Prior using observed the transmission time distribution and dealing with unobserved hosts

Graham Jones, www.indriid.com

2020-04-24, May 1, 2020

1 Introduction

THIS IS A WORK IN PROGRESS

The aims are:

(1) To use information about the distribution of transmission times to provide a prior for the tree.

(2) To describe a model for "filling in" between observed hosts by allowing a chain of zero or more unobserved hosts between pairs of them. This model is crude, and designed to be easy to implement.

(3) Deal with the fact that there may be no observed host which plays the role of an index case for the rest.

2 Model for tree prior

2.1 Notation

Time is measured backwards, as heights. Heights become smaller towards the present.

Data:

- Q_d is a set of zero or more sampling heights $q_d^{(1)}, q_d^{(2)}, \ldots$, for host d. Q denotes all these sets.
- Y_d is the corresponding set of genetic data for host d.
- Exposure intervals $[i_d, r_d]$. All the Q_d lie in this interval. These may be inferred by some other analysis, but are treated as data by BADTRIP.

Parameters in the model:

- τ is the transmission tree topology. There is one node for each host, and one node is the root (ie index case) d'. The originating host for host d is I_d for $d \neq d'$, and $I_{d'} = \emptyset$. The collection of I_d values contain the same information as τ.
- u is a vector of counts $u_d \geq 0$, for $d \neq d'$, providing the number of unobserved hosts between I_d and d.
- g is a vector of transmission heights g_d, for $d \neq d'$, at which the virus was shed from I_d. Viewed from the point of view of an originating host x, denote by C_x the set of zero or more transmission heights $c_x^{(1)}, c_x^{(2)}, \ldots$, from host x, derived from τ and g. Each $c_x^{(j)}$ is a g_y for some host y, where $x = I_y$.
- h is a vector of infection heights h_d for host d. It is the time at which host d is infected, and for $d \neq d'$ we have $h_d < g_d$.

2.2 Prior

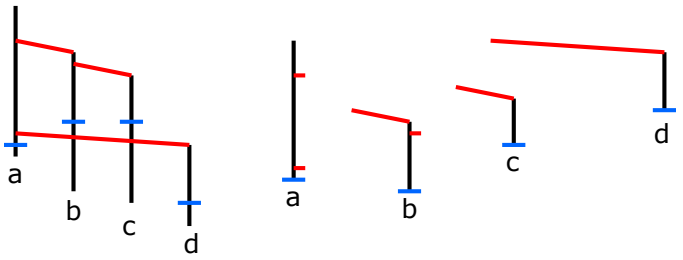

Figure 1: On the left is a small outbreak. Red lines show transmissions, and blue lines show when samples are taken. On the right, the tree is shown decomposed into parts which are assumed to behave independently.

I assume that, given the topology τ and counts u, there is independent behaviour for each host d from the transmission at height g_d to the transmissions from host d at heights in C_d. Figure 1 indicates how a transmission tree is decomposed.

I assume that the prior density for C_d only depends on h_d, not g_d.

I assume that the prior for u is independent of that for τ, and that the u_d are independent of one another.

Then the prior $\Pr(g, h, u, \tau)$ is written as $\Pr(g, h, u, \tau)$

$$\Pr(g, h, u, \tau) = \Pr(g, h|u, \tau) \Pr(u, \tau). \tag{1}$$

In words, this is "the density for heights given the tree structure, multiplied by the probability of the tree structure". The prior probability for τ and u is

$$\Pr(u, \tau) = \Pr(\tau) \Pr(u) = \Pr(\tau) \prod_{d \neq d'} \Pr(u_d), \tag{2}$$

I assume a uniform prior over all possible topologies, so $\Pr(\tau)$ can be ignored in a single MCMC analysis.

Next, $\Pr(g, h|u, \tau)$ is a product of terms, one for each host, as indicated in Figure 1. For non-index cases, the prior density is

$$\prod_{d \neq d'} \Pr(C_d|h_d) \Pr(h_d|g_{I_d}), \tag{3}$$

and the prior density for the index case is

$$\Pr(C_{d'}|h_{d'}) \Pr(h_{d'}). \tag{4}$$

where $\Pr(h_{d'})$ is the prior density for the index case, constrained by $[i_{d'}, r_{d'}]$, and usually diffuse.

For a host d (including d')

$$\Pr(C_d|h_d) = \prod_j f_{tt}(c_d^{(j)} - h_d) \tag{5}$$

where $f_{tt}()$ is the density for the transmission times, based on observed data. See *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing*, Ferretti et al, https://science.sciencemag.org/content/early/2020/04/09/science.abb6936

The remaining pieces of the prior are $\Pr(h_d|g_{I_d})$ and $\Pr(u_d)$. This involves a model for chains of unobserved hosts.

2.3 Chain of unobserved hosts

Consider one host $d \neq d'$. There is a time interval t from g_d to h_d during which u_d unobserved infections occur.

If $u_d = 0$, there is a delay of t, followed by a bottleneck. In this case, there is a density $f_{delay}(t)$ providing the prior for t.

If $u_d > 0$ it is assumed there is a bottleneck followed by u pairs (infection, bottleneck) where each infection has duration following the density $f_{tt}()$. The density for t is then the u_d-fold convolution of f_{tt} times, convolved with the $(u_d + 1)$-fold convolution of f_{delay}.

[For the likelihood, the bottlenecks will be modelled as a drift-only process, and the infections will be modelled as drift-plus-mutation or mutation-only. The intensity of the bottlenecks is independent of the values of t and will probably identical for all the bottlenecks. The infections in one chain of unobserved hosts will all be assumed to be the same duration. For example t might be divided into $(u_d + 1)$ equal durations for delays, and u_d equal durations for infections in proportion to the means of f_{delay} and f_{tt}.]

This is not intended to be very realistic, but keeps things simple by using a single parameter t to model all u_d values. The exact forms of f_{delay} and f_{tt} are to be decided.

The prior $\Pr(u_d)$ for u_d could be a geometric, such as 0.5^{-1-u_d}.

2.4 Unobserved index case

The index case of the true outbreak may or may not be observed. If it is not, there may still be a host among the observed ones which is the source of infection for the others. See Figure 2. An extra unobserved host 0 can be added (a 7th host in scenario (1), a 4th in scenarios (2), (3), and (4)) with an infection height h_0 constrained to be larger than the transmission height g_d of all observed cases d. The properties of the extra host cannot be inferred and will be sampled from the prior subject to the constraint on its infection height.

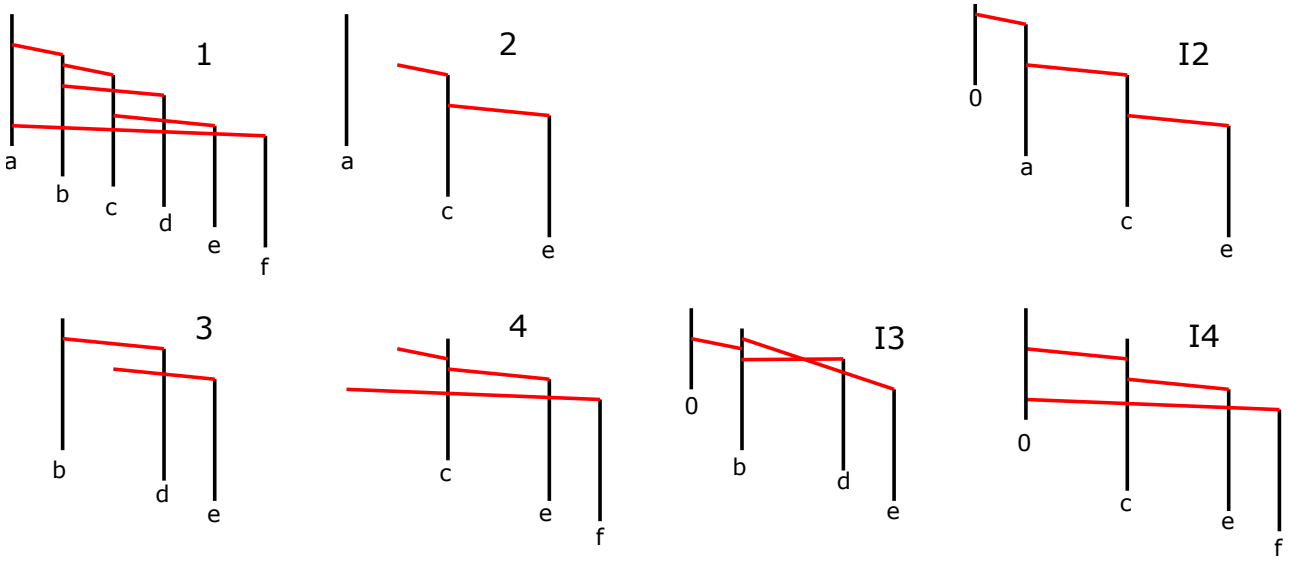

Figure 2: (1) Full outbreak. (2) If a,c,e are observed, the index case is included, and there is an unobserved host between a and c. (3) The index case is not observed, but b plays that role for d and e with an unobserved host between b and e. (4) There are two sources for the observed hosts, one for c and e, another for f. (I2), (I3), (I4) indicate how (2),(3),(4) might be inferred after the addition of an index case, if the inference was as accurate as possible.