Approximate Bayesian Computation for Disease Outbreaks

Sohum Mehrotra, Aining Wang, Zhongwei Peng

Github Repository (https://github.com/nuknuk48/STAT-S610_Project)

Backgrounds

In biological sciences, for many complex problems (e.g. population genetics, ecology and systems biology), it's hard to achieve closed forms of models. Under these circumstances, computer simulations have become a popular method to resolve these problems. In our project, epidemiologists have developed models for disease spread using the real data, which is describing influenza A (H3N2 & H1N1) and influenza B outbreaks in different years and locations. However, the evaluation of the likelihood is computationally intractable and suitable probability models do not exist. The researchers have to find a way to overcome these problems. One commonly used strategy for fitting parameters in models is approximate Bayesian computation (ABC), which is a simulation-based method.

Introduction

In Approximate Bayesian Computation, given a prior distribution $P(\theta)$ of parameter θ and the likelihood function $f(D_0|\theta)$, the goal is to approximate the posterior distribution, $P(\theta|D_0) \propto f(D_0|\theta) * P(\theta)$. The general steps are the following:

(1) Sample the parameter θ^* from prior distribution
(2) Simulate the dataset D^* using distribution $f(D^*|\theta^*)$
(3) Compared the simulated dataset D^* with the observed dataset D_0 using the distance function d and tolerance ϵ to accept or discard the parameter θ^*
(4) Run this process multiple times to get a posterior distribution for parameter θ^*

Project Goals

In our project, our goal is to address whether:

- Different outbreaks of the same strain can be described by the same model of disease spread
- Outbreaks of different molecular strains of the influenza virus can be described by the same model of disease spread
- Reproduce the Figures 3a and 3c from Toni and Stumpf's paper

Methodology

We are trying to generate the posterior distribution for parameters q_c^* and q_h^*, where q_c^* is the probability that a susceptible individual does not get infected from the community, and q_h^* is the probability that a susceptible individual does not get infected within the household. And the ABC function involves the following elements:

- Choose uniform prior distributions for all parameters and model selections

- Use the probability equations from Toni and Stumpf's paper to simulate datasets
- Use the Frobenious norm to build up the distance function

Code Review

We will build two models to address the questions. The first model contains two datasets, $H3N2_{78}$ and $H3N2_{81}$, which are influenza A outbreaks from 1977 to 1978 and from 1980 to 1981 in Tecomseh, Michigan. And the second model contains two datasets, B_{76} and $H1N1_{79}$, which are influenza B outbreaks in 1975-1976 in Seattle and influenza A (H1N1) outbreaks in 1978-1979 in Seattle.

```r
##Model 1
H3N2_78<- matrix(data = c(66,87,25,22,4,13,14,15,9,4,0,4,4,9,1,0,0,4,
                          3,1,0,0,0,1,1,0,0,0,0,0), ncol = 5, byrow=TRUE)

H3N2_81<- matrix(data = c(44,10,0,0,0,0,62,13,9,0,0,0,47,8,2,3,0,0,38,
                          11,7,5,1,0,9,5,3,1,0,1), ncol = 5)

##Model 2
B_76 <- matrix(data= c(9,1,0,0,0,0,12,6,2,0,0,0,18,6,3,1,0,0,9,4,
                       4,3,0,0,4,3,0,2,0,0), ncol=5)
H1N1_79 <- matrix(data= c(15,11,0,0,0,0,12,17,21,0,0,0,4,4,4,5,0,0),
                  ncol = 3)
```

The Data_generator_helper function simulates the probability matrix W using the probability equations below and return a simulated count matrix, which takes three variables below:

- qlist: a list of randomly generated uniform probabilities of q_c^* and q_h^* for two datasets in each model
- susc: the number of susceptible individuals in a household
- index: helps us select which probabilities from qlist to use
- w_{js}: the probability that j out of the s susceptibles in a household become infected

$$w_{js} = \binom{s}{j} w_{jj} (q_c q_h^j)^{s-j}$$

$$w_{0s} = q_c^s \ for \ s = 0, 1, 2, \ldots$$

$$w_{jj} = 1 - \sum_{i=0}^{j-1} w_{ij}$$

```r
Data_generator_helper <- function(qlist, susc, index, data){
  w <- matrix(0, nrow = 6, ncol = susc)
  w[1,] <- sapply(1:susc, function(n,theta){theta^n}, qlist[index])
  ##populating 1st row w0s: w01,w02,w03,w04,w05
  w[2,1] <- 1 - w[1,1]
  ##populating w11
  for(s in 2:susc){ ##populating rows
     for(j in 2:6){
       if(j<=s){
         w[j,s] <- choose(s,j-1)*w[j,j-1]*(qlist[index] * (qlist[index-1]^(j-1)))^(s-j+1)
       }
       else{
         w[j,s] <- 1-sum(w[,s])
         break
       }
     }
  }
  ## size needs to be the sum of the column of the original data
  ## simulate the count matrix by using the simulated probability matrix W
  initSize <- colSums(data)[1]
  countMatrix <- rmultinom(prob=w[,1],size=initSize,n=1)
  for(col in 2:susc){
    nextSize <- colSums(data)[col]
    countCol <- rmultinom(prob=w[,col],size=nextSize,n=1)
    countMatrix <- cbind(countMatrix, countCol)
  }
  return(countMatrix)
}
```

The Data_generator function returns the results of the data_generator_helper function for a selected parameter from qlist. There will be 4 probabilities generated into qlist. These are denoted $q_{c1}, q_{h1}, q_{c2}, q_{h2}$ and will be used to answer our questions listed under **Project Goals**.

If each outbreak has its own unique characteristics, then it can be described by a model that has all four parameters from qlist. On the other hand, if outbreaks of different strains can be described by the same model, then they should share the same epidemiological parameter values.

```
Data_generator <- function(qlist, susc1, susc2, data1, data2) {

  c1 <- Data_generator_helper(qlist, susc1, 2, data1)
  c2 <- Data_generator_helper(qlist, susc2, 4, data2)
  return(list("c1" = c1, "c2" = c2))
}
```

To apply ABC, we need a test statistic to qualify a sample for acceptance or rejection. We use the Frobenious norm to build up the distance function. The Distance function takes in observed data and simulated data for each model, and calculates a test statistic for acceptance or rejection of a sample.

- Use the Frobenious norm to build up the distance function:

$$d(D_0, D^*) = \frac{1}{2}(||D_1 - D^*(q_{h1}, q_{c1})||_F + ||D_2 - D^*(q_{h2}, q_{c2})||_F)$$

$$||A||_F = \sqrt{trace(A^T A)} \text{ which denotes the Frobenious norm of A}$$

```
library(matrixcalc)
Distance <- function(data1, data2, generated_data) {
  return((1/2) * (frobenius.norm(data1 - generated_data$c1) +
                  frobenius.norm(data2 - generated_data$c2)))
}
```

Finally, the ABC function pulls all our previous functions together to output lists of parameters. The inputs are:

- epsilon: a chosen value which is used to accept or reject a simulated sample based on its Distance calculation
- n_samples: the number of times we want the function to repeat
- data1 and data2: observed datasets for each model

```
ABC <- function(epsilon, n_samples, data1, data2) {
  parameters <- list()
  i <- 1
  for(j in 1 : n_samples) {
    qlist <- runif(4) ##qc1 qh1 qc2 qh2
    data <- Data_generator(qlist,ncol(data1),ncol(data2),data1, data2)
    distance <- Distance(data1, data2, data)

    if(distance <= epsilon){
      parameters[[i]] <- qlist
      i = i + 1
    }
  }
  return(parameters)
}
```

We tried several epsilon values and decided on an epsilon value of 50 and 15 being the most favorable.

```
m1 <- ABC(50, 10000, H3N2_78, H3N2_81)
m2 <- ABC(15, 10000, B_76, H1N1_79)

m1 <- as.data.frame(m1)
m1 <- t(m1)
rownames(m1) <- NULL
colnames(m1) <- c("qc1","qh1","qc2","qh2")

m2 <- as.data.frame(m2)
m2 <- t(m2)
rownames(m2) <- NULL
colnames(m2) <- c("qc1","qh1","qc2","qh2")
```

Results

```
plot(m1[,2], m1[,1], col=c("red"), xlim=c(0,1), ylim=c(0,1),
     xlab = "qh",ylab="qc", main = "Different Outbreaks of the Same Strain of the Influenza Virus ")
points(m1[,4], m1[,3], col=c("blue"))
```

```
plot(m2[,2], m2[,1], col=c("red"), xlim=c(0,1), ylim=c(0,1),
          xlab = "qh",ylab="qc", main = "Outbreaks of Different Strains of the Influenza Virus")
points(m2[,4], m2[,3], col=c("blue"))
```

Discussion

For Model 1, we try to answer whether different outbreaks of the same strain can be described by the same model of disease spread. The results obtained are summarized in Figure 1. This figure, unsurprisingly, shows strong evidence in favor of the two-parameter model. And the figure strongly suggest that two outbreaks of the same strain appear to have shared the same epidemiological characteristics.

For Model 2, we try to answer whether outbreaks of different molecular strains of the influenza virus can be described by the same model of disease spread. The results obtained are summarized in Figure 2. This figure shows strong evidence in favor of the four-parameter model. And the figure strongly suggest that two outbreaks of different viral strain have different epidemiological characteristics.

Reference

Toni and Stumpf, “Simulation-based model selection for dynamical systems in systems and population biology”, Bioinformatics (2010) (available at https://academic.oup.com/bioinformatics/article/26/1/104/182571)

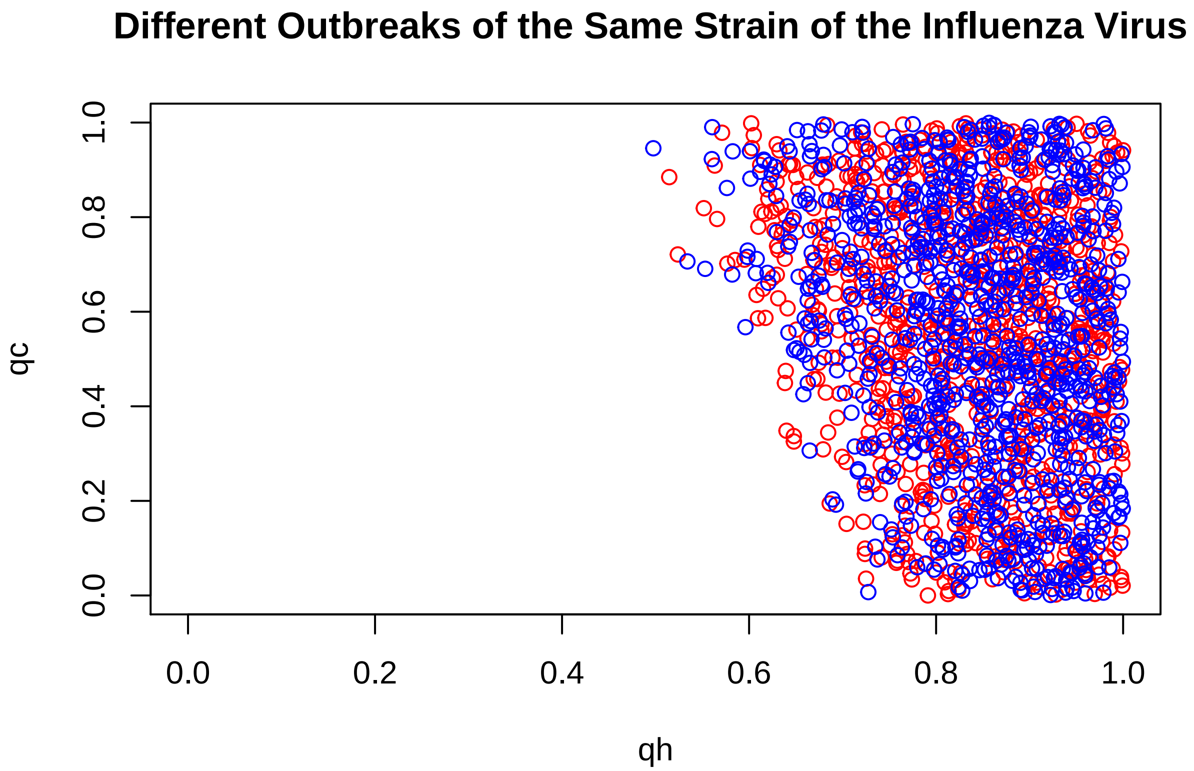

Figure 1: Model 1 Results

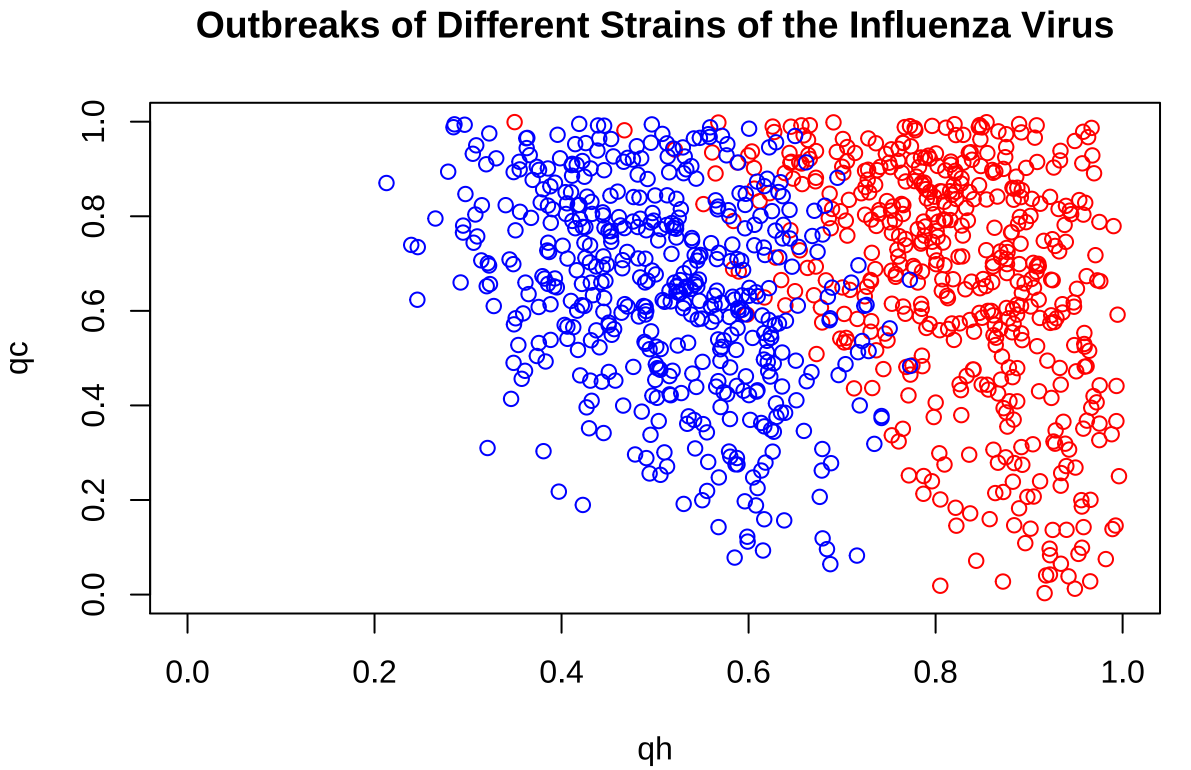

Figure 2: Model 2 Results