ABC Project Write-up

Austin Lesh, Corey Maxedon, Joe Stoica

December 12, 2019

Our objective in this project was to reproduce Figures 3(a) and 3(c) in Toni and Stumpf, "Simulation-based model selection for dynamical systems in systems and population biology", Bioinformatics (2010). More concretely, this looks like modeling the spread of disease: for two influenza outbreaks, we use Approximate Bayseian Computation (ABC) to model the probability that an invidual isn't infected by anyone in their community and the probability that they aren't infected by anyone in their household.

After converting the data in the paper supplement to .csv, we read it into R and dropped the first column (row indices). Each table was then put through a function called `divfun`, which converted the raw data of infected individual counts into proportions, which are usable for our purposes (i.e. probability calculations).

We then set about defining our priors for the four parameters (`qc1`, `qh`, `qc2`, `qh2` – a `qc` and `qh` for each of the Washington and Michigan outbreak data) and our model for data generation. The priors are simply four uniform random variables in the interval [0, 1]. The data generation model, which we call `model (1)`, calculates the probabilities of an outcome (`j` number of individuals infected out of `s` susceptible individuals), defined as follows:

q_c = the probability that a susceptible individual does not get infected from the community

q_h = the probability that a susceptible individual escapes infection within their household

w_{js} = the probability that j out of the s susceptibles in a household become infected, is given by

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

where $w_{0s} = q_c^s$ for $s = 0, 1, 2, \ldots$ and $w_{jj} = 1 - \sum_{i=0}^{j-1} w_{ij}$.

Approximate Bayesian Computation also requires a test statistic for acceptance or rejection of a sample. We use the distance function given in Toni and Stumpf:

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

where

$$||A||_F = \sqrt{trace(A^T A)}$$

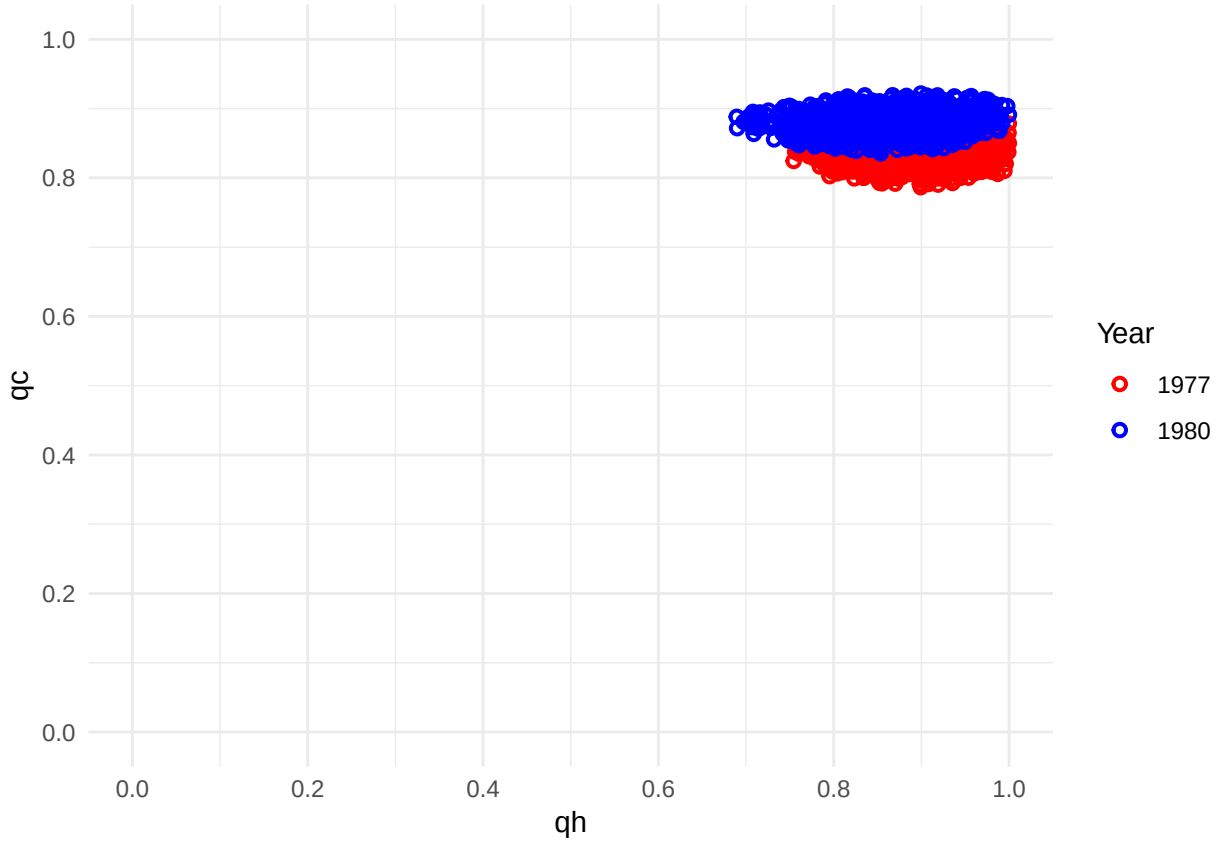
denotes the Frobenious norm of A. We also have:

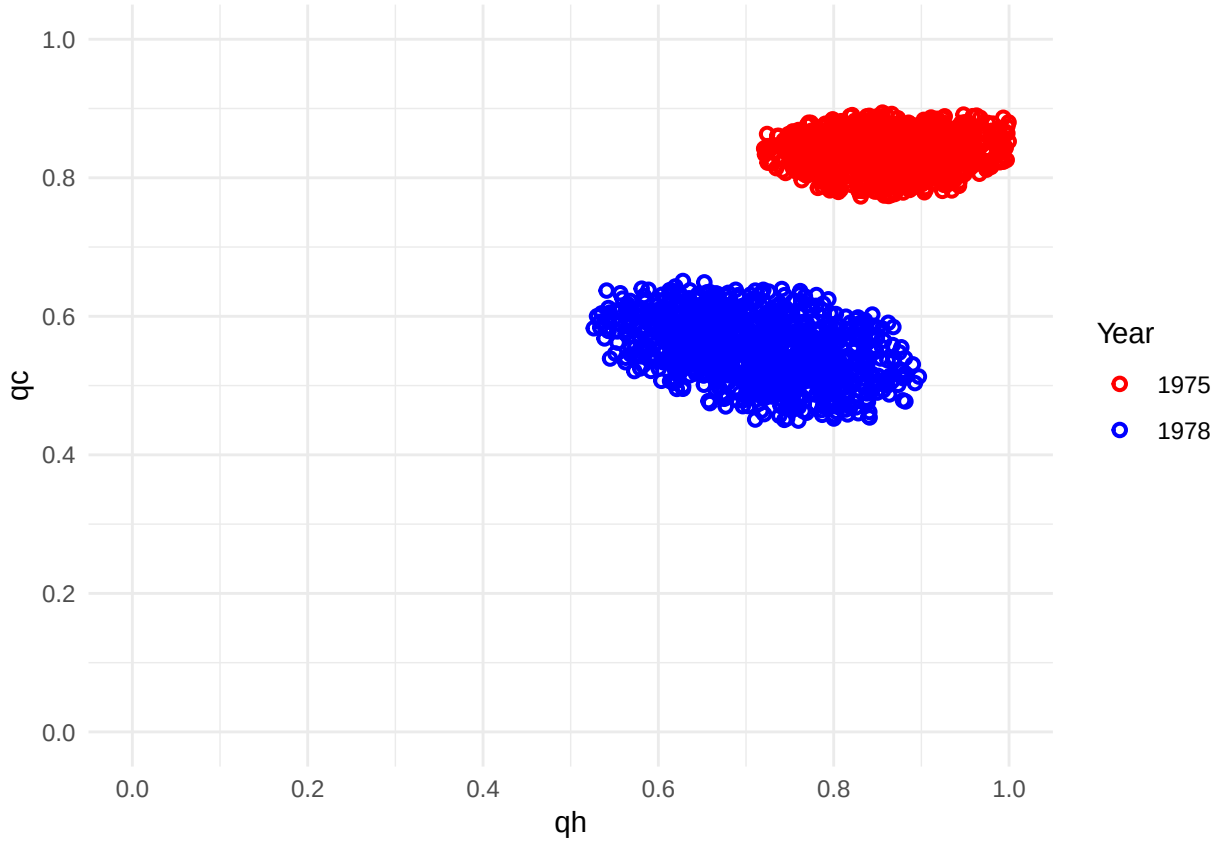
- $D_0 = D_1 \cup D_2$, with
- D_1 being the 1977-98 outbreak
- D_2 being the 1980-81 outbreak
- D^* is the simulation output from model (1).

We coded this function (and the underlying frobenious norm function) into R and used it later to accept or reject a sample if the distance was less than our tolerance level `epsilon`. `Epsilon` itself came from trial and error using the `test_epsilon` function, and we used $\epsilon = 0.3$ when it produced results similar to the plots in the original paper. Allowing a larger epsilon would admit more ABC samples, thus spreading out our posterior data, so we tried to keep it fairly small (0.001 percentile of 1,000 random distances).

Finally, we arrive at our `generate_abc_sample` function which does what its name implies. This function combines the prior distributions, data generating function, and distance function and performs the ABC

algorithm as described in class. We ran this function 1,000 times each for the 1977/1980 data and the 1975/1978 data. The resulting (`qh`, `qc`) pairs were plotted in `ggplot2` to recreate tables 3(a) and 3(c) from the Toni and Stumpf paper, respectively. They can be found below:





Because these match the plots in the original paper, we can assume we have successfully recreated an equivalent posterior distribution.