Extending the transmission flow model

Fan Bu

March 2020

The original model

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}_{(s_i=1)} \mathbb{1}_{(s_j=1)} \mathbb{1}_{(w_{ij}>\zeta)} \tag{1}$$

$$n_{ab} \sim \text{Bin}(z_{ab}, \xi_a \xi_b) \tag{2}$$

$$z_{ab} \sim \text{Poi}(\lambda_{ab}) \tag{3}$$

Here (2) and (3) give us

$$n_{ab} \sim \text{Poi}(\lambda_{ab} \xi_a \xi_b) \tag{4}$$

The parameters of interest are $\lambda = \{\lambda_{ab}\}$ and $\xi = \{\xi_a\}$.

Under the assumption that transmission is independent of sampling, the joint posterior can be decomposed as

$$p(\lambda, \xi \mid \mathbf{n}, \mathbf{s}, \mathbf{X}) \propto p(\mathbf{n} \mid \lambda, \xi) p(\lambda \mid \xi) p(\xi \mid \mathbf{s}, \mathbf{X}), \tag{5}$$

where $p(\xi \mid \mathbf{s}, \mathbf{X})$ is estimated via (Bayesian) logistic regression, $p(\lambda \mid \xi) = p(\lambda)$ is a 2-d Gaussian Process prior under the 1-year age-band stratification.

One potential concern pertains to the thresholding in (1): the threshold ζ is arbitrary, and such thresholding throws away the information from deep sequencing that (at the very least) gives us probabilities of different transmission directions/relationships.

So instead, we can treat n_{ab} as unobserved as well, without losing much of the simplicity and elegance of the original model.

Proposed extension

Suppose the observed data are "weight" vectors for pairs of individuals that represent the proportions of different phylogenies:

$$\mathbf{w}_{ij} = (w_{i \to j}, w_{j \to i}, w_{i-j}, w_{\text{none}})^T, \tag{6}$$

with the four components summing to 1.

Further assume that the ground truth relationship between individuals i and j is one (and only one) of the three cases: 1) i transmitted to j ("$i \to j$"), 2) j transmitted to i ("$j \to i$"), 3) otherwise ("else"). Then their relationship can be coded by a length-3 one-hot vector:

$$\mathbf{T}_{ij} = (T_{i \to j}, T_{j \to i}, T_{\text{else}}). \tag{7}$$

The observed $\mathbf{w}_{ij}$ can be thought of as a random sample from a distribution that depends on $\mathbf{T}_{ij}$,

$$\mathbf{w}_{ij} \sim g(\mathbf{w}_{ij}; \mathbf{T}_{ij}), \tag{8}$$

for example,

$$\mathbf{w}_{ij} \sim \text{Dir}(\alpha(\mathbf{T}_{ij})). \tag{9}$$

This is equivalent to adopting a 3-component Dirichlet mixture model for all the $\mathbf{w}_{ij}$'s, with the $\mathbf{T}_{ij}$'s as labels of mixture components. [1]

Now, the number of "observed" transmission cases from group a to group b is instead

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}_{(s_i=1)} \mathbb{1}_{(s_j=1)} \mathbb{1}_{(T_{i \to j}=1)}. \tag{10}$$

The rest of the original generative model remains unchanged.

The parameters of interest now include $\lambda = \{\lambda_{ab}\}$, $\xi = \{\xi_a\}$, and the α's. In addition, a collection of latent variables $\{\mathbf{T}_{ij}\}$ are introduced.

The MCMC algorithm only needs minor changes, and the steps in each iteration will be something like:

1. Update the α's conditioned on the $\mathbf{w}_{ij}$'s and $\mathbf{T}_{ij}$'s [2];
2. Update the $\mathbf{T}_{ij}$'s conditioned on the $\mathbf{w}_{ij}$'s and α's;
3. Update the n_{ab}'s according to (10);
4. Do everything else in the original MCMC.

[1] Of course this is the simplest choice; we can allow the α's to vary across individual groups as well to accommodate more heterogeneity.

[2] To avoid label switching issues, some common-sense (or informed-by-domain-expertise) constraints should be imposed on the α's such that, e.g., when $T_{i \to j} = 1$ the component $w_{i \to j}$ should be more likely to dominate.