**REU Poster Guide**

**Quick Pitch:** During an outbreak of an emerging infectious disease, such as the recent Zika virus epidemic in South America or Ebola in West Africa, questions about how long an epidemic will last or how many individuals will get infected immediately begin to be asked. In order to answer these questions, we need to estimate epidemiological parameters, such as transmission potential and mean infectious period, for a timely and effective public health response. So how do we do this? The standard practice has become to fit transmission models to incidence data. However, there are two types of incidence data: cumulative, which is the total number of cases to date, and raw, which is the number of cases in a defined reporting period. This begs the question of which type of data should public health practitioners be using to estimate these critically important parameter values. We aim to answer this question by fitting both deterministic and stochastic models to raw and cumulative data in order to assess the potential biases or errors associated with the choice of data. Using the classical Susceptible-Infected-Recovered (SIR) model, we generated 25 realizations of simulated epidemic data, and used different techniques for parameter estimations and fitting simulated-models to the data. We find that when we consider the entire time-series of an epidemic, cumulative and raw data are both be useful in parameter estimation depending on the level of uncertainty we are willing to accept. Cumulative data often under predicts the true incidence values in its simulated fits, but produces wider confidence intervals in parameter estimations. On the other hand, raw data produces much closer simulated model fits, but narrower confidence intervals.

**SIR Model:** Simple compartmental model for infectious disease transmission. Each person in population is either **S**usceptible, **I**nfected, or **R**ecovered. People are able to transition S🡪I and I🡪R. S🡪I involves disease transmission, dependent upon prevalence of infected, contact rates within a population, and probability of disease transmission given contact. Usually quantified by β, the transmission term. I🡪R involves recovery from infection. It is generally observed that the amount of time spent in the infectious class is distributed around a mean value, and from the modeling perspective we quantify γ as the recovery rate (inverse of infectious period). Our model also incorporates demography with births and deaths, as quantified by μ. We also add in some process noise by incorporating random Gamma white-noise in the transmission term. Formulized as a series of differential equations with respect to time.

**POMP:** Partially observed Markov processes consist of an unobserved (hidden) stochastic state process (process model), connected to the data via a model of the observation process (measurement model). Both models are a probability distribution. The process model is Markovian, meaning that the conditional probability of the future states of the process only depends upon the present state, and not on the sequence of events that preceded it. The measurement process depends on the state at the current time. Two operations to perform on the models: simulate (draw random sample from the distribution) or evaluate the density. rprocess is the simulation of process model, which is the stochastic user defined SIR model explained above, dprocess not used, rmeasure is simulation of observation model (we use a Poisson model using reporting probability and real cases state), and dmeasure is evaluation of observation model (Poisson density function).

**Trajectory Matching:** Trajectory matching is the method of fitting a deterministic model to data assuming independent errors (no correlation between consecutive errors in time series data). Trajectory matching uses a deterministic skeleton, which is simply the deterministic differential equations from our SIR model. Trajectory matching then searches parameter space to find parameters under which the likelihood of the data, given the deterministic skeleton, is maximized. Thus only uses the skeleton and dmeasure components of our POMP model. Ascribes all stochasticity to measurement error, given that it excludes process noise.

**Iterated Filtering:** Iterated filtering is a stochastic method of estimating parameters of a partially observed Markov process. It maximizes the likelihood by repeatedly applying a particle filter on a perturbed version of the model at each iteration. Particle filtering is a sequential Monte Carlo algorithm. Given the prediction formula from the Markov property and the filtering formula from Bayes’ theorem, the particle filter uses Monte Carlo techniques to sequentially estimate the integrals in the prediction and filtering distributions (continual random sampling). At each iteration, the model results in the selection of parameter values more consistent with the data. Eventually this should smooth the likelihood surface and as the perturbations diminish, it converges to a maximum likelihood estimate.

**Particle Markov Chain Monte Carlo:** An algorithm for estimating the parameters of a partially observed Markov process. Running pMCMC causes a particle random-walk Metropolis-Hastings algorithm to run. MCMC methods are used to sample from a probability distribution based on constructing a Markov chain (random state space process characterized by probability distribution of next state dependent on only current state) that has desired equilibrium distribution. Metropolis-Hastings algorithm is a MCMC method of obtaining a sequence of random samples from a probability distribution. It continually generates a sequence of sample values so that as more values are produced, they get closer to the desired distribution. Each sample value is either accepted (used in next iteration) or rejected (discarded, and current value is reused in next iteration) based on probability determined by comparing values of the function of the current and candidate sample values with respect to desired distribution. Full Bayesian inference (updates probability as more information becomes available). We use this to sample from the posterior distribution, which is the distribution of unobserved observations conditional on observed data.