Likelihood-based inference for dynamic systems

Edward Ionides
University of Michigan, Department of Statistics

Lecture 1 at Wharton Statistics Department
Tuesday 25th April, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

~~Likelihood-based inference for dynamic systems~~

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

Edward Ionides
University of Michigan, Department of Statistics

Lecture 1 at Wharton Statistics Department
Tuesday 25th April, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

~~Likelihood-based inference for dynamic systems~~

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

Edward Ionides
University of Michigan, Department of Statistics

Lecture 1 at Wharton Statistics Department
Tuesday 25th April, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

How can a title with so many modifiers be of widespread interest?

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).
- The biological importance of stochasticity was emphasized by Darwin and Mendel.

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).
- The biological importance of stochasticity was emphasized by Darwin and Mendel.
- Commonly, a mechanistic model includes some quantities that can't be directly observed.

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).
- The biological importance of stochasticity was emphasized by Darwin and Mendel.
- Commonly, a mechanistic model includes some quantities that can't be directly observed.
- Simulation-based inference leads to practical inference methodology in various applications.

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).
- The biological importance of stochasticity was emphasized by Darwin and Mendel.
- Commonly, a mechanistic model includes some quantities that can't be directly observed.
- Simulation-based inference leads to practical inference methodology in various applications.
- Likelihood-based inference has good statistical properties (Fisher) and is consistent with deductive scientific reasoning (Neyman, Popper).

Full-information likelihood-based inference via simulation for partially observed stochastic mechanistic models of dynamic systems

- "Everything flows." (Heraclitus of Ephesus, circa 500 BC).
- "The ancients esteemed the science of mechanics of greatest importance in the investigation of natural things, and the moderns have endeavoured to subject the phenomena of nature to the laws of mathematics." (Newton, 1687, *Principia Mathematica*).
- The biological importance of stochasticity was emphasized by Darwin and Mendel.
- Commonly, a mechanistic model includes some quantities that can't be directly observed.
- Simulation-based inference leads to practical inference methodology in various applications.
- Likelihood-based inference has good statistical properties (Fisher) and is consistent with deductive scientific reasoning (Neyman, Popper).
- Full-information means working with the likelihood of the entire data, not just summary statistics.

Three motivating data analysis challenges

1. Time series analysis: cholera in Bangladesh.
   - The classic challenge of discovering properties of a nonlinear system from a single long time series.

Three motivating data analysis challenges

1. Time series analysis: cholera in Bangladesh.
   - The classic challenge of discovering properties of a nonlinear system from a single long time series.
2. Panel data analysis: dynamic variation in sexual contact rates.
   - Observations on a collection of units lead to a panel of time series.
   - Analyzed together, the panel strengthens inferences available from any one time series.

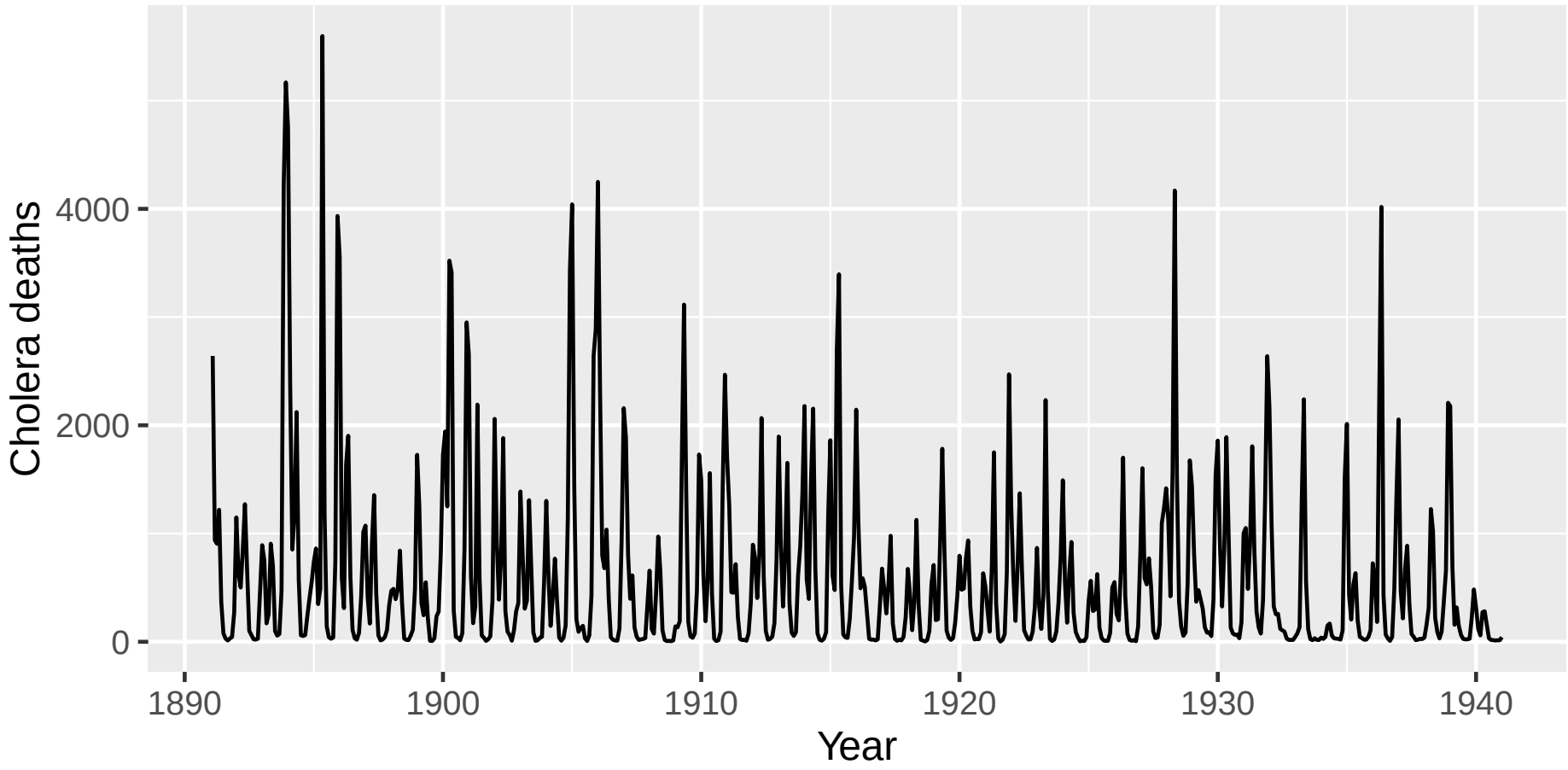
Three motivating data analysis challenges

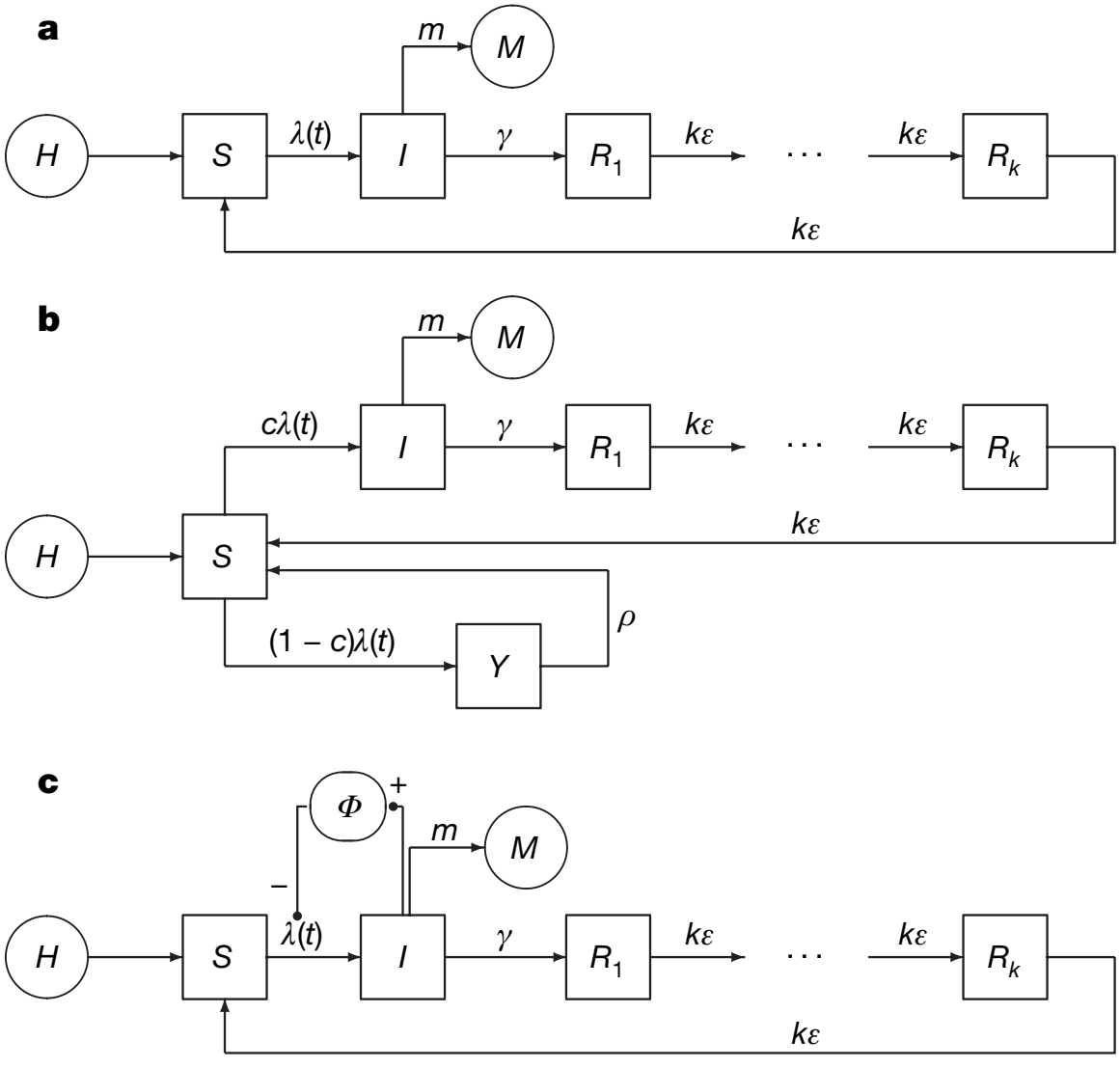
1. Time series analysis: cholera in Bangladesh.
   - The classic challenge of discovering properties of a nonlinear system from a single long time series.
2. Panel data analysis: dynamic variation in sexual contact rates.
   - Observations on a collection of units lead to a panel of time series.
   - Analyzed together, the panel strengthens inferences available from any one time series.
3. Genetic sequence data: HIV transmission within and between demographic groups.
   - Genetic sequences of pathogens can inform transmission relationships between infected hosts. This demonstrates analysis of data having structure differing from time series.

1. Time series analysis: cholera in Bangladesh

- Cholera is severe diarrhea caused by a bacterium, *Vibrio cholerae*. Death from dehydration can result rapidly without medical treatment.
- A cholera epidemic in Haiti, from 2010, has led to over 9,000 deaths (Luquero et al., 2016), comparable to the total Ebola deaths in the 2014–2015 African epidemic.
- Management of all infectious diseases is assisted by quantitative models of transmission dynamics:
  - Zika, drug-resistant bacterial infections in hospitals, malaria, the current global effort to eradicate polio, etc.
  - Diseases of agricultural crops, farm animals and wildlife.
- Models should be confronted with data — statistical analysis!
- An endless source of challenges for interested statisticians.
- Our cholera example will demonstrate that fitting a dynamic model to data can lead to qualitative scientific insights as well as quantitative understanding.

Monthly cholera deaths in Dhaka, Bangladesh, 1891-1940





Competing models

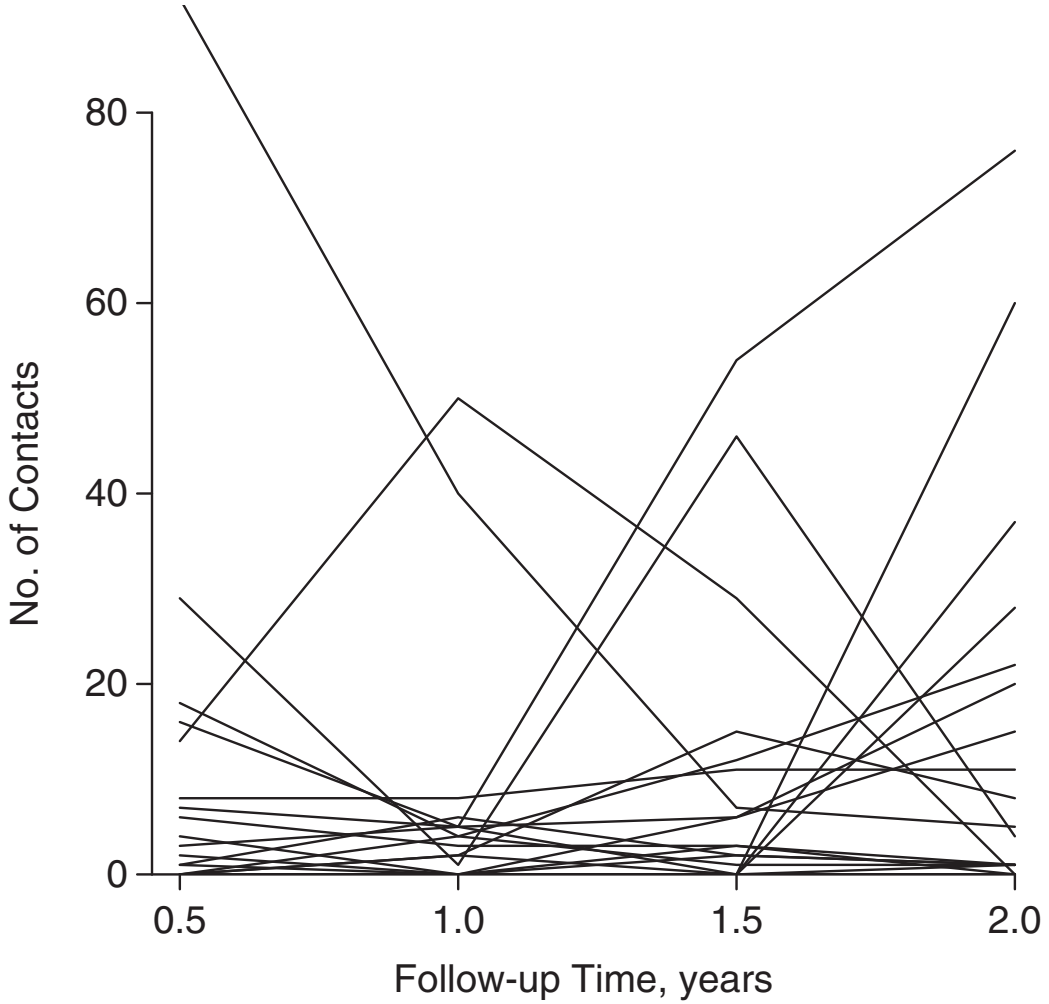
(King et al., 2008)

S	Susceptible
I	Infected
R_j	Recovered
M	Mortality
H	Population size
Y	Asymptomatics in **b**
Φ	Phage in **c**
λ	force of infection
γ	recovery rate
ϵ	loss of immunity
m	cholera mortality

2. Panel data on sexual contacts

- Mathematical models of HIV transmission struggle to explain observed incidence due to the low measured probability of transmission per sexual contact.
- The anomaly can be resolved by models that include individual-level variability in sexual behavior over time.
- This raises the question of whether dynamic variation in individual sexual behavior is a real phenomenon that can be observed and measured.
- We are motivated to construct behavioral models with various heterogeneities, both between individuals and within individuals over time, and see which models best explain available behavioral data.

Total sexual contacts in 6 month intervals



- Time series for 15 units from a panel of 882 gay men who completed a 2 year longitudinal study (Romero-Severson et al., 2015).
- Sexual contacts were reported in various categories: oral, anal, protected, unprotected, etc. Here, we show total reported contacts.

Modeling dynamic variation in sexual contact rates

- Individual i is modeled via a dynamic latent contact rate process,

$$\{X_i(t),\, 0 \le t \le 2\},$$

  giving rise to a measurement process,

$$\{Y_{ij},\, j \in 1{:}4\}.$$

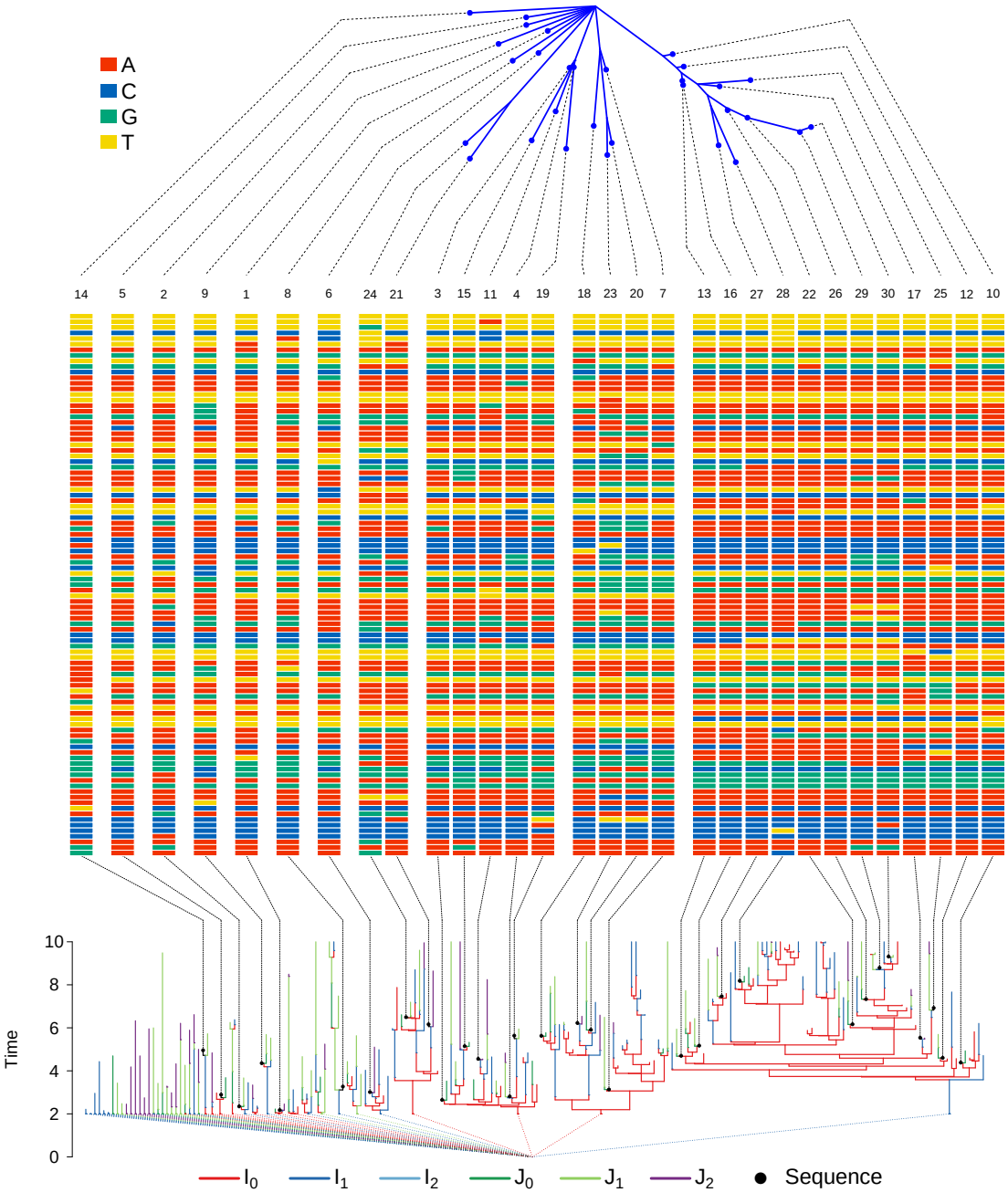
- We will construct models that can explain data,

$$\{y^*_{ij},\, i \in 1{:}882,\, j \in 1{:}4\},$$

  in terms of dynamic variation and/or between-individual heterogeneity and/or overdispersion.
- Which, if any, of these effects are statistically identifiable from available data is an empirical question.
- We look for statistical methodology that can fit flexible classes of scientifically interpretable models.

3. Infectious disease dynamics inferred from genetic data

- Genetic sequences from pathogens can provide information about infectious disease dynamics that may supplement or replace information from other epidemiological observations.
- Traditional incidence data tells who gets infected, but not who transmitted it.
- Genetic sequence data for pathogens is increasingly available.
- Statistically rigorous reconciliation of genetic sequence data with nonlinear, structured population dynamics has been an open problem.
- Formally, the disease dynamics and molecular evolution processes can be jointly modeled. How do we do inference for this complex system?



A simulated HIV epidemic (Smith et al., 2017)

Top: phylogeny of observed sequences.

Middle: simulated sequence data. Actual data are confidential.

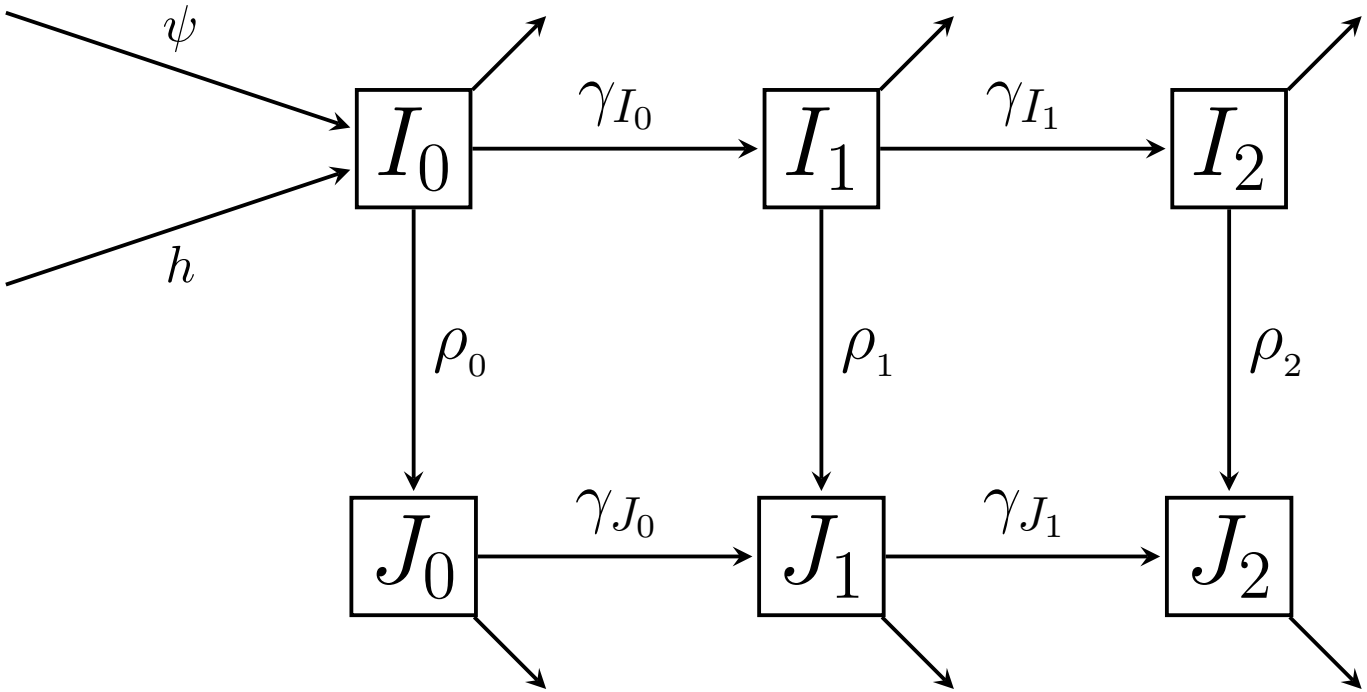
Bottom: Transmission forest for the full epidemic.

red: undiagnosed early infection

blue: undiagnosed chronic infection

green: diagnosed

Model for infection and disease progression



A flow diagram for HIV.

- I_k classes represent undiagnosed infections.
- J_k classes represent diagnosed infections.
- $k = 0, 1, 2$ denotes early, chronic and AIDS stages.
- Infection can come from within, or outside, the study population.

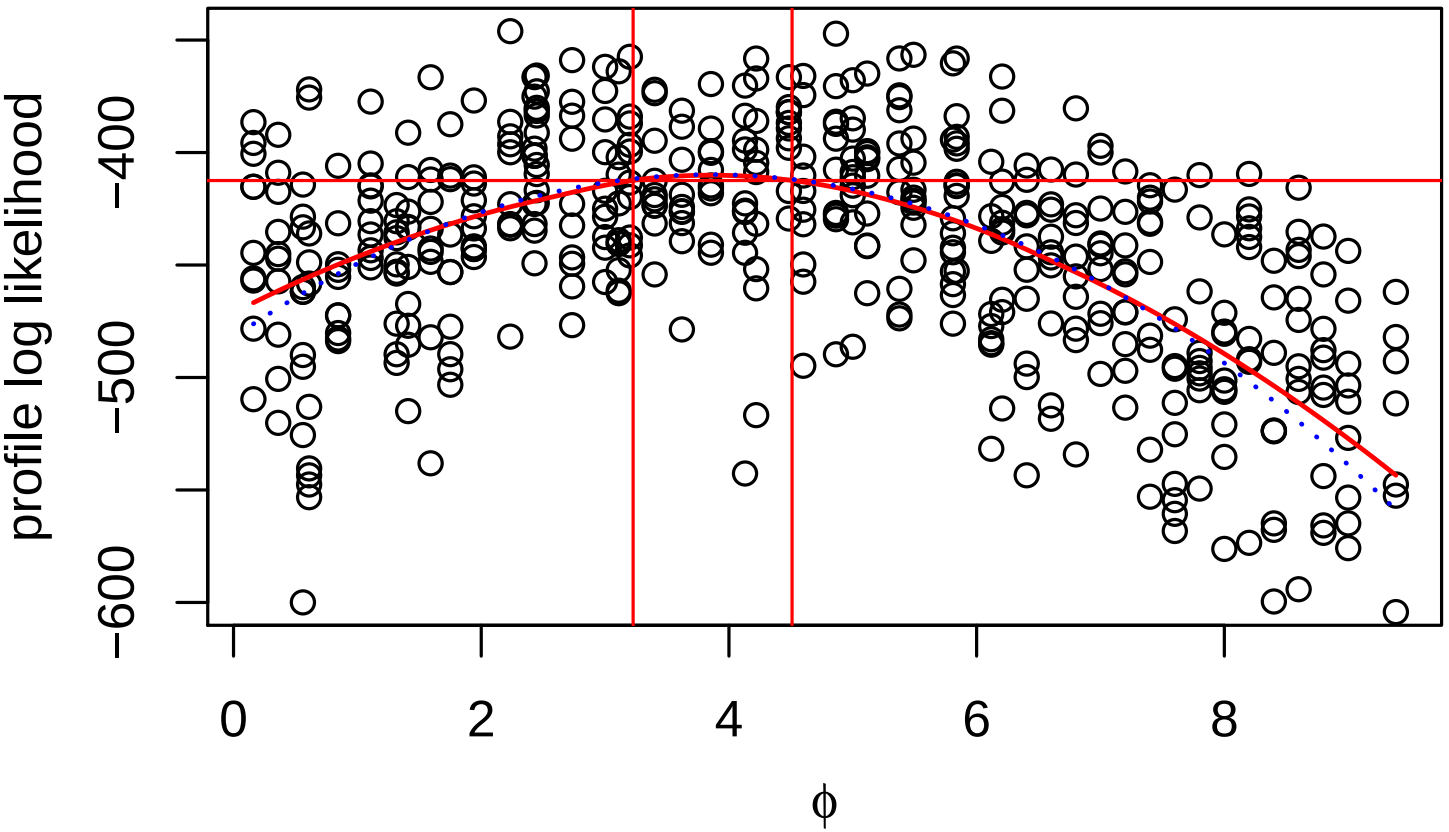
4. Inference for nonlinear mechanistic **spatiotemporal** models

- Many processes of interest happen in both space and time.
  - Movements of all species (animals, pathogens, plant seeds, etc) are basic ecological processes.
  - Business logistics place supply and demand in space and time.
- Spatiotemporal analysis is a generalization of panel data, where units in the panel correspond to spatial locations.
  - For panel analysis, units are modeled as independent.
  - For spatiotemporal analysis, units can have dynamic dependence.
- Spatiotemporal inference is a frontier that is beyond the scope of this seminar series.
- The ideas we develop here can be extended to spatiotemporal analysis (manuscript in preparation).

Key innovations

- New Monte Carlo optimization algorithms facilitate likelihood maximization for large partially observed Markov process (POMP) models: **iterated filtering**.
  - Iterated filtering algorithms optimize the likelihood using a sequence of random parameter perturbations, with decreasing magnitude. Sequential Monte Carlo (SMC) provides a tool for numerical solution to this nonlinear filtering problem.
  - Existing variations on expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) do not scale well for these problems.
  - We are doing parametric inference. The main problem using likelihood or Bayesian methods is computational. If existing methods worked computationally, there would be no problem!
- A new perspective on likelihood-based inference via **Monte Carlo profile likelihood**.

Monte Carlo profile for genetic data on HIV dynamics



- ϕ models HIV transmitted by recently infected, diagnosed individuals.
- The profile confidence interval is constructed by a cutoff that is adjusted for the Monte Carlo variability (Ionides et al., 2016).
  - A proper 95% cutoff is 2.35. Without Monte Carlo error, it is 1.92.
  - Each point took approximately 10 core days to compute.
  - Alternative approaches struggle with Monte Carlo likelihood error of order 100 log units.

References

Ionides, E. L., Breto, C., Park, J., Smith, R. A., and King, A. A. (2016). Monte Carlo profile confidence intervals. *ArXiv:1612.02710*.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454:877–880.

Luquero, F. J., Rondy, M., Boncy, J., Munger, A., Mekaoui, H., Rymshaw, E., Page, A.-L., Toure, B., Degail, M. A., Nicolas, S., et al. (2016). Mortality rates during cholera epidemic, Haiti, 2010–2011. *Emerging Infectious Diseases*, 22(3):410.

Romero-Severson, E., Volz, E., Koopman, J., Leitner, T., and Ionides, E. (2015). Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *American Journal of Epidemiology*, 182:255–262.

Smith, R. A., Ionides, E. L., and King, A. A. (2017). Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Molecular Biology and Evolution*, pre-published online, doi:10.1093/molbev/msx124.