Adding a spatial element to Outbreaker

CHAPTER 1

Introduction

1. Aims

The aim of this project is to see if group structure or contact structure data can be added to the Outbreaker model by NAME HERE to help us infer information about the transmission dynamics between different groups or to further improve the process of inferring the most likely transmission tree for an outbreak. In certain outbreak scenarios it may be possible to separate the cases of infection into two or more distinct groups, the groups could be defined by what ward a patient is on in a hospital, what class a child is in at school or perhaps whether a case is a child or an adult. An interesting question in this scenario might be whether individuals in the same group are more likely to infect each other and less likely to infect individuals in the other group. We will aim to add a framework to the outbreaker model that allows further outbreak parameters to be estimated which may help us answer this question. Additionally, we will see how we can incorporate contract tracing data into outbreaker and how this can be used to help the statistical inference. In this first section I will introduce the Outbreaker package and go through the previous body of work which led up to the development of Outbreaker.

The Outbreaker package uses a Markov chain Monte Carlo process to try and sample from the posterior probability distribution of various parameters and pieces of augmented data. Specifically, it uses DNA sequence data and case data collected from an outbreak along with a generation time distribution and a time-to-collection distribution to infer the likely infector of each case. These likely ancestors can be combined into a transmission tree and if we are fairly confident in the truth of our assembled transmission tree then we can infer further properties about the outbreak from it such as the rate of mutation of nucleotides in the DNA sequence of the pathogen and the effective reproduction numbers of individuals through time.

2. A Brief History

The Outbreaker model builds upon previous methods that use a similar process of assigning a likelihood value to transmission trees (out of potentially thousands of configurations) and then searching for the tree with the maximum likelihood value or using the likelihood of a tree to sample from the posterior distribution of the trees given the data we have collected. The earliest implementation of this approach was Haydon et al, 2002, who proposed a likelihood function for transmission trees that describe the transmission of foot-and-mouth disease between farms in the UK. Their likelihood function for each transmission event is a product of two independent terms. The first term gives a likelihood value based upon whether the timings of the infection between two farms match well. We can assign a value based upon how well the period farm A was infectious during overlaps with the predicted time

period that farm B was infected during. The better these time periods overlap, the more likely we think that this infection event is. The second term gives a likelihood value based on how far apart the two farms are, seeing as animals on different farms do not freely mix (especially during an outbreak) then there is only the possibility of an aerial infection which is more plausible the closer two farms are. For each tree we can consider the likelihood of all of the separate events in the tree and come to an overall likelihood for the tree, Haydon et al proposed an algorithm which would work towards producing the transmission tree with the highest overall likelihood.

As genetic sequencing became faster and cheaper, Cottam et al could expand this model by including a genetic likelihood term. Now that most infected cases could also provide a DNA sequence of the pathogen then the DNA sequences could be compared to see if they can help figure out the ancestor of each case. Infectious diseases mutate quickly and therefore mutations in the DNA sequences can occur during one generation. We can compare the DNA sequences and produce a likelihood that one case is the ancestor of another case depending on how similar the two sampled DNA sequences are, the more similar two sequences are the more likely that one is an ancestor of the other. Cottam et al proceeded by selecting the transmission tree configurations which had the highest genetic likelihoods and then using the previous epidemiological likelihood (based on infection and collection times) to choose 4 final trees which accounted for 95% of the sum of the likelihood for every possible tree. Further alterations to the approach were made by Ypma et al ,2012, who combined the genetic and epidemiological likelihoods into a single term, therefore removing the assumption that the two likelihoods are independent. This is an important assumption to consider because more mutations will occur in a DNA sequence over longer periods of time so we expect that there will be some correlation between the generation times and the number of mutations that we find.

In the same year Morelli et al (2012) used a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior distribution of transmission trees given the collected data. This technique begins with a tree and then moves the suspected ancestors of each case around according to some rules to form a new tree. The likelihood of both of these trees are calculated and the new tree is accepted a sample with a probability based on the ratio of the two likelihoods. The chain and the movement rules are constructed so that the probability of accepting a tree as a sample is equal to the posterior probability of the tree given the data. We can then look at the trees with the highest posterior probabilities or consider the posterior probability that one case is an ancestor of another.

The Outbreaker model is a combination of these approaches, it uses an MCMC approach with independent genetic and epidemiological likelihoods. It also allows for unsampled cases to occur between two cases and a more complex account of the DNA sequence mutations. Unlike the previous approaches it can also be generalised into a package for the programming language R which means it can be run on personal computers by people with less technical computing skills within a reasonable amount of time. In the next section I will briefly introduce Markov Chain Monte Carlo processes before describing the Outbreaker process in more detail.

3. The Outbreaker Model

3.1. Markov Chain Monte Carlo Processes. Markov Chain Monte Carlo processes are a combination of two statistical tools, the easiest of the two is Monte

Carlo methods. Monte Carlo methods aim to approximate values such as the expected value of a probability distribution by using a large number of samples from that distribution. To work out the expected value of a probability distribution analytically we would integrate over every possible value in the distribution multiplied by the probability of it occurring. Using Monte Carlo methods we would approximate this integration by sampling from the probability distribution thousands of times and taking the mean average of the results. Put simply, Monte Carlo is a way of approximating values of interest given lots of the right samples.

Markov Chains play the role of providing these samples, a Markov Chain is a chain of subsequent states where the next state in the chain is decided by probabilities which are only determined by the current state. A simple discrete Markov Chain may have 3 states called A, B and C, if we are currently in state A then whatever state we move to next only depends on the probability of moving from state A to states B or C. If I move to state B then the fact that I was just previously in state A does not play a role in my decision making about what state I will go to next. If we run certain Markov Chains for long periods of time we will find that they settle into a stationary distribution which determines what the probability of the chain being in each state at each time is. If we have a stationary distribution of interest, such as $p(X = A) = p(X = B) = p(X = C) = \frac{1}{3}$ then we can devise a Markov Chain that will have this distribution as its stationary distribution, run it for a while until it converges upon this steady state, and then use the values in the Markov Chain as samples from our distribution of interest.

We can use the Metropolis-Hastings algorithm to easily find Markov Chains that have a specified stationary distribution, therefore our stationary distribution could be very complex and we could still use a fairly simple Markov Chain to sample from it. We can then use these stationary distribution samples in our Monte Carlo methods to make approximations about the distribution. Metropolis-Hastings can also be used in a Bayesian setting by specifying the stationary distribution as a posterior distribution of interest. This means that instead of having to find a posterior distribution analytically we can instead use an MCMC process to sample from it and then make inferences about it. This is what the Outbreaker model does in the specific context of finding the posterior distribution of transmission trees when given outbreak data.

4. Metropolis-Hastings Algorithm

We can use the Metropolis-Hastings algorithm to sample from our target density π as follows:

- Start with a value X_0 that is from the target density, thus $\pi(X) > 0$.
- We then need a transition density $p(x, y)$ where $p(x, \cdot)$ is the probability density of the next possible states of the Markov chain given that the previous state was x. We then sample a value X_1 from the distribution $p(X_0, \cdot)$.
- We then calculate

$$\alpha(X_0, X_1) = min\left(\frac{\pi(X_1)p(X_1, X_0)}{\pi(X_0)p(X_0, X_1)}, 1\right)$$

- We then generate a random variable $U_1 \sim U[0,1]$, if $\alpha(X_0, X_1) > U_1$ then we accept X_1 as a sample from π. If not then we set $X_1 \leftarrow X_0$ and accept this as a sample.
- We repeat this process for thousands of iterations, saving all of the accepted values in a chain. These values are samples from our target density π.

The need to find the values $\pi(X)$ presents a problem because it requires that we know the target distribution which we are trying to sample from and this might not be the case. We can get around this when looking for posterior densities by substituting in the likelihood function which is usually easier to calculate. If our target density is a posterior density of the form $\pi(X|D)$ with parameter X and observed data D then we can write this as

$$\pi(\theta|D) \propto \frac{\pi(D|\theta) \times \pi(\theta)}{\pi(D)}$$

Since D represents fixed data, $\pi(D)$ is a constant and therefore cancels out in the equation for $\alpha(x, y)$ so we are left with

$$\alpha(X_0, X_1) = min\left(\frac{\pi(D|X_0)\pi(X_0)p(X_1, X_0)}{\pi(D|X_1)\pi(X_1)p(X_0, X_1)}\right)$$

Therefore to use Metropolis Hastings to sample from a posterior distribution we only need to be able to construct the likelihood function and calculate values from the prior densities of our parameters.

5. The Outbreaker MCMC Process

Outbreaker uses the Metropolis-Hastings algorithm to sample from the posterior distribution of transmission trees when we have data on an outbreak. There is a transition density that moves around parameters such as the rate of DNA mutation and then accepts or reject the candidate parameter based on the genetic likelihood defined in the outbreaker model. Additionally outbreaker uses augmented data which are pieces of data that are moved around as if they were parameters and accepted or rejected. In the context of outbreaker each case i has an ancestor α_i; a transition density is used to suggest a new candidate ancestor, the likelihood of this potential ancestry is calculated depending on how well the infection time, spatial and DNA sequence data fit together between the cases. In the next section I will introduce the new group structure data and likelihood and describe how it fits into the current outbreaker model.

CHAPTER 2

Methods

1. Group Data and Parameters

In certain outbreak scenarios individuals might belong to two or more different groups, people in these groups could have different levels of contact between members of their own group and members of other groups. This could lead to different rates of transmission events within and between different groups as a whole. One example of this could be groups of patients on different wards of a hospital, if someone on one ward falls ill it seems plausible that they are more likely to transmit this infection to another patient on their own ward rather than a patient on a different ward. If the outbreak spreads through several wards we could potentially use our knowledge of what ward people are on to assess the probability that one patient infected another and help us construct a plausible transmission tree. Another example could be to separate cases into groups based on their age, several studies have looked into how rates of common cold transmission differ from child to child, child to adult, adult to child and adult to adult [CITE THE STUDIES PLZ].

We can represent these differing rates of transmission between l groups as parameters in an l by l transmission rate matrix where the element λ_{12} is the rate of transmission from group 1 to group 2.

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1l} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{l1} & \lambda_{l2} & \cdots & \lambda_{ll} \end{pmatrix}$$

We can then normalise these rates to give us a transmission probability matrix where

$$P_{ij} = \frac{\lambda_{ij}}{\sum_{k=1}^{l} \lambda_{ik}}$$

is the probability that a newly infected case in is group j given that the ancestor is in group i.

To introduce these parameters into the MCMC process we need to propose a transition density (also known as a "move") that will produce candidate transmission rates and a likelihood function that will be used to accept or reject candidate rates depending on how well they fit the data (in this case the data is a record of what group each case is in).

2. Group Likelihood

The existing pseudo-likelihood function of the outbreaker model is composed of the product of the genetic and epidemiological likelihoods. We need to construct a group likelihood function to multiply onto the existing likelihood function, the

group likelihood function will give us a likelihood value for the probabilities in our transmission matrix given the observed and augmented data. Some of the observed and augmented data will not play a role in determining how likely the transmission probabilities are, such as the genetic sequences and timings of each case. The data that will play a role are α_i, the index of the ancestor of case i (and their corresponding group) and κ_i, the number of generations between i and α_i.

To write the group likelihood we can think of the transmission event between case i with group g_i and its immediate ancestor α_i with group $g_{\alpha i}$ as a Bernoulli trial with $P_{g_i g_{\alpha_i}}$ chance of succeeding. We know the group membership of each case and the current candidate ancestor α_i, therefore the likelihood of a transmission event between i and α_i for the candidate transmission rates matrix, denoted L, is given by:

$$\Omega_i^3 = p(g_i = x | \alpha_i, \kappa_i, g_{\alpha_i} = y, L)$$

and the likelihood of a Bernoulli trial is the probability that the event takes place, which can be found by calculating the transmission probability matrix, so for this particular case the likelihood is:

$$\Omega_i^3 = P_{xy}$$

In the Outbreaker model each transmission event is assumed to be independent of other events, therefore the group likelihood for a whole transmission tree is the product of all of the individual likelihoods for each transmission event (or the sum of the group log likelihoods).

$$\Omega^3 = \prod_i \Omega_i^3 = \prod_i P_{g_i g_{\alpha_i}}$$

This likelihood term is multiplied onto the existing likelihood term to give an overall likelihood for a case: $\Omega_i^1 \times \Omega_i^2 \times \Omega_i^3$. This assumes that the group, epidemiological and genetic likelihoods are all independent. This assumption simplifies the likelihood term but in real outbreak data we would expect to see some correlation between the likelihood terms. For example, if transmission rates really were higher within a group than between other groups we might expect that observed DNA sequences are generally more similar between cases in the same group because mutations that occur between two cases in the same group are more likely to stay within that group, therefore distinguishing the DNA sequences from these cases from those belonging to other groups.

3. Transmission Rate Matrix Move

The parameters in the transmission rate matrix are moved one element at a time except for one element on each row which is set to 1 and does not move. The other elements in the row should converge upon a value which is the rate of transmission relative to the constrained rate. New candidate elements are generated from a lognormal distribution:

$$X_{n+1} \sim lognormal(\log(X_n), \sigma)$$

Where σ is determined by a tuning process which increases or decreases its value based upon the proportion of moves being accepted and rejected. This move is symmetrical so the probability of moving from a value y to another value x has the same probability of moving from the value x to the value y. This means that for

the transition density $p(x, y) = p(y, x)$ so these values cancel each other out in the function $\alpha(x, y)$.

CHAPTER 3

Results

CHAPTER 4

Discussion