Epidemiological Data: Parameter Estimation and Pitfalls

Jonathan Waring, Ana Bento, and Pejman Rohani
Odum School of Ecology, University of Georgia, Athens, GA 30606

Introduction

During an outbreak of an emerging infection, there are many epidemiologically important quantities that we need to estimate in order to get a better picture of how severe the epidemic may turn out to be. These include quantities such as the mean infectious period and the transmission potential of the pathogen. Fitting transmission models to incidence reports has become a standard way of achieving quick estimates of these parameters. Practitioners often use cumulative data (total number of infections to date) rather than raw incidence (number of new cases in a defined reporting period). There is evidence to suggest this choice can severely affect our perception of the variability in parameters and hence the certainty in predictions [1]. This project is focused on further elaborating on this problem using simulated epidemic data.

Objective

The aim is to use both deterministic and stochastic models to fit both raw and cumulative data in order to systematically assess the likely biases and errors that result from the choice of data.

Methods

For the purposes of this project, we have encoded a Susceptible-Infected-Recovered (SIR) model as a partially observed Markov process using the R package pomp [2]. POMP models consist of a hidden, stochastic state process that is connected to some set of data via an explicit model of the observation process.

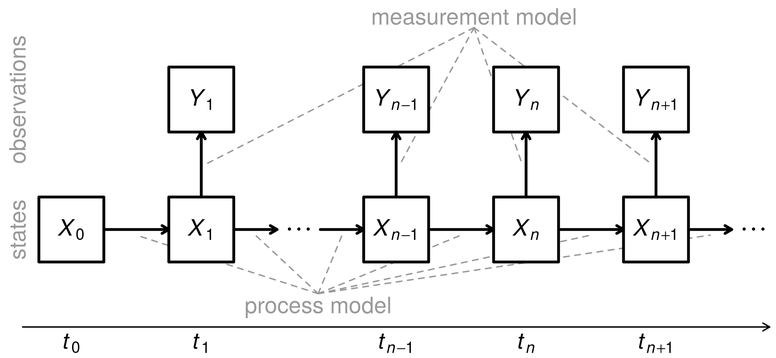

Figure 1: POMP model schematic, showing dependence among model variables.[2]

The state process, X_n, is Markovian, and thus its probability is measured as:

$$Prob[X_n|X_0,\ldots,X_{n-1}, Y_1,\ldots Y_{n-1}] = Prob[X_n|X_{n-1}]$$

The measurement process, Y_n, depends only on the state at the current time:

$$Prob[Y_n|X_0,\ldots,X_n, Y_1,\ldots Y_{n-1}] = Prob[Y_n|X_n]$$

Both for all n = 1,…, N

25 simulated epidemic time-series data sets were generated using the pomp model with key parameters (i.e. Beta (transmission term), gamma (recovery rate), rho (reporting probability), etc.) set at pre-defined values. Both raw and cumulative data were then fit using a deterministic method, trajectory matching, and a stochastic method, iterated filtering.

Trajectory Matching

Trajectory matching is the method of fitting a deterministic model to data assuming independent errors. In the pomp package, the function traj.match searches parameter space to find parameters under which the likelihood of the data, given a deterministic skeleton, is maximized.

Iterated Filtering

A stochastic method for maximizing likelihood of parameters of a partially-observed Markov process is iterated filtering. The iterated filtering algorithm [3] runs a particle filter (sequential Monte Carlo algorithm) at each iteration on a perturbed version of the model, which effectively smooths the likelihood surface.

Results

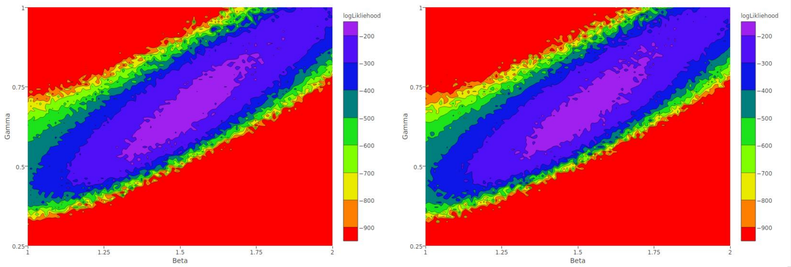

Figure 2: Raw (left) and Cumulative (right) log likelihood contour maps over Beta (transmission term) and Gamma (recovery rate).

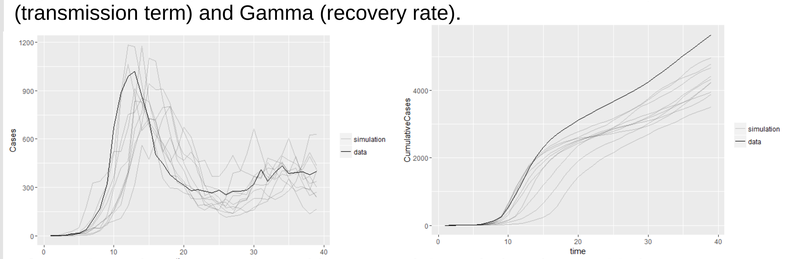

Figure 3: Trajectory matches with Raw and Cumulative data at 38 timestamps.

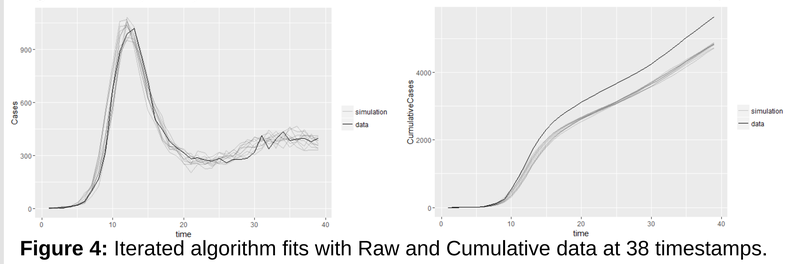

Figure 4: Iterated algorithm fits with Raw and Cumulative data at 38 timestamps.

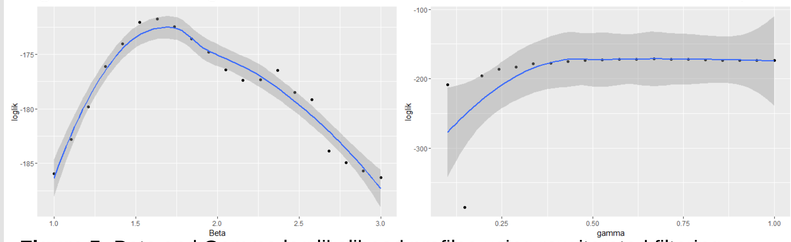

Figure 5: Beta and Gamma log likelihood profiles using raw iterated filtering.

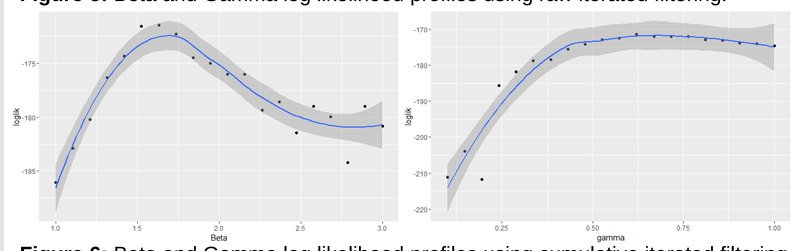

Figure 6: Beta and Gamma log likelihood profiles using cumulative iterated filtering.

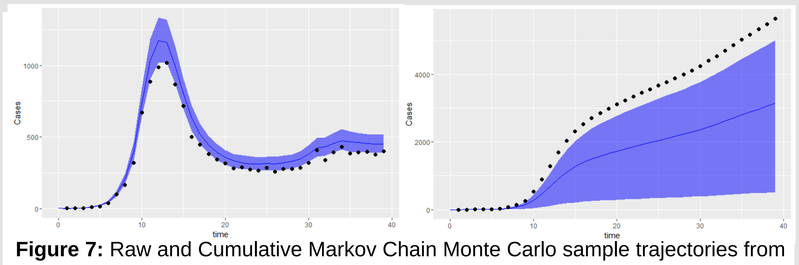

Figure 7: Raw and Cumulative Markov Chain Monte Carlo sample trajectories from posterior distribution with 95% confidence interval at each point.

Conclusions

When one uses deterministic likelihood estimation using the evaluation of the observation model given certain parameter values, it can be seen that raw and cumulative data give very similar results (**Figure 2**). It appears that even when one includes the entire time-series of an epidemic, not just the takeoff, that fitting deterministic models to raw data and cumulative data will quantify uncertainty equally, but cumulative will often under predict the true incidence. Raw models will fit very well to the data (**Figure 3**). Iterated filtering algorithm fits match more closely with the data. However, there is less variance within the simulations, and possibly under quantifies uncertainty in parameter values. Cumulative data is still under predicting the true incidence (**Figure 4**).
The raw beta profile fits show that 65% of points fall in the 95% confidence interval, and the gamma profile fits show 90% of points fall in 95% confidence interval (**Figure 5**). There is a wider confidence interval for the beta and gamma cumulative profile fits, with 70% beta coverage and 80% of gamma coverage (**Figure 6**). When sampling from the posterior using particle Markov chain Monte Carlo, it can be seen that the raw incidence data has a much narrower confidence interval, and thus fails to quantify uncertainty, but cumulative incidence does not match the data as closely (**Figure 7**).

In conclusion, when one considers the entire time-series of an epidemic, cumulative incidence will quantify uncertainty in parameter values better than raw data, but you will get closer fits and closer counts to actual incidence with raw data.

Acknowledgements

This work was sponsored by the NSF Population Biology of Infectious Diseases REU program at the University of Georgia. This material is based upon work supported by the National Science Foundation (NSF award no. 1156707) and the National Institutes of Health (NIH award no. 1U01GM110744-01).

References

1. King AA, de Celles MD, Magpantay FMG, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. Proceedings of the Royal Society B: Biological Sciences. The Royal Society; 2015;282: 20150347–20150347. doi:10.1098/rspb.2015.0347
2. King AA, Ionides EL, Bretó CM, Ellner SP, Ferrari MJ, Kendall BE, et al. pomp: Statistical inference for partially observed Markov processes [Internet]. 2016. Available: http://kingaa.github.io/pomp
3. Ionides EL, Nguyen D, é YA, Stoev S, King AA. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences; 2015;112: 719–724. doi:10.1073/pnas.1410597112