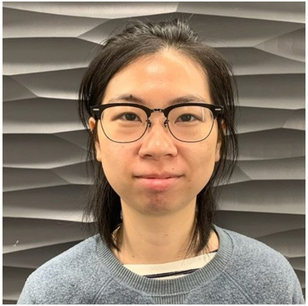
Lesson 1: Introduction to Simulation-based Inference for Epidemiological Dynamics

Qianying (Ruby) Lin   Spencer J. Fox   Zian (Larry) Zhuang

Qianying



- Postdoctoral research associate at Theoretical Biology of LANL
- Infectious disease modeling and phylodynamic theories and methodologies
- "`pomping`" since 2015, currently developing `phylopomp`
- Cat person

Spencer

- Assistant Professor at UGA in department of epidemiology and biostatistics
- Infectious disease modeler and forecaster of emerging infectious diseases
- Fan of the outdoors (started as biologist)
- Worked with pomp for $\sim$8 years (still learning)

Zian

- PhD Student at UCLA's Department of Biostatistics
- Supervisor: Dr. Gang Li
- Research Interests:
  - Infectious disease modeling
  - Causal inference in survival outcomes
- Hobby: Tennis, (Watching) British soccer

Course objectives

1. Demonstrate the utility of partially observed markov processes (POMP) for epidemiological and ecological modeling
2. Provide theoretical underpinnings of statistical inference of POMP models
3. Outline the process of formulating models and coding them in the `pomp` R package
4. Provide hands on experience working with such models and inference methods
5. Highlight research case studies and examples that can be adapted and re-used for future work

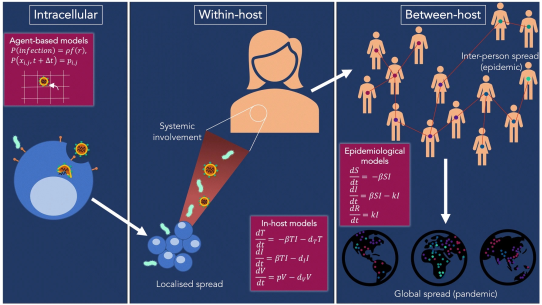
Survey results

- A few key points from the surbey that we want to highlight - maybe move to before the course objective depending on what we think

Objectives for this lesson

- To understand the motivations for simulation-based inference in the study of epidemiological and ecological systems.
- To introduce the class of partially observed Markov process (POMP) models.
- To introduce the `pomp` R package.

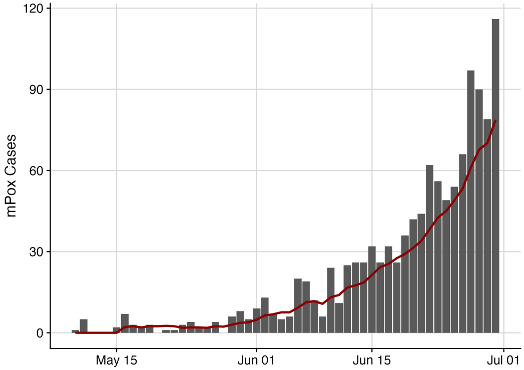
Why is ecological and epidemiological inference difficult?



- Ecological systems are complex, open, nonlinear, nonstationary, and multi-scalar
- We don't fully know the "Laws of Nature" governing the system
- Limited data and many unobserved aspects
- Multiple ways to explain available data
  - Remember herd immunity debate for COVID-19?
  - Does wearing face coverings reduce transmission?

https://link.springer.com/article/10.1007/s40139-020-00213-x

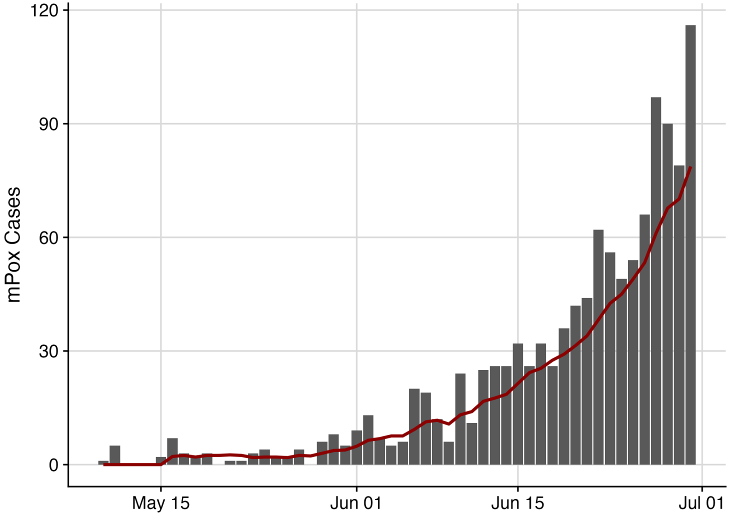
In 2022, Mpox was growing rapidly in the United States



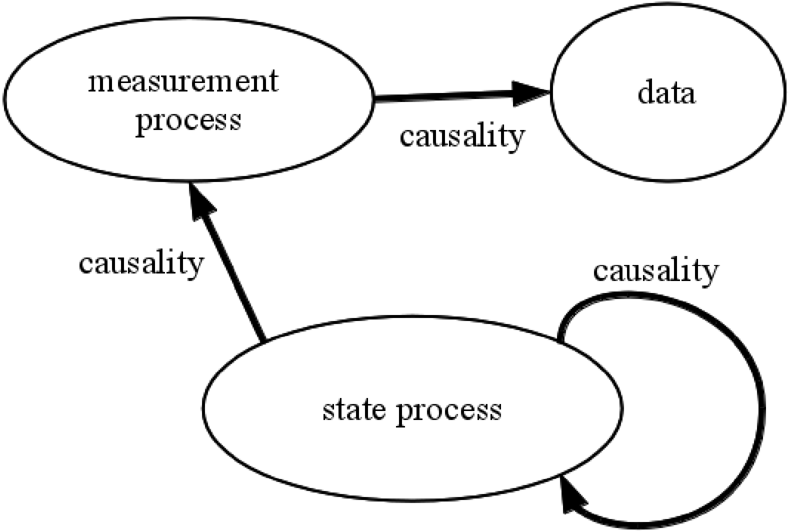
Example questions public health officials had:

1. What is the reproduction number of the virus?
2. What will case counts be over the next 4 weeks?
3. How would vaccination campaigns alter the progression of the epidemic?

We need to understand the data generating process (DGP) - what is driving the observed Mpox case counts?



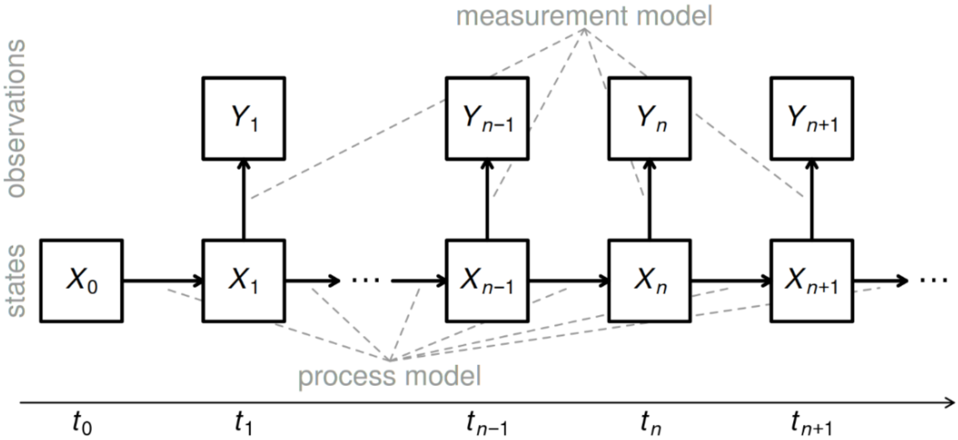
Ecological/Epidemiological DGPs



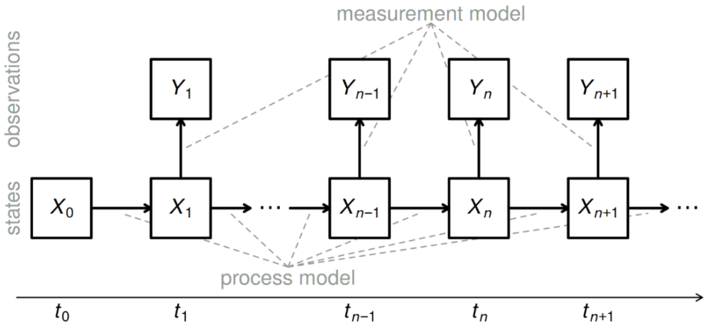
Partially observed Markov process (POMP) models

- A model where observations (data) are dependent and/or generated by a latent (hidden) Markov model
- A Markov model is a stochastic (described by probability distribution) model of a system that assumes that future states depend only on the current state, not on the events that occurred before it
- POMPs are also known as hidden Markov models or state space models
- Data collected at each time step are modeled as noisy, incomplete, and indirect observations of a Markov process
- POMP models can address the ecological/epidemiological inference issues
- Any system of differential equations $dx/dt = f(x)$ is Markovian

Time-based POMP model schematic



How do these apply for the Mpox example?



- Process model could be an epidemiological model (e.g. SEIR ODE)
- Measurement model would be how observations are generated probabilistically (e.g. some fraction of infections are randomly reported)

Three goals for models

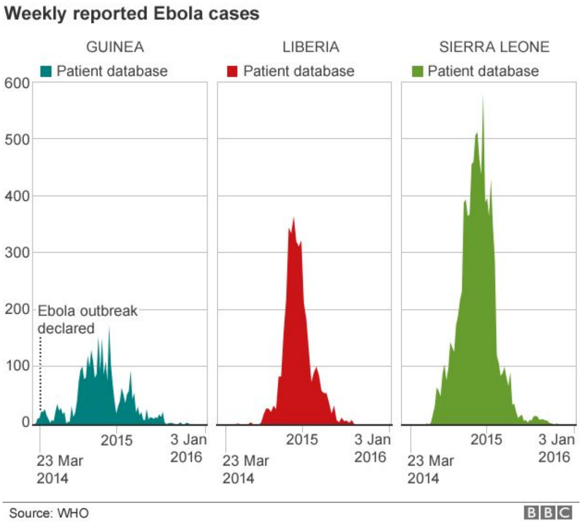
1. Inference - learn about the current or historic variables governing the system
   - What is the reproduction number of the virus?
2. Forecast - predict the values of the future states or observations (usually observations)
   - What will case counts be over the next 4 weeks?
3. Scenario Projections - predict the values of future states or observations under pre-specified scenarios
   - How would vaccination campaigns alter the progression of the epidemic?

Why are POMPs important for answering these questions?

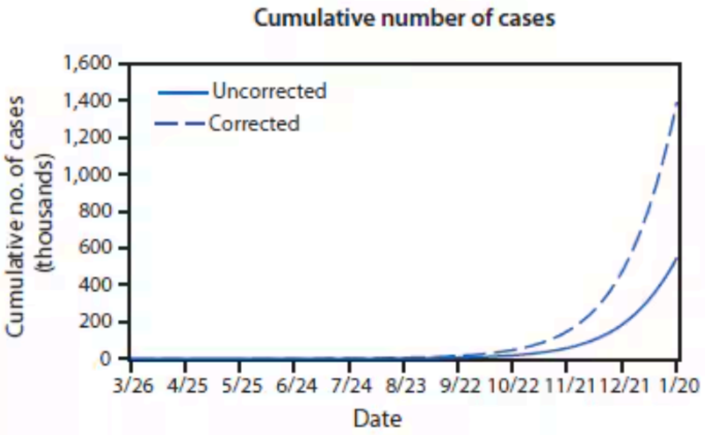
Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola

Aaron A. King[1,2,3,4], Matthieu Domenech de Cellès[1], Felicia M. G. Magpantay[1] and Pejman Rohani[1,2,4]

The 2014-2015 Ebola epidemic was devastating (>28k cases and >11k deaths)



First time models and predictions were highly visible in real-time during an epidemic/pandemic

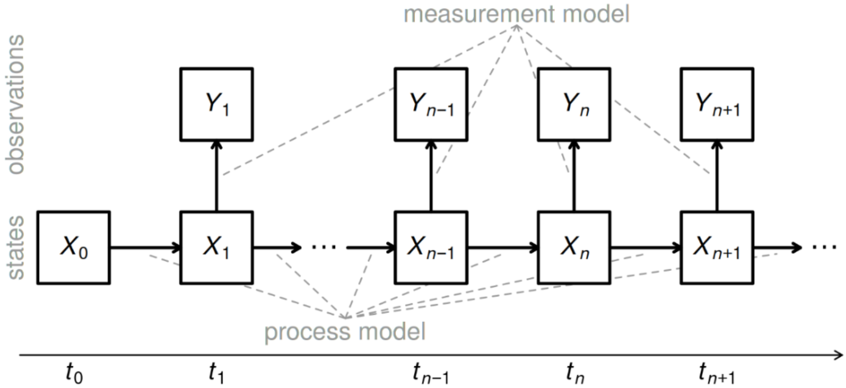


Meltzer et al 2014

King et al set out to show the impact of common modeling errors

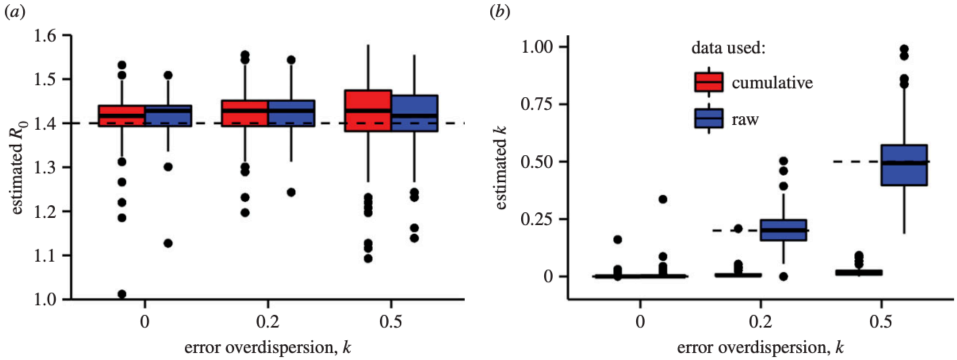
1. Using cumulative data
2. Using a deterministic rather than stochastic process model

Using cumulative (or accumulated) data break assumptions in statistical models regarding the independence of observations



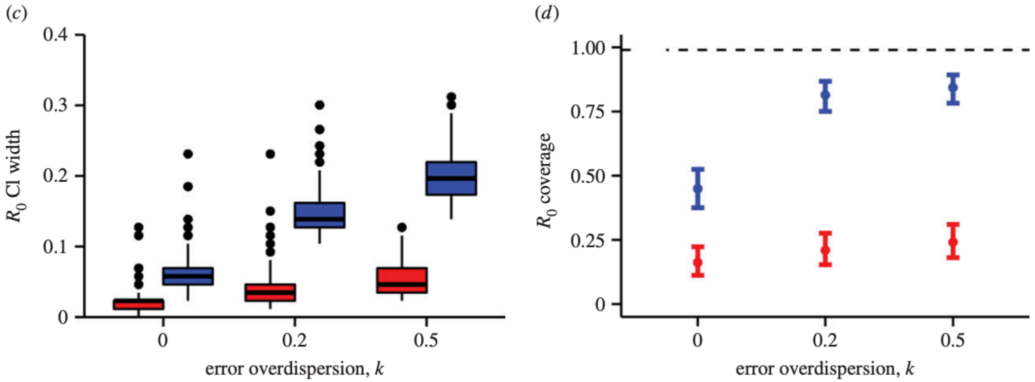
- Y_i assumed to be independent conditioned upon the X_i

Accumulated data leads to incorrect parameter estimates



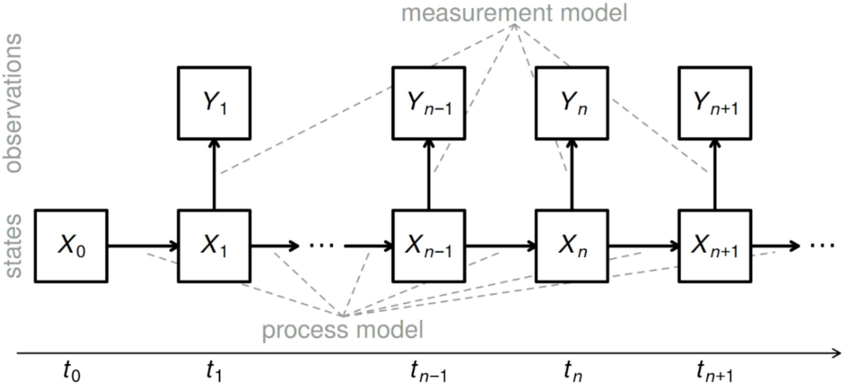
- (a) R_0 estimates similar between data types
- (b) the measurement dispersion parameter incorrect for accumulated counts

Accumulated data leads to overconfidence in parameter estimates

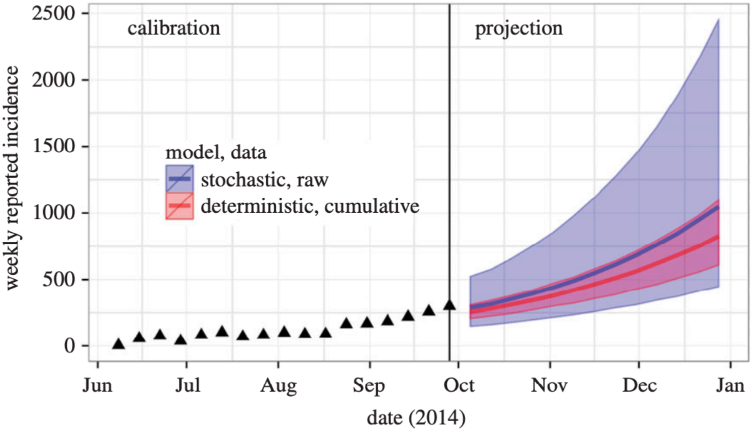


- (c) R_0 confidence intervals narrower when using accumulated data
- (d) the coverage is lower with accumulated data
  - Why is nominal coverage still below expected 99%

Using a deterministic model assigns all discrepancies between model prediction and observations to the measurement model



The two “errors” trickle into overconfidence in forecasting and projections



What is recommended?

- Fit stochastic models to incidence data
- POMPs, `pomp`, and simulated inference are one of the few ways to do so

Recent pomp examples

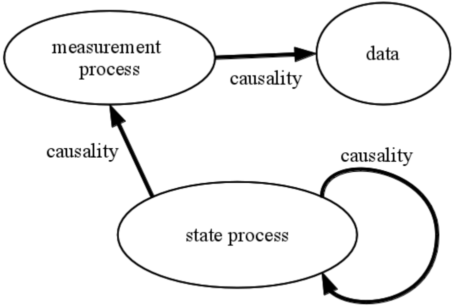
1. Quantifying asymptomatic COVID-19 infections (Subramanian, He, and Pascual 2021)
2. Estimating the effectiveness of non-pharmaceutical interventions for controlling SARS-CoV-2 spread (Shirreff et al. 2022)
3. Using human mobility data to infer epidemiological parameters (Andrade and Duggan 2022)
4. Using mobility data to forecast COVID-19 burden (Fox et al. 2022)
5. Identifying effective strategies to contain mumps spread (Shah et al. 2022)
6. Explaining the resurgence of pertussis (Domenech de Cellès et al. 2018)
7. Explaining strain dynamics in enteroviruses (Pons-Salort and Grassly 2018)
8. Contributions of population heterogeneity to HIV epidemic (Romero-Severson et al. 2015)
9. Estimating the role that adults play in polio transmission (Blake et al. 2014)
10. Relating hydrology to cholera dynamics (Baracchini et al. 2017)

Partially observed Markov process (POMP) models

- Data $y_1^*, \dots, y_N^*$ collected at times $t_1 < \dots < t_N$ are modeled as noisy, incomplete, and indirect observations of a Markov process $\{X(t), t \geq t_0\}$.
- This is a *partially observed Markov process (POMP)* model, also known as a hidden Markov model or a state space model.
- $\{X(t)\}$ is Markov if the history of the process, $\{X(s), s \leq t\}$, is uninformative about the future of the process, $\{X(s), s \geq t\}$, given the current value of the process, $X(t)$.
- If all quantities important for the dynamics of the system are placed in the *state*, $X(t)$, then the Markov property holds by construction.
- Systems with delays can usually be rewritten as Markovian systems, at least approximately.
- An important special case: any system of differential equations $dx/dt = f(x)$ is Markovian.
- POMP models can include all the features desired by Bjørnstad and Grenfell (2001).

Schematic of the structure of a POMP

- Arrows in the following diagram show causal relations.
- A key perspective to keep in mind is that *the model is to be viewed as the process that generated the data*.
- That is: the data are viewed as one realization of the model's stochastic process.



Notation for POMP models

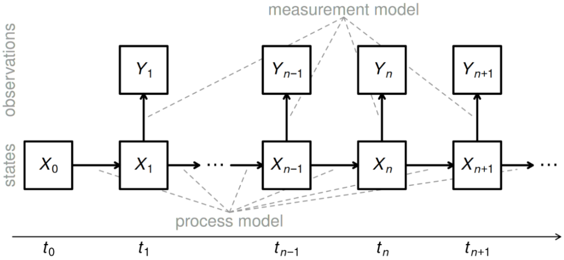
- Write $X_n = X(t_n)$ and $X_{0:N} = (X_0, \dots, X_N)$. Let Y_n be a random variable modeling the observation at time t_n.
- The one-step transition density, $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$, together with the measurement density, $f_{Y_n|X_n}(y_n|x_n;\theta)$ and the initial density, $f_{X_0}(x_0;\theta)$, specify the entire POMP model.
- The joint density $f_{X_{0:N},Y_{1:N}}(x_{0:N}, y_{1:N};\theta)$ can be written as

$$f_{X_0}(x_0;\theta)\,\prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)\, f_{Y_n|X_n}(y_n|x_n;\theta)$$

- The marginal density for $Y_{1:N}$ evaluated at the data, $y^*_{1:N}$, is

$$f_{Y_{1:N}}(y^*_{1:N};\theta) = \int f_{X_{0:N},Y_{1:N}}(x_{0:N}, y^*_{1:N};\theta)\, dx_{0:N}$$

Another POMP model schematic



- The state process, X_n, is Markovian, i.e.,

$$f_{X_n|X_{0:n-1},Y_{1:n-1}}(x_n|x_{0:n-1},y_{1:n-1}) = f_{X_n|X_{n-1}}(x_n|x_{n-1}).$$

- Moreover, Y_n, depends only on the state at that time:

$$f_{Y_n|X_{0:N},Y_{1:n-1}}(y_n|x_{0:n},y_{1:n-1}) = f_{Y_n|X_n}(y_n|x_n), \quad \text{for } n = 1,\dots,N.$$

Moving from math to algorithms for POMP models

We specify some *basic model components* which can be used within algorithms:

- `rprocess`: a draw from $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$
- `dprocess`: evaluation of $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$
- `rmeasure`: a draw from $f_{Y_n|X_n}(y_n|x_n;\theta)$
- `dmeasure`: evaluation of $f_{Y_n|X_n}(y_n|x_n;\theta)$
- `rinit`: a draw from $f_{X_0}(x_0;\theta)$

These basic model components define the specific POMP model under consideration.

What is a simulation-based method?

- Simulating random processes is often much easier than evaluating their transition probabilities.
- In other words, we may be able to write rprocess but not dprocess.
- *Simulation-based* methods require the user to specify rprocess but not dprocess.
- *Plug-and-play*, *likelihood-free* and *equation-free* are alternative terms for "simulation-based" methods.
- Much development of simulation-based statistical methodology has occurred in the past decade.

The pomp package for POMP models

- pomp is an R package for data analysis using partially observed Markov process (POMP) models (King, Nguyen, and Ionides 2016).
- Note the distinction: lower case pomp is a software package; upper case POMP is a class of models.
- pomp builds methodology for POMP models in terms of arbitrary user-specified POMP models.
- pomp provides tools, documentation, and examples to help users specify POMP models.
- pomp provides a platform for modification and sharing of models, data-analysis workflows, and methodological development.

Structure of the pomp package

It is useful to divide the pomp package functionality into different levels:

- Basic model components
- Workhorses
- Elementary POMP algorithms
- Inference algorithms

Basic model components

Basic model components are user-specified procedures that perform the elementary computations that specify a POMP model. There are nine of these:

- `rinit`: simulator for the initial-state distribution, i.e., the distribution of the latent state at time t_0.
- `rprocess` and `dprocess`: simulator and density evaluation procedure, respectively, for the process model.
- `rmeasure` and `dmeasure`: simulator and density evaluation procedure, respectively, for the measurement model.
- `rprior` and `dprior`: simulator and density evaluation procedure, respectively, for the prior distribution.
- `skeleton`: evaluation of a deterministic skeleton.
- `partrans`: parameter transformations.

The scientist must specify whichever of these basic model components are required for the algorithms that the scientist uses.

Workhorses

Workhorses are `R` functions, built into the package, that cause the basic model component procedures to be executed.

- Each basic model component has a corresponding workhorse.
- Effectively, the workhorse is a vectorized wrapper around the basic model component.
- For example, the `rprocess()` function uses code specified by the rprocess model component, constructed via the `rprocess` argument to `pomp()`.
- The rprocess model component specifies how a single trajectory evolves at a single moment of time. The `rprocess()` workhorse combines these computations for arbitrary collections of times and arbitrary numbers of replications.

Elementary POMP algorithms

These are algorithms that interrogate the model or the model/data confrontation without attempting to estimate parameters. There are currently four of these:

- `simulate` performs simulations of the POMP model, i.e., it samples from the joint distribution of latent states and observables.
- `pfilter` runs a sequential Monte Carlo (particle filter) algorithm to compute the likelihood and (optionally) estimate the prediction and filtering distributions of the latent state process.
- `probe` computes one or more uni- or multi-variate summary statistics on both actual and simulated data.
- `spect` estimates the power spectral density functions for the actual and simulated data.

POMP inference algorithms I

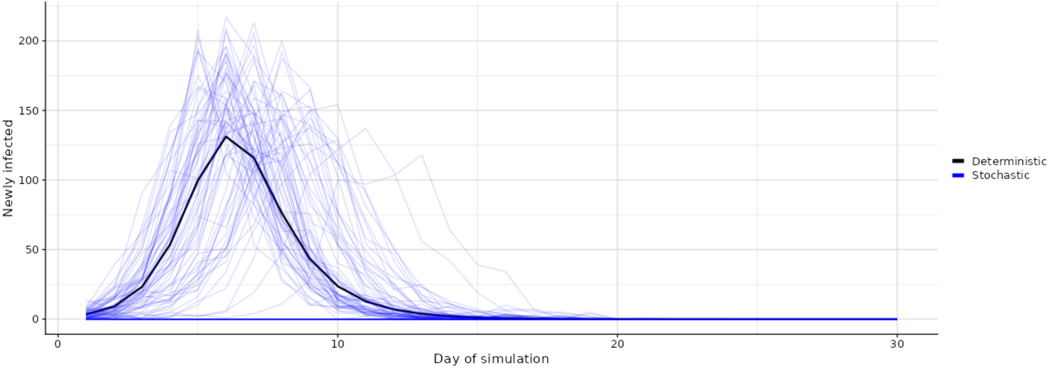
These are procedures that build on the elementary algorithms and are used for estimation of parameters and other inferential tasks. There are currently ten of these:

- `abc`: approximate Bayesian computation
- `bsmc2`: Liu-West algorithm for Bayesian SMC
- `pmcmc`: a particle MCMC algorithm
- `mif2`: iterated filtering (IF2)
- `enkf`, `eakf` ensemble and ensemble adjusted Kalman filters
- `traj_objfun`: trajectory matching
- `spect_objfun`: power spectrum matching
- `probe_objfun`: probe matching
- `nlf_objfun`: nonlinear forecasting

Objective function methods: among the estimation algorithms just listed, four are methods that construct stateful objective functions that can be optimized using general-purpose numerical optimization algorithms such as `optim`, `subplex`, or the optimizers in the nloptr package.

Activity: how do stochastic and deterministic models differ?

- Navigate to: https://spncrfx.shinyapps.io/stochastic-sir/
- Read introduction, play with parameters to understand what factors impact concordance between stochastic and deterministic models



References I

Andrade, Jair, and Jim Duggan. 2022. "Inferring the Effective Reproductive Number from Deterministic and Semi-Deterministic Compartmental Models Using Incidence and Mobility Data." *PLoS Comput Biol* 18 (6): e1010206. https://doi.org/10.1371/journal.pcbi.1010206.

Baracchini, Theo, Aaron A. King, Menno J. Bouma, Xavier Rodó, Enrico Bertuzzo, and Mercedes Pascual. 2017. "Seasonality in Cholera Dynamics: A Rainfall-Driven Model Explains the Wide Range of Patterns in Endemic Areas." *Adv Water Resour* 108C: 357–66. https://doi.org/10.1016/j.advwatres.2016.11.012.

Bjørnstad, O. N., and B. T. Grenfell. 2001. "Noisy Clockwork: Time Series Analysis of Population Fluctuations in Animals." *Science* 293: 638–43. https://doi.org/10.1126/science.1062226.

Blake, Isobel M., Rebecca Martin, Ajay Goel, Nino Khetsuriani, Johannes Everts, Christopher Wolff, Steven Wassilak, R. Bruce Aylward, and Nicholas C. Grassly. 2014. "The Role of Older Children and Adults in Wild Poliovirus Transmission." *Proc Natl Acad Sci* 111 (29): 10604–9. https://doi.org/10.1073/pnas.1323688111.

References II

Domenech de Cellès, Matthieu, Felicia M. G. Magpantay, Aaron A. King, and Pejman Rohani. 2018. "The Impact of Past Vaccination Coverage and Immunity on Pertussis Resurgence." *Sci Transl Med* 10 (434): eaaj1748. https://doi.org/10.1126/scitranslmed.aaj1748.

Fox, Spencer J., Michael Lachmann, Mauricio Tec, Remy Pasco, Spencer Woody, Zhanwei Du, Xutong Wang, et al. 2022. "Real-Time Pandemic Surveillance Using Hospital Admissions and Mobility Data." *Proc Natl Acad Sci* 119 (7): e2111870119. https://doi.org/10.1073/pnas.2111870119.

King, Aaron A., Dao Nguyen, and Edward L. Ionides. 2016. "Statistical Inference for Partially Observed Markov Processes via the R Package Pomp." *J Stat Softw* 69 (12): 1–43. https://doi.org/10.18637/jss.v069.i12.

Pons-Salort, Margarita, and Nicholas C. Grassly. 2018. "Serotype-Specific Immunity Explains the Incidence of Diseases Caused by Human Enteroviruses." *Science* 361 (6404): 800–803. https://doi.org/10.1126/science.aat6777.

References III

Romero-Severson, E. O., E. Volz, J. S. Koopman, T. Leitner, and E. L. Ionides. 2015. "Dynamic Variation in Sexual Contact Rates in a Cohort of HIV-Negative Gay Men." *Am J Epidemiol* 182 (3): 255–62. https://doi.org/10.1093/aje/kwv044.

Shah, Mirai, Gabrielle Ferra, Susan Fitzgerald, Paul J. Barreira, Pardis C. Sabeti, and Andrés Colubri. 2022. "Containing the Spread of Mumps on College Campuses." *R Soc Open Sci* 9 (1): 210948. https://doi.org/10.1098/rsos.210948.

Shirreff, George, Jean-Ralph Zahar, Simon Cauchemez, Laura Temime, and Lulla Opatowski. 2022. "Measuring Basic Reproduction Number to Assess Effects of Nonpharmaceutical Interventions on Nosocomial SARS-CoV-2 Transmission." *Emerg Infect Dis* 28 (7): 1345–54. https://doi.org/10.3201/eid2807.212339.

Subramanian, Rahul, Qixin He, and Mercedes Pascual. 2021. "Quantifying Asymptomatic Infection and Transmission of COVID-19 in New York City Using Observed Cases, Serology, and Testing Capacity." *Proc Natl Acad Sci* 118 (9): e2019716118. https://doi.org/10.1073/pnas.2019716118.

License, acknowledgments, and links

- This lesson is prepared for the Simulation-based Inference for Epidemiological Dynamics module at the Summer Institute in Statistics and Modeling in Infectious Diseases, SISMID.
- The materials build on previous versions of this course and related courses.
-
- Produced with R version 4.3.2 and pomp version 5.10.
- Compiled on 2024-06-13.

Back to Lesson

pomp homepage