(t)/\mathbb{E}[N(t)|N(T) = n]$ converges in probability to 1 as $n \to \infty$. Hence, $N(t)/n$ becomes deterministic in the limit of large population size.

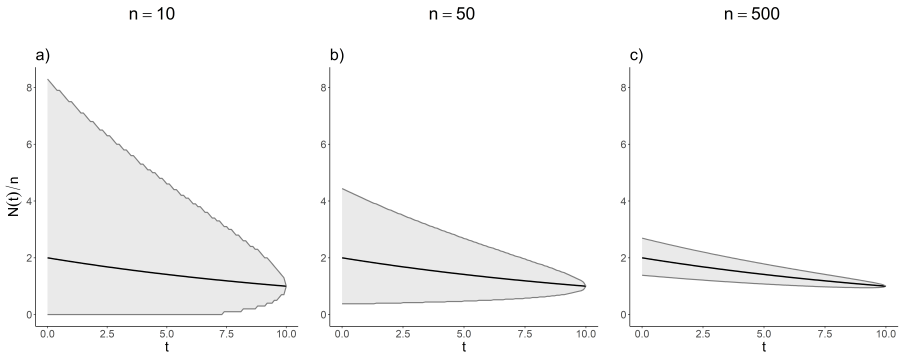

Figure 1: Distribution of the relative population size $N(t)/n$ for the unconditional BD, for different values of n, with fixed $\lambda(t) = 1$ and $r(t) \simeq 0.069$. The black line is the expected value (equal to $\exp \int_t^T r(s)ds$) and the ribbon ranges from the .025 to the .975 quantile. The distribution narrows as n increases, illustrating the convergence in probability of $N(t)/n$ to the deterministic function $\exp \int_t^T r(s)ds$.

Limit of the sample process Z

We now turn to the distribution of Z given $Z(0) = z_0$ (because the KC is expressed conditionally on the sample size, the BD must also be conditioned on the sample size if the two models are to be matched). In Section J.1 of the Appendix, we provide a heuristic suggesting that the pairwise coalescent rate under the BD is given by (see also 24, 33, 34)

$$\omega_{BD}\big(N(t), t\big) = \frac{2\lambda(t)}{N(t)}$$

This rate is random, since $N(t)$ is random under the BD.

Given that $N(t)/\mathbb{E}[N(t)|N(T) = n] \xrightarrow[n\to\infty]{p} 1$, if the BD converges to a KC, it must be to the KC with parameter $\theta(t) = \mathbb{E}[N(t)|N(T) = n]/\big(2\lambda(t)\big)$, whose pairwise coalescent rate is

$$\omega_{KC}(t) = \frac{1}{\theta(t)} = \frac{2\lambda(t)}{\mathbb{E}[N(t)|N(T) = n]}$$

and is deterministic.

Indeed we have that the ratio of the two models' coalescent rates converges to 1 in probability:

$$\frac{\omega_{KC}(t)}{\omega_{BD}\big(N(t)\big)} = \frac{N(t)}{\mathbb{E}[N(t)|N(T) = n]} \xrightarrow[n\to\infty]{p} 1 \tag{1}$$

It is however important to note that the matching KC has a null coalescent rate in the limit $n \to \infty$, since

$$\omega_{KC}(t) = \frac{2\lambda(t)}{\mathbb{E}[N(t)|N(T)=n]} = \frac{2\lambda(t)}{n \exp \int_t^T r(s)ds} \xrightarrow[n\to\infty]{} 0$$

As shown in Figure 2, this limit is not useful, as the limiting process is the process whereby the outcome $\forall t \in [0,T]\ Z(t) = z_0$ has probability 1.

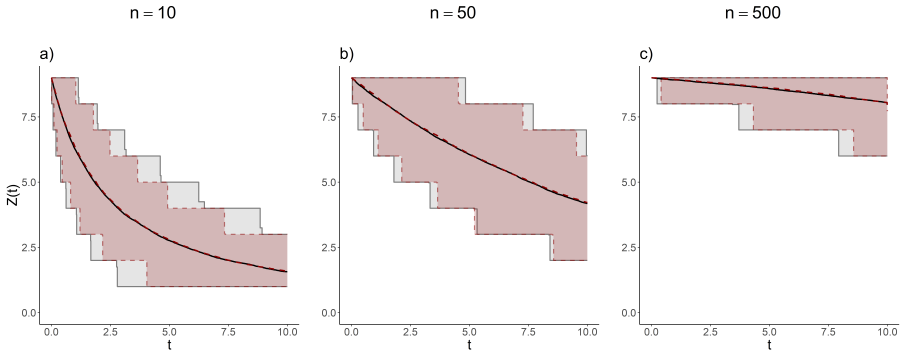

Figure 2: Distribution of the sample process Z given $Z(0) = z_0 = 9$ for the unconditional BD (black solid line and grey ribbon), for different values of n, with fixed $\lambda(t) = 1$ and $r(t) \simeq 0.069$. The thick line is the expected value and the ribbon ranges from the .025 to the .975 quantile. For comparison, the red dashed line and red ribbon show the distribution of Z under a KC with $\theta(t) = \mathbb{E}[N(t)|N(T) = n]/(2\lambda(t))$. We see that, as n increases, although the two distributions get more similar, coalescences are ever rarer, and they would not happen at all in the limit $n \to \infty$.

Summary

There are good arguments to reckon that the pairwise coalescent rate of the BD is given by $2\lambda(t)/N(t)$ at time t (see Section J.1 of the Appendix and 24, 33, 34). Then the obvious problem for matching the BD with a KC is that this rate is random under the BD (since $N(t)$ is random). Yet, the convergence in probability of $N(t)/\mathbb{E}[N(t)|N(T) = n]$ to 1 as $n \to \infty$ could imply that the BD converges to a KC with coalescent rate $2\lambda(t)/\mathbb{E}[N(t)|N(T) = n]$ as $n \to \infty$. Unfortunately, this limiting process is uninteresting since it has a null coalescent rate.

The unconditional BD with fixed $\lambda_1(t)$ and $r(t)$ as $n \to \infty$

We now study the BD parametrized with ρ_1, $\lambda_1(t)$ and $r(t)$ such that

$$\begin{aligned}\lambda(t) &= \lambda_n(t) = \lambda_1(t)n \\ \mu(t) &= \mu_n(t) = \lambda_n(t) - r(t) \\ \rho &= \rho_n = \rho_1/n\end{aligned}$$

with $\rho_1 \in (0, 1]$, and with on $[0, T]$: (i) $\lambda_1(t) > 0$, (ii) $\lambda_1(t)$ and $r(t)$ uniformly continuous and (iii) $\lambda_1(t) \geq r(t)$. We consider this parametrization because the BD coalescent rate, as we shall see below, does not vanish as $n \to \infty$, yielding non-trivial limiting processes.

We show that N/n remains random as $n \to \infty$, and that Z converges in distribution to a random process $\dot{Z}$ which is not a KC.

Limit of the relative population size $N(t)/n$

Lemma D.1 shows that $\forall n \in \mathbb{N}_{>0}$

$$\mathbb{E}\left[\frac{N(t)}{n}\middle| N(T) = n\right] = \exp\int_t^T r(s)ds$$

and Lemma D.2 shows that

$$\mathrm{Var}\left[\frac{N(t)}{n}\middle| N(T) = n\right] \xrightarrow[n\to\infty]{} \exp\left(2\int_t^T r(s)ds\right)\int_t^T \frac{2\lambda_1(s)ds}{\exp\left(\int_s^T r(u)du\right)} > 0$$

In other words, $N(t)/n$ does not converge in probability to its mean but remains random instead (see Figure 3 for an illustration).

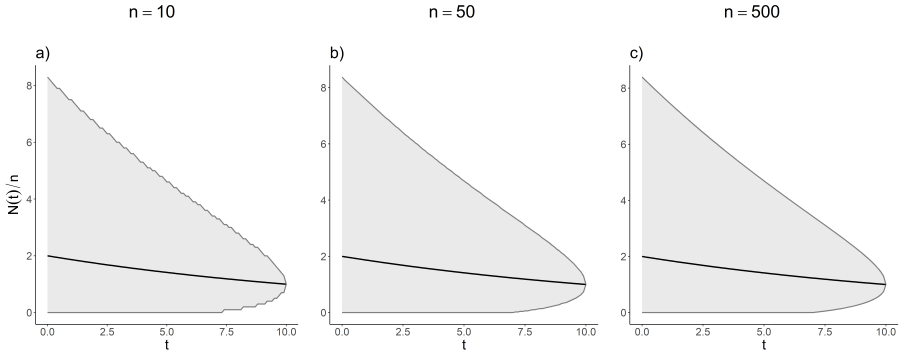

Figure 3: Distribution of the relative population size $N(t)/n$ for the unconditional BD, for different values of n, with fixed $\lambda_1(t) = 0.1$ and $r(t) \simeq 0.069$. The black line is the expected value and the ribbon ranges from the .025 to the .975 quantile. The distribution barely changes as $n \to \infty$, illustrating that it reaches a limit.

We notice that under the current parametrization, the BD coalescent rate does not vanish as $n \to \infty$. Indeed, we have

$$\frac{2\lambda(t)}{N(t)} = \frac{2\lambda_1(t)}{N(t)/n}$$

and $N(t)/n$ is stochastically bounded, since (using Bishop (35)'s Theorem 14.4-1 and the $O_p(\cdot)$ notation defined therein):

$$\begin{aligned}\frac{N(t)}{n} &= \mathbb{E}[N(t)/n|N(T) = n] + O_p\big(\mathrm{Var}[N(t)/n|N(T) = n]^{\frac{1}{2}}\big) \\ &= \exp\int_t^T r(s)ds + O_p\big(\mathrm{Var}[N(t)/n|N(T) = n]^{\frac{1}{2}}\big)\end{aligned}$$

with $\mathrm{Var}[N(t)/n|N(T) = n]$ bounded since it reaches a limit.

Also, the BD coalescent rate remains random as $n \to \infty$, since $\mathrm{Var}[N(t)/n|N(T) = n]$ does not vanish. This suggests that the KC, which has a deterministic coalescent rate, will not be recovered as the limit of Z, as shown in the next section.

Limit of the sample process Z

Lemma D.3 shows that Z converges in distribution to a random process $\dot{Z}$ with density

$$\begin{aligned}f_{\dot{Z}}(z) := \lim_{n\to\infty} f_Z(z) = &\frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!z(T)!}\left(\frac{\rho_1 D(0,T)}{\big(1+\rho_1 B_1(0,T)\big)^2}\right)^{z(T)} \\ &\times \exp\left(\frac{-\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right)\prod_{t\in\alpha(z)}\frac{\lambda_1(t)\rho_1 D(0,t)}{(1+\rho_1 B_1(0,t))^2} \qquad (2)\end{aligned}$$

with $z = \big(z(t)\big)_t$ a valid trajectory of the sample process (i.e. a decreasing integer-valued function with unit decrements), $\alpha(z)$ the set of coalescent times induced by z and with

$$D(t_1, t_2) := \exp \int_{t_1}^{t_2} r(s)ds \qquad B_1(t_1, t_2) := \int_{t_1}^{t_2} \lambda_1(s) D(s, t_2) ds$$

Furthermore, Lemma D.4 in the Appendix shows that this implies that $Z|Z(0) = z_0$ converges to a random process $\dot{Z}|\dot{Z}(0) = z_0$ with density function

$$f_{\dot{Z}}(z|\dot{Z}(0) = z_0) \propto f_{\dot{Z}}(z) \tag{3}$$

Figure 4 shows the distribution of Z given $Z(0) = z_0$ for different values of n.

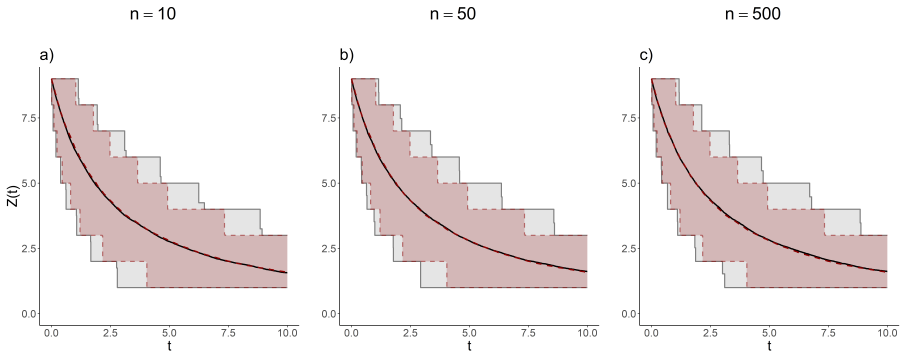

Figure 4: Distribution of the sample process Z given $Z(0) = z_0 = 9$ for the unconditional BD, for different values of n, with fixed $\lambda_1(t) = 0.1$ and $r(t) \simeq 0.069$. The black line is the expected value and the grey ribbon ranges from the .025 to the .975 quantile. The distribution barely changes as $n \to \infty$, illustrating the limit. For comparison, the red dashed line and red ribbon show the distribution of Z under a KC with $\theta = \mathbb{E}[N(t)|N(T) = n]/(2\lambda(t))$. We see that, as n increases, the two distributions do not get more similar, illustrating that the limiting BD distribution of Z is not a KC.

Importantly, the limiting density of Z conditioned on $Z(0) = z_0$ (Eq. 3), is different from the KC density, which is of the form

$$f_Z^{KC}(z) = \exp\left(\int_{t_{z(0)-z(T)}}^{T} \frac{\binom{z(T)}{2}}{\theta(s)ds}\right) \prod_{i=1}^{z(0)-z(T)} \frac{\binom{z(0)-i+1}{2}}{\theta(t_i)} \exp\left(\int_{t_{i-1}}^{t_i} \frac{\binom{z(0)-i+1}{2}}{\theta(s)ds}\right) \tag{4}$$

with $t_0 := 0$ and with $\forall i \in \{1, ..., z(0) - z(T)\}$, $t_i \in \alpha(z)$, $t_i > t_{i-1}$. Indeed, it can be verified that there does not exist a parametrization of $\theta(t)$ in terms of ρ_1, $\lambda_1(t)$ and $r(t)$ such that Eq. (4) reduces into Eq. (3).

Summary

The parametrization considered here is interesting because the BD coalescent rate does not vanish as $n \to \infty$, contrarily to the limit with fixed $\lambda(t)$ and $\mu(t)$ considered in the previous section. However, the BD coalescent rate remains random in the limit, which logically implies that the limiting sample process is not a KC.

The conditional BD with fixed $\lambda_1(t)$ and $r(t)$ as $n \to \infty$

We now consider the same limit as in the previous section, with fixed $\lambda_1(t)$ and $r(t)$ (ρ will be irrelevant here), but we condition N on belonging to some set Ω_n of population trajectories such that $N(t)/n$ converges in probability to some deterministic function $u(t)$ as $n \to \infty$. In this case, the BD coalescent rate is given by

$$\frac{2\lambda(t)}{N(t)} = \frac{2\lambda_1(t)n}{N(t)} = \frac{2\lambda_1(t)}{N(t)/n} \xrightarrow[n\to\infty]{p} \frac{2\lambda_1(t)}{u(t)} > 0,$$

becomes deterministic and does not vanish as $n \to \infty$. As we show below, this means that the limiting sample process of the BD is a non-trivial KC.

For Ω_n we chose the set of population trajectories defined as follows. Given $u(t)$ any function of time such that, on $[0, T]$: (i) $u(t)$ is continuous (ii) $u'(t)$ exists and is bounded, (iii) $u(t) > 0$ and (iv) $u(T) = 1$, we define

$$y(t) = y_n(t) := \lceil u(t)n \rceil$$

We further introduce $g_1 : \frac{T}{g_1} \in \mathbb{N}_{>0}$, define $g := \frac{g_1}{n}$, and define the set Ω_n such that $N \in \Omega_n$ if and only if

$$\forall t \in \{0, g, 2g, ..., T\}\ N(t) = y(t) \tag{5}$$

We call the *conditional BD* the BD process conditioned on $N \in \Omega_n$. This condition means that $N(t)$ is constrained to be equal to $y(t)$ every g units of time, with $g \propto \frac{1}{n}$ (see Figure 5). Notice that all the trajectories in Ω_n that are not valid BD trajectories (BD trajectories are step functions with unit increments/decrements) have probability 0.

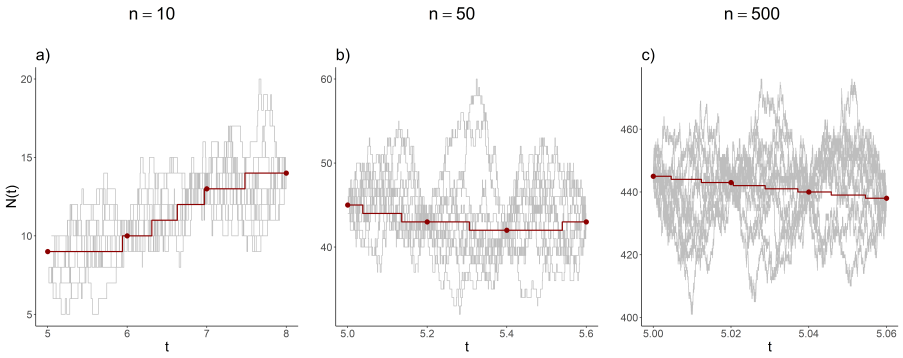

Figure 5: Valid population trajectories belonging to the set Ω_n. The red line is the function $y(t)$, and the grey lines are ten random trajectories belonging to Ω_n, i.e. conditioned on $N(t) = y(t)$ every $g = g_1/n$ units of time. The figure shows the interval $[5, 5 + 3g]$. Red dots delineate the three intervals of length g. In this example $\lambda_1(t) = 0.1$, $r(t) \simeq 0.069$, $g_1 = 10$, $T = 10$ and $u(t)$ is illustrated in Figure 6.

Limit of the relative population size $N(t)/n$

We show in Theorem G.3 in the Appendix that the condition $N \in \Omega_n$ implies that the relative population size $N(t)/n$ converges in probability to $u(t)$ as $n \to \infty$, uniformly on $[0, T]$ (see Figure 6), that is, for any $\epsilon > 0$

$$\inf_{t\in[0,T]} \mathbb{P}\left(\left| \frac{N(t)}{n} - u(t) \right| < \epsilon \;\middle|\; N \in \Omega_n \right) \xrightarrow[n\to\infty]{} 1$$

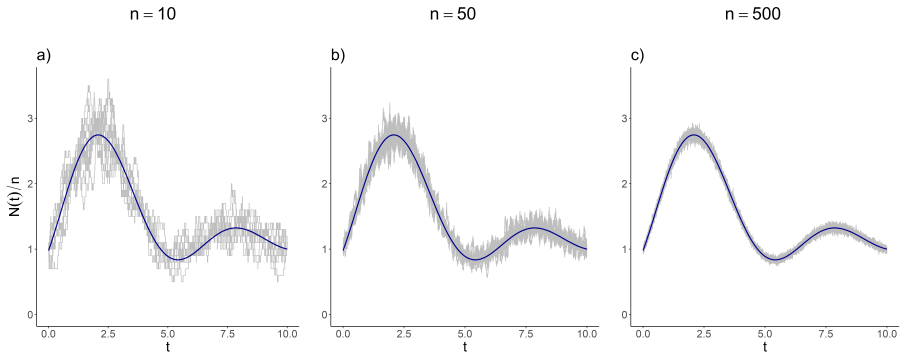

Figure 6: Same example population trajectories belonging to the set Ω_n as in Figure 5, but scaled by n and shown on $[0, T]$, with $T = 10$. The blue line is the function $u(t)$, and the grey lines are ten random trajectories. We see how the trajectories in Ω_n, scaled by n, narrow around $u(t)$ as $n \to \infty$, illustrating the convergence in probability of $N(t)/n \mid N \in \Omega_n$ to $u(t)$.

Limit of the sample process Z

The main result of this study is given in Theorem F.1 (see also Section B of the Appendix for an overview of the proof) and shows that the sample process Z, conditional on $N \in \Omega_n$ and on $Z(0) = z_0$, converges (in terms of finite-dimensional distributions) to the KC with a sample size z_0 and with parameter

$$\theta(t) = \frac{u(t)}{2\lambda_1(t)} \tag{6}$$

See Figure 7 for an example.

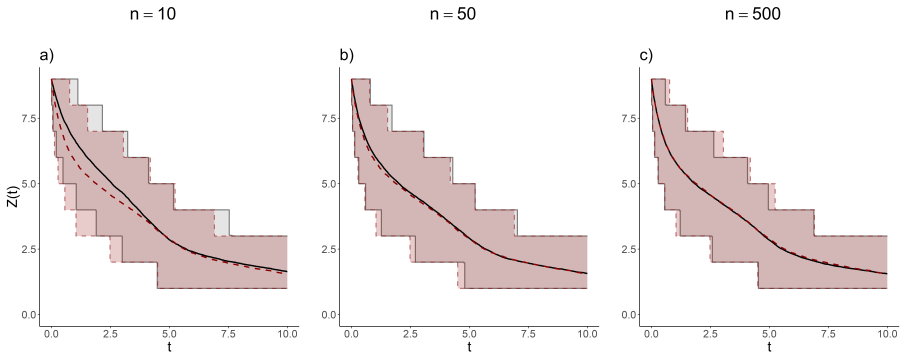

Figure 7: Distribution of the sample process Z given $Z(0) = z_0 = 9$ for the conditional BD (black lines and grey ribbons), for different values of n, with fixed $\lambda_1(t) = 0.1$, $r(t) \simeq 0.069$, $g_1 = 10$ and $u(t)$ as in Figure 6. The thick line is the expected value and the ribbon ranges from the .025 to the .975 quantile. For comparison, the red dashed lines and red ribbons show the limiting distribution of Z, i.e. that of a KC with $\theta(t) = u(t)/(2\lambda_1(t))$. We see that, as n increases, the two distributions get more similar, contrarily to Figure 4, and coalescences do not become rarer as n increases, contrarily to Figure 2.

Importantly, $\theta(t)$ is finite, and since the coalescent rate of the limiting process depends on $u(t)$ and $\lambda_1(t)$, non-trivial limiting processes, which may or may not coalesce by time T, can be studied.

We acknowledge that the conditional BD process considered here, if observed every g units of time (i.e. for $t \in \{0, g, 2g, \dots, T\}$), is in effect a (time-inhomogeneous) Cannings process. It is therefore expected that the limiting process be the KC, given the previous literature which shows that Cannings processes converge to the KC in the limit of large population size (e.g. 6, 19). However, the demonstration that the parameter of the limiting KC is given by the formula in Eq. (6) is far from trivial (see the proofs in the Appendix), and to our knowledge original.

We further acknowledge that our proof is specific to our choice of definition of Ω_n, although we conjecture that our result would hold for Ω_n defined as any set of population trajectories such that $\left(\frac{N(t)}{n} | N \in \Omega_n\right) \xrightarrow[n\to\infty]{p} u(t)$.

Summary

We observe that two ingredients are necessary for the BD to converge to a KC. First, one must condition the population process in such a way that the relative population size $N(t)/n$ converges in probability to a deterministic function $u(t)$ as $n \to \infty$. This ensures that, in the limit, the BD coalescent rate becomes deterministic, like for the KC. Second, the birth rate must be proportional to n (actually our theorems could

probably be generalized to the case where $\lambda_n(t)/n$ reaches a positive limit). This ensures that the BD coalescent rate does not vanish as $n \to \infty$, as the numerator and the denominator of the BD coalescent rate remain of the same order.

In the next section we show how considering a KC with $\theta(t) = u(t)/\big(2\lambda_1(t)\big)$ as the large-population limit of a BD conditioned in such a way that $N(t)/n \xrightarrow[n\to\infty]{p} u(t)$ provides new avenues for phylogenetic inference with the KC.

Application to the phylogenetic study of epidemics

The present section aims to serve as a proof of concept for the practical usefulness of our main result (Theorem F.1 and Eq. 6) for the statistical inference of evolutionary parameters from an observation $z = \big(z(t)\big)_t$ of the sample process. We focus on the epidemiological context and consider using z as data to estimate a pathogen prevalence (i.e. its population size $N(t)$) and effective reproduction number $R_e(t)$.

After reviewing briefly the methods most commonly used in phylogenetic epidemiology, we present a new method ensuing from the main result of this paper. We argue that this new method is robust with regard to the sampling procedure and the data-generating model. We illustrate this robustness by showing examples where the new method is accurate while some of the traditional methods fail.

Classical methods in phylogenetic epidemiology

As explained in the introduction, both the BD and the KC are used for the phylogenetic study of epidemics. Because pathogen samples are collected at various points in time, the BD has been extended to model the sampling procedure. This has given rise to the "birth-death sampling" (BDS) model (29, 30), whereby each individual pathogen is sampled at rate $\psi(t)$. There are two main appeals of the BDS: (i) the parameters of the BDS are readily converted into meaningful epidemiological parameters such as the prevalence or the effective reproduction number (29, 30), and (ii) if the chosen parametrization for the function $\psi(t)$ fits well the observed sampling times, the BDS leverages the information of the sampling times to gain precision (34). On the other hand, the main drawback of the BDS is its lack of robustness to the misspecification of $\psi(t)$ (see 34 and our example below). In contrast to the BD, the KC conditions on sampling times, and should therefore be robust with regard to the sampling procedure. Uses of the KC in epidemiology can be classified into two main approaches. First, the "skyline" approach sets $\theta(t) = f(t)$, with $f(t)$ a piecewise function (e.g. 36–38). Such KC models can fit the data well, provided $f(t)$ has enough free parameters. However, the estimate of θ is hardly interpretable, since f is a phenomenological representation of θ. Importantly, θ is not, in general, proportional to the census population size N (25). In the case of the BD, for instance, $\theta(t) \simeq N(t)/(2\lambda(t))$, so that proportionality holds only under the assumption that the birth rate is constant through time, which is violated by most epidemiological models (e.g. 39). Thus, with this approach, little can

be inferred about demographic or epidemiological parameters (although θ does reflect the genetic diversity and the potential for adaptation, 40). The second category of KC models falls into the "mechanistic" approach, which sets $\theta(t) \propto f(t)$, with $f(t)$ the population size expected under some deterministic epidemiological model (e.g. 28, 36, 39, 41). For instance, Pybus et al. (39) studied the Hepatitis C virus using the KC, setting $f(t)$ to the number of infections predicted by a deterministic SIS model. Contrarily to the skyline approach, and similarly to the BDS approach, the mechanistic approach has the advantage that epidemiological parameters are readily obtained. However, as we show below, it is not robust to the misspecification of the data-generating model (e.g. the estimates obtained assuming an SIR model are inaccurate if the true model was more like an SIS model). Furthermore, the same problem as for the skyline approach remains: θ is not, in general, proportional to N, so that the prevalence may not be inferred. Finally, the KC has been criticized for its assuming that the population size is deterministic (e.g. 24, 42), spawning extensions of the KC to accommodate a random population size (e.g. 43, 44). Given our results, we would instead argue that the KC does not assume that population size is deterministic: it conditions on a given population trajectory (i.e. on one realization of an otherwise random population process) in order to estimate it. In line with this, the KC is known to be the limit of Cannings models, which are Galton-Watson processes (with a random population size) conditioned on a population trajectory (45).

Our new method

We propose a new method for epidemiological phylogenetics, ensuing from our main result (Theorem F.1 and Eq. 6) in that it consists of a KC seen as the large-population limit of a conditional BD. Its purpose is to gain insight into the historical progress of an epidemic from an observation of the sample process. We argue that our method combines the three following advantages: (i) it is robust with regard to the data-generating process (i.e. whether the epidemic unfolded as predicted by an SIS, SIR, SIRV, etc... model), (ii) important epidemiological parameters can be estimated, in particular the prevalence, effective reproduction number, transmission rate and duration of infections, and (iii) it is robust with regard to the sampling procedure (i.e. whether samples are collected homogeneously through time, punctually, erratically...).

Before presenting our new method, we here define a class of models, called BD-type models, because we will argue that our method performs well for all the models of this class. We define a BD-type model as a model whose population process is such that, in any sufficiently small interval of time, every individual of the population can give birth to at most one new individual or can be removed from the population. Moreover, individuals are exchangeable at all times. This body of models notably encompasses the BD model with deterministic per-capita rates which we studied in this paper, and also the stochastic versions of many compartmental models used in epidemiology, which are BD models with random per-capita rates (e.g. in the SIR model, the per-capita birth rate of infectious individuals is proportional to the random number of susceptible individuals).

Finally, we will now define $\tilde{N}(t)$, $\tilde{\lambda}(t)$ and $\tilde{\mu}(t)$, which are phenomenological functions

representing respectively the realized population trajectory, and what we call the realized birth and death rates, and which our method will aim at inferring. For any BD-type model, the population size is given by $N(t) = N(0) - B(t) + D(t)$, where $B(t)$ and $D(t)$ are the random cumulative number of births and deaths, respectively, between the present and time t. To any population at study correspond fixed realizations $N = \nu$, $B = \beta$ and $D = \delta$ of these random processes, which we would like to estimate. To do so, we represent β and δ with differentiable functions $\tilde{B}$ and $\tilde{D}$ and logically we represent ν with a function $\tilde{N} = \tilde{N}(0) - \tilde{B} + \tilde{D}$. We define the realized birth and death rates as $\tilde{\lambda} := \tilde{B}'/\tilde{N}$ and $\tilde{\mu} := \tilde{D}'/\tilde{N}$. Understandably $\tilde{\lambda}(t)$ represents the per-capita birth rate in action at time t, given that the number of births that occurred by time t is $\beta(t) \approx \tilde{B}(t) = \int_0^t \tilde{\lambda}(s)\tilde{N}(s)ds$, and similarly for deaths and $\tilde{\mu}(t)$. We call these rates "realized", because they describe the accumulation of births and deaths in a realization of the population process.

Our method consists of a KC parametrized with

$$\theta(t) = \frac{\tilde{N}(t)}{2\tilde{\lambda}(t)} = \frac{\tilde{N}(0)\exp\int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big)ds}{2\tilde{\lambda}(t)},$$

that is parametrized in terms of important epidemiological parameters: $\tilde{N}$ (the realized population trajectory, or prevalence), $\tilde{\lambda}$ (the realized transmission rate), $\tilde{\mu}$ (the realized inverse of the average duration of infections), and also $\tilde{R}_e = \tilde{\lambda}/\tilde{\mu}$ (the realized effective reproduction number) can be deduced. We will now explain how this parametrization of θ arises and why we think it should be robust.

Robustness with regard to the data-generating model. We have shown in this paper that, for the BD with deterministic rates, after conditioning the population process in such a way that $N(t)/n \xrightarrow{p} u(t)$ as $n \to \infty$, the sample process Z converges to a KC with $\theta(t) = \frac{u(t)}{2\lambda_1(t)}$. Given that $\lambda_1(t) = \lambda(t)/n$ and that $N(t)/n \xrightarrow{p} u(t)$, we have that $\frac{N(t)}{2\lambda(t)} = \frac{N(t)/n}{2\lambda(t)/n} \xrightarrow{p} \frac{u(t)}{2\lambda_1(t)}$. This can be phrased informally as follows. For some function $u(t)$, for sufficiently large n, we should have that a KC with $\theta(t) = \frac{N(t)}{2\lambda(t)}$ is a good approximation to the BD distribution of Z, conditioned on $N(t)/n \approx u(t)$. Hence, our proposed method is a KC parametrized with $\theta(t) = \frac{N(t)}{2\lambda(t)}$. While our result is specific to the BD with deterministic rates, we argue heuristically in Appendix J.1 that our method should perform well for all BD-type models. Very briefly, for all BD-type models, given realizations ν and β of N and B, the probability that a pair of individuals do not coalesce in $[0, t]$ is given by (see notably 46)

$$P(t) = \prod_{i=1}^{\beta(t)} \left(1 - \frac{1}{\binom{\nu(t_i)}{2}}\right)$$

where t_i is the time of the i-th birth from present to past. In Appendix J.1, we suggest that, provided that ν is sufficiently large in $[0, t]$, the following approximation should hold:

$$P(t) \approx \exp\left(-\int_0^t \frac{2\tilde{\lambda}(s)}{\tilde{N}(s)}ds\right)$$

The latter expression is equal to the probability of a pair of individuals not coalescing in $[0, t]$ under a KC with $\theta = \tilde{N}/(2\tilde{\lambda})$ and, importantly, it should hold for all the

BD-type models, given that our heuristic did not use any more assumptions than the absence of multiple births, the exchangeability of individuals, and large population size (see Appendix J.1).

Important epidemiological parameters can be inferred. Without further considerations, fitting a KC with $\theta = \tilde{N}/(2\tilde{\lambda})$, for some parametrizations of the curves $\tilde{N}$ and $\tilde{\lambda}$, would not allow to estimate interesting parameters. Obviously, the scales of the two curves cannot be separately identified, and $\tilde{\mu}$ and $\tilde{R}_e$ cannot be derived. To remedy this, we notice that, given the above definitions of $\tilde{\lambda}$ and $\tilde{\mu}$, then $\tilde{N} = \tilde{N}(0) - \tilde{B} + \tilde{D}$ implies that (see Appendix J.2)

$$\tilde{N}(t) = \tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds$$

This makes $\tilde{\mu}$ appear as a parameter, and consequently $\tilde{R}_e = \tilde{\lambda}/\tilde{\mu}$ too. We thus have θ parametrized in terms of all the parameters that we wish to estimate. Still, these parameters cannot be identified, since an infinite number of combinations of $\tilde{\lambda}$, $\tilde{\mu}$ and $\tilde{N}(0)$ can yield the same θ, and hence the same likelihood. Actually, letting a triple $\big(\tilde{\lambda}, \tilde{\mu}, \tilde{N}(0)\big)$ define a "model", and calling an equivalence class of models a set of models yielding the same θ, we show in Appendix J.3 that the equivalence classes for our method are the same as the equivalence classes defined in Louca and Pennell (47) for the Bernoulli-sampled BD model. Therefore, our method cannot be used without the use of auxiliary data to inform at least one of the curves $\tilde{\lambda}$, $\tilde{\mu}$ or $\tilde{N}$. This must be kept in mind as an important subject of study in the future, but identifiability shall not be our primary concern here. Our purpose is only to show that, provided that an external estimate of $\tilde{\lambda}$, $\tilde{\mu}$ or $\tilde{N}$ can be obtained, the other parameters can be inferred. In the remaining of this section, we will thus assume that $\tilde{\mu}$ is fixed to a value inferred by other means than phylogenetics, and the phylogenetic analysis then comes to bring additional information about the epidemic. Appendix I gives more detail about this, and explains the chosen parametrization for the remaining free curve $\tilde{\lambda}$.

Robustness with regard to the sampling procedure. Because the KC conditions on sampling times, it does not need to assume a particular sampling procedure. This contrasts with the BDS approach, which is sensitive to the misspecification of the sampling procedure.

Comparison of our new method with classical methods

In sum, in the previous section, we argued that our method of using a KC parametrized with

$$\theta(t) = \frac{\tilde{N}(t)}{2\tilde{\lambda}(t)} = \frac{\tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds}{2\tilde{\lambda}(t)}$$

provides estimators of important epidemiological quantities such as the prevalence $\tilde{N}$, or the effective reproduction number $\tilde{R}_e$, in a way that is robust with regard to the data-generating process and the sampling procedure. To properly test our claims, a

dedicated simulation study will be warranted. Here, we merely want to illustrate our words with a few worked examples. We disclose that we have actively searched for examples of failures of the methods other than ours, and we do not claim that these methods fail in general. We assert that in the process of looking for examples, we have not come across a case where our method was inaccurate.

We present the results of three pairs of analyses. For each pair, the same simulated datasets are analysed with (i) our method and (ii) one of the classical methods (i.e. the skyline KC, the mechanistic KC, and the BDS approaches described above). For each pair of analyses, we simulated a single population trajectory, and then simulated 100 realizations of the sample process, conditional on the simulated population trajectory. Each dataset was analyzed with the two methods being compared (i.e. parameter estimates were obtained by maximum likelihood). All the details of the simulations and inferences are given in Appendices H and I.

For the first comparison, we simulated the population trajectory using an SIS model whose parameters have been chosen so that the predicted curve for θ is not proportional to N, because of important temporal variations of the birth rate. Indeed, with an SIS model, the per-capita birth rate is proportional to the number of susceptibles, which decreases during the course of the epidemic. As a result, the predicted curve for θ is shifted in time with respect to N. In this case, the estimate of θ obtained with the skyline KC method should not be an accurate estimate of the relative population trajectory. Figure 8 shows that our method is accurate for estimating both N and R_e (in the sense that the range of estimated curves overall covers the true curves). In comparison, the skyline KC method provides an accurate estimate of θ, which is however not an accurate estimate of the relative population trajectory. In general, θ cannot be conflated with the relative population size. For instance, lockdown measures immediately decrease the transmission rate (i.e. the realized birth rate), but the resulting decrease in prevalence is delayed. Probably the skyline KC method would reveal the immediate decrease of θ, but would be unable to measure the delay of the effect of the lockdown on prevalence. Finally, the skyline KC method has the drawback that it does not provide an estimate of R_e. The precision is lower towards the present with both methods, due to less frequent coalescences in the recent past in the datasets analyzed.

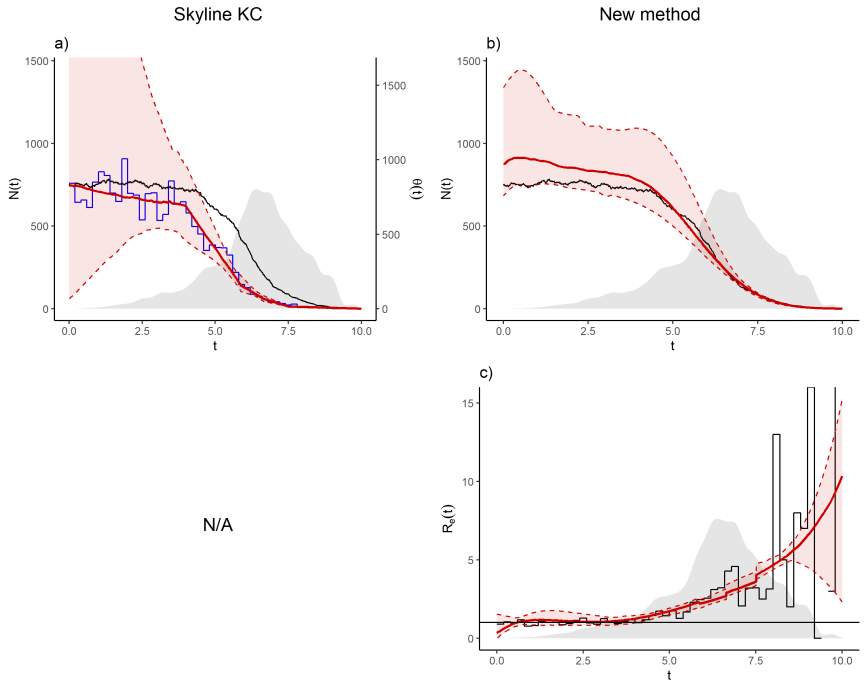

Figure 8: Comparison of the skyline KC approach with our new method. The population trajectory was simulated under an SIS model. The sampling procedure assumed a fixed per-capita sampling rate. a) Estimate of θ with the skyline KC method. Black curve (left y-axis), true population trajectory. Blue curve (right y-axis), true (realized) value of θ. Red solid curve, median of estimates for θ across the 100 inferences. Red ribbon, .025 to the .975 percentiles of the θ estimates. If $N \propto \theta$ is assumed, the estimate of θ may be taken for an estimate of the relative population trajectory. In this example, it is inaccurate because $N \propto \theta$ is untrue. b) Estimate of the population trajectory with our method. Black curve, true population trajectory. Red curve, median of the estimates for N. Red ribbon, .025 to the .975 percentiles of these estimates. c) Estimate of the effective reproduction number with our method. Black curve, true (realized) value of R_e. Red curve, median of the estimates for R_e. Red ribbon, .025 to the .975 percentiles. R_e cannot be estimated with the skyline KC approach. The gray shaded areas in all panels are the distribution of the coalescence times. The death rate curve was fixed to a pre-estimated value for inference with our method.

For the second comparison, the population trajectory was again simulated with an SIS model. Population trajectories typical of the SIS model reach a plateau when new infections and recoveries balance out. In contrast, typical curves of the SIR model reach a peak and then decrease. Importantly, there is no parametrization of an SIR

model that predicts population trajectories with a plateau. Therefore, we expect that an SIR-based mechanistic KC approach, which assumes that θ is proportional to a deterministic SIR population trajectory, performs inaccurately. Figure 8 shows that the estimates obtained with our method are again accurate. The estimates of N follow closely the true curve, which reaches a plateau towards the present (Fig. 8b). The plateau further shows in the estimates of R_e, which decrease to about 1 near the present (Fig. 8d). For the SIR mechanistic approach, we fixed the proportionality constant between N and θ to 1, as we found that this constant, the time of origin of the process, and the size of the whole population (i.e. the combined number of susceptible, infectious and recovered individuals) were not separately identifiable. Therefore, we assessed whether the estimates for θ reflect the relative population trajectory. We see that the estimated θ curves show a clear peak around time 3 (Figure 8a). This further shows in the estimates for R_e, which tell with confidence that R_e goes below 1 around time 3 (Figure 8d). Based on this result, an investigator would conclude that the epidemic is presently recessing, which is incorrect. Of course, in practice, an investigator would carry out model selection if they do not know that their study system approximately obeys a particular model. But delineating a good set of candidate models may not be an easy task, in particular if exogenous factors induce temporal variations of the epidemiological parameters. In contrast, since our approach is valid for a large set of models, it does not require model selection.

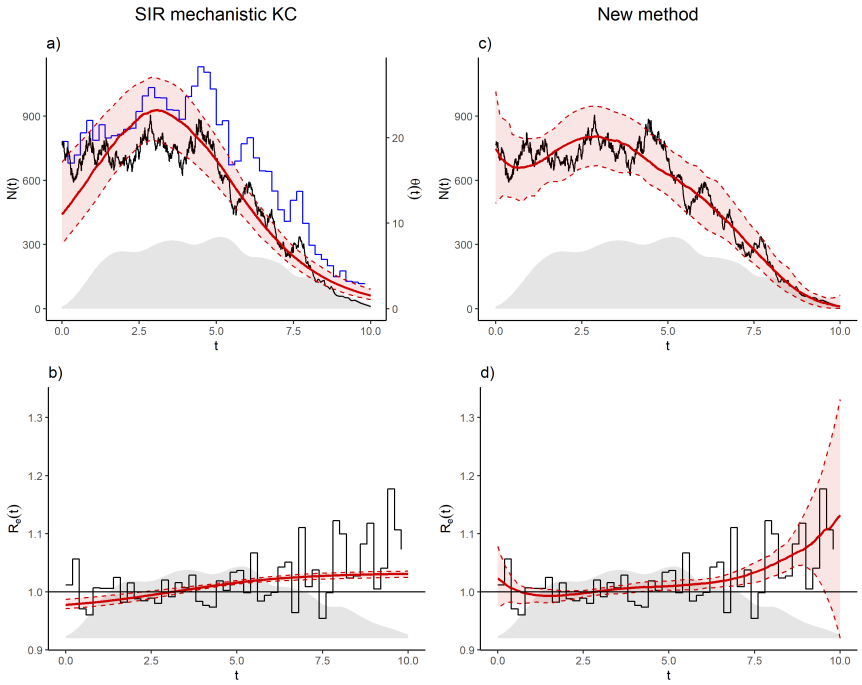

Figure 9: Comparison of the SIR mechanistic KC approach with our new method. The population trajectory was simulated under an SIS model. The sampling procedure assumed a fixed per-capita sampling rate. This figure reads as Figure 8. The death rate curve was fixed to the same pre-estimated value for both methods (see Appendix I).

For the third comparison, we used the same simulated population trajectory as for the second comparison. However, instead of drawing sampling times based on a constant sampling rate, we drew a fixed number of samples at three fixed times: 71 samples at time 0.1, 66 samples at time 0.5, and 26 samples at time 1. This sampling procedure mimics punctual sampling campaigns carried out for other purposes than phylogenetics. Such a distribution of sampling times is at odds with the expectations of a constant-rate-based sampling procedure. As a consequence, we expect the BDS method to fail if it assumes a constant sampling rate. Figure 10 shows that our method still performs accurately, for estimating both N and R_e. The precision decreases rapidly looking into the past, as the sampling procedure used in this example yielded coalescent times concentrated in the recent past. In contrast, the BDS parametrized with a constant sampling rate ($\psi(t) = \psi$) performs very inaccurately, inferring an exponential growth that accelerates dramatically in the recent past, a scenario which is totally different from the truth. We put this failure down to the erroneous specification of the sampling procedure in the BDS, as the same analyses with the data of Figure 9 (wherein sampling times were drawn based on a constant rate) work very well (see Appendix K). We acknowledge that a thorough investigator

would not analyze data sampled from three points in time with a model assuming a constant sampling rate, but this example highlights the fact that the misspecification of the sampling procedure when using the BDS can have dramatic effects on the results. In contrast, our method, being based on the KC, conditions on the sampling times and does not assume a sampling procedure.

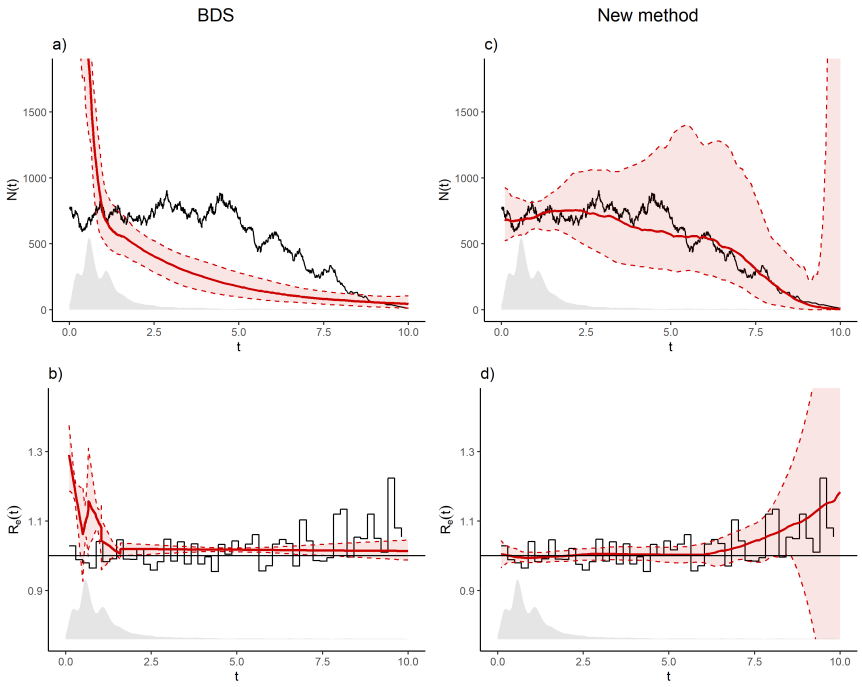

Figure 10: Comparison of the BDS approach with our new method. The population trajectory was simulated under an SIS model. The sampling procedure assumed that fixed numbers of samples were taken at three fixed points in time. This figure reads as Figure 9, except that a) here shows the estimates of the population trajectory rather than estimates of θ, since θ is not a parameter of the BDS model. Appendix I explains how an estimate for the population trajectory is obtained from estimates of λ, μ and ψ under the BDS. The death rate curve was fixed to the same pre-estimated value for both methods (see Appendix I).

Conclusion

This paper outlines a mathematical bridge between the BD and the KC by showing that the KC is the large-population limit of a BD conditioned on a given population trajectory. Earlier studies have searched to link the BD and the KC, and have found the same formula as us for the BD coalescent rate (24, 33, 34, 48). However, the present

study is the first, to our knowledge, to formally demonstrate this formula and to show that, besides the assumption of a large population size, it is required that the BD be conditioned on a population trajectory to obtain a KC. This is a reminder that the KC does not assume that the population size is deterministic: it conditions on a given population trajectory. We further demonstrate how the mathematical relationship between the two models may be leveraged to improve phylogenetic inference with the KC, in particular in the context of pathogen phylodynamics: we present a new method which may be able to estimate accurately important epidemiological parameters. Like its BD pendant (the BDS model), this new method requires external information to identify the parameters, and is intended to be used in conjunction with auxiliary epidemiological data. In principle, this new method should be robust with regard to the data-generating model and the sampling procedure. Simulation studies dedicated to testing these claims will be warranted.

Acknowledgments

This work was supported by French Agence Nationale de la Recherche through the CoCoAlSeq project (ANR-19-CE45-0012). This is the contribution ISEM 2024-XXX of the Institut des Sciences de l'Evolution de Montpellier. We thank François Bienvenu for insightful discussions.

References

[1] B Rannala. Gene genealogy in a population of variable size. *Heredity*, 78(4): 417–423, 1997.

[2] A J Drummond, A Rambaut, B Shapiro, and O G Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005.

[3] S Möller, L du Plessis, and T Stadler. Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):4200–4205, 2018.

[4] T Stadler. Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26:1203–1219, 2013.

[5] D G Kendall. On the generalized " Birth-and-Death " process. *The Annals of Mathematical Statistics*, 19(1):1–15, 1948.

[6] J F C Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[7] S Nee. Birth-death models in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 37:1–17, 2006.

[8] F L Condamine, J Rolland, and H Morlon. Macroevolutionary perspectives to environmental change. *Ecology Letters*, 16:72–85, 2013.

[9] R A Pyron and F T Burbrink. Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. *Trends in Ecology and Evolution*, 28(12):729–736, 2013.

[10] D L Rabosky, J Chang, P O Title, P F Cowman, L Sallan, M Friedman, K Kaschner, C Garilao, T J Near, M Coll, and M E Alfaro. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559:392–395, 2018.

[11] J Rolland, F L Condamine, F Jiguet, and H Morlon. Faster Speciation and Reduced Extinction in the Tropics Contribute to the Mammalian Latitudinal Diversity Gradient. *PLOS Biology*, 12(1):e1001775, 2014.

[12] H Morlon, T L Parsons, and J B Plotkin. Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences*, 108(39): 16327–16332, 2011.

[13] B T Kopperud, A F Magee, and S Höhna. Rapidly changing speciation and extinction rates can be inferred in spite of nonidentifiability. *Proceedings of the National Academy of Sciences of the United States of America*, 120(7):e2208851120, 2023.

[14] J Cornuault, B H Warren, J AM Bertrand, B Milá, C Thébaud, and P Heeb. Timing and number of colonizations but not diversification rates affect diversity patterns in hemosporidian lineages on a remote oceanic archipelago. *American Naturalist*, 182(6):820–833, 2013.

[15] L Sorenson, F Santini, and M E Alfaro. The effect of habitat on modern shark diversification. *Journal of Evolutionary Biology*, 27(8):1536–1548, 2014.

[16] D L Rabosky. Speciation rate and the diversity of fishes in freshwaters and the oceans. *Journal of Biogeography*, 47(6):1207–1217, 2020.

[17] M Nordborg. Coalescent Theory. In DJ Balding, M Bishop, and C Cannings, editors, *Handbook of Statistical Genetics*. John Wiley & Sons Ltd, 2004.

[18] M Stephens. Inference Under the Coalescent. *Handbook of Statistical Genetics: Third Edition*, 2:878–908, 2008.

[19] M Möhle. Robustness Results for the Coalescent. *Journal of Applied Probability*, 35(2):438–447, 1998.

[20] N Marchi, F Schlichta, and L Excoffier. Demographic inference. *Current Biology*, 31:267–281, 2021.

[21] A MacPherson, S Louca, A McLaughlin, J B Joy, and M W Pennell. Unifying Phylogenetic Birth–Death Models in Epidemiology and Macroevolution. *Systematic Biology*, 71(1):172–189, 2021.

[22] J Cornuault and I Sanmartín. A road map for phylogenetic models of species trees. *Molecular Phylogenetics and Evolution*, 173:107483, 2022.

[23] H Morlon, M D Potts, and J B Plotkin. Inferring the dynamics of diversification: A coalescent approach. *PLoS Biology*, 8(9):e1000493, 2010.

[24] T Stadler, T G Vaughan, A Gavryushkin, S Guindon, D Kühnert, G E Leventhal, and A J Drummond. How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics? *Proceedings of the Royal Society B: Biological Sciences*, 282:20150420, 2015.

[25] L A Featherstone, J M Zhang, T G Vaughan, and S Duchene. Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications. *Virus Evolution*, 8(1):1–12, 2022.

[26] O G Pybus, A J Drummond, T Nakano, B H Robertson, and A Rambaut. The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach. *Molecular Biology and Evolution*, 20(3):381–387, 2003.

[27] K Koelle and D A Rasmussen. Rates of coalescence for common epidemiological models at equilibrium. *Journal of The Royal Society Interface*, 9(70):997–1007, 2012.

[28] B Dearlove and D J Wilson. Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368:20120314., 2013.

[29] T Stadler, R Kouyos, V VonWy, S Yerly, J Böni, P Bürgisser, T Klimkait, B Joos, P Rieder, D Xie, H F Günthard, A J Drummond, and S Bonhoeffer. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, 29(1):347–357, 2012.

[30] T Stadler, D Kühnert, S Bonhoeffer, and A J Drummond. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America*, 110(1):228–233, 2013.

[31] D Kühnert, T Stadler, T G Vaughan, and A J Drummond. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *Journal of the Royal Society Interface*, 11:20131106, 2014.

[32] T Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009.

[33] E M Volz. Complex Population Dynamics and the Coalescent Under Neutrality. *Genetics*, 190:187–201, 2012.

[34] E M Volz and S D W Frost. Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of the Royal Society Interface*, 11:20140945, 2014.

[35] Y M M Bishop, S E Fienberg, and P W Holland. Asymptotic methods. In *Discrete Multivariate Analysis*, pages 458–485. MIT Press, 1975.

[36] C V F Carrington, J E Foster, O G Pybus, S N Bennett, and E C Holmes. Invasion and Maintenance of Dengue Virus Type 2 and Type 4 in the Americas. *Journal of Virology*, 79(23):14680–14687, 2005.

[37] A Rambaut, O G Pybus, M I Nelson, C Viboud, J K Taubenberger, and E C Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453:615–619, 2008.

[38] F Aldunate, F Gámbaro, A Fajardo, M Soñora, and J Cristina. Evidence of increasing diversification of Zika virus strains isolated in the American continent. *Journal of Medical Virology*, 89(12):2059–2063, 2017.

[39] O G Pybus, M A Charleston, S Gupta, A Rambaut, E C Holmes, and P H Harvey. The epidemic behavior of the hepatitis C virus. *Science*, 292(5525):2323–2325, 2001.

[40] T I Gossmann, P D Keightley, and A Eyre-Walker. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and Evolution*, 4(5):658–667, 2012.

[41] C Fraser, C A Donnelly, S Cauchemez, W P Hanage, M D Van Kerkhove, T D Hollingsworth, J Griffin, R F Baggaley, H E Jenkins, E J Lyons, T Jombart, W R Hinsley, N C Grassly, F Balloux, A C Ghani, N M Ferguson, A Rambaut, O G Pybus, H Lopez-Gatell, C M Alpuche-Aranda, I B Chapela, E P Zavala, D Ma. Espejo Guevara, F Checchi, E Garcia, S Hugonnet, and C Roth. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934): 1557–1561, 2009.

[42] V Boskova, S Bonhoeffer, and T Stadler. Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models. *PLOS Computational Biology*, 10(11):e1003913, 2014.

[43] D A Rasmussen, O Ratmann, and K Koelle. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLOS Computational Biology*, 7(8):e1002136, 2011.

[44] D A Rasmussen, E M Volz, and K Koelle. Phylodynamic Inference for Structured Epidemiological Models. *PLoS Comput Biol*, 10(4):1003570, 2014.

[45] A Lambert. Population dynamics and random genealogies. *Stochastic Models*, 24:45–163, 2008.

[46] F F Crespo, D Posada, and C Wiuf. Coalescent models derived from birth–death processes. *Theoretical Population Biology*, 142:1–11, 2021.

[47] S Louca and M W Pennell. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505, 2020.

[48] E M Volz, K Koelle, and T Bedford. Viral Phylodynamics. *PLoS Computational Biology*, 9(3):e1002947, 2013.

[49] T Stadler. How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology*, 62(2):321–329, 2013.

[50] P J Davis. Leonhard Euler's Integral: A Historical Profile of the Gamma Function: In Memoriam: Milton Abramowitz. *The American Mathematical Monthly*, 66(10): 849–869, 1959.

[51] R A Doney. One-sided local large deviation and renewal theorems in the case of inÿnite mean. *Probability Theory and Related Fields*, 107:451–465, 1997.

[52] S Louca, A McLaughlin, A MacPherson, J B Joy, and M W Pennell. Fundamental Identifiability Limits in Molecular Epidemiology. *Molecular Biology and Evolution*, 38(9):4010–4024, 2021.

Appendix

After introducing the definitions and notations needed for our proofs in Section A, we present in Section B an overview of the proof of the central result of this paper (Theorem F.1). Sections C through to E derive the proofs needed for Theorem F.1. In Section C we provide general results about the BD. In Section D we provide results about the scaled BD (i.e. parametrized such that $\lambda(t)$ is proportional to n). In Section E, we provide results about the scaled BD, conditioned on $Z(0) = z_0$ and on the population process taking certain values. Section F provides the formal statement of the central result of the paper, along with its proof. Then, Section G gives the proof of the convergence in probability of $N(t)/n$ to $u(t)$, for the conditional scaled BD.

The rest of the appendix gives the details of the phylogenetic inferences carried out with the BDS, the skyline KC approach, the SIR-based mechanistic approach, and our new method. Section H gives details about the data simulation procedure. Section I gives details about the inference methods. Finally, Section J provides details about our method.

A Preliminaries

The following functions will be used frequently:

$$D(t_1, t_2) := \exp \int_{t_1}^{t_2} r(t)dt$$
$$B(t_1, t_2) := \int_{t_1}^{t_2} \lambda(t)D(t, t_2)dt$$
$$B_1(t_1, t_2) := \frac{B(t_1, t_2)}{n} = \int_{t_1}^{t_2} \lambda_1(t)D(t, t_2)dt$$

We use the $O(\cdot)$ and $o(\cdot)$ notations as follows (where y may represent several variables).

- $h(x, y) = O(f(x))$ as $x \to 0$ (resp. $x \to \infty$) means that there exist a constant

$C > 0$ and a number X such that $\forall x \in (0, X]$ (resp. $\forall x \geq X$)

$$\sup_y |h(x,y)| \leq Cf(x)$$

- $h(x,y) = O_y(f(x))$ as $x \to 0$ (resp. $x \to \infty$) means that, for fixed y, there exist a constant $C_y > 0$ and a number X_y such that $\forall x \in (0, X_y]$ (resp. $\forall x \geq X_y$)

$$|h(x,y)| \leq C_y f(x)$$

- $h(x,y) = o(f(x))$ as $x \to 0$ (resp. $x \to \infty$) means that for any choice of $C > 0$, there exists a number X such that $\forall x \in (0, X]$ (resp. $\forall x \geq X$)

$$\sup_y |h(x,y)| \leq Cf(x)$$

- $h(x,y) = o_y(f(x))$ as $x \to 0$ (resp. $x \to \infty$) means that, for fixed y, for any choice of $C > 0$, there exists a number X_y such that $\forall x \in (0, X_y]$ (resp. $\forall x \geq X_y$)

$$|h(x,y)| \leq Cf(x)$$

Importantly, all the variables on which a function $O(f(x))$ or $o(f(x))$ depends (besides x) are listed as indices, so that the absence from the list of indices of any one variable means that the function does not depend on the variable. NB: the index p is reserved for the notation $O_p(\cdot)/o_p(\cdot)$, which means order in probability.

The BD process

Time (t) runs backwards, from present to past, with $t = 0$ the present.

A BD process starts at time T and is parametrized with a birth rate $\lambda(t)$, a growth rate $r(t) \leq \lambda(t)$ and an initial number of individuals n. The death rate is given by $\mu(t) = \lambda(t) - r(t)$.

The BD process unravels in continuous time and is defined by the probability distribution of the number ζ_t of offspring at time $t - dt$ of one individual at time t:

$$\begin{aligned}
&\mathbb{P}(\zeta_t = 0) = \mu(t)dt + o(dt) \\
&\mathbb{P}(\zeta_t = 1) = 1 - \big(\lambda(t) + \mu(t)\big)dt + o(dt) \\
&\mathbb{P}(\zeta_t = 2) = \lambda(t)dt + o(dt) \\
&\mathbb{P}(\zeta_t = r) = o(dt) \qquad\qquad (r > 2)
\end{aligned}$$

as $dt \to 0$.

The offspring distribution is the same for all individuals: the individuals are exchangeable.

Finally, at the present, a random number $Z(0)$ of individuals are sampled. We will specify which sampling procedure is assumed where relevant.

Random outcomes

Counting individuals induces the random **population process** $N := (N(t))_{t\in[0,T]}$, where $N(t)$ is the number of individuals (the population size) at time t. The initial state is $N(T) = n$.

Labelling the initial n individuals, and keeping track of which individuals give birth and die, induces the complete forest, that is a totally ordered set of n ordered time trees. Removing from the complete forest the vertices and edges that do not have any descendants in the sample then induces the **sample forest**, denoted $\Psi := (\Gamma_i)_{i=1,\dots,n}$, where Γ_i is the ordered time tree generated by individual i, whose leaves are individuals in the sample. Note that Γ_i may be the order-zero tree if individual i has no descendants in the sample.

Labelling the $Z(0)$ sampled individuals and keeping track (from the present to time T) of the ancestry relationships among sampled individuals and their ancestors induces the random **ancestral process**, denoted $\mathcal{R} := (\mathcal{R}(t))_{t\in[0,T]}$. $\mathcal{R}(t)$ is the equivalence relation which contains the pair (i, j) if and only if the i-th and the j-th sampled individuals have a common ancestor at time t. The initial state is $\mathcal{R}(0) = \big\{(i, i)|i \in \{1, \dots, Z(0)\}\big\}$. Notice that $\mathcal{R}$ can be represented as a forest of $|\mathcal{R}(T)|$ leaf-labelled time trees.

We denote $Z(t)$ the number of ancestors at time t of the $Z(0)$ sampled individuals. $Z(t)$ is the number of edges in Ψ at time t, as well as the number of equivalence classes in $\mathcal{R}(t)$. We call the curve $Z := (Z(t))_{t\in[0,T]}$ the **sample process**.

We denote $\alpha(Z)$ the set of **coalescent times** induced by Z, that is the times when $Z(t)$ decreases (when two ancestors coalesce into one). We further denote $\bar{\alpha}(\Gamma_i)$ the node ages of Γ_i. Then, given a realisation $z := (z(t))_{t\in[0,T]}$ of the sample process, there exist not one, but a set $S_n(z)$ of forests that induce z. Specifically, a given forest $\psi = (\gamma_i)_{i=1,\dots,n}$ belongs to $S_n(z)$ if and only if $\cup_{i=1,\dots,n}\bar{\alpha}(\gamma_i) = \alpha(z)$ (i.e. $S_n(z)$ is the set of forests with node ages matching the coalescent times of z). Similarly, a number of different realisations of the ancestral process induce the same value z, all of which have in common that $\mathcal{R}(0) = \big\{(i, i)|i \in \{1, \dots, z(0)\}\big\}$ and that transitions between different states occur at times $\alpha(z)$.

In this paper we are interested in the distribution of Z, but the above definitions of the other types of random outcomes (N, Ψ and $\mathcal{R}$) will be needed to derive results about Z. Notably, our main focus is the distribution of Z under some condition about N. We will come across Ψ because the literature about the BD, especially in phylogenetics, often focuses on the distribution of Ψ (oftentimes with $n = 1$, in which case $\Psi = \Gamma_1$ is a tree, e.g. (12, 47, 49)). And we will come across $\mathcal{R}$ because many results about the Kingman coalescent concern the ancestral process (e.g. (6, 19)).

B Overview of the proof of the convergence of the conditional BD to the Kingman coalescent

The central result of this paper is that, for fixed $\lambda_1(t) := \lambda(t)/n$ and $r(t)$, the sample process Z of the BD, conditioned on $Z(0) = z_0$ and on $N \in \Omega_n$ (with Ω_n a set of population trajectories defined in such a way that $N \in \Omega_n$ implies that $N(t)/n \to u(t)$ as $n \to \infty$), converges to a Kingman coalescent with z_0 samples and with parameter

$$\theta(t) = \frac{u(t)}{2\lambda_1(t)}$$

This result is formally enunciated in Theorem F.1 of Section F below, wherein sufficient conditions on $\lambda_1(t)$, $r(t)$ and $u(t)$ for Theorem F.1 to apply are provided.

Here, we only briefly explain the rationale leading to Theorem F.1. The following is not a mathematical proof, and is only intended to provide an intuition of how Theorem F.1 arises.

B.1 The conditional BD process is a Cannings model

The definition of Ω_n is such that the condition $N \in \Omega_n$ fixes the population size every g units of time, with $g \propto 1/n$. Specifically, $\forall t \in \{0, g, 2g, \ldots, T\},\ N(t) = y(t)$. Hence, the conditional BD process reduces into a discrete-time model with fixed population sizes. Furthermore, the individuals in a BD process are exchangeable. These are the characteristics of a (time-heterogeneous) Cannings model. Therefore, the conditional BD process, when observed at the discrete times $t \in \{0, g, 2g, \ldots, T\}$, is a Cannings model. We can then use one of the existing theorems that show the convergence of Cannings models to the KC (hereafter a convergence theorem).

B.2 Convergence to the KC

We then have to verify that the conditional BD process verifies the conditions of application of a convergence theorem. The most important such condition concerns the probability $c_n(t)$ of coalescence of two individuals in $[t-g, t]$, defined $\forall t \in \{g, 2g, \ldots, T\}$ as

$$c_n(t) := \mathbb{P}\big(Z(t) = 1 | Z(t-g) = 2, N(t) = y(t), N(t-g) = y(t-g)\big)$$

The expression of this probability is given by Möhle (19):

$$\begin{aligned} c_n(t) = &\frac{y(t)}{y(t-g)(y(t-g)-1)} \\ &\times \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1) | N(t) = y(t), N(t-g) = y(t-g)\Big] \end{aligned}$$

where ν_{gt1} is the number of descendants at time $t-g$ of an individual existing at time t.

The central condition for convergence theorems to apply is that $c_n(t)/g$ reach a limit as $n \to \infty$. This limit is the coalescent rate of the limiting KC process, that is $1/\theta(t)$. For the conditional BD model, we find that

$$\frac{1}{\theta(t)} = \lim_{n\to\infty} c_n(t)/g = \frac{2\lambda_1(t)}{u(t)}$$

To obtain this limit, we showed that (Lemma E.6)

$$\frac{1}{g} \cdot \frac{y(t)}{y(t-g)(y(t-g)-1)} \xrightarrow[n\to\infty]{} \frac{1}{u(t)g_1}$$

which was easily done.

Then we showed that (Lemma E.7)

$$\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)|N(t)=y(t), N(t-g)=y(t-g)\Big] \xrightarrow[n\to\infty]{} 2\lambda_1(t)g_1$$

which, to us, was much more involved and necessitated a long list of preliminary lemmas (in Sections C-E).

C General results concerning the birth-death

In this section we consider a BD process parametrized with $\lambda(t)$, $r(t)$ and n. A Bernoulli sampling procedure with probability ρ is assumed where relevant. It is not assumed that the parameters are scaled by n, as opposed to the scaled BD process considered in Section D, nor that the population process is conditioned, like in Section E. The goal of this section is to establish some notation and derive useful formulae for the following sections.

C.1 Number of (sampled) descendants of one individual

In this section we give the distribution of the number of sampled individuals $Z(0)$ given $N(T) = 1$, along with the distribution of the population size $N(t_1)$, given a population size $N(t_2) = 1$ at an earlier time. The provided formulae are known (see (47)), and our goal here is to translate them into our notation.

Lemma C.1. *Assuming a Bernoulli sampling procedure with sampling probability ρ,*

the distribution of $Z(0)$ given $N(T) = 1$ is

$$\mathbb{P}\big(Z(0) = 0 | N(T) = 1\big) = 1 - \frac{\rho D(0,T)}{1 + \rho B(0,T)}$$
$$\mathbb{P}\big(Z(0) = k | N(T) = 1\big) = \frac{\rho D(0,T)}{\big(1 + \rho B(0,T)\big)^2} \left(\frac{\rho B(0,T)}{1 + \rho B(0,T)}\right)^{k-1} \qquad k \geq 1$$

Proof. This distribution is given by Louca and Pennel ((47), Eq. 65). After setting $\tau_0 = T$ in Louca's and Pennel's expression, rearranging, and using the definitions of D and B given earlier, we obtain the formula of the present lemma. □

Lemma C.2. *The distribution of $N(t_1)$ given $N(t_2) = 1$, with $t_1, t_2 \in [0,T]$, $t_2 \geq t_1$ is*

$$\mathbb{P}\big(N(t_1) = 0 | N(t_2) = 1\big) = 1 - \frac{D(t_1,t_2)}{1 + B(t_1,t_2)}$$
$$\mathbb{P}\big(N(t_1) = k | N(t_2) = 1\big) = \frac{D(t_1,t_2)}{\big(1 + B(t_1,t_2)\big)^2} \left(\frac{B(t_1,t_2)}{1 + B(t_1,t_2)}\right)^{k-1}, \qquad k \geq 1$$

Proof. If $\rho = 1$, then $Z(0) = N(0)$ with probability 1. Therefore, Lemma C.1 with $\rho = 1$ gives $\mathbb{P}\big(N(0) | N(T) = 1\big)$. Then, setting $T = t_2 - t_1$ in the formula of Lemma C.1, and changing variables from t to $t' = t + t_1$ yields the formula of the present lemma. □

C.2 Number of sampled descendants of several individuals

In this section we give the distribution of the number of sampled descendants $Z(0)$ of $N(T) = n$ individuals (a generalization of Lemma C.1).

Lemma C.3. *Assuming a Bernoulli sampling procedure with sampling probability ρ, we have*

$$\mathbb{P}(Z(0) = z_0 | N(T) = n)$$
$$= \begin{cases} Q_0^n, & z_0 = 0 \\ \sum_{z_T=1}^{z_0} \binom{n}{z_T} Q_0^{n-z_T} \binom{z_0-1}{z_0-z_T} \left(1 - \frac{Q_1}{1-Q_0}\right)^{z_0-z_T} Q_1^{z_T}, & z_0 \geq 1 \end{cases}$$

with $Q_k \equiv Q_k^\rho(T) := \mathbb{P}(Z(0) = k | N(T) = n)$.

Proof. For $z_0 = 0$, trivially,

$$\mathbb{P}(Z(0) = 0 | N(T) = n) = Q_0^n$$

For $z_0 \geq 1$, denoting L_i the number of sampled descendants of individual i existing at time T, we have

$$\begin{aligned}
&\mathbb{P}(Z(0) = z_0, Z(T) = z_T | N(T) = n) \\
&= \mathbb{P}(Z(T) = z_T | N(T) = n)\, \mathbb{P}(Z(0) = z_0 | Z(T) = z_T) \\
&= \binom{n}{z_T} (1 - Q_0)^{z_T} Q_0^{n - z_T}\, \mathbb{P}\Big(\sum_{i=1}^{z_T} L_i = z_0 | \forall i\ L_i \geq 1 \Big) \\
&= \binom{n}{z_T} (1 - Q_0)^{z_T} Q_0^{n - z_T}\, \mathbb{P}\Big(\sum_{i=1}^{z_T} (L_i - 1) = z_0 - z_T | \forall i\ L_i - 1 \geq 0 \Big)
\end{aligned}$$

Using Lemma C.1, and noticing that $\forall i \in \{1, \dots, z_T\}$ $\mathbb{P}(L_i = l_i) = \mathbb{P}(Z(0) = l_i | N(T) = 1)$, we have $\forall i \in \{1, \dots, z_T\}$

$$\begin{aligned}
\mathbb{P}(L_i - 1 = k | L_i - 1 \geq 0) &= \mathbb{P}(L_i = k + 1 | L_i \geq 1) \\
&= \frac{\mathbb{P}(L_i = k + 1)}{1 - \mathbb{P}(L_i = 0)} \\
&= \frac{\mathbb{P}(Z(0) = k + 1 | N(T) = 1)}{1 - \mathbb{P}(Z(0) = 0 | N(T) = 1)} \\
&= \frac{Q_1}{1 - Q_0} \left(1 - \frac{Q_1}{1 - Q_0} \right)^{k + 1 - 1} \\
&= \frac{Q_1}{1 - Q_0} \left(1 - \frac{Q_1}{1 - Q_0} \right)^{k}
\end{aligned}$$

showing that $L_i - 1 | L_i - 1 \geq 0$ for $i = 1, \dots, z_T$ are iid. geometric random variables with parameter $\frac{Q_1}{1 - Q_0}$. Given that the sum of z_T geometric variables with parameter $\frac{Q_1}{1 - Q_0}$ is a negative-binomial random variable with parameters z_T the number of successes and $\frac{Q_1}{1 - Q_0}$ the probability of success, we have that

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^{z_T} (L_i - 1) = z_0 - z_T | \forall i\ L_i - 1 \geq 0 \Big) \\
&= \binom{z_0 - z_T + z_T - 1}{z_0 - z_T} \left(1 - \frac{Q_1}{1 - Q_0} \right)^{z_0 - z_T} \left(\frac{Q_1}{1 - Q_0} \right)^{z_T} \\
&= \binom{z_0 - 1}{z_0 - z_T} \left(1 - \frac{Q_1}{1 - Q_0} \right)^{z_0 - z_T} \left(\frac{Q_1}{1 - Q_0} \right)^{z_T}
\end{aligned}$$

It then follows that, for $z_0 \geq 1$,

$$\begin{aligned}
&\mathbb{P}(Z(0) = z_0 | N(T) = n) \\
&= \sum_{z_T=1}^{z_0} \mathbb{P}(Z(0) = z_0, Z(T) = z_T | N(T) = n) \\
&= \sum_{z_T=1}^{z_0} \binom{n}{z_T} (1-Q_0)^{z_T} Q_0^{n-z_T} \binom{z_0-1}{z_0-z_T} \left(1 - \frac{Q_1}{1-Q_0}\right)^{z_0-z_T} \left(\frac{Q_1}{1-Q_0}\right)^{z_T} \\
&= \sum_{z_T=1}^{z_0} \binom{n}{z_T} Q_0^{n-z_T} \binom{z_0-1}{z_0-z_T} \left(1 - \frac{Q_1}{1-Q_0}\right)^{z_0-z_T} Q_1^{z_T}
\end{aligned}$$

□

C.3 Mean and variance of the population size $N(t)$

Kendall (5) gave the formulae for the mean and variance of $N(t)$ of a BD process in forward time. In this paper we always use backward time, and the purpose of this section is to translate Kendall's formulae from forward to backward time.

Lemma C.4. *For a BD process starting at time T before the present, the formulae for the mean and variance of $N(t)$ in backward time are*

$$\begin{aligned}
\mathbb{E}\big[N(t)|N(T)=1\big] &= \exp\left(\int_t^T r(s)ds\right) \\
\mathbb{E}\big[N(t)|N(T)=n\big] &= n\exp\left(\int_t^T r(s)ds\right) \\
\mathrm{Var}\big[N(t)|N(T)=1\big] &= \exp\left(2\int_t^T r(s)ds\right)\int_t^T \frac{2\lambda(s)-r(s)}{\exp\left(\int_s^T r(u)du\right)}ds \\
\mathrm{Var}\big[N(t)|N(T)=n\big] &= n\exp\left(2\int_t^T r(s)ds\right)\int_t^T \frac{2\lambda(s)-r(s)}{\exp\left(\int_s^T r(u)du\right)}ds
\end{aligned}$$

Proof. We denote q the forward time, and t the backward time. We set the origin of q at backward time T, so that $q = T - t$ and $t = T - q$. This yields $dq = -dt$.

Our functions $\lambda(t)$ and $r(t)$ take backward time as their argument. We define $\tilde{\lambda}(q) := \lambda(T-q)$ and $\tilde{r}(q) := r(T-q)$ their equivalent in forward time. We also have $\mu(t) = \lambda(t) - \mu(t)$ and $\tilde{\mu}(q) = \tilde{\lambda}(q) - \tilde{r}(q) = \mu(T-q)$.

We denote

$$\begin{aligned}
M(T,t) &:= \mathbb{E}\big[N(t)|N(T)=1\big] \\
W(T,t) &:= \mathrm{Var}\big[N(t)|N(T)=1\big]
\end{aligned}$$

and $\tilde{M}(q) := M(T, T-q)$ and $\tilde{W}(q) := W(T, T-q)$ their forward-time equivalents. The functions $\tilde{M}(q)$ and $\tilde{W}(q)$ are those given by Kendall (5), and our purpose is here to derive $M(T,t)$ and $W(T,t)$.

Below we will use τ as a specific value of backward time and thus save the symbol t for inside the integrals. This way it is clear that when there is dq inside the integral, we integrate over forward time, and when there is dt we integrate over backward time. When integrals are nested, we also use ds for backward time.

From Kendall ((5), Eq. 13), we have for $\tau \in [0, T]$:

$$\begin{aligned} M(T,\tau) &= \tilde{M}(T-\tau) \\ &= \exp \int_0^{T-\tau} \tilde{r}(q)dq \\ &= \exp \int_0^{T-\tau} r(T-q)dq \end{aligned} \tag{7}$$

We proceed to a change of variable from q to t, giving

$$\begin{aligned} \mathbb{E}\big[N(\tau)|N(T)=1\big] =: M(T,\tau) &= \exp \int_T^{\tau} -r(t)dt \\ &= \exp \int_{\tau}^{T} r(t)dt \end{aligned} \tag{8}$$

From Kendall ((5), Eq. 14c), we have

$$\begin{aligned} W(T,\tau) &= \tilde{W}(T-\tau) \\ &= \tilde{M}(T-\tau)^2 \int_0^{T-\tau} \frac{\big(\tilde{\lambda}(q)+\tilde{\mu}(q)\big)}{\tilde{M}(q)} dq \\ &= M(T,\tau)^2 \int_0^{T-\tau} \frac{\big(\lambda(T-q)+\mu(T-q)\big)}{M(T,T-q)} dq \end{aligned} \tag{9}$$

We proceed to a change of variable from q to t in the integral, giving

$$\begin{aligned} \mathrm{Var}\big[N(\tau)|N(T)=1\big] =: W(T,\tau) &= M(T,\tau)^2 \int_T^{\tau} -\frac{\big(\lambda(t)+\mu(t)\big)}{M(T,t)} dt \\ &= M(T,\tau)^2 \int_{\tau}^{T} \frac{\big(\lambda(t)+\mu(t)\big)}{M(T,t)} dt \\ &= \exp\left(2\int_{\tau}^{T} r(t)dt\right) \int_{\tau}^{T} \frac{\big(2\lambda(t)-r(t)\big)}{\exp\big(\int_t^T r(s)ds\big)} dt \end{aligned} \tag{10}$$

Then, because the number of descendants of different individuals existing at time T

are independent, we have

$$\mathbb{E}\big[N(\tau)|N(T)=n\big] = n\exp\int_{\tau}^{T} r(t)dt$$
$$\mathrm{Var}\big[N(\tau)|N(T)=n\big] = n\exp\left(2\int_{\tau}^{T} r(t)dt\right)\int_{\tau}^{T}\frac{\big(2\lambda(t)-r(t)\big)}{\exp\big(\int_{t}^{T} r(s)ds\big)}dt$$

□

C.4 Convergence of the relative population size $N(t)/n$

This section gives the proof that, under a BD model with fixed $\lambda(t)$ and $r(t)$, the relative population size $N(t)/n$ converges in probability to $\int_t^T r(s)ds$ as $n\to\infty$.

Lemma C.5. *Given fixed functions $\lambda(t)$ and $r(t)$ bounded on $[0,T]$, the relative population size $N(t)/n$ given $N(T)=n$ converges uniformly in probability to $\int_t^T r(s)ds$ as $n\to\infty$.*

Proof. As per Bishop (35)'s Theorem 14.4-1, using Bishop's $O_p(\cdot)$ and $o_p(\cdot)$ notation, we have that

$$\left(\frac{N(t)}{n}\Big|N(T)=n\right) = \mathbb{E}\left[\frac{N(t)}{n}\Big|N(T)=n\right] + O_p\left(\mathrm{Var}\left[\frac{N(t)}{n}\Big|N(T)=n\right]^{\frac{1}{2}}\right)$$

Using Lemma C.4 we have

$$\mathbb{E}\left[\frac{N(t)}{n}\Big|N(T)=n\right] = \frac{1}{n}\cdot\mathbb{E}\big[N(t)|N(T)=n\big] = \exp\left(\int_t^T r(s)ds\right)$$

and

$$\begin{aligned}\mathrm{Var}\left[\frac{N(t)}{n}\Big|N(T)=n\right] &= \frac{1}{n^2}\cdot\mathrm{Var}\big[N(t)|N(T)=n\big]\\ &= \frac{1}{n}\exp\left(2\int_t^T r(s)ds\right)\int_t^T\frac{2\lambda(s)-r(s)}{\exp\big(\int_s^T r(u)du\big)}ds\end{aligned}$$

Provided that $r(t)$ and $\lambda(t)$ are bounded on $[0,T]$, we have

$$\mathrm{Var}\left[\frac{N(t)}{n}\Big|N(T)=n\right] = O(1/n)$$

We thus obtain

$$\begin{aligned}\left(\frac{N(t)}{n}\Big|N(T)=n\right) &= \exp\left(\int_t^T r(s)ds\right) + O_p(1/\sqrt{n})\\ &= \exp\left(\int_t^T r(s)ds\right) + o_p(1)\end{aligned}$$

which concludes the proof.

□

C.5 Unconditional distribution of the sample process Z

The following lemma provides the unconditional BD density $f_Z(z|N(T)=n)$. This density has been established elsewhere (see (12, 47)) for $n=1$. Here, we establish the general formula for any n.

Lemma C.6. *Assuming a Bernoulli sampling procedure with sampling probability ρ, the density of Z given $N(T)=n$ is*

$$f_Z\big(z|N(T)=n\big) = \frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!}\binom{n}{z(T)} Q_0^\rho(T)^{n-z(T)} Q_1^\rho(T)^{z(T)} \prod_{t\in\alpha(z)} \lambda(t) Q_1^\rho(t)$$

with

$$Q_k^\rho(t) := \mathbb{P}\big(Z(0)=k|N(t)=1\big)$$

and with $\alpha(z)$ the set of coalescent times induced by z.

Proof. We recall that a BD process starting at time T with n labelled individuals and where $Z(0)$ individuals have been sampled induces a sample forest $\Psi := (\Gamma_i)_{i=1,\dots,n}$, with Γ_i the ordered time tree comprising the ancestors of the sample descending from individual i existing at time T. Given that the n trees composing the sample forest are iid. under the Bernoulli-sampled BD, it follows that

$$f_\Psi(\psi) = \prod_{i=1}^n f_\Gamma(\gamma_i)$$

with γ_i a given value of Γ_i and $\psi = (\gamma_i)_{i=1,\dots,n}$ a given value of Ψ.

Consider a realisation $Z=z$ of the sample process and recall that $S_n(z)$ is the set of sample forests with n trees that induce $Z=z$, i.e. all the sample forests with n trees, with $z(0)$ leaves and with branching times $\alpha(z)$. We have

$$f_Z\big(z|N(T)=n\big) = \sum_{\psi\in S_n(z)} f_\Psi(\psi) = \sum_{\psi\in S_n(z)} \prod_{i=1}^n f_\Gamma(\gamma_i)$$

Denote L_i the number of leaves of Γ_i, denote $\Xi_0(\Psi)$ the set of initial individuals without sampled descendants (i.e. $i\in\Xi_0(\Psi)$ iff $L_i=0$), and $\Xi_{>0}(\Psi)$ the set of initial individuals with sampled descendants (i.e. $i\in\Xi_{>0}(\Psi)$ iff $L_i>0$). Notice that, for

some realization z, and for some forest $\psi \in S_n(z)$, $\mathrm{Card}\big(\Xi_0(\psi)\big) = n - z(T)$ and $\mathrm{Card}\big(\Xi_{>0}(\psi)\big) = z(T)$. We thus have

$$\begin{aligned} f_Z\big(z|N(T)=n\big) &= \sum_{\psi\in S_n(z)} \left(\prod_{i\in\Xi_0(\psi)} f_\Gamma(\gamma_i)\right) \times \left(\prod_{i\in\Xi_{>0}(\psi)} f_\Gamma(\gamma_i)\right) \\ &= \sum_{\psi\in S_n(z)} \left(\prod_{i\in\Xi_0(\psi)} Q_0^\rho(T)\right) \times \left(\prod_{i\in\Xi_{>0}(\psi)} f_\Gamma(\gamma_i)\right) \\ &= \sum_{\psi\in S_n(z)} Q_0^\rho(T)^{n-z(T)} \times \prod_{i\in\Xi_{>0}(\psi)} f_\Gamma(\gamma_i) \end{aligned}$$

Recalling that $\bar{\alpha}(\gamma)$ denotes the set of branching times of tree γ, Louca and Pennel ((47), Eq. 16 in the Supporting Information) give the density

$$f_\Gamma(\gamma|L>0) = \frac{Q_1^\rho(T)}{1-Q_0^\rho(T)} \prod_{t\in\bar{\alpha}(\gamma)} \lambda(t) Q_1^\rho(t)$$

since $\rho\Psi(0,t)$ and $E(t)$ in Louca's and Pennel's notation are respectively equivalent to $Q_1^\rho(t)$ and $Q_0^\rho(t)$ in our notation. Then we have that $\forall i \in \Xi_{>0}(\psi)$

$$\begin{aligned} f_\Gamma(\gamma_i) &= f_\Gamma(\gamma_i|L>0)\mathbb{P}(L>0) \\ &= f_\Gamma(\gamma_i|L>0)(1-Q_0^\rho(T)) \\ &= Q_1^\rho(T) \prod_{t\in\bar{\alpha}(\gamma_i)} \lambda(t) Q_1^\rho(t) \end{aligned}$$

Recalling that $\psi \in S_n(z)$ implies that $\bigcup_{i\in\Xi_{>0}(\psi)} \bar{\alpha}(\gamma_i) = \alpha(z)$, we then have

$$\begin{aligned} f_Z\big(z|N(T)=n\big) &= \sum_{\psi\in S_n(z)} Q_0^\rho(T)^{n-z(T)} \times \prod_{i\in\Xi_{>0}(\psi)} Q_1^\rho(T) \prod_{t\in\bar{\alpha}(\gamma_i)} \lambda(t) Q_1^\rho(t) \\ &= \sum_{\psi\in S_n(z)} Q_0^\rho(T)^{n-z(T)} \times Q_1^\rho(T)^{z(T)} \prod_{i\in\Xi_{>0}(\psi)} \prod_{t\in\bar{\alpha}(\gamma_i)} \lambda(t) Q_1^\rho(t) \\ &= \sum_{\psi\in S_n(z)} Q_0^\rho(T)^{n-z(T)} \times Q_1^\rho(T)^{z(T)} \prod_{t\in\alpha(z)} \lambda(t) Q_1^\rho(t) \\ &= Q_0^\rho(T)^{n-z(T)} \times Q_1^\rho(T)^{z(T)} \prod_{t\in\alpha(z)} \lambda(t) Q_1^\rho(t) \sum_{\psi\in S_n(z)} 1 \end{aligned}$$

There then only remains to count the number of elements in $S_n(z)$:

$$\sum_{\psi\in S_n(z)} 1 = \binom{n}{z(T)} \frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!}$$

where $\binom{n}{z(T)}$ is the number of ways of choosing the $z(T)$ individuals with sampled descendants among n individuals, and where $\frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!}$ is the number of orderings of a forest of $z(T)$ trees (each with at least one leaf) with a total of $z(0)$ leaves.

□

D Results for the scaled BD model

In this section, we consider the scaled BD process, parametrized with n, $\lambda_1(t)$ and $r(t)$ such that

$$\lambda(t) = \lambda_n(t) = \lambda_1(t)n$$
$$\mu(t) = \mu_n(t) = \lambda_n(t) - r(t)$$

We recall that we have set the conditions that on $[0, T]$: (i) $\lambda_1(t) > 0$, (ii) $\lambda_1(t)$ and $r(t)$ are uniformly continuous (and thus bounded above and below) and (iii) $\lambda_1(t) \geq r(t)$.

Where relevant, a Bernoulli sampling with probability $\rho = \rho_1/n$ is assumed.

Contrarily to Section E, it is not assumed that the population process is conditioned.

D.1 Results for the distribution of $N(t)$ and $Z(t)$ in the limit $n \to \infty$

The purpose of this section is to derive the limits of certain quantities as $n \to \infty$, and it can be verified that these limits do not depend on n, as they should. All limits may indeed depend only on $r(t)$, $\lambda_1(t)$, $D(t_1, t_2)$, $B_1(t_1, t_2)$, ρ_1 and on some realization z of Z, all of which are independent of n.

Lemma D.1. *The expected relative population size is given by,* $\forall n \in \mathbb{N}_{>0}$

$$\mathbb{E}\left[\frac{N(t)}{n}\middle| N(T) = n\right] = \exp \int_t^T r(s)ds$$

Proof. This lemma's claim follows trivially from Lemma C.4. □

Lemma D.2. *The variance of the relative population size reaches the following limit*

$$\mathrm{Var}\left[\frac{N(t)}{n}\middle| N(T) = n\right] \xrightarrow[n\to\infty]{} \exp\left(2\int_t^T r(s)ds\right) \int_t^T \frac{2\lambda_1(s)ds}{\exp\left(\int_s^T r(u)du\right)}$$

Proof. Using Lemma C.4, we have

$$\begin{aligned}
&\mathrm{Var}\left[\frac{N(t)}{n}\middle|N(T)=n\right] \\
&\quad= \frac{1}{n^2}\cdot \mathrm{Var}\big[N(t)|N(T)=n\big] \\
&\quad= \frac{1}{n^2}\cdot n\exp\left(2\int_t^T r(s)ds\right)\int_t^T \frac{\big(2\lambda(s)-r(s)\big)}{\exp\big(\int_s^T r(u)du\big)}ds \\
&\quad= \exp\left(2\int_t^T r(s)ds\right)\left(\frac{1}{n}\int_t^T \frac{2\lambda_1(s)n\,ds}{\exp\big(\int_s^T r(u)du\big)} - \frac{1}{n}\int_t^T \frac{r(s)ds}{\exp\big(\int_s^T r(u)du\big)}\right) \\
&\quad= \exp\left(2\int_t^T r(s)ds\right)\left(\int_t^T \frac{2\lambda_1(s)ds}{\exp\big(\int_s^T r(u)du\big)} - \frac{1}{n}\int_t^T \frac{r(s)ds}{\exp\big(\int_s^T r(u)du\big)}\right) \\
&\quad\xrightarrow[n\to\infty]{} \exp\left(2\int_t^T r(s)ds\right)\int_t^T \frac{2\lambda_1(s)ds}{\exp\big(\int_s^T r(u)du\big)}
\end{aligned}$$

□

Lemma D.3. *Assuming a Bernoulli sampling procedure with sampling probability $\rho = \frac{\rho_1}{n}$, the density of Z given $N(T)=n$ reaches the following limit as $n\to\infty$.*

$$\begin{aligned}
&\forall z : z(0) > 0, \\
&f_Z(z|N(T)=n) \xrightarrow[n\to\infty]{} \frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!z(T)!}\left(\frac{\rho_1 D(0,T)}{\big(1+\rho_1 B_1(0,T)\big)^2}\right)^{z(T)} \\
&\qquad\qquad \times \exp\left(\frac{-\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right)\prod_{t\in\alpha(z)}\frac{\lambda_1(t)\rho_1 D(0,t)}{(1+\rho_1 B_1(0,t))^2}
\end{aligned}$$

with $\alpha(z)$ the set of coalescent times induced by z.

Proof. By Lemma C.6,

$$\begin{aligned}
&f_Z\big(z|N(T)=n\big) \\
&= \frac{\big(z(0)-1\big)!}{\big(z(T)-1\big)!}\binom{n}{z(T)}Q_0^\rho(T)^{n-z(T)}Q_1^\rho(T)^{z(T)}\prod_{t\in\alpha(z)}\lambda(t)Q_1^\rho(t)
\end{aligned}$$

with

$$Q_k^\rho(t) := \mathbb{P}\big(Z(0)=k|N(t)=1\big)$$

We will now show that

$$Q_0^\rho(T)^{n-z(T)} \xrightarrow[n\to\infty]{} \exp\left(\frac{-\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right) \tag{11}$$

$$\binom{n}{z(T)}Q_1^\rho(T)^{z(T)} \xrightarrow[n\to\infty]{} \frac{1}{z(T)!}\left(\frac{\rho_1 D(0,T)}{(1+\rho_1 B_1(0,T))^2}\right)^{z(T)} \tag{12}$$

$$\lambda(t)Q_1^\rho(t) = \frac{\lambda_1(t)\rho_1 D(0,t)}{(1+\rho_1 B_1(0,t))^2}$$

hence proving the lemma.

Notice that

$$\begin{aligned}\rho B(0,t) &= \rho \int_0^t \lambda(s) D(s,t) ds = \frac{\rho_1}{n} \int_0^t \lambda_1(s) n D(s,t) ds \\ &= \rho_1 \int_0^t \lambda_1(s) D(s,t) ds = \rho_1 B_1(0,t)\end{aligned}$$

Using Lemma C.1,

$$\begin{aligned}Q_0^\rho(T)^{n-z(T)} =& \left(1 - \frac{\rho D(0,T)}{1+\rho B(0,T)}\right)^{n-z(T)} \\ =& \left(1 - \frac{1}{n} \cdot \frac{\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right)^{n-z(T)} \\ \xrightarrow[n\to\infty]{}& \exp\left(\frac{-\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right)\end{aligned}$$

$$\begin{aligned}\binom{n}{z(T)} Q_1^\rho(T)^{z(T)} &= \binom{n}{z(T)} \left(\frac{\rho D(0,T)}{(1+\rho B(0,T))^2}\right)^{z(T)} \\ &= \binom{n}{z(T)} \frac{1}{n^{z(T)}} \left(\frac{\rho_1 D(0,T)}{(1+\rho_1 B_1(0,T))^2}\right)^{z(T)} \\ &= \frac{1}{z(T)!} \cdot \frac{\left(\frac{n-z(T)+1}{n}\right)^{z(T)}}{\frac{(n-z(T))!(n-z(T)+1)^{z(T)}}{n!}} \left(\frac{\rho_1 D(0,T)}{(1+\rho_1 B_1(0,T))^2}\right)^{z(T)}\end{aligned}$$

Trivially,

$$\left(\frac{n-z(T)+1}{n}\right)^{z(T)} \xrightarrow[n\to\infty]{} 1$$

Also, using Euler's infinite-product definition of the factorial (see (50)):

$$\lim_{n\to\infty} \frac{(n-z(T))!(n-z(T)+1)^{z(T)}}{n!} = \lim_{\substack{m=n-z(T) \\ m\to\infty}} \frac{m!(m+1)^{z(T)}}{(m+z(T))!} = 1$$

Therefore,

$$\binom{n}{z(T)} Q_1^\rho(T)^{z(T)} \xrightarrow[n\to\infty]{} \frac{1}{z(T)!} \left(\frac{\rho_1 D(0,T)}{(1+\rho_1 B_1(0,T))^2}\right)^{z(T)}$$

Finally, using Lemma C.1,

$$\lambda(t)Q_1^\rho(t) = \lambda(t)\frac{\rho D(0,t)}{(1+\rho B(0,t))^2} = \frac{\lambda_1(t)\rho_1 D(0,t)}{(1+\rho_1 B_1(0,t))^2}$$

is independent of n. □

Lemma D.4. *Assuming a Bernoulli sampling procedure with sampling probability $\rho = \frac{\rho_1}{n}$, the density of Z given $N(T) = n$ and $Z(0) = z_0 \in \mathbb{N}_{>0}$ reaches a non-zero limit as $n \to \infty$.*

Proof.

$$\forall z : z(0) = z_0 \in \mathbb{N}_{>0} \qquad f_Z(z|N(T) = n, Z(0) = z_0) = \frac{f_Z(z|N(T) = n)}{\mathbb{P}(Z(0) = z_0|N(T) = n)}$$

By Lemma D.3, we know that $f_Z(z|N(T) = n)$ reaches a non-zero limit as $n \to \infty$. There therefore remains to prove that, for a given z_0, $\mathbb{P}(Z(0) = z_0|N(T) = n)$ reaches a non-zero limit to prove this lemma.

We write $Q_k \equiv Q_k^\rho(T) := \mathbb{P}(Z(0) = k|N(T) = n)$.

Using Lemma C.3, we have for $z_0 \geq 1$

$$\begin{aligned}&\mathbb{P}(Z(0) = z_0|N(T) = n)\\ &= \sum_{z_T=1}^{z_0} \binom{n}{z_T} Q_0^{n-z_T} \binom{z_0-1}{z_0-z_T} \left(1 - \frac{Q_1}{1-Q_0}\right)^{z_0-z_T} Q_1^{z_T}\end{aligned}$$

We notice that (using Lemma C.1)

$$1-\frac{Q_1}{1-Q_0} = 1-\frac{\frac{\rho D(0,T)}{(1+\rho B(0,T))^2}}{\frac{\rho D(0,T)}{1+\rho B(0,T)}} = 1-\frac{1}{1+\rho B(0,T)} = 1-\frac{1}{1+\rho_1 B_1(0,T)} = \frac{\rho_1 B_1(0,T)}{1+\rho_1 B_1(0,T)}$$

is independent of n. Then, plugging in the limits as $n \to \infty$ of $\binom{n}{z_T} Q_1^{z_T}$ and $Q_0^{n-z_T}$ (given in Eqs. 11,12), we obtain

$$\begin{aligned}&\mathbb{P}(Z(0) = z_0|N(T) = n)\\ &\qquad \xrightarrow[n\to\infty]{} \sum_{z_T=1}^{z_0} \left\{ \binom{z_0-1}{z_0-z_T} \left(\frac{\rho_1 B_1(0,T)}{1+\rho_1 B_1(0,T)}\right)^{z_0-z_T} \right.\\ &\qquad\qquad \left. \times \exp\left(\frac{-\rho_1 D(0,T)}{1+\rho_1 B_1(0,T)}\right) \frac{1}{z_T!} \left(\frac{\rho_1 D(0,T)}{(1+\rho_1 B_1(0,T))^2}\right)^{z_T} \right\}\end{aligned}$$

□

D.2 Results for the distribution of the number of descendants of one individual after g units of time.

In this section we derive a number of results that will be important for the study of the conditional scaled BD tackled in Section E. In particular, since the condition considered in Section E constrains the population size every $g = g_1/n$ units of time, we will need some results about the distribution of the number of descendants at time $t-g$ of a single individual existing at time t. Given some number $N(t)$ of individuals existing at time t, we denote ν_{gti} the random number of descendants at time $t-g$ of the i-th individual (with $i = 1, \dots, N(t)$). Given that under the BD process the ν_{gti}'s are iid., we consider here the distribution of ν_{gt1} without loss of generality.

Given that $g = g_1/n$ and $n \in \mathbb{N}_{>0}$, the values of interest of g are comprised in the set $\{\frac{g_1}{1}, \frac{g_1}{2}, \frac{g_1}{3}, \dots\}$, and the domain of t is $[g, T]$. However, since the BD unravels in continuous time, the distribution of ν_{gt1} is defined continuously for $g \in (0, g_1]$ and for $t \in [g, T]$. Furthermore, as we will see, the limiting distribution of ν_{gt1} as $g \to 0$ exists, so we will consider $[0, g_1]$ as the domain of g, and $[g, T]$ as the domain of t.

Notice that in this section we will derive some results in the limit $g \to 0$, which is equivalent to the limit $n \to \infty$, the focus of this paper, since $g := g_1/n$. Notice that our notation does not always explicitly mention the dependency of variables or functions on g (equivalently on n), so as not to encumber formulae.

Lemma D.5. *The distribution of ν_{gt1} is given by*

$$\mathbb{P}\big(\nu_{gt1} = 0\big) = 1 - \frac{D(t-g,t)}{1+B(t-g,t)}$$
$$\mathbb{P}\big(\nu_{gt1} = k\big) = \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\left(\frac{B(t-g,t)}{1+B(t-g,t)}\right)^{k-1} \qquad k \geq 1$$

with mean $A(g,t)$ and variance $V(g,t)$ given by

$$A(g,t) = D(t-g,t)$$
$$V(g,t) = 2A(g,t)B(t-g,t) + A(g,t) - A(g,t)^2$$

Proof. Given that ν_{gt1} is equivalent by definition to $N(t-g)|N(t) = 1$, the distribution of ν_{gt1} is obtained immediately by setting $t_1 = t-g$ and $t_2 = t$ in Lemma C.2.

Then we have

$$\begin{aligned}
A(g,t) &= \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\cdot\sum_{k=1}^{\infty}k\bigg(\frac{B(t-g,t)}{1+B(t-g,t)}\bigg)^{k-1}\\
&= \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\cdot\frac{1}{\Big(1-\frac{B(t-g,t)}{1+B(t-g,t)}\Big)^2}\\
&= D(t-g,t)
\end{aligned}$$

$$\begin{aligned}
V(g,t) &= \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\cdot\sum_{k=1}^{\infty}k^2\bigg(\frac{B(t-g,t)}{1+B(t-g,t)}\bigg)^{k-1} - A(g,t)^2\\
&= \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\cdot\frac{1+\frac{B(t-g,t)}{1+B(t-g,t)}}{\Big(1-\frac{B(t-g,t)}{1+B(t-g,t)}\Big)^3} - A(g,t)^2\\
&= D(t-g,t)\big(1+2B(t-g,t)\big) - A(g,t)^2\\
&= 2A(g,t)B(t-g,t) + A(g,t) - A(g,t)^2
\end{aligned}$$

□

Lemma D.6. *Defining ν_{0t1} as the random variable with distribution*

$$\mathbb{P}\big(\nu_{0t1}=k\big) := \lim_{g\to 0}\mathbb{P}\big(\nu_{gt1}=k\big),$$

ν_{0t1} has distribution

$$\begin{aligned}
\mathbb{P}\big(\nu_{0t1}=0\big) &= 1-\frac{1}{1+\lambda_1(t)g_1}\\
\mathbb{P}\big(\nu_{0t1}=k\big) &= \frac{1}{\big(1+\lambda_1(t)g_1\big)^2}\bigg(\frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1}\bigg)^{k-1} \qquad\qquad k\geq 1
\end{aligned}$$

with mean $A(0,t)$ and variance $V(0,t)$ given by

$$\begin{aligned}
A(0,t) &= 1\\
V(0,t) &= 2\lambda_1(t)g_1
\end{aligned}$$

Proof. We note that, since $\lambda_1(t)$ and $r(t)$ are continuous on $[0,T]$,

$$R(t) := \int_0^t r(s)ds \qquad\qquad \text{and} \qquad\qquad \Lambda_1(t) := \int_0^t \lambda_1(s)ds$$

are continuous and differentiable on $[0,T]$.

Therefore, by the mean-value theorem, $\exists s\in[t-g,t]$ such that

$$\begin{aligned}
R(t)-R(t-g) &= r(s)g\\
\Lambda_1(t)-\Lambda_1(t-g) &= \lambda_1(s)g
\end{aligned}$$

Hence, $\exists s \in [t-g, t]$ such that

$$\begin{aligned} D(t-g,t) &= \exp \int_{t-g}^{t} r(u)du = \exp\big(R(t) - R(t-g)\big) \\ &= \exp\big(r(s)g\big) = 1 + O(r(s)g) \\ &= 1 + O(g) \xrightarrow[g\to 0]{} 1 \end{aligned}$$

where we have used the fact that $O(r(s)g) = O(g)$ since $r(t)$ is bounded on $[0, T]$.

Also, $\exists s \in [t-g, t]$ such that

$$\begin{aligned} B(t-g,t) &= \int_{t-g}^{t} \lambda(u) D(u,t) du = \int_{t-g}^{t} \lambda(u)\big(1 + O(g)\big)du \\ &= \big(1 + O(g)\big) \int_{t-g}^{t} \lambda(u)du = \big(1 + O(g)\big)\frac{g_1}{g} \int_{t-g}^{t} \lambda_1(u)du \\ &= \big(1 + O(g)\big)\frac{g_1}{g}\big(\Lambda_1(t) - \Lambda_1(t-g)\big) \\ &= \big(1 + O(g)\big)\frac{g_1}{g}\big(\lambda_1(s)g\big) \\ &= \big(1 + O(g)\big) g_1 \lambda_1(s) \end{aligned}$$

Then, given that $\lambda_1(t)$ is uniformly continuous, $\lambda_1(s) = \lambda_1(t) + o(1)$ as $t - s \to 0$, or equivalently as $g \to 0$, since $t - s \leq g$. Hence,

$$\begin{aligned} B(t-g,t) &= \big(1 + O(g)\big) g_1 (\lambda_1(t) + o(1)) \\ &= \lambda_1(t) g_1 + o(1) \\ &\xrightarrow[g\to 0]{} \lambda_1(t) g_1 \end{aligned}$$

where we have used the fact that $\lambda_1(t)O(g) = O(g)$ since $\lambda_1(t)$ is bounded on $[0, T]$.

Plugging the limiting values of $D(t-g,t)$ and $B(t-g,t)$ into the expressions of Lemma D.5 yields this lemma's statement. □

Lemma D.7.

$$\mathbb{P}\big(\nu_{gt1} = k\big) = \mathbb{P}\big(\nu_{0t1} = k\big) + o_k(1)$$

as $g \to 0$.

Proof. In the proof of Lemma D.6 we have seen that

$$\begin{aligned} D(t-g,t) &= 1 + O(g) \\ B(t-g,t) &= \lambda_1(t) g_1 + o(1) \end{aligned}$$

as $g \to 0$.

Hence, using the fact that $\lambda_1(t)$ is bounded on $[0,T]$, for $k=0$

$$\begin{aligned}\mathbb{P}\big(\nu_{gt1}=0\big) &= 1-\frac{D(t-g,t)}{1+B(t-g,t)} = 1-\frac{1+O(g)}{1+\lambda_1(t)g_1+o(1)}\\ &= 1-\frac{1}{1+\lambda_1(t)g_1}+o(1)\\ &= \mathbb{P}\big(\nu_{0t1}=0\big)+o(1)\end{aligned}$$

and for $k\geq 1$

$$\begin{aligned}\mathbb{P}\big(\nu_{gt1}=k\big) &= \frac{D(t-g,t)}{\big(1+B(t-g,t)\big)^2}\left(\frac{B(t-g,t)}{1+B(t-g,t)}\right)^{k-1}\\ &= \frac{1+O(g)}{\big(1+\lambda_1(t)g_1+o(1)\big)^2}\left(\frac{\lambda_1(t)g_1+o(1)}{1+\lambda_1(t)g_1+o(1)}\right)^{k-1}\\ &= \frac{1+O(g)}{\big(1+\lambda_1(t)g_1\big)^2+o(1)}\left(\frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1}+o(1)\right)^{k-1}\\ &= \left(\frac{1}{\big(1+\lambda_1(t)g_1\big)^2}+o(1)\right)\left(\frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1}+o(1)\right)^{k-1}\\ &= \left(\frac{1}{\big(1+\lambda_1(t)g_1\big)^2}+o(1)\right)\left(\left(\frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1}\right)^{k-1}+(k-1)o(1)\right)\\ &= \frac{1}{\big(1+\lambda_1(t)g_1\big)^2}\left(\frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1}\right)^{k-1}+o_k(1)\\ &= \mathbb{P}\big(\nu_{0t1}=k\big)+o_k(1)\end{aligned}$$

□

Lemma D.8.

$$A(g,t) = 1+O(g)$$

as $g\to 0$.

Proof. By Lemma D.5, $A(g,t)=D(t-g,t)$. Furthermore, in the proof to Lemma D.6, we have seen that $D(t-g,t)=1+O(g)$. □

Lemma D.9. *Given*

$$\eta(g,t) := 1-\frac{p(1)}{1-p(0)}$$

with $p(k):=\mathbb{P}(\nu_{gt1}=k)$*, there exist* $\check{\eta}>0$ *and* $\hat{\eta}<1$ *such that* $\forall g\in[0,g_1]$ *and* $\forall t\in[g,T]$

$$\check{\eta}\leq\eta(g,t)\leq\hat{\eta}$$

Proof. By Lemma D.5, for $g \in (0, g_1]$

$$\eta(g,t) = \frac{B(t-g,t)}{1+B(t-g,t)}$$

with

$$B(t-g,t) := \int_{t-g}^{t} \lambda_1(t)\frac{g_1}{g} D(s,t)ds$$

and with

$$D(s,t) := \exp \int_{s}^{t} r(u)du$$

First we will show that $\exists \check{D} > 0,\ \exists \hat{D} < \infty$ such that $\forall g \in (0, g_1],\ \forall t \in [g, T]$ and $\forall s \in [0, t]$

$$\check{D} \leq D(s,t) \leq \hat{D} \tag{13}$$

Denote $\check{r} := \inf_{t\in[0,T]} r(t)$ and $\hat{r} := \sup_{t\in[0,T]} r(t)$. Both $\check{r}$ and $\hat{r}$ exist since $r(t)$ is bounded on $[0, T]$.

The following will hold $\forall g \in (0, g_1]$, $\forall t \in [g, T]$ and $\forall s \in [0, t]$.

We have

$$D(s,t) \geq \exp \int_{s}^{t} \check{r}du = \exp\left(\check{r}(t-s)\right)$$

If $\check{r} < 0$, then

$$\exp\left(\check{r}(t-s)\right) \geq \exp(\check{r}T) = \exp(-|\check{r}|T)$$

If $\check{r} \geq 0$, then

$$\exp\left(\check{r}(t-s)\right) \geq \exp(0) \geq \exp(-|\check{r}|T)$$

Hence,

$$D(s,t) \geq \exp(-|\check{r}|T) =: \check{D} > 0$$

We also have

$$D(s,t) \leq \exp \int_{s}^{t} \hat{r}du = \exp\left(\hat{r}(t-s)\right)$$

If $\hat{r} \leq 0$, then

$$\exp\left(\hat{r}(t-s)\right) \leq \exp(0) \leq \exp(|\hat{r}|T)$$

If $\hat{r} > 0$, then

$$\exp\left(\hat{r}(t-s)\right) \leq \exp(\hat{r}T) = \exp(|\hat{r}|T)$$

Hence,

$$D(s,t) \leq \exp(|\hat{r}|T) =: \hat{D} < \infty$$

Now we will show that $\exists \check{B} > 0,\ \exists \hat{B} < \infty$ such that $\forall g \in (0, g_1]$ and $\forall t \in [g, T]$

$$\check{B} \leq B(t-g,t) \leq \hat{B} \tag{14}$$

Denote $\check{\lambda}_1 := \inf_{t\in[0,T]} \lambda_1(t)$ and $\hat{\lambda}_1 := \sup_{t\in[0,T]} \lambda_1(t)$. Given that $\lambda_1(t)$ is bounded and > 0 on $[0,T]$, we have $0 < \check{\lambda}_1 \le \hat{\lambda}_1 < \infty$.

We have $\forall g \in (0, g_1]$ and $\forall t \in [g,T]$

$$B(t-g,t) \ge \int_{t-g}^{t} \check{\lambda}_1 \frac{g_1}{g} \check{D} ds = \check{\lambda}_1 g_1 \check{D} =: \check{B} > 0$$

and

$$B(t-g,t) \le \int_{t-g}^{t} \hat{\lambda}_1 \frac{g_1}{g} \hat{D} ds = \hat{\lambda}_1 g_1 \hat{D} =: \hat{B} < \infty$$

From this it follows that $\forall g \in (0, g_1]$ and $\forall t \in [g,T]$

$$0 < \frac{\check{B}}{1+\check{B}} \le \eta(g,t) \le \frac{\hat{B}}{1+\hat{B}} < 1$$

For $g = 0$, using Lemma D.6, we have $\eta(0,t) = \frac{\lambda_1(t) g_1}{1+\lambda_1(t) g_1}$. Hence, $\forall t \in [0,T]$

$$0 < \frac{\check{\lambda}_1 g_1}{1+\check{\lambda}_1 g_1} \le \eta(0,t) \le \frac{\hat{\lambda}_1 g_1}{1+\hat{\lambda}_1 g_1} < 1$$

Hence, setting

$$\check{\eta} := \min\left(\frac{\check{B}}{1+\check{B}}, \frac{\check{\lambda}_1 g_1}{1+\check{\lambda}_1 g_1} \right) > 0$$
$$\hat{\eta} := \max\left(\frac{\hat{B}}{1+\hat{B}}, \frac{\hat{\lambda}_1 g_1}{1+\hat{\lambda}_1 g_1} \right) < 1$$

we obtain the lemma's claim.

□

Lemma D.10. *There exists $\check{V} > 0$ such that $\forall g \in [0, g_1]$ and $\forall t \in [g,T]$*

$$V(g,t) \ge \check{V}$$

Proof. The following will hold $\forall g \in (0, g_1]$ and $\forall t \in [g,T]$.

By Lemma D.5, we have

$$A(g,t) = D(t-g,t) = \exp \int_{t-g}^{t} r(s) ds$$
$$V(g,t) = A(g,t)\big(2B(t-g,t) + 1 - A(g,t)\big)$$

with

$$B(t-g,t) = \int_{t-g}^{t} \lambda(s) D(s,t) ds$$

By Eq. (13), there exists $\check{D} > 0$ such that, $\forall s \in [0,t]$

$$D(s,t) \geq \check{D}$$

Hence, since $t - g \in [0,t]$, we have that

$$A(g,t) = D(t-g,t) \geq \check{D} > 0$$

Noticing that

$$\frac{\partial}{\partial s} D(s,t) = -r(s) D(s,t)$$

we have

$$\begin{aligned} 1 - A(g,t) &= 1 - D(t-g,t) = D(t,t) - D(t-g,t) \\ &= \int_{t-g}^{t} \frac{\partial}{\partial s} D(s,t) ds = \int_{t-g}^{t} -r(s) D(s,t) ds \end{aligned}$$

We thus have

$$\begin{aligned} 2B(t-g,t) + 1 - A(g,t) &= B(t-g,t) + B(t-g,t) + 1 - A(g,t) \\ &= B(t-g,t) + \int_{t-g}^{t} \lambda(s) D(s,t) ds - \int_{t-g}^{t} r(s) D(s,t) ds \\ &= B(t-g,t) + \int_{t-g}^{t} \big(\lambda(s) - r(s)\big) D(s,t) ds \end{aligned}$$

Given that $\lambda(s) - r(s) \geq 0$, that $D(s,t) > 0$, and using Eq. (14), we have

$$2B(t-g,t) + 1 - A(g,t) \geq B(t-g,t) \geq \check{B} > 0$$

Therefore, $\forall g \in (0, g_1]$ and $\forall t \in [g,T]$

$$V(g,t) \geq \check{D}\check{B} > 0$$

Then, for $g = 0$ and $\forall t \in [0,T]$, by Lemma D.6 we have

$$V(0,t) = 2\lambda_1(t) g_1 \geq 2\check{\lambda}_1 g_1 > 0$$

where $\check{\lambda}_1 := \inf_{t \in [0,T]} \lambda_1(t) > 0$ since $\lambda_1(t)$ is bounded and > 0 on $[0,T]$.

Hence, $\forall g \in [0, g_1]$ and $\forall t \in [g,T]$

$$V(g,t) \geq \min(\check{D}\check{B}, 2\check{\lambda}_1 g_1) =: \check{V} > 0$$

□

Lemma D.11. *For all $k \in \mathbb{N}_{>0}$, there exists a sequence $(M_{km})_m$ such that $\forall g \in [0, g_1]$ and $\forall t \in [g, T]$*

$$\mathbb{P}\big(\nu_{gt1} = 0\big) \leq M_{k0} \qquad m^k \mathbb{P}\big(\nu_{gt1} = m\big) \leq M_{km} \qquad M_k := \sum_{m=0}^{\infty} M_{km} < \infty$$

Proof. Let $M_{k0} = 1$ and obviously $\forall g \in [0, g_1]$ and $\forall t \in [g, T]$

$$\mathbb{P}\big(\nu_{gt1} = 0\big) \leq M_{k0}$$

For $m \geq 1$, by Lemmas D.5 and D.6, $\forall g \in [0, g_1]$ and $\forall t \in [g, T]$

$$\mathbb{P}(\nu_{gt1} = m) = \mathbb{P}(\nu_{gt1} = 1)\eta(g,t)^{m-1}$$

with

$$\eta(g,t) = 1 - \frac{\mathbb{P}(\nu_{gt1} = 1)}{1 - \mathbb{P}(\nu_{gt1} = 0)}$$

By Lemma D.9, $\forall g \in [0, g_1]$ and $\forall t \in [g, T]$

$$0 < \check{\eta} \leq \eta(g,t) \leq \hat{\eta} < 1$$

Hence,

$$m^k \mathbb{P}(\nu_{gt1} = m) = m^k \mathbb{P}(\nu_{gt1} = 1)\eta(g,t)^{m-1} \leq m^k \hat{\eta}^{m-1} =: M_{km}$$

We then have

$$M_k := \sum_{m=0}^{\infty} M_{km} = 1 + \frac{1}{\hat{\eta}} \sum_{m=1}^{\infty} m^k \hat{\eta}^m = 1 + \frac{\mathrm{Li}_{-k}(\hat{\eta})}{\hat{\eta}}$$

where $\mathrm{Li}_{-k}(z)$ is the polylogarithm of order $-k$ of z, which exists for $|z| < 1$. Given that $|\hat{\eta}| < 1$, then $\mathrm{Li}_{-k}(\hat{\eta})$ exists. Given that $\hat{\eta} > 0$, then

$$M_k < \infty$$

for all $k \in \mathbb{N}_{>0}$. □

Corollary D.11.1. *For all $k \in \mathbb{N}_{>0}$*

$$\sup_{g \in [0, g_1]} \sup_{t \in [g, T]} \mathbb{E}[\nu_{gt1}^k] < \infty$$

Proof. $\forall k \in \mathbb{N}_{>0}$, $\forall g \in [0, g_1]$ and $\forall t \in [g, T]$

$$\mathbb{E}[\nu_{gt1}^k] = \sum_{m=1}^{\infty} m^k \mathbb{P}(\nu_{gt1} = 1)\eta^{m-1} \leq \sum_{m=1}^{\infty} m^k \hat{\eta}^{m-1} = M_k - 1 < \infty$$

□

Lemma D.12. *There exist $\hat{V} < \infty$ and $\hat{X} < \infty$ such that, $\forall g \in [0, g_1]$, $\forall t \in [g, T]$*

$$V(g,t) \leq \hat{V} \qquad\qquad X(g,t) := \mathbb{E}\big[|\nu_{gt1} - \mathbb{E}\nu_{gt1}|^3\big] \leq \hat{X}$$

Proof. Define the L_p norm of random variable J as $||J||_p := \mathbb{E}\big[|J|^p\big]^{1/p}$. Notice that

$$||\mathbb{E}\nu_{gt1}||_p = \mathbb{E}\big[\mathbb{E}[\nu_{gt1}]^p\big]^{1/p} = \big(\mathbb{E}[\nu_{gt1}]^p\big)^{1/p} = \mathbb{E}\nu_{gt1}$$

By Minkowski's inequality, we have

$$||\nu_{gt1} - \mathbb{E}\nu_{gt1}||_p \leq ||\nu_{gt1}||_p + ||-\mathbb{E}\nu_{gt1}||_p = ||\nu_{gt1}||_p + ||\mathbb{E}\nu_{gt1}||_p = ||\nu_{gt1}||_p + \mathbb{E}\nu_{gt1}$$

Since $\nu_{gt1} \geq 0$ and x^p is convex for $x \geq 0$, by Jensen's inequality we have

$$\begin{aligned}\mathbb{E}[\nu_{gt1}]^p &\leq \mathbb{E}[\nu_{gt1}^p] \\ \mathbb{E}\nu_{gt1} &\leq \mathbb{E}[\nu_{gt1}^p]^{1/p} = ||\nu_{gt1}||_p\end{aligned}$$

Hence,

$$||\nu_{gt1} - \mathbb{E}\nu_{gt1}||_p \leq ||\nu_{gt1}||_p + \mathbb{E}\nu_{gt1} \leq 2||\nu_{gt1}||_p$$

Therefore,

$$V(g,t) = \mathbb{E}\big[|\nu_{gt1} - \mathbb{E}\nu_{gt1}|^2\big] = \Big(||\nu_{gt1} - \mathbb{E}\nu_{gt1}||_2\Big)^2 \leq \Big(2||\nu_{gt1}||_2\Big)^2 = 4\mathbb{E}\big[\nu_{gt1}^2\big]$$

$$X(g,t) = \mathbb{E}\big[|\nu_{gt1} - \mathbb{E}\nu_{gt1}|^3\big] = \Big(||\nu_{gt1} - \mathbb{E}\nu_{gt1}||_3\Big)^3 \leq \Big(2||\nu_{gt1}||_3\Big)^3 = 8\mathbb{E}\big[\nu_{gt1}^3\big]$$

Using Corollary D.11.1, we can set

$$\begin{aligned}\hat{V} &= 4 \cdot \sup_{g\in[0,g_1]} \sup_{t\in[g,T]} \mathbb{E}\big[\nu_{gt1}^2\big] < \infty \\ \hat{X} &= 8 \cdot \sup_{g\in[0,g_1]} \sup_{t\in[g,T]} \mathbb{E}\big[\nu_{gt1}^3\big] < \infty\end{aligned}$$

which completes the proof. □

Lemma D.13. *Let $\varphi(g,t,s) := \mathbb{E}\big[e^{is\nu_{gt1}}\big]$ be the characteristic function of ν_{gt1}. We have*

$$\varphi(g,t,s) = p(0) + p(1) \cdot \frac{e^{is}}{1 - \eta(g,t)e^{is}}$$

and

$$\big|\varphi(g,t,s)\big|^2 = p(0)^2 + \frac{2p(0)p(1)\big(\cos(s) - \eta(g,t)\big) + p(1)^2}{1 - 2\eta(g,t)\cos(s) + \eta(g,t)^2}$$

with $p(k) := \mathbb{P}\big(\nu_{gt1} = k\big)$ and $\eta(g,t) := 1 - \frac{p(1)}{1-p(0)}$.

Proof. For simplicity, let $\eta = \eta(g,t)$ in this proof.

By Lemmas D.5 and D.6 we have for $k \geq 1$

$$p(k) = p(1)\eta^{k-1}$$

We then have

$$\begin{aligned}
\varphi(g,t,s) &= \sum_{k=0}^{\infty} e^{isk} p(k) = p(0) + \sum_{k=1}^{\infty} e^{isk} p(k) \\
&= p(0) + \sum_{k=1}^{\infty} e^{isk} p(1)\eta^{k-1} \\
&= p(0) + p(1)e^{is} \sum_{k=1}^{\infty} e^{is(k-1)} \eta^{k-1} \\
&= p(0) + p(1)e^{is} \sum_{k=0}^{\infty} \left(e^{is}\eta\right)^{k}
\end{aligned}$$

We have $|e^{is}| \leq 1$ and by Lemma D.9 $|\eta| < 1$, so that $|e^{is}\eta| < 1$. We can thus use the formula for convergent geometric series and obtain

$$\varphi(g,t,s) = p(0) + p(1)e^{is} \cdot \frac{1}{1 - \eta e^{is}}$$

Then, using Euler's formula and doing some rearrangements, we obtain

$$\varphi(g,t,s) = p(0) + \frac{p(1)\big(\cos(s) - \eta\big)}{1 - 2\eta\cos(s) + \eta^2} + i\frac{p(1)\sin(s)}{1 - 2\eta\cos(s) + \eta^2}$$

Then we obtain

$$\begin{aligned}
\big|\varphi(g,t,s)\big|^2 &= \left(p(0) + \frac{p(1)\big(\cos(s) - \eta\big)}{1 - 2\eta\cos(s) + \eta^2}\right)^2 + \left(\frac{p(1)\sin(s)}{1 - 2\eta\cos(s) + \eta^2}\right)^2 \\
&= p(0)^2 + \frac{2p(0)p(1)\big(\cos(s) - \eta\big) + p(1)^2}{1 - 2\eta\cos(s) + \eta^2}
\end{aligned}$$

after some rearrangements.

□

Lemma D.14. *For any $l \in (0,\pi]$, $\forall s \in [l,\pi]$*

$$\big|\varphi(g,t,s)\big| \leq \hat{\varphi}_0 + o(1)$$

as $g \to 0$, with $\hat{\varphi}_0 < 1$.

Proof. First we show that

$$\hat{\varphi}_0^2 := \sup_{t\in[0,T]} \sup_{s\in[l,\pi]} \big|\varphi(0,t,s)\big|^2 < 1$$

From Lemma D.13 we have

$$\big|\varphi(0,t,s)\big|^2 = p_0(0)^2 + \frac{2p_0(0)p_0(1)\big(\cos(s)-\eta_0\big) + p_0(1)^2}{1-2\eta_0\cos(s)+\eta_0^2}$$

with (see Lemma D.6)

$$\begin{aligned} p_0(0) &:= \mathbb{P}\big(\nu_{0t1}=0\big) = \frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1} \\ p_0(1) &:= \mathbb{P}\big(\nu_{0t1}=1\big) = \frac{1}{\big(1+\lambda_1(t)g_1\big)^2} \\ \eta_0 = \eta(0,t) &:= 1 - \frac{p_0(1)}{1-p_0(0)} = \frac{\lambda_1(t)g_1}{1+\lambda_1(t)g_1} \end{aligned}$$

We have

$$\frac{\partial}{\partial s}\big|\varphi(0,t,s)\big|^2 = -\sin(s)\cdot\frac{2p_0(1)\big(p_0(0)(1-\eta_0^2)+p_0(1)\eta_0\big)}{\big(1-2\eta_0\cos(s)+\eta_0^2\big)^2}$$

By Lemma D.9, $\forall t\in[0,T]$, $0<\eta_0<1$, which immediately implies that $p_0(1)>0$. These then imply that $\forall s\in(0,\pi)$

$$\frac{\partial}{\partial s}\big|\varphi(0,t,s)\big|^2 < 0$$

so that, $\forall t\in[0,T]$, $\big|\varphi(0,t,s)\big|^2$ strictly decreases with s on $[0,\pi]$, from a maximum at $s=0$, which is easily verified to be 1. We thus have $\forall t\in[0,T]$

$$\sup_{s\in[l,\pi]}\big|\varphi(0,t,s)\big|^2 = \big|\varphi(0,t,l)\big|^2 < \big|\varphi(0,t,0)\big|^2 = 1$$

Given that $\lambda_1(t)$ is continuous on $[0,T]$, then $\big|\varphi(0,t,l)\big|^2$ is continuous on $[0,T]$. Consequently, by the extreme-value theorem, there exists $\tau\in[0,T]$ such that

$$\hat{\varphi}_0^2 := \sup_{t\in[0,T]}\sup_{s\in[l,\pi]}\big|\varphi(0,t,s)\big|^2 = \sup_{t\in[0,T]}\big|\varphi(0,t,l)\big|^2 = \big|\varphi(0,\tau,l)\big|^2 < \big|\varphi(0,\tau,0)\big|^2 = 1$$

Now we proceed to the demonstration of the lemma's claim. We recall that by Lemma D.13

$$\big|\varphi(g,t,s)\big|^2 = p(0)^2 + \frac{2p(0)p(1)\big(\cos(s)-\eta\big)+p(1)^2}{1-2\eta\cos(s)+\eta^2}$$

with $p(r) := \mathbb{P}(\nu_{gt1}=r)$ and $\eta=\eta(g,t) := 1-\frac{p(1)}{1-p(0)}$.

By Lemma D.7 we have that

$$p(0) = p_0(0) + o(1)$$
$$p(1) = p_0(1) + o(1)$$

as $g \to 0$, which induces

$$\eta = 1 - \frac{p_0(1) + o(1)}{1 - p_0(0) + o(1)} = \eta_0 + o(1)$$

as $g \to 0$.

We then obtain, $\forall s \in [l, \pi]$

$$\begin{aligned}
\big|\varphi(g,t,s)\big|^2 &= \big(p_0(0) + o(1)\big)^2 \\
&\quad + \frac{2\big(p_0(0) + o(1)\big)\big(p_0(1) + o(1)\big)\big(\cos(s) - \eta_0 + o(1)\big) + \big(p_0(1) + o(1)\big)^2}{1 - 2\big(\eta_0 + o(1)\big)\cos(s) + \big(\eta_0 + o(1)\big)^2} \\
&= p_0(0)^2 + \frac{2p_0(0)p_0(1)\big(\cos(s) - \eta_0\big) + p_0(1)^2}{1 - 2\eta_0\cos(s) + \eta_0^2} + o(1) \\
&= \big|\varphi(0,t,s)\big|^2 + o(1) \\
&\leq \hat{\varphi}_0^2 + o(1)
\end{aligned}$$

as $g \to 0$.

Therefore, given that $\sqrt{C + x} = \sqrt{C} + O(x)$ as $x \to 0$, we have

$$\big|\varphi(g,t,s)\big| \leq \hat{\varphi}_0 + o(1)$$

as $g \to 0$. □

E Results for the conditional scaled BD process

In this section, we consider the scaled BD process, parametrized with n, $\lambda_1(t)$ and $r(t)$ such that

$$\lambda(t) = \lambda_n(t) = \lambda_1(t)n$$
$$\mu(t) = \mu_n(t) = \lambda_n(t) - r(t)$$

We recall that we have set the conditions that on $[0, T]$: (i) $\lambda_1(t) > 0$, (ii) $\lambda_1(t)$ and $r(t)$ are uniformly continuous (and thus bounded above and below) and (iii) $\lambda_1(t) \geq r(t)$.

We consider this process conditioned on $Z(0) = z_0$ and on $N \in \Omega_n$, with Ω_n defined such that $N \in \Omega_n$ if and only if, for $g_1 : \frac{T}{g_1} \in \mathbb{N}_{>0}$ and with $g := \frac{g_1}{n}$

$$\forall t \in \{0, g, 2g, ..., T\}\ N(t) = y_n(t)$$

where

$$y(t) = y_n(t) := \lceil u(t)n \rceil$$

with $u(t)$ any function of time such that, on $[0, T]$: (i) $u(t)$ is continuous (ii) $u'(t)$ exists and is bounded, (iii) $u(t) > 0$ and (iv) $u(T) = 1$.

We notice that, given that $u(t)$ is bounded on $[0, T]$, then $y(t) = O(n)$ as $n \to \infty$, or equivalently $y(t) = O(1/g)$ as $g \to 0$.

E.1 Pairwise coalescent probability

We denote $c_n(t)$ the "pairwise coalescent probability", that is the probability that two ancestors at time $t-g$ have a common ancestor at time t, under the condition $N \in \Omega_n$:

$$c_n(t) := \mathbb{P}\big(Z(t) = 1 | Z(t-g) = 2, N \in \Omega_n\big)$$

for $t \in \{g, 2g, \ldots, T\}$.

In the time interval $[t-g, t]$, the conditional scaled BD is represented by the random vector $\nu_{gt} = (\nu_{gti})_i,\ i = 1, \ldots, y(t)$, of fixed sum $\sum_{i=1}^{y(t)} \nu_{gti} = N(t-g) = y(t-g)$. (Hence the condition $N \in \Omega_n$ reduces to the condition $N(t-g) = y(t-g) \wedge N(t) = y(t)$ in the interval $[t-g, t]$.)

We note two properties of the conditional scaled BD:

1. Because of the Markov property of the BD process, and the fact that individuals are exchangeable, ν_{gt} is conditionally independent of ν_{gs} for any $s \geq t+g$, given $N(t) = y(t)$.
2. Because the individuals are exchangeable, then, given ν_{gt}, any assignment of the $y(t-g)$ offspring to the $y(t)$ parents are equally likely (provided these assignments are compatible with ν_{gt}).

These two properties are those assumed by Möhle in (19), allowing us to apply Möhle's results to the conditional scaled BD. In particular, Möhle gives the expression for $c_n(t)$ (denoted c_r in his paper):

$$\begin{aligned} & c_n(t) \\ &= \frac{1}{y(t-g)(y(t-g)-1)} \sum_{i=1}^{y(t)} \mathbb{E}\Big[\nu_{gti}(\nu_{gti}-1) | N(t-g) = y(t-g), N(t) = y(t)\Big] \\ &= \frac{1}{y(t-g)(y(t-g)-1)} \sum_{i=1}^{y(t)} \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1) | N(t-g) = y(t-g), N(t) = y(t)\Big] \\ &= \frac{y(t)}{y(t-g)(y(t-g)-1)} \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1) | N(t-g) = y(t-g), N(t) = y(t)\Big] \end{aligned} \tag{15}$$

where we have used the fact that $\forall i$

$$\mathbb{E}\Big[\nu_{gti}(\nu_{gti}-1)|N(t-g)=y(t-g),N(t)=y(t)\Big] \\ =\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)|N(t-g)=y(t-g),N(t)=y(t)\Big]$$

since individuals are exchangeable.

Notice that, since the distribution of ν_{gt1}, $y(t)$ and $y(t-g)$ are all defined continuously for $t\in[g,T]$, $c_n(t)$ is also actually defined continuously on $[g,T]$. Thus, for simplicity, we shall consider below that the function $c_n(t)$ has a continuous domain, although ultimately we are only interested in the discrete domain $t\in\{g,2g,\ldots,T\}$.

E.2 Limit as $n\to\infty$ of the discrete-time pairwise coalescent rate of the conditional scaled BD

We define the "discrete-time pairwise coalescent rate" as

$$h_n(t):=\frac{c_n(t)}{g}=c_n(t)\frac{n}{g_1}$$

which measures the (average) density of a coalescence event over $[t-g,t]$.

The purpose of this section is to derive the "limiting coalescent rate", defined as

$$h(t):=\lim_{n\to\infty}h_n(t)$$

which will be central for showing that the conditional scaled BD converges to the Kingman coalescent.

Specifically, we will show that (Lemma E.8 below)

$$h(t)=\frac{2\lambda_1(t)}{u(t)}$$

and that

$$\sup_{t\in[g,T]}|h_n(t)-h(t)|\xrightarrow[n\to\infty]{}0$$

E.2.1 Preliminary results

In this section we derive some results necessary for studying the limit of $h_n(t)$.

Lemma E.1.

$$\Big|\sqrt{V(g,t)y(t)}\cdot\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big) - \phi\Big(\frac{y(t-g)-y(t)A(g,t)}{\sqrt{V(g,t)y(t)}}\Big)\Big| = O(\sqrt{g})$$

$$\Big|\sqrt{V(g,t)\big(y(t)-1\big)}\cdot\mathbb{P}\big(N(t-g)=y(t-g)-r|N(t)=y(t)-1\big) - \phi\Big(\frac{y(t-g)-r-\big(y(t)-1\big)A(g,t)}{\sqrt{V(g,t)\big(y(t)-1\big)}}\Big)\Big| = O(\sqrt{g})$$

as $g\to 0$.

Proof. Obviously,

$$\Big|\sqrt{V(g,t)y(t)}\cdot\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big) - \phi\Big(\frac{y(t-g)-y(t)A(g,t)}{\sqrt{V(g,t)y(t)}}\Big)\Big| \le \sup_k\Big|\sqrt{V(g,t)y(t)}\cdot\mathbb{P}\big(N(t-g)=k|N(t)=y(t)\big) - \phi\Big(\frac{k-y(t)A(g,t)}{\sqrt{V(g,t)y(t)}}\Big)\Big|$$

We will show that the RHS of this inequality is $O(\sqrt{g})$ as $g\to 0$, hence proving the first claim of the present lemma.

By Doney's Lemma 3 in (51),

$$\sup_k\Big|\mathbb{P}\big(N(t-g)=k|N(t)=y(t)\big) - \frac{1}{\sqrt{V(g,t)y(t)}}\phi\Big(\frac{k-y(t)A(g,t)}{\sqrt{V(g,t)y(t)}}\Big)\Big| \le \frac{CX(g,t)}{y(t)V(g,t)^2} + d(g,t)$$

where C is an absolute constant, where $X(g,t):=\mathbb{E}\big[|\nu_{gt1}-\mathbb{E}\nu_{gt1}|^3\big]$ and where

$$d(g,t):=2\int_{l(g,t)}^{\pi}\exp\big(-y(t)(1-|\varphi(g,t,s)|)\big)ds$$

with $l(g,t)=\frac{V(g,t)}{4X(g,t)}$ and with $\varphi(g,t,s)$ the characteristic function of ν_{gt1}.

Multiplying by $y(t)\sqrt{V(g,t)}$, we obtain

$$\sqrt{y(t)}\cdot\sup_k\Big|\sqrt{V(g,t)y(t)}\cdot\mathbb{P}\big(N(t-g)=k|N(t)=y(t)\big) - \phi\Big(\frac{k-y(t)A(g,t)}{\sqrt{V(g,t)y(t)}}\Big)\Big| \le \frac{CX(g,t)}{V(g,t)^{3/2}} + y(t)\sqrt{V(g,t)}\cdot d(g,t) \le \frac{C\hat{X}}{\check{V}^{3/2}} + y(t)\sqrt{V(g,t)}\cdot d(g,t) \quad (16)$$

with (see Lemmas D.10 and D.12)

$$0 < \check{V} \leq \inf_{g\in[0,g_1]} \inf_{t\in[g,T]} V(g,t)$$
$$\sup_{g\in[0,g_1]} \sup_{t\in[g,T]} X(g,t) \leq \hat{X} < \infty$$

Given that $\sqrt{y(t)} = O(1/\sqrt{g})$, we need to show that the two terms of the RHS of Eq. (16) are bounded above by a function that is $O(1)$ as $g \to 0$ to prove the lemma's first claim. This is the case for the first term, since it is an absolute constant. We will now proceed to show that this is also the case for the second term.

We note that

$$l(g,t) \geq \frac{\check{V}}{4\hat{X}} =: l > 0$$

We thus have

$$y(t)\sqrt{V(g,t)} \cdot d(g,t) \leq 2y(t)\sqrt{V(g,t)} \int_l^\pi \exp\big(-y(t)(1-|\varphi(g,t,s)|)\big)ds$$

given that the integrand is non-negative.

If $l > \pi$, the integral is non-positive, whatever the values of g or t, and therefore

$$y(t)\sqrt{V(g,t)} \cdot d(g,t) \leq 0 = O(1)$$

as $g \to 0$.

If $l \leq \pi$, by Lemma D.14, we have that $|\varphi(g,t,s)| \leq \hat{\varphi}_0 + o(1)$ as $g \to 0$, with $\hat{\varphi}_0 < 1$, provided $s \in [l,\pi]$. Hence,

$$\begin{aligned} y(t)\sqrt{V(g,t)} \cdot d(g,t) &\leq 2y(t)\sqrt{\hat{V}} \int_l^\pi \exp\Big(-y(t)\big(1-\hat{\varphi}_0+o(1)\big)\Big)ds \\ &= 2y(t)\sqrt{\hat{V}}(\pi-l)\exp\big(o(1)\big)\exp\big(-y(t)(1-\hat{\varphi}_0)\big) \\ &= 2y(t)\sqrt{\hat{V}}(\pi-l)\big(1+o(1)\big)\exp\big(-y(t)(1-\hat{\varphi}_0)\big) \\ &= O(1)y(t)\exp\big(-y(t)(1-\hat{\varphi}_0)\big) \end{aligned}$$

as $g \to 0$, with (see Lemma D.12)

$$\sup_{g\in[0,g_1]} \sup_{t\in[g,T]} V(g,t) \leq \hat{V} < \infty$$

It is then straightforward to verify that, given that $1-\hat{\varphi}_0 > 0$,

$$y(t)\exp\big(-y(t)(1-\hat{\varphi}_0)\big) \leq \frac{\exp(-1)}{1-\hat{\varphi}_0} = O(1)$$

as $g \to 0$.

The second claim of the lemma is proven in exactly the same way. □

Lemma E.2.

$$\begin{aligned} & y(t-g) - A(g,t)y(t) = O(1) \\ \forall m \quad & y(t-g) - m - A(g,t)\big(y(t)-1\big) = O_m(1) \end{aligned}$$

as $g \to 0$.

Proof. We recall that

$$y(t) := \left\lceil \frac{u(t)g_1}{g} \right\rceil$$

with $u(t) > 0$ bounded on $[0,T]$ and with $u'(t)$ bounded on $[0,T]$.

First we show that $y(t-g) - y(t) = O(1)$ as $g \to 0$.

We denote

$$\epsilon_{t,g} := y(t) - \frac{u(t)g_1}{g} = \left\lceil \frac{u(t)g_1}{g} \right\rceil - \frac{u(t)g_1}{g}$$

$$\epsilon_{t-g,g} := y(t-g) - \frac{u(t-g)g_1}{g} = \left\lceil \frac{u(t-g)g_1}{g} \right\rceil - \frac{u(t-g)g_1}{g}$$

and note that $\epsilon_{t,g}, \epsilon_{t-g,g} \in [0,1)$.

We then have

$$\begin{aligned} |y(t-g) - y(t)| &= \left| \frac{u(t-g)g_1}{g} - \frac{u(t)g_1}{g} + \epsilon_{t-g,g} - \epsilon_{t,g} \right| \\ &\le \left| \frac{u(t-g) - u(t)}{g} \right| g_1 + |\epsilon_{t-g,g} - \epsilon_{t,g}| \end{aligned}$$

Then, because $u'(t)$ is bounded, there exists $C > 0$ such that for all $g \in (0, g_1]$ and $t \in [g,T]$

$$\left| \frac{u(t-g) - u(t)}{g} \right| \le C$$

Also, $|\epsilon_{t-g,g} - \epsilon_{t,g}| < 1$, so that for all $g \in (0,g_1]$ and $t \in [g,T]$

$$|y(t-g) - y(t)| \le Cg_1 + 1$$

Hence,

$$y(t-g) - y(t) = O(1)$$

as $g \to 0$.

Now we show that $y(t-g) - y(t)A(g,t) = O(1)$ as $g \to 0$.

Using Lemma D.8

$$\begin{aligned} y(t-g) - y(t)A(g,t) &= y(t-g) - y(t) + y(t)\big(1 - A(g,t)\big) \\ &= O(1) + y(t)O(g) \end{aligned}$$

as $g \to 0$. Then, given that $u(t)$ is bounded on $[0,T]$, $y(t) = O(1/g)$ as $g \to 0$, yielding

$$\begin{aligned} y(t-g) - y(t)A(g,t) &= O(1) + O(1/g)O(g) \\ &= O(1) \end{aligned}$$

as $g \to 0$.

Similarly,

$$\begin{aligned} y(t-g) - m - A(g,t)\big(y(t) - 1\big) &= y(t-g) - y(t)A(g,t) + A(g,t) - m \\ &= O(1) + 1 + O(g) - m \\ &= O_m(1) \end{aligned}$$

as $g \to 0$.

□

Lemma E.3. $\forall m \in \mathbb{N}$

$$\frac{\mathbb{P}\big(N(t-g) = y(t-g) - m | N(t) = y(t) - 1\big)}{\mathbb{P}\big(N(t-g) = y(t-g) | N(t) = y(t)\big)} = 1 + O_m\big(\sqrt{g}\big)$$

as $g \to 0$.

Proof. In this proof, when we use the $O(\cdot)$ and $o(\cdot)$ notation, "as $g \to 0$" is implicit.

For conciseness we write

$$P(i,j) := \mathbb{P}\big(N(t-g) = j | N(t) = i\big)$$

We have

$$\begin{aligned} &\frac{P\big(y(t)-1, y(t-g)-m\big)}{P\big(y(t), y(t-g)\big)} \\ = \frac{\sqrt{y(t)}}{\sqrt{y(t)-1}} &\cdot \frac{\sqrt{(y(t)-1)V(g,t)} \cdot P\big(y(t)-1, y(t-g)-m\big)}{\sqrt{y(t)V(g,t)} \cdot P\big(y(t), y(t-g)\big)} \end{aligned}$$

We have

$$\frac{y(t)}{y(t)-1} = \frac{u(t)\frac{g_1}{g} + \epsilon_{t,g}}{u(t)\frac{g_1}{g} + \epsilon_{t,g} - 1} = \frac{1 + g\frac{\epsilon_{t,g}}{u(t)g_1}}{1 + g\frac{\epsilon_{t,g}-1}{u(t)g_1}}$$

Given that $u(t)$ is continuous and > 0 on $[0,T]$, $\inf_{t\in[0,T]} u(t) > 0$. Furthermore, $\epsilon_{t,g} \in [0,1)$. We thus have

$$\frac{y(t)}{y(t)-1} = \frac{1+O(g)}{1+O(g)} = 1+O(g)$$

Then, given that $\sqrt{1+x} = 1+O(x)$ as $x \to 0$, we have

$$\sqrt{\frac{y(t)}{y(t)-1}} = 1+O(g) \tag{17}$$

Then, using Lemma E.1, we have

$$\begin{aligned}
&\sqrt{(y(t)-1)V(g,t)}\cdot P\big(y(t)-1, y(t-g)-m\big)\\
&= \sqrt{(y(t)-1)V(g,t)}\cdot P\big(y(t)-1, y(t-g)-m\big) - \phi\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)\\
&\quad + \phi\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)\\
&= O(\sqrt{g}) + \phi\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)
\end{aligned}$$

By Lemma D.10 we have that $\forall g \in [0,g_1]$, $\forall t \in [g,T]$, $V(g,t) \geq \check{V} > 0$, by Lemma E.2 we have that $y(t-g)-m-(y(t)-1)A(g,t) = O_m(1)$, and also $y(t) = O(1/g)$ and $1/\big(y(t)-1\big) = O(g)$. Therefore

$$\begin{aligned}
\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)^2 &\leq \frac{1}{\check{V}\big(y(t)-1\big)}\Big(y(t-g)-m-(y(t)-1)A(g,t)\Big)^2\\
&= \frac{1}{\check{V}}O(g)O_m(1) = O_m(g)
\end{aligned}$$

Then,

$$\begin{aligned}
&\phi\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)\\
&= \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\Big(\frac{y(t-g)-m-(y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\Big)^2\Big)\\
&= \frac{1}{\sqrt{2\pi}}\exp\big(O_m(g)\big) = \frac{1}{\sqrt{2\pi}}\big(1+O_m(g)\big) = \frac{1}{\sqrt{2\pi}} + O_m(g)
\end{aligned}$$

This yields

$$\begin{aligned}
\sqrt{(y(t)-1)V(g,t)}\cdot P\big(y(t)-1, y(t-g)-m\big) &= O(\sqrt{g}) + \frac{1}{\sqrt{2\pi}} + O_m(g)\\
&= \frac{1}{\sqrt{2\pi}} + O_m(\sqrt{g})
\end{aligned}$$

Noting that by Lemma E.2 $y(t-g) - A(g,t)y(t) = O(1)$, by the same rationale we obtain

$$\sqrt{y(t)V(g,t)} \cdot P\big(y(t), y(t-g)\big) = \frac{1}{\sqrt{2\pi}} + O(\sqrt{g}) \tag{18}$$

Finally, we get

$$\frac{P\big(y(t)-1, y(t-g)-m\big)}{P\big(y(t), y(t-g)\big)} = \big(1+O(g)\big) \cdot \frac{\frac{1}{\sqrt{2\pi}} + O_m(\sqrt{g})}{\frac{1}{\sqrt{2\pi}} + O(\sqrt{g})} = 1 + O_m(\sqrt{g})$$

□

Lemma E.4.

$$\frac{\mathbb{P}\big(N(t-g) = y(t-g) - m | N(t) = y(t) - 1\big)}{\mathbb{P}\big(N(t-g) = y(t-g) | N(t) = y(t)\big)} = O(1)$$

as $g \to 0$.

Proof. For conciseness we write

$$P(i,j) := \mathbb{P}\big(N(t-g) = j | N(t) = i\big)$$

We have

$$\phi\left(\frac{y(t-g) - m - (y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\right) \leq \phi(0) = O(1)$$

as $g \to 0$.

Using this result and Lemma E.1 yields

$$\begin{aligned}
&\sqrt{(y(t)-1)V(g,t)} \cdot P\big(y(t)-1, y(t-g)-m\big) \\
&= \sqrt{(y(t)-1)V(g,t)} \cdot P\big(y(t)-1, y(t-g)-m\big) - \phi\left(\frac{y(t-g) - m - (y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\right) \\
&\quad + \phi\left(\frac{y(t-g) - m - (y(t)-1)A(g,t)}{\sqrt{(y(t)-1)V(g,t)}}\right) \\
&= O(\sqrt{g}) + O(1) \\
&= O(1)
\end{aligned}$$

as $g \to 0$.

Then, using Eqs. (17,18), we obtain

$$\begin{aligned}
&\frac{P\big(y(t)-1, y(t-g)-m\big)}{P\big(y(t), y(t-g)\big)} \\
&= \frac{\sqrt{y(t)}}{\sqrt{y(t)-1}} \cdot \frac{\sqrt{(y(t)-1)V(g,t)} \cdot P\big(y(t)-1, y(t-g)-m\big)}{\sqrt{y(t)V(g,t)} \cdot P\big(y(t), y(t-g)\big)} \\
&= \big(1+O(g)\big) \cdot \frac{O(1)}{\frac{1}{\sqrt{2\pi}} + O(\sqrt{g})} \\
&= \big(1+O(g)\big) \cdot O(1) \\
&= O(1)
\end{aligned}$$

as $g \to 0$.

□

Theorem E.5.

$$\sup_{t\in[g,T]} \Big| \mathbb{E}\big[\nu_{gt1}(\nu_{gt1}-1) | N(t-g) = y(t-g), N(t) = y(t)\big] - \mathbb{E}\big[\nu_{0t1}(\nu_{0t1}-1)\big] \Big| \xrightarrow[g\to 0]{} 0$$

Proof. To simplify notation, we write

$$\begin{aligned}
P_1(m) &:= \mathbb{P}\big(N(t-g) = y(t-g) - m | N(t) = y(t) - 1\big) \\
P_2 &:= \mathbb{P}\big(N(t-g) = y(t-g) | N(t) = y(t)\big) \\
p(m) &:= \mathbb{P}(\nu_{gt1} = m) \\
p_0(m) &:= \mathbb{P}(\nu_{0t1} = m)
\end{aligned}$$

and we write $\sup_t$ for $\sup_{t\in[g,T]}$ and $\sup_g$ for $\sup_{g\in(0,g_1]}$, unless specified otherwise.

We have

$$\begin{aligned}
&\mathbb{E}\big[\nu_{gt1}(\nu_{gt1}-1) | N(t-g) = y(t-g), N(t) = y(t)\big] \\
&\qquad = \sum_{m=0}^{y(t-g)} m(m-1)p(m)\frac{P_1(m)}{P_2} \\
&\qquad = \sum_{m=0}^{\infty} m(m-1)p(m)\frac{P_1(m)}{P_2}
\end{aligned}$$

where the last line is obtained after noticing that $\forall m > y(t-g),\ P_1(m) = 0$.

Furthermore,

$$\mathbb{E}\big[\nu_{0t1}(\nu_{0t1}-1)\big] = \sum_{m=0}^{\infty} m(m-1)p_0(m)$$

Hence, the present theorem's statement is equivalent to

$$\lim_{g\to 0}\sup_t \left| \sum_{m=0}^{\infty} m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right| = 0$$

First we note that

$$\begin{aligned}
&\lim_{g\to 0}\sup_t \left| \sum_{m=0}^{\infty} m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right| \\
&\leq \lim_{g\to 0} \sum_{m=0}^{\infty} \sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right|
\end{aligned}$$

Using Lemmas D.11 and E.4, we have

$$\begin{aligned}
&\sup_{g,t} \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right| \\
&\leq \sup_{g,t} \left| m(m-1)p(m)\frac{P_1(m)}{P_2}\right| + \sup_{g,t}\left| m(m-1)p_0(m)\right| \\
&\leq \sup_{g,t} \left| m^2 p(m)\frac{P_1(m)}{P_2}\right| + \sup_{g,t}\left| m^2 p_0(m)\right| \\
&\leq M_{2m}\sup_{g,t}\left|\frac{P_1(m)}{P_2}\right| + M_{2m} \\
&= M_{2m}O(1)
\end{aligned}$$

By the definition of $O(1)$, there exist $C>0$ and $G\in(0,g_1]$ such that

$$\sup_{g\in(0,G]}\sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right| \leq M_{2m}C$$

Furthermore, by Lemma D.11 we have

$$\sum_{m=0}^{\infty} M_{2m}C = C\sum_{m=0}^{\infty} M_{2m} < \infty$$

The two last equations together imply, by the Weierstrass M-test, that the series

$$\sum_{m=0}^{\infty}\sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m)\right|$$

converges uniformly for $g \in (0, G]$. By the Moore-Osgood theorem, this allows us to swap the limits and write

$$\begin{aligned}\lim_{g\to 0} \sum_{m=0}^{\infty} \sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m) \right| \\ = \sum_{m=0}^{\infty} \lim_{g\to 0} \sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m) \right|\end{aligned}$$

Now we will show that each term of this last sum is equal to 0, hence proving the theorem.

We recall that, by Lemma D.7,

$$p(m) = p_0(m) + o_m(1)$$

as $g \to 0$.

We recall that, by Lemma E.3,

$$\frac{P_1(m)}{P_2} = 1 + O_m(\sqrt{g})$$

as $g \to 0$.

We thus have

$$\begin{aligned}m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m) \\ &= m(m-1)\left(p(m)\frac{P_1(m)}{P_2} - p_0(m) \right) \\ &= m(m-1)\left(\big(p_0(m) + o_m(1)\big)\big(1 + O_m(\sqrt{g})\big) - p_0(m) \right) \\ &= m(m-1)o_m(1)\end{aligned}$$

as $g \to 0$.

The last line is equivalent to

$$\lim_{g\to 0} \sup_t \left| m(m-1)p(m)\frac{P_1(m)}{P_2} - m(m-1)p_0(m) \right| = 0$$

for any fixed m. □

E.2.2 Actual limit

This section gives the limit of $h_n(t)$ as $n \to \infty$, in the form of Lemma E.8.

We define

$$U_n(t) := \frac{y(t)}{y(t-g)(y(t-g)-1)} \cdot \frac{n}{g_1}$$
$$E_n(t) := \mathbb{E}\big[\nu_{gt1}(\nu_{gt1}-1) | N(t-g) = y(t-g), N(t) = y(t)\big]$$

so that

$$h_n(t) = U_n(t)E_n(t)$$

Lemma E.6.

$$U_n(t) = \frac{1}{u(t)g_1} + O(1/n)$$

as $n \to \infty$.

Proof. Denote

$$\epsilon_{t,g} := y(t) - u(t)n$$
$$\epsilon_{t-g,g} := y(t-g) - u(t-g)n$$

and note that $\epsilon_{t,g}, \epsilon_{t-g,g} \in [0, 1)$.

Recalling that $u(t)$ is bounded on $[0, T]$, we have

$$\begin{aligned}
U_n(t) &:= \frac{y(t)}{y(t-g)\big(y(t-g)-1\big)} \cdot \frac{n}{g_1} \\
&= \frac{u(t)n + \epsilon_{t,g}}{\big(u(t-g)n + \epsilon_{t-g,g}\big)\big(u(t-g)n + \epsilon_{t-g,g} - 1\big)} \cdot \frac{n}{g_1} \\
&= \frac{u(t) + \epsilon_{t,g}/n}{\big(u(t-g) + \epsilon_{t-g,g}/n\big)\big(u(t-g) + (\epsilon_{t-g,g} - 1)/n\big)} \cdot \frac{1}{g_1} \\
&= \frac{u(t) + O(1/n)}{\big(u(t-g) + O(1/n)\big)\big(u(t-g) + O(1/n)\big)} \cdot \frac{1}{g_1} \\
&= \frac{u(t) + O(1/n)}{u(t-g)^2 + O(1/n)} \cdot \frac{1}{g_1} \\
&= \frac{u(t)}{u(t-g)^2} \cdot \frac{1}{g_1} + O(1/n) \\
&= \left(\frac{u(t) - u(t-g)}{u(t-g)} + 1\right)^2 \cdot \frac{1}{g_1 u(t)} + O(1/n)
\end{aligned}$$

Also, because $u'(t)$ is bounded on $[0, T]$, then $u(t) - u(t-g) = O(g)$ as $g \to 0$, and equivalently $u(t) - u(t-g) = O(1/n)$ as $n \to \infty$. Thus, because $u(t)$ is bounded and > 0 on $[0, T]$, we have

$$\frac{u(t) - u(t-g)}{u(t-g)} = O(1/n)$$

as $n \to \infty$. Hence,

$$\begin{aligned} U_n(t) &= \big(O(1/n) + 1\big)^2 \cdot \frac{1}{g_1 u(t)} + O(1/n) \\ &= \frac{1}{g_1 u(t)} + O(1/n) \end{aligned}$$

as $n \to \infty$. □

Lemma E.7.

$$E_n(t) = 2\lambda_1(t) g_1 + o(1)$$

as $n \to \infty$.

Proof. By Theorem E.5, we have that

$$\begin{aligned} &\sup_{t \in [g,T]} \Big| \mathbb{E}\big[\nu_{gt1}(\nu_{gt1} - 1) | N(t-g) = y(t-g), N(t) = y(t)\big] - \mathbb{E}\big[\nu_{0t1}(\nu_{0t1} - 1)\big] \Big| \\ &\qquad =: \sup_{t \in [g,T]} \Big| E_n(t) - \mathbb{E}\big[\nu_{0t1}(\nu_{0t1} - 1)\big] \Big| \xrightarrow[n \to \infty]{} 0 \end{aligned}$$

(since $n \to \infty$ is equivalent to $g \to 0$), which writes equivalently

$$E_n(t) = \mathbb{E}\big[\nu_{0t1}(\nu_{0t1} - 1)\big] + o(1)$$

as $n \to \infty$. Using Lemma D.6, we obtain

$$E_n(t) = \mathrm{Var}[\nu_{0t1}] + \mathbb{E}\nu_{0t1}(\mathbb{E}\nu_{0t1} - 1) + o(1) = 2\lambda_1(t) g_1 + o(1)$$

as $n \to \infty$. □

Lemma E.8.

$$h_n(t) = h(t) + o(1)$$

as $n \to \infty$, *with*

$$h(t) := \lim_{n \to \infty} h_n(t) = \frac{2\lambda_1(t)}{u(t)}$$

Proof. Using the facts that $u(t)$ is bounded and > 0 and that $\lambda_1(t)$ is bounded on $[0, T]$, and using Lemmas E.6 and E.7, we have

$$\begin{aligned} h_n(t) &= U_n(t) E_n(t) \\ &= \left(\frac{1}{g_1 u(t)} + O(1/n) \right) \big(2\lambda_1(t) g_1 + o(1) \big) \\ &= \frac{2\lambda_1(t)}{u(t)} + o(1) \end{aligned}$$

□

Corollary E.8.1.

$$c_n(t) = O(1/n)$$

as $n \to \infty$.

Proof. We recall that

$$h_n(t) := \frac{c_n(t)}{g} = c_n(t)\frac{n}{g_1}$$

Hence by Lemma E.8 we have

$$\begin{aligned} h_n(t) &= h(t) + o(1) \\ c_n(t)\frac{n}{g_1} &= h(t) + o(1) \\ c_n(t) &= \frac{g_1}{n}h(t) + o(1/n) \end{aligned}$$

Given that $\lambda_1(t)$ is bounded on $[0, T]$ and that $u(t)$ is bounded and > 0 on $[0, T]$, then

$$h(t) = \frac{2\lambda_1(t)}{u(t)}$$

is bounded on $[0, T]$. Therefore,

$$c_n(t) = O(1/n)$$

□

F Convergence of the sample process of the conditional scaled BD to the Kingman coalescent

Theorem F.1 below constitutes the main result of this paper, about the convergence of the conditional scaled BD to the Kingman coalescent. The proof to Theorem F.1 relies on lemmas and theorems derived in earlier sections of this appendix. The proofs to these previous lemmas and theorems are valid conditionally on a series of assumptions, which we recall here as sufficient conditions for the applicability of Theorem F.1.

Sufficient conditions for the applicability of Theorem F.1. We recall that the scaled BD process is a BD process starting at time T with $N(T) = n$ individuals and parametrized with $\lambda_1(t)$ and $r(t)$ such that

$$\begin{aligned} \lambda(t) &= \lambda_n(t) = \lambda_1(t)n \\ \mu(t) &= \mu_n(t) = \lambda_n(t) - r(t) \end{aligned}$$

The conditional scaled BD process is the scaled BD process conditioned on $Z(0) = z_0$ and on $N \in \Omega_n$, with Ω_n a set of population trajectories defined such that $N \in \Omega_n$ if and only if, for $g_1 : \frac{T}{g_1} \in \mathbb{N}_{>0}$ and with $g := \frac{g_1}{n}$,

$$\forall t \in \{0, g, 2g, ..., T\}\ N(t) = y_n(t)$$

where

$$y(t) = y_n(t) := \lceil u(t)n \rceil$$

The following conditions on $\lambda_1(t)$, $r(t)$ and $u(t)$ have been assumed throughout earlier sections of this appendix and constitute sufficient conditions for the applicability of Theorem F.1:

- $\forall t \in [0, T],\ \lambda_1(t) > 0$
- $\lambda_1(t)$ is uniformly continuous on $[0, T]$
- $r(t)$ is uniformly continuous on $[0, T]$
- $\forall t \in [0, T],\ \lambda_1(t) \geq r(t)$
- $u(T) = 1$
- $\forall t \in [0, T],\ u(t) > 0$
- $u(t)$ is continuous on $[0, T]$
- $\forall t \in [0, T],\ u'(t)$ exists
- $u'(t)$ is bounded on $[0, T]$

Introduction to Theorem F.1. Consider the discrete-time ancestral process $\big(\mathcal{R}(\lceil t/g \rceil)\big)_{t\in[0,T]}$ of the conditional scaled BD. Because of the condition $Z(0) = z_0$,

$$\mathcal{R}(0) = \big\{(i, i)|i \in \{1, \ldots, z_0\}\big\} =: \Delta$$

Let $\big(\mathcal{K}(t)\big)_{t\in[0,T]}$ denote the ancestral process of the Kingman coalescent with $\theta(t) = \frac{u(t)}{2\lambda_1(t)}$, restricted to $t \in [0, T]$, with initial state $\mathcal{K}(0) = \Delta$.

Theorem F.1. *$\big(\mathcal{R}(\lceil t/g \rceil)\big)_{t\in[0,T]}$ converges to $\big(\mathcal{K}(t)\big)_{t\in[0,T]}$ as $n \to \infty$ in terms of finite-dimensional distributions.*

Proof. Define

$$\Lambda(t) = \int_0^t \frac{2\lambda_1(s)}{u(s)} ds$$

Notice that, since the integrand is strictly positive, Λ strictly increases, and its inverse Λ^{-1} exists and strictly increases.

Consider the time scaling $x = \Lambda(t)$. We have

$$\big(\mathcal{R}(\lceil t/g \rceil)\big)_{t\in[0,T]} = \big(\mathcal{R}(\lceil \Lambda^{-1}(x)/g \rceil)\big)_{x\in[0,\Lambda(T)]}$$

and

$$\big(\mathcal{K}(t)\big)_{t\in[0,T]} = \big(\mathcal{K}(\Lambda^{-1}(x))\big)_{x\in[0,\Lambda(T)]} = \big(\mathcal{K}^1(x)\big)_{x\in[0,\Lambda(T)]}$$

where $\big(K^1(x)\big)_x$ denotes the ancestral process of the Kingman coalescent with $\theta(x) = 1$.

By Möhle's Theorem 1 in (19), if the conditions stated below are verified, then $\big(\mathcal{R}(\lceil \Lambda^{-1}(x)/g \rceil)\big)_{x\in[0,\Lambda(T)]}$ converges to $\big(\mathcal{K}^1(x)\big)_{x\in[0,\Lambda(T)]}$ as $n \to \infty$ in terms of finite-dimensional distributions, which proves the present theorem.

We thus need to prove that the conditions of application of Möhle's theorem are satisfied by the conditional scaled BD. These conditions read as follows in our notation.

- **Condition 1** For all $x \in [0, \Lambda(T)]$

$$\lim_{n\to\infty} \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} c_n(mg) = x$$

- **Condition 2** For all $x \in [0, \Lambda(T)]$

$$\lim_{n\to\infty} \sup_{m \le \lceil \Lambda^{-1}(x)/g \rceil} c_n(mg) = 0$$

  which is implied in our case by the following stronger condition (given that $c_n(t)$ is defined continuously on $[g, T]$):

$$\lim_{n\to\infty} \sup_{t\in[g,T]} c_n(t) = 0$$

- **Condition 3** For all $x \in [0, \Lambda(T)]$ and for all $k \in \mathbb{N}_{>0}$

$$0 = \lim_{n\to\infty} \sup_{m \le \lceil \Lambda^{-1}(x)/g \rceil} \Bigg(\frac{1}{y\big((m-1)g\big)^3 c_n(mg)} \times \sum_{i=1}^{y(mg)} \mathbb{E}\Big[\nu_{g,mg,i}(\nu_{g,mg,i} - 1)\nu_{g,mg,i}^k | N\big((m-1)g\big) = y\big((m-1)g\big), N\big(mg\big) = y\big(mg\big)\Big] \Bigg)$$

  which is implied in our case by the following stronger condition. For all $k \in \mathbb{N}_{>0}$:

$$0 = \lim_{n\to\infty} \sup_{t\in[g,T]} \Bigg(\frac{1}{y(t-g)^3 c_n(t)} \times \sum_{i=1}^{y(t)} \mathbb{E}\Big[\nu_{gti}(\nu_{gti} - 1)\nu_{gti}^k | N(t-g) = y(t-g), N(t) = y(t)\Big] \Bigg)$$

- **Condition 4** For all $x \in [0, \Lambda(T)]$

$$0 = \lim_{n\to\infty} \sup_{m \le \lceil \Lambda^{-1}(x)/g \rceil} \Bigg(\frac{1}{y\big((m-1)g\big)^4 c_n(mg)} \times \sum_{i,j=1}^{y(mg)} \mathbb{E}\Big[\nu_{g,mg,i}(\nu_{g,mg,i}-1)\nu^2_{g,mg,j} | N\big((m-1)g\big) = y\big((m-1)g\big), N\big(mg\big) = y\big(mg\big)\Big]\Bigg)$$

which is implied in our case by the following stronger condition:

$$0 = \lim_{n\to\infty} \sup_{t \in [g,T]} \Bigg(\frac{1}{y(t-g)^4 c_n(t)} \times \sum_{i,j=1}^{y(t)} \mathbb{E}\Big[\nu_{gti}(\nu_{gti}-1)\nu^2_{gtj} | N(t-g) = y(t-g), N(t) = y(t)\Big]\Bigg)$$

Proof of Condition 1.

Notice that, since $\Lambda^{-1}(x)$ strictly increases, for $x \in [0, \Lambda(T)]$, $\Lambda^{-1}(x) \le \Lambda^{-1}(\Lambda(T)) = T$ is bounded.

By Lemma E.8, with $h(t) = \frac{2\lambda_1(t)}{u(t)}$, we have

$$h_n(t) = h(t) + o(1)$$

as $n \to \infty$ (or, equivalently, as $g \to 0$).

Then, recalling that $h_n(t) := c_n(t)/g$, we have for all $x \in [0, \Lambda(T)]$

$$\begin{aligned}
\sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} c_n(mg) &= \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} h_n(mg)g \\
&= \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} \big(h(mg) + o(1)\big)g \quad (\text{as } g \to 0) \\
&= \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} h(mg)g + \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} o(1)g \\
&= \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} h(mg)g + \lceil \Lambda^{-1}(x)/g \rceil o(1)g \\
&= \sum_{m=1}^{\lceil \Lambda^{-1}(x)/g \rceil} h(mg)g + o(1) \quad (\text{because } \Lambda^{-1}(x) \text{ is bounded}) \\
&\xrightarrow[g\to 0]{} \int_0^{\Lambda^{-1}(x)} h(t)dt = \int_0^{\Lambda^{-1}(x)} \frac{2\lambda_1(t)}{u(t)} dt = \Lambda(\Lambda^{-1}(x)) = x
\end{aligned}$$

which proves **Condition 1**, recalling that $g \to 0$ is equivalent to $n \to \infty$.

Proof of Condition 2.

Condition 2 follows immediately from Corollary E.8.1.

Proof of Condition 3.

Given that individuals are exchangeable, we have

$$\begin{aligned}
&\frac{1}{y(t-g)^3 c_n(t)} \cdot \sum_{i=1}^{y(t)} \mathbb{E}\Big[\nu_{gti}(\nu_{gti}-1)\nu_{gti}^k | N(t-g)=y(t-g), N(t)=y(t)\Big] \\
&= \frac{1}{y(t-g)^3 c_n(t)} \cdot \sum_{i=1}^{y(t)} \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^k | N(t-g)=y(t-g), N(t)=y(t)\Big] \\
&= \frac{y(t)}{y(t-g)^3 c_n(t)} \cdot \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^k | N(t-g)=y(t-g), N(t)=y(t)\Big]
\end{aligned}$$

Given that $y(t) = O(n)$ and that, by Corollary E.8.1, $c_n(t) = O(1/n)$, we have

$$\frac{y(t)}{y(t-g)^3 c_n(t)} = O(1/n)$$

as $n \to \infty$.

Therefore, there remains to show that

$$\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^k | N(t-g)=y(t-g), N(t)=y(t)\Big] = o_k(n)$$

as $n \to \infty$ to prove **Condition 3**.

We have

$$\begin{aligned}
&\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^k | N(t-g)=y(t-g), N(t)=y(t)\Big] \\
&= \sum_{m=0}^{y(t-g)} m(m-1)m^k \mathbb{P}(\nu_{gt1}=m) \frac{\mathbb{P}\big(N(t-g)=y(t-g)-m | N(t)=y(t)-1\big)}{\mathbb{P}\big(N(t-g)=y(t-g) | N(t)=y(t)\big)}
\end{aligned}$$

Lemma E.4 shows that

$$\frac{\mathbb{P}\big(N(t-g)=y(t-g)-m | N(t)=y(t)-1\big)}{\mathbb{P}\big(N(t-g)=y(t-g) | N(t)=y(t)\big)} = O(1)$$

as $n \to \infty$. Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^k | N(t-g)=y(t-g), N(t)=y(t)\Big] \\
&= O(1) \sum_{m=0}^{y(t-g)} m(m-1)m^k \mathbb{P}(\nu_{gt1}=m) \\
&\leq O(1) \sum_{m=0}^{\infty} m^{k+2} \mathbb{P}(\nu_{gt1}=m) \\
&= O(1)\mathbb{E}\big[\nu_{gt1}^{k+2}\big]
\end{aligned}$$

as $n \to \infty$. By Corollary D.11.1, for any given k,

$$\sup_{g\in[0,g_1]} \sup_{t\in[g,T]} \mathbb{E}\big[\nu_{gt1}^{k+2}\big] < \infty$$

Hence,

$$\mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt1}^{k}|N(t-g)=y(t-g),N(t)=y(t)\Big] = O_k(1) = o_k(n)$$

as $n \to \infty$.

Proof of Condition 4.

Given that individuals are exchangeable, we have

$$\begin{aligned}
&\frac{1}{y(t-g)^4 c_n(t)} \sum_{i,j=1}^{y(t)} \mathbb{E}\Big[\nu_{gti}(\nu_{gti}-1)\nu_{gtj}^2|N(t-g)=y(t-g),N(t)=y(t)\Big] \\
&= \frac{1}{y(t-g)^4 c_n(t)} \sum_{i,j=1}^{y(t)} \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt2}^2|N(t-g)=y(t-g),N(t)=y(t)\Big] \\
&= \frac{y(t)^2}{y(t-g)^4 c_n(t)} \cdot \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt2}^2|N(t-g)=y(t-g),N(t)=y(t)\Big]
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\frac{y(t)^2}{y(t-g)^4 c_n(t)} \cdot \mathbb{E}\Big[\nu_{gt1}(\nu_{gt1}-1)\nu_{gt2}^2|N(t-g)=y(t-g),N(t)=y(t)\Big] \\
&\le \frac{y(t)^2}{y(t-g)^4 c_n(t)} \cdot \mathbb{E}\Big[\big(\nu_{gt1}(\nu_{gt1}-1)\big)^2|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2} \\
&\qquad \times \mathbb{E}\Big[\nu_{gt2}^4|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2}
\end{aligned}$$

By the same arguments as for **Condition 3**, we have that

$$\frac{y(t)^2}{y(t-g)^4 c_n(t)} = O(1/n)$$

as $n \to \infty$, and there thus remains to show that

$$\begin{aligned}
&\mathbb{E}\Big[\big(\nu_{gt1}(\nu_{gt1}-1)\big)^2|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2} \\
&\qquad \times \mathbb{E}\Big[\nu_{gt2}^4|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2} = o(n)
\end{aligned}$$

as $n \to \infty$.

Using the same arguments as for **Condition 3**, we have that

$$\begin{aligned}
&\mathbb{E}\Big[\big(\nu_{gt1}(\nu_{gt1}-1)\big)^2|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2} \\
&\times \mathbb{E}\Big[\nu_{gt2}^4|N(t-g)=y(t-g),N(t)=y(t)\Big]^{1/2} = O(1) = o(n)
\end{aligned}$$

as $n \to \infty$.

□

G Convergence of the relative population size of the conditional scaled BD to the function $u(t)$

Theorem G.3 below shows that, under the conditioned scaled BD, the relative population size $N(t)/n$ converges in probability to the function $u(t)$ as $n \to \infty$. This gives a meaning to the function $u(t)$: it is the relative population size of the conditional BD when population size is large. This meaning is particularly useful for interpreting the parameter $u(t)$ in statistical inference with the conditional BD, as the equivalence $u(t) \approx N(t)/n$ can be assumed for sufficiently large n.

Lemma G.1. *For any time $x \in [0, T]$*

$$\mathrm{Var}\left[\frac{N(x)}{n}\Big|N \in \Omega_n\right] = O(1/\sqrt{n})$$

as $n \to \infty$.

Proof. In this proof, all $O(\cdot)$'s are as $n \to \infty$.

For some fixed $g \in (0, g_1]$, for any $t \in \{g, 2g, \ldots, T\}$ and for all $x \in (t-g, t]$, because of the Markov property of the BD, we have that $\big(N(x)|N \in \Omega_n\big) = \big(N(x)|N(t-g) = y(t-g), N(t) = y(t)\big)$, and therefore

$$\mathrm{Var}\left[\frac{N(x)}{n}\Big|N \in \Omega_n\right] = \mathrm{Var}\left[\frac{N(x)}{n}\Big|N(t-g) = y(t-g), N(t) = y(t)\right]$$

By the law of total variance, we have

$$\begin{aligned}\mathrm{Var}\left[\frac{N(x)}{n}\Big|N(t) = y(t)\right] = \mathbb{E}\left[\mathrm{Var}\left[\frac{N(x)}{n}\Big|N(t-g), N(t) = y(t)\right]\right] \\ + \mathrm{Var}\left[\mathbb{E}\left[\frac{N(x)}{n}\Big|N(t-g), N(t) = y(t)\right]\right]\end{aligned}$$

so that

$$\mathbb{E}\left[\mathrm{Var}\left[\frac{N(x)}{n}\Big|N(t-g), N(t) = y(t)\right]\right] \leq \mathrm{Var}\left[\frac{N(x)}{n}\Big|N(t) = y(t)\right]$$

We then have

$$\begin{aligned}
&\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N\in\Omega_n\Big]\ \mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)\\
&\quad=\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N(t-g)=y(t-g),N(t)=y(t)\Big]\ \mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)\\
&\quad\le\sum_{k=0}^{\infty}\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N(t-g)=k,N(t)=y(t)\Big]\ \mathbb{P}\big(N(t-g)=k|N(t)=y(t)\big)\\
&\quad=\mathbb{E}\Big[\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N(t-g),N(t)=y(t)\Big]\Big]\\
&\quad\le\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N(t)=y(t)\Big]
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N\in\Omega_n\Big]&\le\frac{\mathrm{Var}\Big[\frac{N(x)}{n}\big|N(t)=y(t)\Big]}{\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)}\\
&=\frac{1}{n^2}\cdot\frac{\mathrm{Var}\big[N(x)|N(t)=y(t)\big]}{\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)}
\end{aligned}\tag{19}$$

Eq. (19) holds for some fixed $g\in(0,g_1]$, for any $t\in\{g,2g,\dots,T\}$ and for all $x\in(t-g,t]$. However, for some fixed time $x\in[0,T]$, as $n\to\infty$ and $g\to0$, eventually $x\notin(t-g,t]$. Therefore, for some fixed time $x\in(0,T]$, we define

$$\alpha_n=\alpha_n(x):=\min\big\{t\in\{g,2g,\dots,T\}|t\ge x\big\}$$

so that $\forall n\in\mathbb{N}_{>0}$, $x\in(\alpha_n-g,\alpha_n]$ and the following holds $\forall n\in\mathbb{N}_{>0}$ and $\forall x\in(0,T]$

$$\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N\in\Omega_n\Big]\le\frac{1}{n^2}\cdot\frac{\mathrm{Var}\big[N(x)|N(\alpha_n)=y(\alpha_n)\big]}{\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)}$$

We then write

$$\begin{aligned}
&\mathrm{Var}\Big[\frac{N(x)}{n}\Big|N\in\Omega_n\Big]\\
&\quad\le\frac{1}{n^2}\cdot\frac{\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathrm{Var}\big[N(x)|N(\alpha_n)=y(\alpha_n)\big]}{\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)}
\end{aligned}$$

By Lemma D.12, we have that $\sqrt{V(g,t)}=O(1)$, and also $\sqrt{y(t)}=O(\sqrt{n})$. In the following of this proof, we will show that $\forall x\in(0,T]$

$$\mathrm{Var}\big[N(x)|N(\alpha_n)=y(\alpha_n)\big]=O(n)$$

and that

$$\Big(\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)\Big)^{-1}=O(1)$$

which concludes the proof for $x \in (0, T]$. Then, for $x = 0$, we trivially have that

$$\mathrm{Var}\left[\frac{N(0)}{n}\Big|N \in \Omega_n\right] = \mathrm{Var}\left[\frac{N(0)}{n}\Big|N(0) = y(0)\right] = 0$$

By Lemma C.4, we have that

$$\begin{aligned}&\mathrm{Var}\big[N(x)|N(\alpha_n) = y(\alpha_n)\big]\\ &\quad = y(\alpha_n)\exp\left(2\int_x^{\alpha_n} r(s)ds\right)\int_x^{\alpha_n}\frac{2\lambda(s) - r(s)}{\exp\big(\int_s^{\alpha_n} r(u)du\big)}ds\end{aligned}$$

Noticing that $\alpha_n - x \leq g = O(1/n)$ by definition of α_n, we have that

$$2\int_x^{\alpha_n} r(s)ds \leq 2\int_x^{\alpha_n} \hat{r}ds = 2\hat{r}(\alpha_n - x) = O(1/n)$$

with $\hat{r} := \sup_{t\in[0,T]} r(t)$, so that $\exp\big(2\int_x^{\alpha_n} r(s)ds\big) = 1 + O(1/n) = O(1)$.

Similarly, for $s \in [x, \alpha_n]$ (implying that $\alpha_n - s = O(1/n)$), we have

$$-\int_s^{\alpha_n} r(u)du \leq -\int_x^{\alpha_n} \check{r}du = -\check{r}(\alpha_n - s) = O(1/n)$$

with $\check{r} := \inf_{t\in[0,T]} r(t)$, so that $\exp\big(-\int_s^{\alpha_n} r(u)du\big) = 1 + O(1/n) = O(1)$.

Then,

$$\begin{aligned}\int_x^{\alpha_n}\frac{2\lambda(s)}{\exp\big(\int_s^{\alpha_n} r(u)du\big)}ds &= n\int_x^{\alpha_n}\frac{2\lambda_1(s)}{\exp\big(\int_s^{\alpha_n} r(u)du\big)}ds\\ &\leq n\int_x^{\alpha_n} 2\hat{\lambda}_1 O(1)ds\\ &= n(\alpha_n - x)2\hat{\lambda}_1 O(1)\\ &= nO(1/n)O(1)\\ &= O(1)\end{aligned}$$

with $\hat{\lambda}_1 := \sup_{t\in[0,T]} \lambda_1(t)$. Also,

$$\begin{aligned}-\int_x^{\alpha_n}\frac{r(s)}{\exp\big(\int_s^{\alpha_n} r(u)du\big)}ds &\leq -\int_x^{\alpha_n}\check{r}O(1)ds\\ &= -(\alpha_n - x)\check{r}O(1)\\ &= O(1/n)O(1)\\ &= O(1/n)\end{aligned}$$

Given that $y(\alpha_n) = O(n)$, we thus conclude that

$$\mathrm{Var}\big[N(x)|N(\alpha_n) = y(\alpha_n)\big] = O(n)O(1)\big(O(1) + O(1/n)\big) = O(n) \tag{20}$$

Finally, in the proof to Lemma E.3, we have see that (Eq. 18)

$$\begin{aligned}\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)\\ =\frac{1}{\sqrt{2\pi}}+O(1/\sqrt{n})\end{aligned}$$

so that

$$\begin{aligned}\left(\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)\right)^{-1}\\ =\sqrt{2\pi}+O(1/\sqrt{n})=O(1)\end{aligned}\tag{21}$$

□

Lemma G.2. *For any time $x\in[0,T]$*

$$\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)=O(1/\sqrt[4]{n})$$

as $n\to\infty$.

Proof. In this proof, all $O(\cdot)$'s are as $n\to\infty$.

For some fixed $g\in(0,g_1]$, for any $t\in\{g,2g,\dots,T\}$ and for all $x\in(t-g,t]$, because of the Markov property of the BD, we have that $\big(N(x)|N\in\Omega_n\big)=\big(N(x)|N(t-g)=y(t-g),N(t)=y(t)\big)$, and therefore

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)\right|\\ &\quad=\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N(t-g)=y(t-g),N(t)=y(t)\right]-u(x)\right|\end{aligned}$$

We then have (using the triangle and Jensen's inequalities)

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)\right|\\ &\quad=\left|\mathbb{E}\left[\frac{N(x)}{n}-u(x)\;\middle|\;N\in\Omega_n\right]\right|\\ &\quad\le\mathbb{E}\left[\left|\frac{N(x)}{n}-u(x)\right|\;\middle|\;N(t-g)=y(t-g),N(t)=y(t)\right]\\ &\quad\le\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2\;\middle|\;N(t-g)=y(t-g),N(t)=y(t)\right]^{\frac{1}{2}}\end{aligned}$$

Then, by the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2 \;\middle|\; N(t)=y(t)\right] \\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2 \;\middle|\; N(t-g),N(t)=y(t)\right]\right] \\
&\quad=\sum_{k=0}^{\infty}\left\{\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2 \;\middle|\; N(t-g)=k,N(t)=y(t)\right]\right. \\
&\qquad\left.\times\,\mathbb{P}\big(N(t-g)=k|N(t)=y(t)\big)\right\} \\
&\quad\geq\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2 \;\middle|\; N(t-g)=y(t-g),N(t)=y(t)\right] \\
&\qquad\times\,\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)\right| \\
&\qquad\leq\left(\frac{\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2\;\middle|\;N(t)=y(t)\right]}{\mathbb{P}\big(N(t-g)=y(t-g)|N(t)=y(t)\big)}\right)^{\frac{1}{2}}
\end{aligned}\tag{22}$$

Eq. (22) holds for some fixed $g\in(0,g_1]$, for any $t\in\{g,2g,\ldots,T\}$ and for all $x\in(t-g,t]$. However, for some fixed time $x\in[0,T]$, as $n\to\infty$ and $g\to 0$, eventually $x\notin(t-g,t]$. Therefore, for some fixed time $x\in(0,T]$, we define

$$\alpha_n=\alpha_n(x):=\min\big\{t\in\{g,2g,\ldots,T\}|t\geq x\big\}$$

so that $\forall n\in\mathbb{N}_{>0}$, $x\in(\alpha_n-g,\alpha_n]$ and the following holds $\forall n\in\mathbb{N}_{>0}$ and $\forall x\in(0,T]$

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)\right| \\
&\qquad\leq\left(\frac{\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2\;\middle|\;N(\alpha_n)=y(\alpha_n)\right]}{\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)}\right)^{\frac{1}{2}}
\end{aligned}$$

We then write

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{N(x)}{n}\;\middle|\;N\in\Omega_n\right]-u(x)\right| \\
&\quad\leq\left(\frac{\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{E}\left[\left(\frac{N(x)}{n}-u(x)\right)^2\;\middle|\;N(\alpha_n)=y(\alpha_n)\right]}{\sqrt{y(\alpha_n)V(g,\alpha_n)}\cdot\mathbb{P}\big(N(\alpha_n-g)=y(\alpha_n-g)|N(\alpha_n)=y(\alpha_n)\big)}\right)^{\frac{1}{2}}
\end{aligned}$$

By Lemma D.12, we have that $\sqrt{V(g,t)} = O(1)$, and also $\sqrt{y(t)} = O(\sqrt{n})$. We have also seen in the proof to Lemma G.1 that (Eq. 21)

$$\left(\sqrt{y(\alpha_n)V(g,\alpha_n)} \cdot \mathbb{P}\big(N(\alpha_n - g) = y(\alpha_n - g) | N(\alpha_n) = y(\alpha_n)\big)\right)^{-1} = O(1)$$

In the following of this proof, we will show that $\forall x \in (0, T]$

$$\mathbb{E}\left[\left(\frac{N(x)}{n} - u(x)\right)^2 \;\middle|\; N(\alpha_n) = y(\alpha_n)\right] = O(1/n)$$

which concludes the proof for $x \in (0, T]$. The case for $x = 0$ will be dealt with at the end.

We have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{N(x)}{n} - u(x)\right)^2 \;\middle|\; N(\alpha_n) = y(\alpha_n)\right] \\
&= \mathrm{Var}\left[\frac{N(x)}{n} - u(x) \;\middle|\; N(\alpha_n) = y(\alpha_n)\right] - \mathbb{E}\left[\frac{N(x)}{n} - u(x) \;\middle|\; N(\alpha_n) = y(\alpha_n)\right]^2 \\
&= \frac{1}{n^2}\mathrm{Var}\Big[N(x) \;\Big|\; N(\alpha_n) = y(\alpha_n)\Big] - \frac{1}{n^2}\left(\mathbb{E}\Big[N(x) \;\Big|\; N(\alpha_n) = y(\alpha_n)\Big] - nu(x)\right)^2
\end{aligned}$$

In the proof to Lemma G.1 we have seen that (Eq. 20)

$$\mathrm{Var}\Big[N(x) \;\Big|\; N(\alpha_n) = y(\alpha_n)\Big] = O(n)$$

Using Lemma C.4, we have

$$\mathbb{E}\Big[N(x) \;\Big|\; N(\alpha_n) = y(\alpha_n)\Big] - nu(x) = y(\alpha_n) \exp\left(\int_x^{\alpha_n} r(s)ds\right) - nu(x)$$

We recall that $y(\alpha_n) = \lceil nu(\alpha_n) \rceil = nu(\alpha_n) + \epsilon$ with $0 \leq \epsilon < 1$ and that $u(t)$ is bounded on $[0, T]$. We have seen in the proof to Lemma G.1 that $\exp\big(\int_x^{\alpha_n} r(s)ds\big) = 1 + O(1/n)$. Furthermore, because $u'(t)$ is bounded, $\big(u(t) - u(x)\big)/(t - x) = O(1)$. Also, by definition of α_n, $\alpha_n - x < g = O(1/n)$. We then have

$$\begin{aligned}
\mathbb{E}\Big[N(x) \;\Big|\; N(\alpha_n) = y(\alpha_n)\Big] - nu(x) &= \big(nu(\alpha_n) + \epsilon\big)\big(1 + O(1/n)\big) - nu(x) \\
&= nu(\alpha_n) - nu(x) + O(1) \\
&= n(\alpha_n - x)\frac{u(\alpha_n) - u(x)}{\alpha_n - x} + O(1) \\
&= O(1)\frac{u(\alpha_n) - u(x)}{\alpha_n - x} + O(1) \\
&= O(1)
\end{aligned}$$

We thus obtain

$$\mathbb{E}\left[\left(\frac{N(x)}{n} - u(x)\right)^2 \;\middle|\; N(\alpha_n) = y(\alpha_n)\right] = \frac{1}{n^2}O(n) - \frac{1}{n^2}O(1) = O(1/n)$$

concluding the proof for $x \in (0, T]$.

For $x = 0$, with $\epsilon := y(0) - nu(0) = \lceil nu(0) \rceil - nu(0) \in [0, 1)$, we have $\forall n \in \mathbb{N}_{>0}$

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{N(0)}{n} \;\middle|\; N \in \Omega_n\right] - u(0)\right| \\
&\quad = \left|\mathbb{E}\left[\frac{N(0)}{n} \;\middle|\; N(0) = y(0)\right] - u(0)\right| \\
&\quad = \left|\frac{y(0)}{n} - u(0)\right| \\
&\quad = \left|\frac{1}{n}\big(nu(0) + \epsilon\big) - u(0)\right| \\
&\quad = \left|u(0) + \frac{\epsilon}{n} - u(0)\right| \\
&\quad = O(1/n) = O(1/\sqrt[4]{n})
\end{aligned}$$

□

Theorem G.3. $\left(\frac{N(t)}{n} | N \in \Omega_n\right)$ *converges uniformly in probability to* $u(t)$ *as* $n \to \infty$.

Proof. Using Bishop's $O_p(\cdot)$ and $o_p(\cdot)$ notation, we have by Bishop (35)'s Theorem 14.4-1 that

$$\left(\frac{N(t)}{n} | N \in \Omega_n\right) = \mathbb{E}\left[\frac{N(t)}{n}\Big| N \in \Omega_n\right] + O_p\left(\mathrm{Var}\left[\frac{N(t)}{n}\Big| N \in \Omega_n\right]^{\frac{1}{2}}\right)$$

Given that, by Lemma G.1,

$$\mathrm{Var}\left[\frac{N(t)}{n}\Big| N \in \Omega_n\right] = O(1/\sqrt{n})$$

and that by Lemma G.2

$$\mathbb{E}\left[\frac{N(t)}{n}\Big| N \in \Omega_n\right] = u(t) + O(1/\sqrt[4]{n})$$

we have

$$\left(\frac{N(t)}{n} | N \in \Omega_n\right) = u(t) + O(1/\sqrt[4]{n}) + O_p(1/\sqrt[4]{n}) = u(t) + o(1) + o_p(1)$$

which concludes the proof.

□

H Simulation procedure

For each comparison of two methods presented in the main text, we simulated 100 datasets, each consisting of a realization $z = \big(z(t)\big)_{t\in[0,T]}$ of the sample process Z.

A simulation of 100 datasets for one comparison of two methods proceeded in three steps:

1. Simulate a single realization of the population process N, denoted $\nu = \big(\nu(t)\big)_{t\in[0,T]}$. ν is a left-continuous step function with unit increments/decrements.
2. Given ν, a set of sampling times is drawn, represented by an increasing step function $s(t)$. The function $s(t)$ represents the cumulative number of samples taken between times 0 and t. If, at time $t = x$, k samples were taken, then $s(t)$ increases by k at time x. We sample 100 sets of sampling times: $\{s_i(t)\}_{i=1,\dots,100}$.
3. Given ν and $s_i(t)$, we draw a set of coalescence times, represented by an increasing step function $c_i(t)$. The function $c_i(t)$ represents the cumulative number of coalescences between times 0 and t. There cannot be two concomitant coalescences, since our procedure to simulate ν is such that several concomitant births cannot happen. We sample 100 sets of coalescence times, one for each simulated $s_i(t)$, yielding the set $\{c_i(t)\}_{i=1,\dots,100}$.

The i-th realization of the sample process is obtained as $z_i(t) = s_i(t) - c_i(t)$.

Simulation of a population trajectory In all three comparisons presented in the main text, we simulated the population process of an SIS model. The deterministic SIS model obeys the following system of ODEs (in backward time):

$$S'(t) = \beta\frac{I(t)}{X}S(t) - \mu I(t)$$
$$I'(t) = -\beta\frac{S(t)}{X}I(t) + \mu I(t)$$

where $S(t)$ is the number of susceptibles, $I(t)$ is the number of infectious individuals, and $X = S(t) + I(t)$ is the size of the whole population, assumed constant. β is the transmission rate and μ the recovery rate. Because samples of the pathogen are taken only from the population of infectious individuals, the number of infectious individuals represents our population process: $I(t) \equiv N(t)$. Because $S(t) = X - I(t)$, obviously, the above system of ODEs can be reduced to a single equation. We thus have:

$$I'(t) = -\beta\frac{X - I(t)}{X}I(t) + \mu I(t)$$

The stochastic version of the SIS model is obtained by noting that new infections arise at rate $\lambda\big(I(t)\big) = \beta\frac{X-I(t)}{X}$ and infections terminate at rate μ, per infectious individual. Hence, $I(t)$ follows a density-dependent BD model. This model can be simulated using a Gillespie algorithm, proceeding forward in time as follows. At time t, the total rate

of events is given by $\eta\big(I(t)\big) = \lambda\big(I(t)\big)I(t) + \mu I(t)$, so the waiting time $w\big(I(t)\big)$ until the next event is an exponential random variable with rate $\eta\big(I(t)\big)$. Drawing $w\big(I(t)\big)$ gives the time $t' = t - w\big(I(t)\big)$ when the next event happens. If $t' < 0$, the simulation is finished. If $t' > 0$, then the type of event $e(t')$ is a Bernoulli random variable with probability $\lambda\big(I(t)\big)I(t)/\eta\big(I(t)\big)$ that the event is a birth. Looking forward in time, if $e(t') = 1$, then I is increased by 1 at time t', and if $e(t') = 0$, I is decreased by 1 at time t'. Starting at time T from the initial condition $I(T)$ and proceeding by drawing births and deaths as outlined above yields a complete population trajectory on $[0, T]$, denoted ν.

Simulation of sampling times In the comparisons of Figures 8 and 9, sampling times were drawn by assuming a constant per-capita sampling rate ψ. Individuals were placed back into the population after sampling. Hence, the rate-based sampling procedure is a time-inhomogeneous Poisson point process, which can be simulated as follows. Denote t_k, with $k = 1, \ldots, d$ the time of the k-th event (birth or death) from present to past as per ν (i.e. the set $\{t_k\}_{k=1,\ldots,d}$ are the points of discontinuity of ν), and define $t_0 = 0$ and $t_{d+1} = T$. In $[t_k, t_{k+1}]$, the population size is constant and equal to $\nu\big(t_{k+1}\big)$. Hence, in $[t_k, t_{k+1}]$, the total sampling rate is constant and equal to $\Psi_k = \psi\nu\big(t_{k+1}\big)$. Therefore, the waiting times in between sampling events in $[t_k, t_{k+1}]$ are exponentially distributed with rate Ψ_k. Waiting times are drawn iteratively in $[t_k, t_{k+1}]$ until the sum of all waiting times is greater than $t_{k+1} - t_k$, rejecting the last waiting time. The procedure is repeated for $k = 0, \ldots, d$, yielding a set $\{x_j\}_j$ of sampling times. Then, $s(t)$ is constructed as the step function that increases by one when $t \in \{x_j\}_j$. We repeated the procedure 100 times, yielding $\{s_i(t)\}_{i=1,\ldots,100}$.

In the comparison of Figure 10, we deterministically took 71 samples at time 0.1, 66 samples at time 0.5, and 26 samples at time 1.0, for all 100 datasets.

Simulation of coalescence times Denote $\{b_k\}_{k=1,\ldots,B}$ the times of births as per ν (i.e. all the time points at which $\nu(t)$ decreases), with b_k the time of the k-th birth from present to past, and B the total number of births in $[0, T]$. Define $b_0 = 0$ and $b_{B+1} = T$. For a given sampling times function $s_i(t)$, denote γ_{ik} a random variable which is 1 if a coalescence happens at time b_k and 0 otherwise. Denote $\Gamma_{ik} = \sum_{j=1}^{k-1} \gamma_{ik}$ and define $\Gamma_{i0} = 0$. γ_{ik} is Bernoulli distributed with probability $\binom{s_i(b_k)-\Gamma_{ik}}{2}/\binom{\nu(b_k)}{2}$. The γ_{ik}'s are drawn iteratively for $k = 1, \ldots, B$. Then, $c_i(t)$ is constructed as the step function $c_i(t) = \Gamma_{ik}$ for $t \in [b_k, b_{k+1}]$.

Parameters used for the three sets of simulations For all simulations, $T = 10$. The population trajectory of Figure 8 was simulated with the parameters $X = 1000$, $\beta = 2$ and $\mu = 0.5$, with initial condition $N(T) = 1$. The sampling times were drawn based on a rate $\psi = 0.05$. The population trajectory of Figure 9 was simulated with the parameters $X = 10,000$, $\beta \approx 20.02$ and $\mu \approx 18.62$, with initial condition $N(T) = 10$. The sampling times were drawn based on a rate $\psi = 0.06$. For Figure 10, the same population trajectory was used as for Figure 9, but the sampling times were set deterministically.

Estimates of the death rate As explained below, our method and the BDS method cannot identify all the parameters they use. Hence, phylogenetic inference with either method requires that we fix some parameter(s) to a pre-estimated value during inference. In epidemiology, the death rate is probably the easiest parameter to estimate from auxiliary data, as the death rate is the inverse of the expected duration of infections. Hence, we chose to fix the death rate curve to a value summarizing the temporal variation of the realized death rate, an information that is contained in the realized population trajectory ν. We proceeded as follows.

Above we have detailed how we simulated ν using a Gillespie algorithm, using $\lambda\big(I(t)\big)I(t)+\mu I(t)$ as the total rate of events, and drawing waiting times in between events in an exponential distribution. Hence, switching to inference mode, we can estimate $\zeta = \lambda + \mu$ by fitting an exponential distribution with rate $\zeta\nu(t)$ to the waiting times of ν (i.e. the periods of times when ν is constant). We divided the interval $[0, T]$ into five time intervals (defined such that each interval contained 20% of the coalescence times of the 100 simulated realizations of the sample process), and obtained five estimates of ζ: $\{\hat{\zeta}_i\}_{i=1,\dots,5}$. Then, counting the numbers of deaths and births in interval i as per ν, we obtained an estimate $\hat{\mu}_i$ for the i-th interval as

$$\hat{\mu}_i = \hat{\zeta}_i \cdot \frac{\#\text{ deaths in interval i}}{\#\text{ deaths in interval i} + \#\text{ births in interval i}}$$

I Parametrization of the various inference methods

I.1 Fixing the death rate curve

For our method, as explained in Section J.3 below, the parameters are not separately identifiable unless a curve is fixed. The same is true for the BDS method (52). Therefore, we fixed the death rate curve to a pre-estimated value for both these methods. In Section H, we explained how we derived estimates of the death rate in five time intervals. We denote these estimates $\{\hat{\mu}_i\}_{i=1,\dots,5}$. During inference with our method or the BDS, we fixed the death rate curve to the following pairwise constant function:

$$m(t) = \hat{\mu}_i \quad \text{if } t \in [t_i, t_{i+1})$$

with $\{t_i\}_{i=2,\dots,5}$ the times separating the five intervals, $t_1 = 0$ and $t_6 = T$.

I.2 Parametrization of our method

We fixed $\tilde{\mu}(t) = m(t)$.

Rather than parametrizing $\tilde{\lambda}$ and deduce $\tilde{N}$, we chose to parametrize $\tilde{N}$ and deduce $\tilde{\lambda}$, as this resulted in smoother curves overall. We parametrized $\tilde{N}$ as

$$\tilde{N}(t) = \exp P(t)$$

with $P(t)$ a polynomial of degree 6.

Then solving

$$\exp P(t) = \tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds$$

for $\tilde{\lambda}$ yields

$$\tilde{\lambda}(t) = \tilde{\mu}(t) - P'(t)$$

In sum, we parametrized θ as

$$\theta(t) = \frac{\tilde{N}(t)}{2\tilde{\lambda}(t)} = \frac{\tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds}{2\tilde{\lambda}(t)} = \frac{\exp P(t)}{2\big(m(t) - P'(t)\big)}$$

The estimated curve for $\tilde{N}$ is $\exp \hat{P}$ and the estimated curve for $\tilde{\lambda}$ is $m - \hat{P}'$.

There are seven free parameters: the coefficients of $P(t)$.

I.3 Parametrization of the BDS method

We fixed $\tilde{\mu}(t) = m(t)$.

We parametrized λ as

$$\lambda(t) = a_i + b_i t \quad \text{if } t \in [t_i, t_{i+1})$$

with $b_i = \frac{\lambda_{t_{i+1}} - \lambda_{t_i}}{t_{i+1} - t_i}$, $a_i = \lambda_{t_i} - b_i t_i$, with $\{t_i\}_{i=2,\ldots,5}$ delineating the same five intervals used to define $m(t)$, and with $t_1 = 0$ and $t_6 = T$. In other words, $\lambda(t)$ is a continuous piecewise linear function passing through the points $\{(t_i, \lambda_{t_i})\}_{i=1,\ldots,6}$.

The sampling rate was fixed to a constant: $\psi(t) = \psi$.

There are seven free parameters: $\{\lambda_{t_i})\}_{i=1,\ldots,6}$ and ψ.

It can be verified that the BDS model parametrized in this way is identifiable, i.e. no two sets of parameters yield models in the same equivalence class (congruence class *sensu* Louca et al. (52)). However, to any model parametrized as outlined above, there exist infinitely many other congruent BDS models (with other parametrizations of the curves λ and ψ).

The likelihood of this model is implemented in the R package *castor* (function fit_hbds_model_parametric()). We used $\kappa = 1$ (i.e. sampled individuals are put back in the population).

For the BDS, we used the deterministic population size as an estimate of the realized population size, calculated as

$$N(t) = N(0) \exp \int_0^t \big(\mu(s) - \lambda(s)\big) ds$$

With an estimate of λ and with $\mu = m$, we can only calculate the relative population size. To obtain an estimate of the absolute population size, we must derive an estimate of $N(0)$. We proceeded as follows. Given that we used $\kappa = 1$, our model assumes that sampling does not impact population size. Hence, the total deterministic sampling rate is $\psi N(t)$, and the deterministic sampling procedure of the BDS is a time-inhomogeneous Poisson point process (PPP). The waiting times in between sampling events thus contain information about the absolute value of $N(t)$, and fitting a PPP to the sampling waiting times allows to estimate $N(0)$. Specifically, given estimates $\hat{\lambda}$ and $\hat{\psi}$, the probability density of a sampling event at time t_1 given the last sampling event was at time $t_2 \geq t_1$ is

$$\hat{\psi}N(0)\exp\left(\int_0^{t_1}\big(m(t)-\hat{\lambda}(t)\big)dt\right)\exp\left(-\int_{t_1}^{t_2}dt\hat{\psi}N(0)\exp\int_0^{t}\big(m(s)-\hat{\lambda}(s)\big)ds\right)$$

I.4 Parametrization of the skyline KC method

With the skyline KC method, the only curve to parametrize is θ. We parametrized θ as

$$\theta(t) = a_i + b_i t \quad \text{if } t \in [t_i, t_{i+1})$$

with $b_i = \frac{\theta_{t_{i+1}} - \theta_{t_i}}{t_{i+1} - t_i}$, $a_i = \theta_{t_i} - b_i t_i$, with $\{t_i\}_{i=2,\ldots,5}$ delineating the same five intervals used to define $m(t)$ for our method, and with $t_1 = 0$ and $t_6 = T$. In other words, $\theta(t)$ is a continuous piecewise linear function passing through the points $\{(t_i, \theta_{t_i})\}_{i=1,\ldots,6}$.

There is no death rate curve in the skyline KC method, so we do not use the pre-estimated curve $m(t)$. However, to allow a fair comparison of the skyline KC with our method, we use the same time intervals to define the pieces of $\theta(t)$ in the skyline KC method as we use to define $\tilde{\lambda}(t)$ and $\tilde{\mu}(t)$ in our method.

I.5 Parametrization of the SIR-based mechanistic KC method

The SIR model is defined by the following system of ODEs (expressed in forward time, denoted τ):

$$\begin{aligned} S'(\tau) &= -\beta\frac{I(\tau)}{X}S(\tau) \\ I'(\tau) &= \beta\frac{S(\tau)}{X}I(\tau) - \mu I(\tau) \\ R'(\tau) &= \mu I(\tau) \end{aligned}$$

where $S(\tau)$ is the number of susceptibles, $I(\tau)$ is the number of infectious individuals, $R(\tau)$ is the number of removed individuals, and $X = S(\tau) + I(\tau) + R(\tau)$ is the total population size. We use for initial conditions $S(0) = X - 1$, $I(0) = 1$ and $R(0) = 0$.

We integrated numerically the above system of equations from $\tau = 0$ to $\tau = T' \geq T$, where T' is the (backward) time of origin of the SIR process. This yields a numerical

solution for $I(\tau)$ in forward time, that is $I_t(t) = I(T' - t)$ in backward time. The phylogeny being made up of only infectious individuals, $I_t(t)$ is the equivalent of what we call "population size" in this paper: $I_t(t) \equiv N(t)$.

With the SIR-based mechanistic approach, we parametrize θ as

$$\theta(t) = kI_t(t) \equiv kN(t)$$

with k a proportionality constant.

The parameters of the SIR-based mechanistic approach are X, β, μ, T' and k. We found that X, T' and k are at best weakly identifiable, so we fixed $k = 1$. Furthermore, to allow a fair comparison with our method, we fixed μ to the mean of $\{\mu_i\}_{i=1,\dots,5}$. This way, both methods are informed about the true death rate. There results a model with three free parameters: X, β and T'.

J Details on our new method

J.1 Coalescence rate of BD-type models

Let ν be a realization of the population process N and β be the corresponding curve of the cumulative number of births. Define $x := \sup_{s\in[0,t]} 1/\nu(s)$ so that $1/\nu(s) = O(x)$ as $x \to 0$. Every use of the O/o notation in this section is as $x \to 0$.

We consider continuous and differentiable functions $\tilde{B}$ and $\tilde{N}$ as phenomenological representations of β and ν, respectively. We define the realized birth rate $\tilde{\lambda} = \tilde{B}'/\tilde{N}$.

We make the following assumptions, $\forall s \in [0, t]$:

1. $\tilde{B}(s)/\beta(s) = 1 + o(1)$
2. $\tilde{N}(s)/\nu(s) = 1 + o(1)$
3. $\nu(s) = O(1/x)$
4. $\tilde{N}'(s)/\tilde{N}(s) = O(1)$
5. $\tilde{\lambda}(s)/\tilde{N}(s) = O(1)$

The first two assumptions mean that we represent the realizations β and ν by some phenomenological functions $\tilde{B}$ and $\tilde{N}$ that are proportionally closer to β and ν as we consider greater population sizes (as $x \to 0$). The third assumption implies that $\sup_{s\in[0,t]} \nu(s)/\inf_{s\in[0,t]} \nu(s)$ is bounded, and therefore the same is true for $\tilde{N}$. The fourth assumption means that $\tilde{N}$, and therefore ν, does not vary faster than an exponential function. The fifth assumption indicates that the per-capita birth rate is at most of the same order as the population size.

For all BD-type models, the probability $P(t)$ of a pair of individuals not coalescing in $[0,t]$ is given exactly by

$$P(t) = \prod_{i=1}^{\beta(t)} \left(1 - \frac{1}{\binom{\nu(t_i)}{2}}\right) \tag{23}$$

Indeed, at time t_i when the i-th birth happened, there were $\nu(t_i)$ individuals in the population and therefore $\binom{\nu(t_i)}{2}$ pairs of individuals. The probability that a specific pair of individuals coalesces at time t_i is the probability that these two individuals are the two offspring of the i-th birth event, that is $1/\binom{\nu(t_i)}{2}$. At all times when a birth did not happen, coalescences have probability 0. The above equality is true only for models whereby births spawn only two offspring. For instance, if at time t_i the birth spawned k offspring, then the probability of coalescence of a specific pair of lineages would be $\binom{k}{2}/\binom{\nu(t_i)}{2}$, because there are $\binom{k}{2}$ pairs in the whole population that coalesce at time t_i.

Our purpose here is to show that the large-population limit (as $x \to 0$) of $P(t)$ is the corresponding probability under the KC with $\theta = \tilde{N}/(2\tilde{\lambda})$.

Now let's write

$$\begin{aligned} -\ln P(t) &= -\sum_{i=1}^{\beta(t)} \ln\left(1 - \frac{1}{\binom{\nu(t_i)}{2}}\right) \\ &= -\sum_{j=0}^{\lceil t/x\rceil - 1} \sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} \ln\left(1 - \frac{1}{\binom{\nu(t_{jk})}{2}}\right) \end{aligned}$$

where $\forall j \in \{0,\ldots,\lceil t/x\rceil - 1\},\ t_j = jx$ and $t_{\lceil t/x\rceil} = t$ and where t_{jk} is the time of the k-th birth in $[t_j, t_{j+1}]$.

The following will be for all $s \in [0,t]$.

We have

$$\begin{aligned} \frac{1}{\binom{\nu(s)}{2}} &= \frac{2}{\nu(s)^2\big(1 - 1/\nu(s)\big)} = \frac{2}{\nu(s)^2\big(1 + O(x)\big)} \\ &= \frac{2}{\nu(s)^2}\big(1 + O(x)\big) = \frac{2}{\nu(s)^2} + O(x^3) \end{aligned}$$

Also,

$$\left(\frac{\tilde{N}(s)}{\nu(s)}\right)^2 = \big(1 + o(1)\big)^2 = 1 + o(1)$$

Also,

$$\frac{1}{\tilde{N}(s)} = \frac{1}{\nu(s)}\frac{1}{\frac{\tilde{N}(s)}{\nu(s)}} = O(x)\frac{1}{1 + o(1)} = O(x)\big(1 + o(1)\big) = O(x)$$

Hence,

$$\frac{1}{\binom{\nu(s)}{2}} = \left(\frac{\tilde{N}(s)}{\nu(s)}\right)^2 \frac{2}{\tilde{N}(s)^2} + O(x^3) = \frac{2}{\tilde{N}(s)^2}\big(1+o(1)\big) + O(x^3)$$
$$= \frac{2}{\tilde{N}(s)^2} + \frac{2}{\tilde{N}(s)^2}o(1) + O(x^3) = \frac{2}{\tilde{N}(s)^2} + o(x^2)$$

Hence,

$$-\ln P(t) = -\sum_{j=0}^{\lceil t/x\rceil - 1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} \ln\left(1 - \frac{2}{\tilde{N}(t_{jk})^2} + o(x^2)\right)$$

Given that $-\frac{2}{\tilde{N}(s)^2} + o(x^2) = O(x^2)$ and that $\ln(1+z) = z + O(z^2)$, we have

$$-\ln P(t) = -\sum_{j=0}^{\lceil t/x\rceil - 1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} \left(-\frac{2}{\tilde{N}(t_{jk})^2} + O(x^4)\right)$$
$$= \sum_{j=0}^{\lceil t/x\rceil - 1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} \frac{2}{\tilde{N}(t_{jk})^2} + \sum_{j=0}^{\lceil t/x\rceil - 1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} O(x^4) \qquad (24)$$

We now deal with the first term of the RHS of the previous equation.

By Taylor's theorem and given that $\tilde{N}'/\tilde{N} = O(1)$ we have that

$$\tilde{N}(t_{jk}) = \tilde{N}(t_j) + \tilde{N}'(t_j)O(x) + o(x)$$
$$\frac{\tilde{N}(t_{jk})}{\tilde{N}(t_j)} = 1 + \frac{\tilde{N}'(t_j)}{\tilde{N}(t_j)}O(x) + o(x) = 1 + O(x)$$
$$\left(\frac{\tilde{N}(t_{jk})}{\tilde{N}(t_j)}\right)^2 = \big(1+O(x)\big)^2 = 1 + O(x)$$
$$\left(\frac{\tilde{N}(t_j)}{\tilde{N}(t_{jk})}\right)^2 = \frac{1}{1+O(x)} = 1 + O(x)$$

We also have that

$$\beta(t_{j+1}) - \beta(t_j) = \frac{\beta(t_{j+1})}{B(t_{j+1})}B(t_{j+1}) - \frac{\beta(t_j)}{B(t_j)}B(t_j)$$
$$= \big(1+o(1)\big)\big(B(t_{j+1}) - B(t_j)\big)$$

We thus have

$$\begin{aligned}
&\sum_{j=0}^{\lceil t/x\rceil-1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)}\frac{2}{\tilde{N}(t_{jk})^2}\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)}\frac{2}{\tilde{N}(t_j)^2}\big(1+O(x)\big)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\big(\beta(t_{j+1})-\beta(t_j)\big)}{\tilde{N}(t_j)^2}\big(1+O(x)\big)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\big(\tilde{B}(t_{j+1})-\tilde{B}(t_j)\big)\big(1+o(1)\big)}{\tilde{N}(t_j)^2}\big(1+O(x)\big)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\big(\tilde{B}(t_{j+1})-\tilde{B}(t_j)\big)}{\tilde{N}(t_j)^2}\big(1+o(1)\big)
\end{aligned}$$

Then, by Taylor's theorem we have that

$$\tilde{B}(t_{j+1})-\tilde{B}(t_j)=\tilde{B}'(t_j)x+o(x)=\tilde{\lambda}(t_j)\tilde{N}(t_j)x+o(x)$$

We obtain

$$
\begin{aligned}
&\sum_{j=0}^{\lceil t/x\rceil-1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)}\frac{2}{\tilde{N}(t_{jk})^2}\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\big(\tilde{\lambda}(t_j)\tilde{N}(t_j)x+o(x)\big)}{\tilde{N}(t_j)^2}\big(1+o(1)\big)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)\tilde{N}(t_j)x}{\tilde{N}(t_j)^2}\big(1+o(1)\big)+\sum_{j=0}^{\lceil t/x\rceil-1}\frac{o(x)}{\tilde{N}(t_j)^2}\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)\tilde{N}(t_j)x}{\tilde{N}(t_j)^2}\big(1+o(1)\big)+\sum_{j=0}^{\lceil t/x\rceil-1}o(x^3)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)\tilde{N}(t_j)x}{\tilde{N}(t_j)^2}\big(1+o(1)\big)+o(x^2)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)x}{\tilde{N}(t_j)}+\Big(\sum_{j=0}^{\lceil t/x\rceil-1}\frac{\tilde{\lambda}(t_j)o(x)}{\tilde{N}(t_j)}\Big)+o(x^2)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)x}{\tilde{N}(t_j)}+\Big(\sum_{j=0}^{\lceil t/x\rceil-1}o(x)\Big)+o(x^2)\\
&\quad=\sum_{j=0}^{\lceil t/x\rceil-1}\frac{2\tilde{\lambda}(t_j)x}{\tilde{N}(t_j)}+o(1)\\
&\quad=\int_0^t\frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)}+O(x)+o(1)\\
&\quad=\int_0^t\frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)}+o(1)
\end{aligned}
$$

Now we deal with the second term of Eq. (24). We have

$$\begin{aligned}
-\ln P(t) &= \sum_{j=0}^{\lceil t/x\rceil-1}\sum_{k=1}^{\beta(t_{j+1})-\beta(t_j)} O(x^4)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} \beta(t_{j+1})-\beta(t_j)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} \big(\tilde{B}(t_{j+1})-\tilde{B}(t_j)\big)\big(1+o(1)\big)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} \big(\tilde{B}(t_{j+1})-\tilde{B}(t_j)\big)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} \tilde{\lambda}(t_j)\tilde{N}(t_j)x+o(x)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} \frac{\tilde{\lambda}(t_j)}{\tilde{N}(t_j)}\tilde{N}(t_j)^2x+o(x)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} O(1)O(1/x^2)O(x)+o(x)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} O(1/x)+o(x)\\
&= O(x^4)\sum_{j=0}^{\lceil t/x\rceil-1} O(1/x)\\
&= O(x^4)O(1/x)O(1/x)\\
&= O(x^2)
\end{aligned}$$

Putting the two terms back into Eq. (24) yields

$$-\ln P(t) = \int_0^t \frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)} + o(1) + O(x^2) = \int_0^t \frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)} + o(1)$$

And therefore

$$P(t) = \exp\left(-\int_0^t \frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)} + o(1)\right) = \exp\left(-\int_0^t \frac{2\tilde{\lambda}(s)ds}{\tilde{N}(s)}\right) + o(1)$$

Thus, in the large-population limit $x \to 0$, the probability $P(t)$ that a pair of lineages do not coalesce is the corresponding probability under a KC with $\theta = \frac{\tilde{N}}{2\tilde{\lambda}}$. Notice that this limit is only interesting if it is further assumed that $\tilde{N}/\tilde{\lambda} = O(1)$, otherwise the coalescence rate goes to 0 as $x \to 0$. Provided the assumptions mentioned at the

beginning of this section are met, this result applies to all BD-type models, since the initial equation (Eq. 23) is true for all such models (but untrue as soon as multiple concomitant births can happen or if individuals are not exchangeable).

We do not consider the above as a formal demonstration of the convergence of BD-type models to the KC, because it focuses on the probability of coalescence of two individuals (i.e. when $Z(t) = 2$). In contrast, a formal proof must consider the case $Z(t) = k$ and in our particular case here, we would need to show that in the limit coalescences happen at rate $\binom{k}{2}\frac{2\tilde{\lambda}(t)}{\tilde{N}(t)}$ when $Z(t) = k$. This is more involved, and the existing theorems of convergence to the KC, such as that of Möhle (19) that we used in Theorem F.1, cannot be used as they assume that the curve of the cumulative number of births is unknown, unlike in this section where it is fixed.

In sum, Theorem F.1 constitutes our formal proof that a BD model with deterministic rates converges to a KC, while we consider the present section as a heuristic that suggests that the convergence should apply to all BD-type models.

J.2 Expression of the realized population trajectory

Here we explain why we have parametrized $\tilde{N}$ as the following function of the realized rates $\tilde{\lambda}$ and $\tilde{\mu}$:

$$\tilde{N}(t) = \tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds \tag{25}$$

We recall that $N(t)$ is the random population size, $B(t)$ the random cumulative number of births from the present to time t, and $D(t)$ the random cumulative number of deaths. We consider specific realizations of these random processes, which we denote ν for N, β for B, and δ for D. These step functions are represented phenomenologically by continuous and differentiable functions $\tilde{N}$ for ν, $\tilde{B}$ for β, and $\tilde{D}$ for δ. We have defined the realized rates as $\tilde{\lambda} = \tilde{B}'/\tilde{N}$ and $\tilde{\mu} = \tilde{D}'/\tilde{N}$.

Obviously, the random processes N, B and D are linked by the equation $N(t) = N(0) - B(t) + D(t)$, which is also verified by any specific set of realizations: $\nu(t) = \nu(0) - \beta(t) + \delta(t)$. Hence, the phenomenological representations of these realizations should also respect this relation: $\tilde{N}(t) = \tilde{N}(0) - \tilde{B}(t) + \tilde{D}(t)$. By the definitions of the realized rates, given that $\tilde{B}(0) = \tilde{D}(0) = 0$, we have that $\tilde{B}(t) = \int_0^t \tilde{\lambda}(s)\tilde{N}(s)ds$ and $\tilde{D}(t) = \int_0^t \tilde{\mu}(s)\tilde{N}(s)ds$. We thus obtain

$$\begin{aligned}
\tilde{N}(t) &= \tilde{N}(0) - \int_0^t \tilde{\lambda}(s)\tilde{N}(s)ds + \int_0^t \tilde{\mu}(s)\tilde{N}(s)ds \\
\tilde{N}'(t) &= -\tilde{\lambda}(t)\tilde{N}(t) + \tilde{\mu}(t)\tilde{N}(t) \\
\frac{\tilde{N}'(t)}{\tilde{N}(t)} &= -\tilde{\lambda}(t) + \tilde{\mu}(t)
\end{aligned}$$

Integrating on both sides yields Eq. (25).

J.3 Identifiability of the parameters of our new method

J.3.1 Correspondence between the equivalence classes for our method, and those of the Bernoulli-sampled BD model

Our method is a KC with

$$\theta(t) = \frac{\tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds}{2\tilde{\lambda}(t)} \tag{26}$$

The function θ is known to be identifiable. However, there are infinitely many models, each defined by a triple $\big(\tilde{\mu}, \tilde{\lambda}, \tilde{N}(0)\big)$, that yield the same θ. All the models yielding the same θ are called "equivalent" (they yield the same KC likelihood, for any dataset), and an equivalence class is defined by a function θ. For a given dataset with $Z(0) = z_0$, the equivalence classes of our model are the same as the equivalence classes defined by Louca and Pennell (47) for the BD model with Bernoulli sampling.

Specifically, consider a Bernoulli-sampled BD model, with birth rate $\lambda(t)$, death rate $\mu(t)$ and sampling probability ρ. As per Eq. (2) in (47), the deterministic population size is defined as

$$N(t) = N(0) \exp \int_0^t \big(\mu(s) - \lambda(s)\big) ds$$

with $N(0) = z_0/\rho$.

We can then define θ for a Bernoulli-sampled BD model as

$$\theta(t) = \frac{N(t)}{2\lambda(t)} \tag{27}$$

Then, as per Eqs. (3,8) in (47), we have

$$N(t) = \frac{M(t)}{1 - E(t)}$$
$$\lambda(t) = \frac{\lambda_p(t)}{1 - E(t)}$$

where M is the "deterministic LTT plot", λ_p is the "pulled birth rate", and E the probability that a lineage existing at time t has no sampled descendants.

We then obtain

$$\theta(t) = \frac{M(t)}{2\lambda_p(t)}$$

Then, as per Eq. (7) in (47),

$$\lambda_p(t) = -\frac{M'(t)}{M(t)}$$

yielding

$$\theta(t) = \frac{M(t)^2}{2M'(t)} \tag{28}$$

Importantly, all the Bernoulli-sample BD models in a given equivalence class *sensu* Louca and Pennell (47) yield the same function M, and hence the same function θ. It is also easily verified that, for a given function θ, there is a single function M that satisfies Eq. (28). Hence, θ defines the equivalence classes of Louca and Pennell (47) in the same way that M or λ_p do.

In conclusion, according to the equivalences $\tilde{\lambda} \equiv \lambda$, $\tilde{\mu} \equiv \mu$ and $\tilde{N}(0) \equiv N(0)$ between the parameters of our model and those of the Bernoulli-sampled BD, the equivalence classes of our model and those of the Bernoulli-sampled BD model are the same and defined by θ, the latter being defined by Eq. (26) in our model, and by Eq. (27) in the Bernoulli-sampled BD model.

J.3.2 Identifiability of the chosen parametrization of our method

Due to the identifiability issues outlined above, we have to bring external information about some parameter(s) of our model. We chose to fix $\tilde{\mu} = m$. We have also parametrized our model with a polynomial P, parametrizing $\tilde{N}$ as $\tilde{N} = \exp P$, which yields $\tilde{\lambda} = m - P'$. We think that our model thus parametrized is identifiable, i.e. there are no two different parametrizations of the polynomial P yielding the same θ.

However, to any one model parametrized with $\tilde{N} = \exp P$ and $\tilde{\mu} = m$, correspond infinitely many equivalent models with $\tilde{\mu} = m$ and with $\tilde{N}$ of a different form than $\tilde{N} = \exp P$.

Specifically, given $\tilde{\mu} = m$, and given that $\tilde{N}(t) = \tilde{N}(0) \exp \int_0^t \big(\tilde{\mu}(s) - \tilde{\lambda}(s)\big) ds$, we have that $\tilde{\lambda} = m - \tilde{N}'/\tilde{N}$.

Hence, for a given $P = P^{(0)}$, yielding $\tilde{N} = \tilde{N}^{(0)}$ and $\theta = \theta^{(0)}$, all the models with $\tilde{N}$ such that

$$\theta = \frac{\tilde{N}}{2(m - \tilde{N}'/\tilde{N})} = \theta^{(0)}$$

are equivalent to the model with $\tilde{N}^{(0)} = \exp P^{(0)}$. Solving the previous equation for $\tilde{N}$ yields

$$\tilde{N}(t) = \frac{\exp \int_0^t m(s) ds}{\frac{1}{\tilde{N}(0)} + \int_0^t \frac{ds}{2\theta^{(0)}(s)} \exp \int_0^s m(u) du} \tag{29}$$

In the above equation, $\tilde{N}(t)$ has one free parameter: $\tilde{N}(0)$. Hence, there are as many equivalent models to the model with $\tilde{N}^{(0)} = \exp P^{(0)}$ as there are valid values of $\tilde{N}(0)$. A valid value of $\tilde{N}(0)$ is such that $\tilde{N}(0) \geq 0$, and also such that the resulting birth rate $\tilde{\lambda} = m - \tilde{N}'/\tilde{N}$ is positive on $[0, T]$.

Future studies will be necessary to determine if the set of equivalent models are very different from a given estimated model with $\tilde{N} = \exp P$. In the meantime, we present in Figures 11-13 below the equivalent models for the analyses carried out in the main text, and note that the equivalent models differ significantly from the estimated models only in the time periods where few coalescences were observed.

Finally, for future uses of our method, we note that an estimate of either $\tilde{N}$ or $\tilde{\lambda}$ at a single time point completely identifies the model.

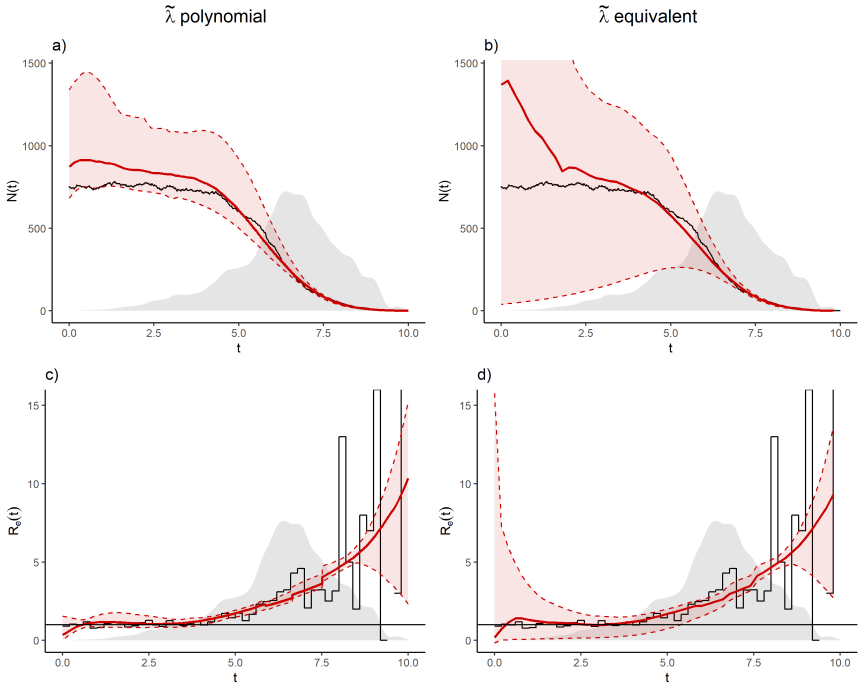

Figure 11: Equivalent models for the datasets of Figure 8. a,c) Estimates of $\tilde{N}$ and $\tilde{R}_e$ with our method, i.e. with $\tilde{N} = \exp P$ and P a polynomial of degree 6 (same as Figure 8). The black line is the true value, the red solid line is the median of estimates and the red ribbon ranges from the 0.025 to the 0.975 percentile of estimates. For each estimate obtained with the constraint $\tilde{N} = \exp P$, we obtained a set of equivalent estimates of $\tilde{N}$, using Eq. (29) with $\tilde{N}(0) \in [\hat{\tilde{N}}(0)/20, 20\hat{\tilde{N}}(0)]$ (panel b). From these equivalent estimates of $\tilde{N}$, we derived equivalent estimates for $\tilde{\lambda}$ as $\tilde{\lambda} = m - \tilde{N}'/\tilde{N}$, and equivalent estimates of $\tilde{R}_e$ as $\tilde{R}_e = \tilde{\lambda}/m$ (panel d). We see that considering equivalent models decreases precision significantly in the most recent period when coalescences events are rarer.

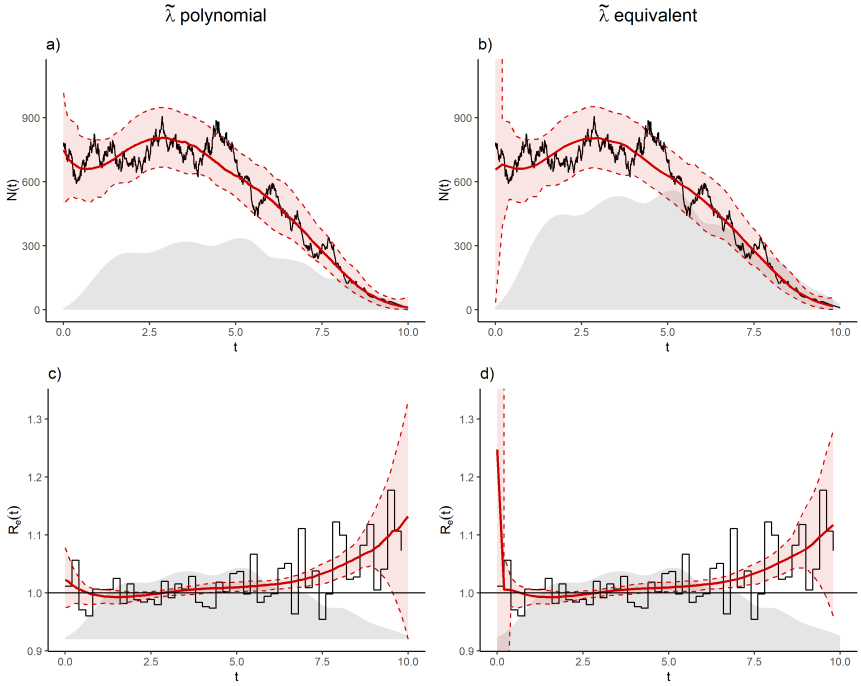

Figure 12: Equivalent models for the datasets of Figure 9. This figure reads as Figure 11. We see that, contrarily to the example of Figure 11, here, considering the equivalent models does not lower precision, except in the very recent past.

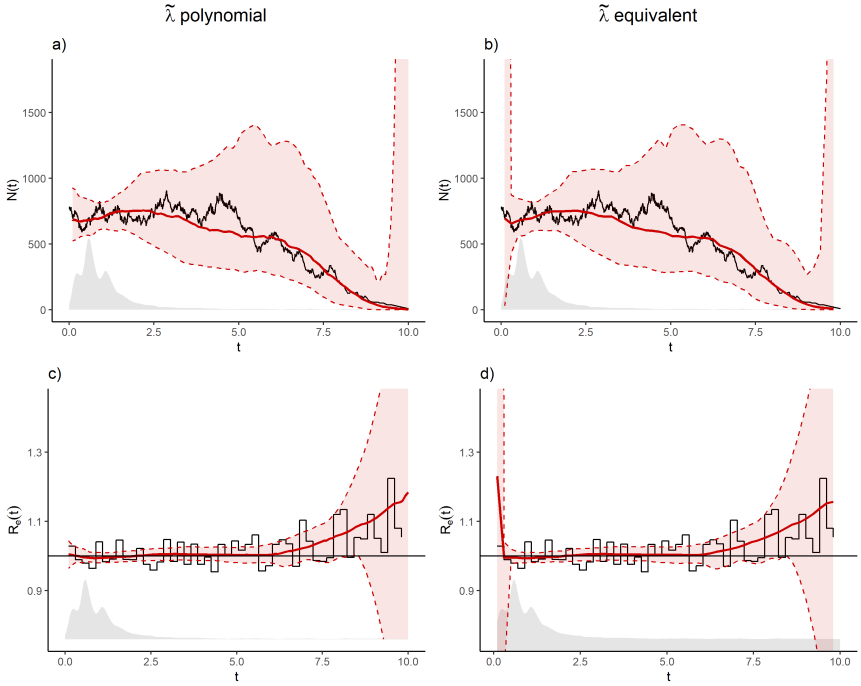

Figure 13: Equivalent models for the datasets of Figure 10. This figure reads as Figure 11. Like for the example of Figure 12, precision is lowered only in the very recent past.

K BDS inference with correct sampling rate

In Figure 10 of the main text, we showed the result of fitting the BDS with a constant sampling rate $\psi(t) = \psi$ to datasets where the samples were taken at three determined points in time. The BDS performed very badly on such datasets, and we reckon that this is due to the fact that the simulated sampling procedure is at odds with the fitted sampling procedure. In Figure 14 below, we show the results of fitting the constant-sampling-rate BDS to datasets simulated with a constant sampling rate (i.e. the sampling procedure assumed in inference is the correct one). We see that indeed the BDS performs accurately in this case.

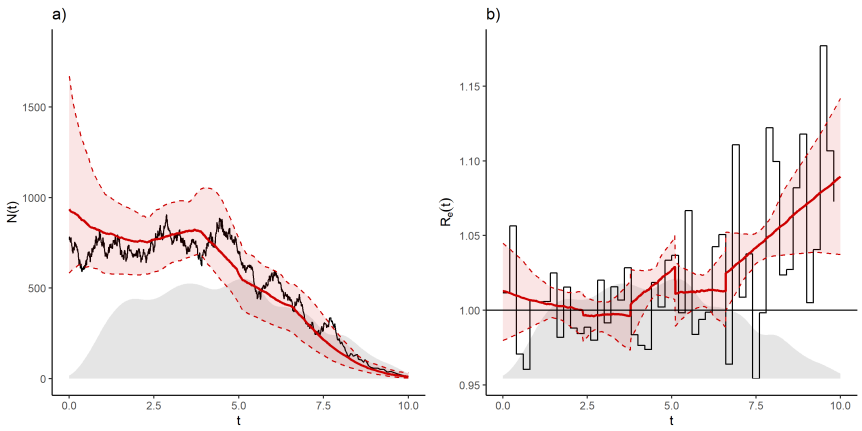

Figure 14: Inference with the BDS method (assuming a constant sampling rate) for datasets for which sample times were drawn according to a constant sampling rate. a) Estimates of the population size. Black curve, the true population size. Red line, median of estimates. The red ribbon ranges from the .025 to the .975 percentiles of the estimates. b) Estimates of the effective reproduction number, reads as a). The simulated population trajectory is the same here as in Figure 10 in the main text. The difference is that here the sampling procedure assumed in inference (i.e. a constant sampling rate) is the one used for simulations. We observe that the BDS performs well when it assumes the correct sampling procedure.