Epidemiological Data: Parameter Estimation and Pitfalls

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

Introduction

During an outbreak, epidemiologically important parameters need to be quantified and estimated to better understand and potentially put in place timely response strategies. 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 attaining quick real-time estimates of these parameters. Cumulative incidence data (total number of infections to date) is often used rather than raw incidence (number of new cases in a defined reporting period). Evidence suggest this choice can critically affect our perception of the variability in parameters and hence the uncertainty in predictions [1]. This project focuses on further elaborating on this problem using simulated epidemic data.

Objective

- Fit deterministic and stochastic models to raw and cumulative data
- Systematically assess the biases and errors that result from the choice of data

Methods

- Susceptible-Infected-Recovered (SIR) model
- Encoded as a partially observed Markov process (POMP) 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.

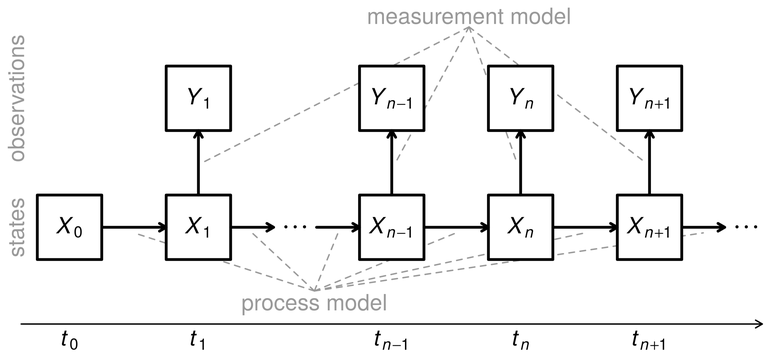

Figure 1: POMP model schematic, showing dependence among model variables.[2]

- 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

- Parameters used: β (transmission term), γ (recovery rate), ρ (reporting probability)
- Trajectory matching and iterated filtering were fitted to both raw and cumulative data

Trajectory Matching

- Fits a deterministic model to data assuming independent errors.

Iterated Filtering

- Stochastic method for maximizing likelihood of parameters of POMP process

Particle Markov Chain Monte Carlo

- Full-information Bayesian inference using pMCMC.

Results

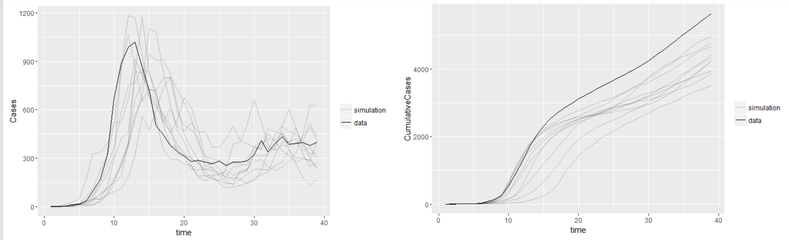

Figure 2: Trajectory matches with Raw and Cumulative data.

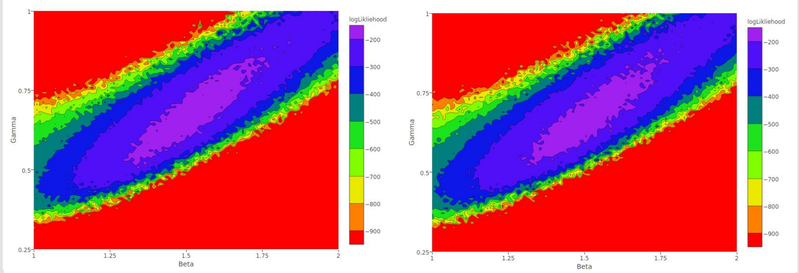

Figure 3: Raw (left) and Cumulative (right) log likelihood contour maps over β (transmission term) and γ (recovery rate).

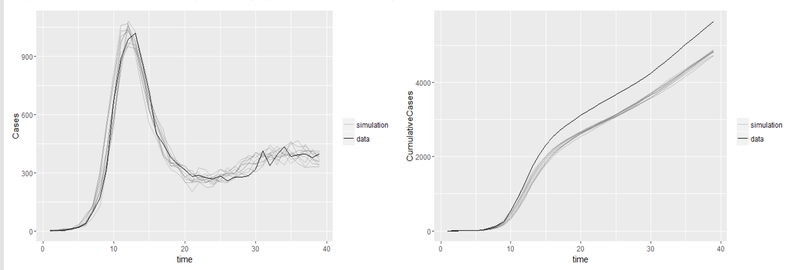

Figure 4: Iterated algorithm fits with Raw and Cumulative data.

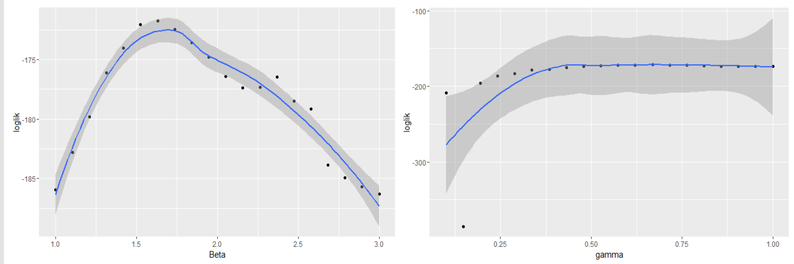

Figure 5: β and γ log likelihood profiles using raw iterated filtering.

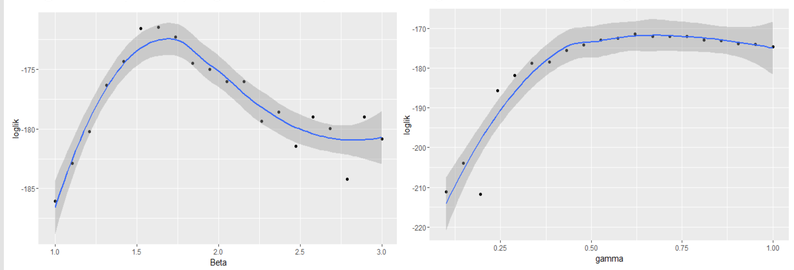

Figure 6: β and γ log likelihood profiles using cumulative iterated filtering.

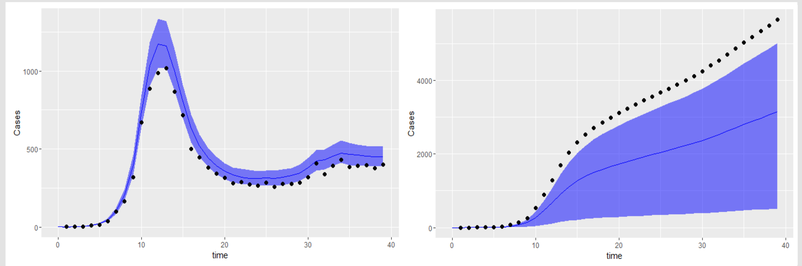

Figure 7: Raw and Cumulative particle Markov Chain Monte Carlo (pMCMC) sample trajectories from posterior distribution with 95% confidence interval at each point.

Discussion

- When we consider the entire time-series of an epidemic, cumulative and raw data will both be useful in parameter estimation depending on how much uncertainty we are willing to accept.

- Fitting deterministic models to raw and cumulative data will result in comparable variance in the trajectory-matched simulations (**Figure 2**)
- Cumulative will often under predict the true incidence, whereas raw models will fit very well to the data (**Figure 2**).
- Deterministic likelihood estimations are similar for both raw and cumulative data (**Figure 3**).
- Iterated filtering algorithm is a better fit to the data when comparing with **Figures 2 and 4**.
- 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**).
- β profiling for the raw data show 65% of points falling in the 95% CI (**Figure 5**).
- γ profiling for the raw data show 90% of points falling in 95% CI (**Figure 5**).
- We observe wider confidence intervals for the β and γ profiles using cumulative data, with 70% and 80% of points falling in the 95% CI respectively (**Figure 6**).
- Posterior sampling with pMCMC shows raw incidence data having a much narrower confidence interval than cumulative data (**Figure 7**).
- B and γ estimations using pMCMC for both raw and cumulative data generated similar results (plots not shown)

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