EXACT PHYLODYNAMICS VIA STRUCTURED MARKOV GENEALOGY PROCESSES

AARON A. KING, QIANYING LIN, AND EDWARD L. IONIDES

ABSTRACT.

1. Introduction.

Problem of phylodynamics. Factorization of problem into two subproblems.

Relation to previous work. Existing methods (Volz et al., 2009; Stadler, 2010). Large-population, small sample-size approximations.

Extension of previous results (King et al., 2022). Broader class of state-spaces. Accommodating discrete structure.

Classes of Markov processes. Utility and flexibility of Markov assumptions.

Population process induces Markov history and genealogy processes. Using these, we derive equations for the likelihood of a genealogy conditional on the history. We then integrate out the history to obtain nonlinear filtering equations, the solution of which yields the likelihood. These readily lend themselves to a family of sequential Monte Carlo algorithms for computing the likelihood. We demonstrate with several examples.

In the following, we show a Markov population process of the kind that is a staple in epidemiology induces a Markov process on the space of genealogies. We then show how one can comput the likelihood of a given genealogy.

2. From population processes to genealogy processes.

2.1. Population processes.

Motivating examples: compartmental models. Wide variety of models. Linear chain trick. Migration, superspreading, competition between strains.

Another perspective on the Markov processes is to be had from its Markov state transition diagram (Fig. 2).

Mathematical notation. Denote the underlying probability space by $(\Omega, \mathcal{B}, \mathbb{P})$. We will assume that our population process is a time-inhomogeneous Markov jump process, $\mathcal{X}_t$, parameterized by time $t \in \mathbb{R}_+ \coloneqq \{t \in \mathbb{R} \mid t \geqslant 0\}$ and taking values in some space $\mathbb{X}$. In earlier work (King et al., 2022), we limited ourselves to the case $\mathbb{X} = \mathbb{Z}^d$, but here we assume only that $\mathbb{X}$ is a complete metric space with a countable dense subset, *i.e.,* a Polish space. The population process is completely specified by its initial-state distribution, p_0, and its transition rates α. In particular, we suppose that

$$\mathbb{P}\left[\mathcal{X}_0 = x\right] = p_0(x), \tag{1}$$

Date: July 5, 2023.

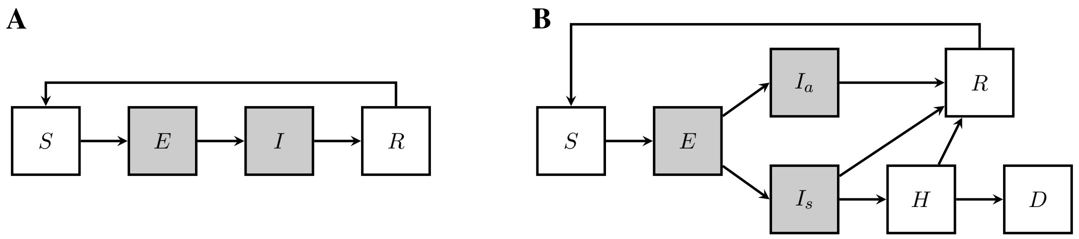

FIGURE 1. Examples of compartmental models. Demes are shaded. [Perhaps another one or two examples here?] [We could add dots to the deme compartments to signify individuals. . . .]

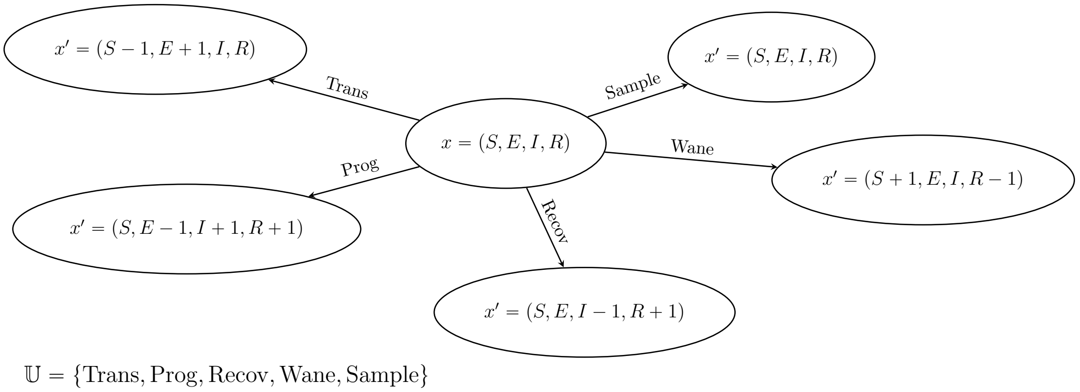

FIGURE 2. Markov state transition diagram for an SEIR model. The state is characterized by four numbers, S, E, I, and R. From a given state x, there are five possible kinds of events $x \mapsto x'$ as shown. From the point of view of the induced genealogy process, Trans (transmission) is of birth type, Prog (progression) is of migration type, and Recov (recovery) is of death type, while Wane (loss or waning of immunity) is of neutral type. Note that, in this formulation, when a Sample (sampling) event occurs, the state does not change.

for some choice of initial distribution, p_0. For any $t \in \mathbb{R}_+$, $x, x' \in \mathbb{X}$, we think of the quantity $\alpha(t, x, x')$ as the instantaneous hazard of a jump from x to x'. More precisely, the transition rates have the following properties:

$$\alpha(t, x, x') \geqslant 0, \qquad \int_{\mathbb{X}} \alpha(t, x, x')\,\mathrm{d}x' < \infty,$$

for all $t \in \mathbb{R}_+$ and $x, x' \in \mathbb{X}$. Henceforth, we understand that integrals are taken over all of $\mathbb{X}$ unless otherwise specified. Let $\mathcal{N}_t$ be the number of jumps that $\mathcal{X}$ has taken by time t. We assume that $\mathcal{N}_t$ is a simple counting process so that

$$\mathbb{P}\left[\mathcal{N}_{t+\Delta} = n+1 \mid \mathcal{N}_t = n\right] = \Delta \int \alpha(t, x, x')\,\mathrm{d}x' + o(\Delta), \qquad \mathbb{P}\left[\mathcal{N}_{t+\Delta} > n+1 \mid \mathcal{N}_t = n\right] = o(\Delta),$$

$$\mathbb{P}\left[\mathcal{X}_{t+\Delta} \in \mathcal{E} \mid \mathcal{X}_t = x, \mathcal{N}_{t+\Delta} - \mathcal{N}_t = 1\right] = \frac{\int_{\mathcal{E}} \alpha(t, x, x')\,\mathrm{d}x'}{\int \alpha(t, x, x')\,\mathrm{d}x'} + \frac{o(\Delta)}{\Delta}.$$

[Do we need the last term?] We will further assume that $\mathcal{X}_t$ is non-explosive, *i.e.*, , that $\mathbb{P}\left[\mathcal{N}_t < \infty\right] = 1$ for all t. [Is this equivalent to non-explosivity? Or merely an implication?]

The above may be compactly summarized by stating that if $w(t,x)$ satisfies the Kolmogorov forward equation (KFE),

$$\frac{\partial w}{\partial t}(t,x) = \int w(t,x')\,\alpha(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\alpha(t,x,x')\,\mathrm{d}x', \tag{2}$$

and if, moreover, $w(0,x) = p_0(x)$, then $\int_{\mathcal{E}} w(t,x)\,\mathrm{d}x = \mathbb{P}\left[\mathcal{X}_t \in \mathcal{E}\right]$ for every measurable $\mathcal{E} \subseteq \mathbb{X}$. Eq. 2 is sometimes called the *master equation* for $\mathcal{X}_t$.

Without loss of generality, one can assume, as we do here, that the sample paths $t \mapsto \mathcal{X}_t(\omega)$ for $\omega \in \Omega$ are right-continuous with left limits (càdlàg). In fact, all of the processes we will describe in this paper will be taken to be càdlàg, and we will frequently need to refer to the left-limit. Accordingly, if $z = z(\omega,t)$ is any càdlàg random function, we define

$$\widetilde{z}(\omega,t) \coloneqq \begin{cases} \lim_{t'\uparrow t} z(\omega,t'), & t > 0, \\ z(\omega,0), & t = 0. \end{cases}$$

Note that $\widetilde{z}$ is thus left-continuous with right limits.

Structured populations, demes. In an *unstructured* Markov population process, every lineage is exactly like every other. King et al. (2022) showed how every such process induces an unstructured Markov genealogy process. Here, our aim is to expand the theory considerably by allowing our population of lineages to have discrete structure. In particular, we suppose that there are a countable set of subpopulations that differ in their vital rates, but within each of which, individual lineages are statistically identical. We call these subpopulations *demes*, and use the symbol $\mathbb{D}$ to denote an index set for them.

For any $i \in \mathbb{D}$, we let $n_i(\mathcal{X}_t)$ denote the number of lineages present in deme i at time t, *i.e.,* the *occupancy* of deme i. Thus $n(\mathcal{X}_t) \in \mathbb{Z}_+^{\mathbb{D}}$ is the vector of deme occupancies.

Jump marks. In the following, it will be useful to break the jumps into distinct categories. For this purpose, we let $\mathbb{U}$ be a countable set of jump *marks* such that

$$\alpha(t,x,x') = \sum_u \alpha_u(t,x,x').$$

In Fig. 2, we use the marks to distinguish biologically distinct events. Here and in the following, sums over u are taken over the whole of $\mathbb{U}$ unless otherwise indicated.

Let us define the *jump mark* process, $\mathcal{U}_t$, to be the mark of the latest jump as of time t. As usual, we take the sample paths, $t \mapsto \mathcal{U}_t(\omega)$ for $\omega \in \Omega$, to be càdlàg. Observe that $\mathcal{U}_t$ is a random, but not a Markov, process.

2.2. Examples.

SEIRS model.

SIIR model.

Linear birth-death model.

Moran model and the Kingman coalescent.

2.3. The history, inventory, and genealogy processes.

History process. For $\omega \in \Omega$, $t \mapsto (\mathcal{X}_t(\omega), \mathcal{U}_t(\omega))$ is a càdlàg function of time. Let the *history process*, $\mathcal{H}_t$, be the restriction of this random function to the interval $[0, t]$. Note that $\mathcal{H}_t$ is a Markov process.

The non-explosiveness assumption implies that a.s., for every t, there is a finite, increasing sequence of random jump times $T_t \coloneqq (t_k)_{k=1}^{K_t}$; the length K_t of this sequence is itself random. However, conditional on $\mathcal{H}_t$, T_t and K_t, together with the mark process $\mathcal{U}_t$, and the population process $\mathcal{X}_t$ are deterministic. Using these, one can write down an explicit probability measure for $\mathcal{H}_t$:

$$\begin{aligned}\pi^{\mathcal{H}}(\mathrm{d}\mathcal{H}_t) =& \pi^{\mathcal{H}}(\mathrm{d}T_t, \mathrm{d}\mathcal{X}_t, K_t, \mathcal{U}_t)\\ =& p_0(\mathcal{X}_0)\,\mathrm{d}x_0 \prod_{k=1}^{K_t} \left(\alpha_{\mathcal{U}_{t_k}}\left(t_k, \tilde{\mathcal{X}}_{t_k}, \mathcal{X}_{t_k}\right)\,\mathrm{d}x_k\,\mathrm{d}t_k\right)\\ &\times \exp\left(-\sum_u \int_0^t \int \alpha_u(t', \mathcal{X}_{t'}, x')\mathrm{d}x'\mathrm{d}t'\right).\end{aligned} \tag{3}$$

Inventories. Our goal in this paper is to probabilistically characterize how the genealogical relationships among lineages evolve through time. Accordingly, we develop some notation for this purpose. To begin with, we assign to each lineage a unique number $j \in \mathbb{Z}_+$. This can be done in any fashion, so long as no two lineages ever receive the same number. For example, when a new lineage arises, we can assign it the smallest integer that has not yet been assigned. We will define the *inventory process*, $\mathcal{I}_t$, so that, for every lineage $j \in \mathbb{Z}_+$, $\mathcal{I}_t(j)$ is the deme in which j is found. However, when t is before the birth or after the death of lineage j, then clearly $\mathcal{I}_t(j) \notin \mathbb{D}$. We say in this case that lineage j is in the *underdeme*, which we denote using the symbol $\eth$, so that we can write $\mathcal{I}_t(j) = \eth$ and define $\overline{\mathbb{D}} \coloneqq \mathbb{D} \cup \{\eth\}$. With this definition, $\mathcal{I} : \mathbb{R}_+ \times \mathbb{Z}_+ \to \overline{\mathbb{D}}$.

The birth and death times of lineage j are therefore

$$t_j^b = \min\{t | \mathcal{I}_t(j) \neq \eth\} \qquad \text{and} \qquad t_j^d = \sup\{t | \mathcal{I}_t(j) \neq \eth\},$$

respectively. Observe that $n_i(\mathcal{X}_t) = |\{j \mid \mathcal{I}_t(j) = i\}|$ for all $t \in \mathbb{R}_+$ and $i \in \mathbb{D}$. Note also that n does not count the inhabitants of the underdeme.

Jump types. Different kinds of events that occur for the population process can have different kinds of effects on the inventory process, and indeed not every jump affects $\mathcal{I}_t$ at all. From the point of view of the inventory process, there are five distinct *types* of jumps, which we enumerate here.

1. Birth-type events result in the branching of one or more new lineages, each from some existing lineage. If j is one of the new lineages, we use the expression $\mathsf{Anc}(j)$ to refer to its ancestor. Examples of birth-type events include transmission events, speciations, and actual births. It is not assumed that all new lineages arising from a birth event share the same ancestor.
2. Death-type events result in the extinction of one or more lineages. Examples include recovery, death of a host, and species extinctions.
3. Migration-type events result in the movement of a lineage from one deme to another. Spatial movements, changes in behavior, and progression of an infection can all be represented as migration-type events.
4. Sample-type events result in the collection of a sample from a lineage but do not in themselves affect the inventory process.
5. Neutral-type events result in no change to any of the lineages.

Fig. 2 depicts an example with all five of these types. It is not necessary that a jump fall into just one of these types. It is allowable, for instance to have compound jumps that fall into more than one category. For example, Sample-death-type events, in which a lineage is simultaneously sampled and removed, have been used, as have birth-death events in which one lineage reproduces at the same moment that another dies. The theory presented here places no restrictions on these events.

However, we do impose the restriction that the *production*, *i.e.,* the deme-specific number of lineages emerging from the event, be constant for all jumps of a given mark. To be precise, the *production* is defined to be a function $r : \mathbb{U} \times \mathbb{D} \to \mathbb{Z}_+$, such that r_u^i lineages of deme i emerge from each event of mark u. We write $r_u = (r_u^i)_{i\in\mathbb{D}}$. Note that we do not count lineages that die as a result of the event. Also, it is important to note that the parent lineage or lineages, if they survive the event, are always counted among the emergent lineages.

Because different kinds of events may differ not only in the number of offspring they engender, but also in the number of parent lineages, and the distribution of offspring among parents, there is also assumed to be a deterministic function Q_u, for $u \in \mathbb{U}$, (described below) that expresses these properties.

Inventory process. The structure of the state space for the inventory process, $\mathcal{I}_t$, has already been described. It remains to define its stochastic dynamics. The $\mathcal{I}_t$ process follows the population process $\mathcal{X}_t$ in that jumps in $\mathcal{I}_t$ do not occur except when jumps in $\mathcal{X}_t$ occur: $\widetilde{\mathcal{I}}_t \neq \mathcal{I}_t$ implies $\widetilde{\mathcal{N}}_t \neq \mathcal{N}_t$.

At jumps of birth type and mark u, one or more random parents are selected from the appropriate deme(s). The appropriate number of offspring are created in each deme.

Genealogies. King et al. (2022), showed how an unstructured population process induces a process on the space of genealogies. Although we now treat a more general case, the construction is much the same, so we abbreviate the presentation. Readers wishing for more detail should consult the earlier paper (King et al., 2022).

Formally, we define a *genealogy*, $\mathcal{G}$, to be a finite sequence of *internal nodes*, together with a time. The time is denoted $T(\mathcal{G})$ and the number of nodes is $K(\mathcal{G})$. The k-th node is $\mathsf{p}_k(\mathcal{G})$ and we write $\mathsf{p} \in \mathcal{G}$ if p is one of the nodes of $\mathcal{G}$. Each $\mathsf{p} \in \mathcal{G}$ has a creation-time, $T(\mathsf{p})$ and a parent, $\mathsf{Anc}(\mathsf{p})$. Root nodes are distinguished by being their own parents: p is a root if and only if $\mathsf{Anc}(\mathsf{p}) = \mathsf{p}$. Every node also has one or more descendants called *children*. There are three kinds of children: (i) *Internal nodes*. Children nodes must always be later than their parents: $T(\mathsf{Anc}(\mathsf{p})) < T(\mathsf{p})$. (ii) *Tip nodes*, which represent extant lineages. In particular, every lineage alive at time t is the child of some node. (iii) *Sample nodes*. When a sample is collected at time t, a new node, p is added with $T(\mathsf{p}) = t$ and the sample as child.

Genealogy processes. The population process induces a stochastic process on the space of genealogies. In particular, at each event in the population process, one or more of the following changes happen to the genealogy, according to the type of the event:

(a) A birth-type event at time t results in the creation of one new internal node for each parent lineage. In particular, if j is one of the parent lineages, and p is the new node, then $T(\mathsf{p}) = t$, $\mathsf{Anc}(\mathsf{p})$ is the node in which j lay prior to the event. The children of p include all the new lineages that branched from j, as well as j itself.

(b) In a death-type event, all the lineages j that die are removed. Nodes without children are then recursively removed.

(c) In a migration-type event, one node is added for each migrating lineage; each one takes the migrating lineage as child. The ancestor of the new node is the node in which the migrating lineage lay before the event.

FIGURE 3. Illustration of genealogy processes. [Similar to that of King et al. (2022) but with multiple demes represented.]

(d) At a sample-type event, one new node is introduced for each sampled lineage. Each one takes the sampled lineage as child, along with the sample node. The ancestor of the node is that in which the sampled lineage lay before the vent.
(e) At a neutral-type event, no change is made to the genealogy.
(f) Finally, events of compound type are readily accommodated by combining the rules just stated.

Pruning.

Lineages, event-counter, deme-residence. Let $\mathsf{L} = \mathsf{L}(\mathcal{G})$ denote the finite set of all samples represented in genealogy $\mathcal{G}$. Let $\lll$ be any ordering of L that is compatible with ancestry. That is, if $j, j' \in \mathsf{L}$ are such that j is ancestral to j', then we must have $j \lll j'$. Using this ordering, we can uniquely associate each point on a genealogical tree with the least of those lineages that descend from that point. In particular, any lineage $j \in \mathsf{L}$, corresponding to a sample taken at time t^s_j, can be traced backward from node to node until either it coalesces with some lesser lineage at some time $t^o_j > 0$ or a root is reached (in which case, we define $t^o_j = 0$). Each node encountered along the way represents a genealogical event from which j emerges. Moreover, at each time $t \in [t^o_j, t^s_j)$, lineage j is in precisely one of the demes $\mathbb{D}$. However, for $t \notin [t^o_j, t^s_j)$, lineage j does not exist. To express this, we again say that lineage j is in the *underdeme*, which we denote using the symbol $\eth$. We define $\overline{\mathbb{D}} \coloneqq \mathbb{D} \cup \{\eth\}$.

It will be useful to define a function that captures all the relevant features of a pruned genealogy. Accordingly, let $\mathbb{Y} = \mathbb{Z}_+ \times \overline{\mathbb{D}} \times \mathsf{L}$ and define $y : \mathbb{R}_+ \times \mathsf{L} \to \mathbb{Y}$ so that, for $j \in \mathsf{L}$ and $t \in \mathbb{R}_+$:

(a) $\mathsf{ct}(y_j(t)) \in \mathbb{Z}_+$ is a counting process which increases by 1 at each event along lineage j.
(b) $\mathsf{deme}(y_j(t)) \in \overline{\mathbb{D}}$ indicates in which deme lineage j lies at time t. In particular $\mathsf{deme}(y_j(t)) = \eth$ if $t \notin [t^o_j, t^s_j)$.
(c) $\mathsf{anc}(y_j(t)) \in \mathsf{L}$ indicates the lineage in which the ancestors of lineage j are found. In particular, $\mathsf{anc}(y_j(t)) = j$ for $t \geqslant t^o_j$, but $\mathsf{anc}(y_j(t))$ is well defined for all j and t. [Can define this in a way similar to that of the original Kingman's coalescent....]

One can verify that y is càdlàg. Define $\widetilde{y}(t) = \lim_{t' \uparrow t} y(t')$.

To visualize these functions, one can make a correspondence between demes and colors. Then a pruned genealogy is visualized as a tree with colored branches. Knowing the function $\mathsf{deme}(y)$ is equivalent to knowing the coloring, while $\mathsf{ct}(y)$ determines the locations of events in the genealogy and $\mathsf{anc}(y)$ determines the topology. Note in particular that $y_j(t) \neq \widetilde{y}_j(t)$ if and only if t is the time of an event from which lineage j emerges.

Lineage count, saturation. In the following, we will find that we need to count the deme-specific numbers of lineages present at a given time. Accordingly, for any $\mathsf{L}' \subseteq \mathsf{L}$, $\eta \in (\mathbb{Z}_+ \times \overline{\mathbb{D}} \times \mathsf{L}')^{\mathsf{L}'}$, and $i \in \mathbb{D}$, let us define

$$\ell_i(\eta) \coloneqq \left|\left\{ j \in \mathsf{L}' \mid \mathsf{deme}(\eta_j) = i \right\}\right| \in \mathbb{Z}_+, \qquad \ell(\eta) \coloneqq (\ell_i(\eta))_{i \in \mathbb{D}} \in \mathbb{Z}_+^{\mathbb{D}}.$$

Note that lineages j for which $\eta_j = \eth$ are not counted. With this definition, it follows that $\ell_i(y(t))$ is the number of lineages in deme i at time t.

We will also have occasion to refer to the deme-specific number of lineages emerging from a given event. Accordingly, for $\mathsf{L}' \subseteq \mathsf{L}$, $\eta, \eta' \in (\mathbb{Z}_+ \times \overline{\mathbb{D}} \times \mathsf{L}')^{\mathsf{L}'}$, and $i \in \mathbb{D}$, let us define

$$s_i(\eta,\eta') := \left|\left\{j \in \mathsf{L}' \mid \mathsf{deme}(\eta_j) = i \ \ \& \ \ \mathsf{ct}(\eta'_j) = \mathsf{ct}(\eta_j) + 1\right\}\right|, \qquad s(\eta,\eta') := (s_i(\eta,\eta'))_{i\in\mathbb{D}} \in \mathbb{Z}_+^{\mathbb{D}}.$$

With this definition, $s_i(\widetilde{y}(t), y(t))$ is the number of lineages in deme i that emerge from an event at time t and that, if t is not an event time, then $s(\widetilde{y}(t), y(t)) = 0$.

2.4. Obscuration.

3. Results.

Definition 1. For $n, r, \ell, n \in \mathbb{Z}_+^{\mathbb{D}}$, define the multivariable binomial coefficient by

$$\binom{n}{r} := \prod_{i\in\mathbb{D}} \binom{n_i}{r_i}.$$

Using this, we define the *binomial ratio*

$$\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} := \frac{\dbinom{n-\ell}{r-s}}{\dbinom{n}{r}} \in [0,1].$$

In consequence of the Chu-Vandermonde identity, we have

$$\sum_{s\in\mathbb{Z}_+^{\mathbb{D}}} \begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \binom{\ell}{s} = 1 \tag{4}$$

Likelihood of a pruned genealogy. Now certain population-process events at certain times are incompatible with any given pruned genealogy. Example: a sample event where no sample is seen, or a migration event involving a change of deme not observed in the pruned genealogy. Define the function $Q_u(\eta,\eta') = 1$ if a change $\eta \to \eta'$ at time t is compatible with an event of type u at that time, and $Q_u = 0$ otherwise. In other words, Q_u is the indicator function for the condition that there exist $\omega \in \Omega$, $t \in \mathbb{R}_+$, and $k \in \mathbb{Z}_+$ such that $t = T_k(\omega)$, $u = U_k(\omega)$, $\widetilde{y}(t,\omega) = \eta$, and $y(t,\omega) = \eta'$. Using this, we define

$$\phi_u(\xi,\eta,\eta') := \begin{pmatrix} n(\xi) & \ell(\eta') \\ r_u & s(\eta,\eta') \end{pmatrix} Q_u(\eta,\eta').$$

Theorem 1.

$$\mathbb{P}\left[\mathcal{P}_t|\mathcal{H}_t\right] = \prod_{k=1}^{K} \phi_u\left(\mathcal{X}_{t_k}, \widetilde{y}(t_k), y(t_k)\right) = \prod_{k=1}^{K} \sum_{y_k} \pi(y_{k-1},y_k)\, \frac{\phi_u\left(\mathcal{X}_{t_k}, y_{k-1}, y_k\right)}{\pi(y_{k-1},y_k)}\, \kappa^{\mathcal{P}_t}(t_k, y_{k-1}, y_k),$$

where $\kappa^{\mathcal{P}_t}(t,y,y') = \mathbb{1}\{\widetilde{y}^{\mathcal{P}_t}(t) = y, y^{\mathcal{P}_t}(t) = y'\}$. *[Note that we can just sum over $\mathcal{P}_t$ and interchange sums and products (using conditional independence) to get Thm 2!]*

Proof. Two approaches possible:

1. Conditional on $\mathcal{H}_t$, the likelihood of $\mathcal{P}_t$ is the product of p_{jk}, where j ranges over lineages and k over events. Interchanging the order of integration yields the expression given.
2. Use the filter equations developed in the Appendix.

□

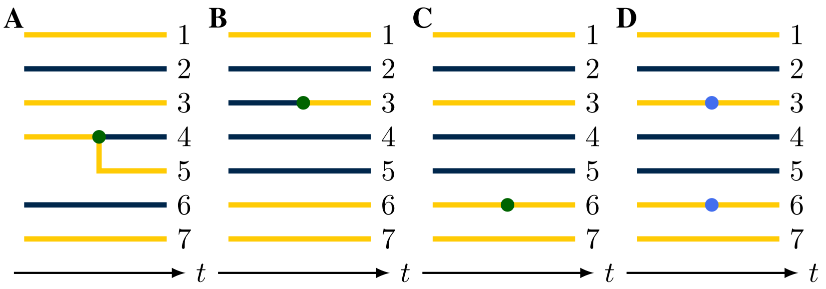

FIGURE 4.

$$\begin{tikzcd}[column sep=large]
(\ell=0,s=0) \arrow[r, "\frac{n-r}{n}"] \arrow[dr, "\frac{r}{n}"] & (1,0) \arrow[r, "\frac{n-r-1}{n-1}"] \arrow[dr, "\frac{r}{n-1}"] & (2,0) \arrow[r, "\frac{n-r-2}{n-2}"] \arrow[dr, "\frac{r}{n-2}"] & (3,0) \arrow[r, "\frac{n-r-3}{n-3}"] \arrow[dr, "\frac{r}{n-3}"] & (4,0) \\
 & (1,1) \arrow[r, "\frac{n-r}{n-1}"] \arrow[dr, "\frac{r-1}{n-1}"] & (2,1) \arrow[r, "\frac{n-r-1}{n-2}"] \arrow[dr, "\frac{r-1}{n-2}"] & (3,1) \arrow[r, "\frac{n-r-2}{n-3}"] \arrow[dr, "\frac{r-1}{n-3}"] & (4,1) \\
 & & (2,2) \arrow[r, "\frac{n-r}{n-2}"] \arrow[dr, "\frac{r-2}{n-2}"] & (3,2) \arrow[r, "\frac{n-r-1}{n-3}"] \arrow[dr, "\frac{r-2}{n-3}"] & (4,2) \\
 & & & (3,3) \arrow[r, "\frac{n-r}{n-3}"] \arrow[dr, "\frac{r-3}{n-3}"] & (4,3) \\
 & & & & (4,4)
\end{tikzcd}$$

$$\frac{r!}{(r-s)!}\cdot\frac{(n-\ell)!}{n!}\cdot\frac{(n-r)!}{(n-r-\ell+s)!}=\frac{r!(n-r)!}{n!}\cdot\frac{(n-\ell)!}{(n-\ell-r+s)!(r-s)!}=\begin{pmatrix} n & \ell \\ r & s \end{pmatrix}$$

Suppose we have a pruned genealogy $\mathcal{P}^*$, defined on the time-interval $[0,T]$, with event times $0=t_0<t_1<\dots<t_n=T$. From the theorem, we have, for $t\notin\mathsf{event}(\mathcal{P}^*)$,

$$\frac{\partial w}{\partial t}=\sum_u\int w(t,x')\,\alpha_u(t,x',x)\,\phi_u\big(x,y(t),y(t)\big)\,\mathrm{d}x'-\sum_u\int w(t,x)\,\alpha_u(t,x,x')\,\mathrm{d}x'. \tag{5}$$

At event times, $t\in\mathsf{event}(\mathcal{P}^*)$, one has

$$w(t,x)=\sum_u\int\widetilde{w}(t,x')\,\frac{\alpha_u(t_k,x',x)}{\mu}\,\phi_u\big(x,\widetilde{y}(t),y(t)\big)\,\mathrm{d}x', \tag{6}$$

In addition, there are the initial and boundary conditions

$$w(0,x)=p_0(x),\qquad w(t,x)=0\quad\text{whenever}\quad n(x)<\ell(y(t)). \tag{7}$$

Likelihood of an obscured genealogy.

Theorem 2. *State the filter equation definition.*

Proof. Two approaches possible:

(1) Use the filter equations developed in the Appendix to compute $\mathbb{E}\left[\mathbb{1}\{\mathsf{obs}(\mathcal{P}_t) = \mathcal{V}_t\}\right]$. In effect, we compute $\mathbb{P}\left[\mathcal{P}_t\right]$ and then sum over paintings y.

(2) Change the order of the summation: $\mathbb{E}\left[\mathbb{1}\{\mathsf{obs}(\mathcal{P}_t) = \mathcal{V}_t\}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{\mathsf{obs}(\mathcal{P}_t) = \mathcal{V}_t\} \mid \mathcal{H}_t\right]\right]$. That is, conditional on $\mathcal{H}_t$, the likelihood of $\mathcal{V}_t$ is sum over all $\mathcal{P}_t$ such that $\mathsf{obs}(\mathcal{P}_t) = \mathcal{V}_t$. Therefore, any importance sampling scheme for $\mathcal{P}_t$ will do. In particular, we choose a scheme that paints the tree forward in time, driven by a mimic of $\mathcal{X}_t$. We show that this is equivalent to the specific filter equation.

    It may be useful in this to first show that $\mathbb{P}\left[\mathcal{P}_t | \mathcal{H}_t\right]$ can be expressed as an expectation over y, where y is constrained to be equal $y^{\mathcal{P}_t}$. Then, show that the sum of the $\mathcal{P}_t$-constraint indicator functions is equal to the $\mathcal{V}_t$-constraint indicator function.

□

Let $\kappa^{\mathcal{V}}$ be the indicator function for the condition that a pruned genealogy is compatible with a given obscured genealogy. In particular, given an obscured genealogy $\mathcal{V}$, set $\kappa^{\mathcal{V}}_u(t,x,x',\eta,\eta') = \mathbb{1}\{\exists \mathcal{P}$ s.t. $\kappa^{\mathcal{P}}_u(t,x,x',y,y') = 1$ & ob

4. Examples.

4.1. SEIRS.

Jumps: $\mathbb{U} = \{\mathrm{Inf}, \mathrm{Prog}, \mathrm{Recov}, \mathrm{Wane}, \mathrm{Birth}, \mathrm{Death}_{\mathrm{S}}, \mathrm{Death}_{\mathrm{E}}, \mathrm{Death}_{\mathrm{I}}, \mathrm{Death}_{\mathrm{R}}, \mathrm{Sample}\}$.

Demes: $\mathbb{D} = \{\mathrm{E}, \mathrm{I}\}$.

Jump rates:

- $\alpha_{\mathrm{Inf}}(t,x,x') = \beta(t)\,\frac{x^{\mathrm{S}} x^{\mathrm{I}}}{N(t)}\,\mathbb{1}\{x' = x + (-1,1,0,0)\}$
- $\alpha_{\mathrm{Prog}}(x,x') = \rho\,x^{\mathrm{E}}\,\mathbb{1}\{x' = x + (0,-1,1,0)\}$
- $\alpha_{\mathrm{Recov}}(x,x') = \gamma\,x^{\mathrm{I}}\,\mathbb{1}\{x' = x + (0,0,-1,1)\}$
- $\alpha_{\mathrm{Wane}}(x,x') = \upsilon\,x^{\mathrm{R}}\,\mathbb{1}\{x' = x + (1,0,0,-1)\}$
- $\alpha_{\mathrm{Sample}}(t,x,x') = \psi\,x^{I}\,\mathbb{1}\{x' = x\}$
- $\alpha_{\mathrm{Birth}}(t,x,x') = B(t)\,\mathbb{1}\{x' = x + (1,0,0,0)\}$
- $\alpha_{\mathrm{Death}_k}(x,x') = \mu\,x^{k}\,\mathbb{1}\{x'^{j} = x^{j} - \delta_{jk}\}$, $k \in \{\mathrm{S},\mathrm{E},\mathrm{I},\mathrm{R}\}$

5. Discussion.

References.

King, A. A., Lin, Q., & Ionides, E. L. (2022) Markov genealogy processes. *Theoretical Population Biology* **143**:77–91.

Stadler, T. (2010) Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* **267**:396–404.

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., & Frost, S. D. W. (2009) Phylodynamics of infectious disease epidemics. *Genetics* **183**:1421–1430.

A. Filter equations.

Explicit expressions for the probability densities that will arise in the following are not always available. Hence, we will develop some technology for manipulating them that avoids the need for explicit expressions.

Definition 2. Suppose X_t is a continuous-time Markov process with Kolmogorov forward equation (KFE)

$$\frac{\partial w}{\partial t} = \int w(t,x')\,\beta(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,\mathrm{d}x', \tag{8}$$

Suppose that $B(t,x,x') > 0$ and $\lambda(t,x) \in \mathbb{R}$ are given functions. We say that the equation

$$\frac{\partial w}{\partial t} = \int w(t,x')\,\beta(t,x',x)\,B(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,\mathrm{d}x' - \lambda(t,x)\,w(t,x) \tag{9}$$

is the *filter equation* with *driver* X_t, *boost* B, and *decay* λ.

Filter equations afford a convenient means of computing the likelihood of a given sequence of events. This is facilitated by the following

Lemma 1. *Eq. 9 is satisfied by* $w(t,x) = \int_0^\infty v\,u(t,x,v)\,\mathrm{d}v$, *where* u *satisfies the KFE*

$$\begin{aligned}\frac{\partial u}{\partial t} = &\int u(t,x',v')\,\beta(t,x',x)\,\delta(v,B(t,x',x)\,v')\,\mathrm{d}x'\,\mathrm{d}v'\\ &- \int u(t,x,v)\,\beta(t,x,x')\,\delta(v',B(t,x,x')\,v)\,\mathrm{d}x'\,\mathrm{d}v' + \tfrac{\partial}{\partial v}\left[\lambda(t,x)\,v\,u(t,x,v)\right].\end{aligned} \tag{10}$$

Here, $\delta(v,v')$ *is the familiar Dirac* δ.

Proof.

$$\begin{aligned}\frac{\partial w}{\partial t} &= \int v\,\frac{\partial u}{\partial t}(t,x,v)\,\mathrm{d}v\\ &= \int v\,u(t,x',v')\,\beta(t,x',x)\,\delta(v,B(t,x',x)v')\,\mathrm{d}v\,\mathrm{d}x'\,\mathrm{d}v'\\ &\qquad - \int v\,u(t,x,v)\,\beta(t,x,x')\,\delta(v',B(t,x,x')v)\,\mathrm{d}v\,\mathrm{d}x'\,\mathrm{d}v'\\ &\qquad + \int v\,\tfrac{\partial}{\partial v}\left[\lambda(t,x)\,v\,u(t,x,v)\right]\,\mathrm{d}v.\end{aligned}$$

Evaluating the first integral with respect to v, the second with respect to v', and the third by parts, we obtain

$$\begin{aligned}\frac{\partial w}{\partial t} = &\int v'\,u(t,x',v')\,\beta(t,x',x)\,B(t,x',x)\,\mathrm{d}v'\,\mathrm{d}x' - \int v\,u(t,x,v)\,\beta(t,x,x')\,\mathrm{d}v\,\mathrm{d}x'\\ &- \lambda(t,x)\int v\,u(t,x,v)\,\mathrm{d}v,\end{aligned}$$

which is simplified to obtain Eq. 9. □

We recognize in Eq. 10 the KFE of a certain process (X_t, V_t). In particular, X_t is the driver with KFE Eq. 8. The V_t process has jumps wherever X_t does, such that when X_t jumps from x to x', V_t jumps by the multiplicative factor $B(t,x,x')$. Between jumps, V_t decays deterministically and exponentially at rate $\lambda(t,x)$. If we view V_t as a weight, then Lemma 1 says that the V_t-weighted average of X_t evolves according to Eq. 9. This motivates the following result, which effectively allows boosts of zero.

Proposition 3. *Suppose X_t is a continuous-time Markov process with state space $\mathbb{X}$ and KFE as in Eq. 8. Let H_t be its history process. That is, for $\omega \in \Omega$ and $t \in \mathbb{R}_+$, $H_t(\omega) : [0,t] \to \mathbb{X}$ such that, for $t' \in [0,t]$, $H_t(\omega)(t') = X_{t'}(\omega)$. Moreover, there are random variables $K \in \mathbb{Z}_+$ and $t_k \in (0,t]$, $k = 1,\dots,K$ such that $\widetilde{X}_t = X_t$ whenever $t \neq t_k$ for all k. Suppose F is a real-valued function of H_t such that*

$$F(H_t) = \prod_k Q\Big(t_k, \widetilde{H}(t_k), H(t_k)\Big)\, B\Big(t_k, \widetilde{H}(t_k), H(t_k)\Big),$$

for some given measurable functions $B > 0$, $Q \in \{0,1\}$. Suppose w satisfies the filter equation

$$\begin{aligned}\frac{\partial w}{\partial t}(t,x) = \int w(t,x')\,\beta(t,x',x)\,Q(t,x',x)\,B(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,Q(t,x,x')\,\mathrm{d}x' \\ - \int w(t,x)\,\beta(t,x,x')\,\Big(1 - Q(t,x,x')\Big)\,\mathrm{d}x'.\end{aligned}$$

Then $\mathbb{E}\left[F(H_t)\right] = \int w(t,x)\mathrm{d}x$.

Proof. Apply Lemma 1 with the driver generated by the rate functions $\beta(t,x,x')\,Q(t,x,x')$. □

An important special case is that of a deterministic driving process. The following result is established by routine calculation.

Proposition 4. *Suppose $X : [0,T] \to \mathbb{X}$ is a deterministic, piecewise constant, càdlàg function. Let E be the set of its jump times. Then the KFE for X_t is Eq. 8 with $\beta(t,x,x') = \sum_{e\in E} \delta(t,e)\delta(x',X_e)$. With this driver, the filter equation (Eq. 9) becomes*

$$\begin{aligned}\frac{\partial w}{\partial t} &= -\lambda(t,x)\,w(t,x) \quad \textit{for} \quad t \notin E, \\ w(e,x) &= \delta(x,X_e)\int \widetilde{w}(e,x')\,B(e,x',X_e)\,\mathrm{d}x' \quad \textit{for} \quad e \in E.\end{aligned}$$

The results so far allow us to compute expectations over random jumps and jump times. We will have occasion to compute marginal expectations in the situation where some jump times are known. The following result handles this case. [Not sure if this is stated correctly.]

Proposition 5. *Suppose X_t is a continuous-time Markov process with state space $\mathbb{X}$ and KFE as in Eq. 2. Let H_t be its history process and $E_t = \mathsf{event}(H_t)$, the set of its jump times to time t. Suppose $0 < t_1 < \cdots < t_N \leqslant t$ are fixed times and set $O = \{t_1,\dots,t_N\}$. Suppose F is an $\mathbb{R}_+$-valued function of H_t and O such that*

$$F(H_t,O) = \prod_{e\in E_t} Q\left(e,\widetilde{H}(e),H(e)\right)\, B\left(e,\widetilde{H}(e),H(e)\right) \times \prod_{e\in O} R\left(e,\widetilde{H}(e),H(e)\right)\, C\left(e,\widetilde{H}(e),H(e)\right),$$

for some given measurable functions $B, C > 0$, $Q, R \in \{0,1\}$. Suppose w satisfies the filter equation

$$\begin{aligned}\frac{\partial w}{\partial t}(t,x) = \int w(t,x')\,\beta(t,x',x)\,Q(t,x',x)\,B(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,Q(t,x,x')\,\mathrm{d}x' \\ - \sum_{e\in O}\int w(t,x')\,\delta(t,e)\;R(e,x',x)\,C(e,x',x)\,\mathrm{d}x' - \sum_{e\in O}\int w(t,x)\,\delta(t,e)\;R(e,x,x')\,\mathrm{d}x' \\ - \int w(t,x)\,\beta(t,x,x')\;\left(1 - Q(t,x,x')\right)\;\mathrm{d}x'.\end{aligned}$$

Then $\mathbb{E}\left[F(H_t,O)\right] = \int w(t,x)\mathrm{d}x$.

Proof. Apply Lemma 1 with the driver generated by $\beta(t,x,x') = \alpha(t,x,x') + \sum_{e\in O} \delta(t,e)\,\pi(x,x')$. □

A. A. King, Department of Ecology & Evolutionary Biology, Center for the Study of Complex Systems, and Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA

Email address: kingaa@umich.edu

URL: https://kinglab.eeb.lsa.umich.edu/

Q.-Y. Lin, Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM XXXXX USA

E. L. Ionides, Department of Statistics University of Michigan, Ann Arbor, MI 48109 USA