Dynamic modeling and inference for ecological systems

IOE 899: Seminar in Industrial and Operations Engineering

September 14, 2011

Edward Ionides

The University of Michigan, Department of Statistics

Collaborators

Anindya Bhadra (University of Michigan, Statistics)

Menno Bouma (London School of Hygiene and Tropical Medicine)

Carles Bretó (Universidad Carlos III de Madrid, Statistics)

Daihai He (McMaster University, Mathematics & Statistics)

Aaron King (Univ. of Michigan, Ecology & Evolutionary Biology)

Karina Laneri (Institut Català de Ciències del Clima, Barcelona, Spain)

Mercedes Pascual (Univ. of Michigan, Ecology & Evolutionary Biology)

Outline

1. Overview of time series analysis for ecological systems.
2. Some practical considerations: relationship between statistical methodology and software.
3. The *plug-and-play* property.
4. Iterated filtering: theory and methodology.
5. Case studies: malaria & measles.
6. Outstanding challenges.

Why study inference for dynamic models in ecology?

- Long time series of fluctuating abundances provide an opportunity to test ecological theories of the relationships driving the system.

  e.g., To what extent is a herbivore population constrained by predators, food resources, or disease?

- Forecasting and parameter estimation are of some interest. But a primary concern is to identify the roles of population dynamics (i.e., reproduction & foodchains), evolutionary processes, and environmental covariates. **BASIC SCIENCE.**

- Humans are increasingly responsible for managing ecosystems. This requires quantitative understanding of ecological relationships and the potential effect of interventions. **ENGINEERING.**

Infectious diseases as ecological systems

- Good spatio-temporal data are available for many human diseases.
- The 20th century saw some successes for vaccination and drug treatment. But the limitations also became evident.
  - ◇ Emerging infectious diseases (SARS; HIV/AIDS; H5N1 influenza "bird flu")
  - ◇ New strains and drug resistance (H1N1 "swine flu"; MRSA "the hospital super-bug"; tuberculosis; malaria)
- Controlling human/livestock/wildlife diseases involves understanding the pathogen-host ecological dynamics.

Methodological problem: Inference for partially observed nonlinear stochastic dynamic systems

- A research area for 60yrs (initially inspired by rocket control theory).
- Difficulties are computational: Bayesian, likelihood and pattern-matching methods are all tricky to implement. Customized approximations and model-specific methods have been needed.
- Linearization & small noise asymptotics are questionable (or worse) for many biological systems.
- No general-purpose software has been available.
  - ◇ WinBUGS performs poorly on these models.
  - ◇ **pomp**, an R package for partially observed Markov processes (POMPs), is being developed.

Partially Observed Markov Process (POMP) models

The unobserved Markov state process is denoted $X(t)$. For observation times $t_1, \ldots, t_N$ we write $X_n = X(t_n)$. The observable variables $Y_1, \ldots, Y_N$ are conditionally independent given $X_1, \ldots, X_N$. The model depends on an unknown parameter vector θ.

- To think algorithmically, we define some function calls:

  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 $f_{Y_n|X_n}(y_n \,|\, x_n\,;\theta)$

Plug-and-play inference for POMP models

- An algorithm operating on a POMP is **plug-and-play** if it involves calls to **`rprocess`** but not to **`dprocess`**. In this case, numerical solution of sample paths is a 'black box' which is plugged into the software.

- Bayesian plug-and-play:
  1. Artificial parameter evolution (Liu and West, 2001)
  2. Approximate Bayesian computation (Sisson et al, *PNAS*, 2007)
  3. Particle MCMC (Andrieu et al, *JRSSB*, 2010)

- Non-Bayesian plug-and-play:
  4. Simulation-based prediction rules (Kendall et al, *Ecology*, 1999)
  5. Iterated filtering (Ionides et al, *PNAS*, 2006)

Plug-and-play is a VERY USEFUL PROPERTY for investigating scientific models.

Classification of methodologies by required operations

	`rprocess`	`dprocess`	`rmeasure`	`dmeasure`
Iterated filtering	✔	✗	✗	✔
Liu-West SMC	✔	✗	✗	✔
EM via SMC	✔	✔	✗	✔
MCMC	✗	✔	✗	✔
Nonlinear forecasting	✔	✗	✔	✗
Particle MCMC	✔	✗	✗	✔
Probe matching	✔	✗	✔	✗

- The usual workhorses of statistical computation (EM and MCMC) are not plug-and-play.
- Nonlinear forecasting and probe matching are simulation-based techniques developed by scientists, likely due to the inapplicability of textbook statistical techniques

Plug-and-play in other settings

- **Optimization**. Methods requiring only evaluation of the objective function to be optimized are sometimes called **gradient-free**. This is the same concept as plug-and-play: the code to evaluate the objective function can be *plugged into* the optimizer.

- **Complex systems**. Methods to study the behavior of large simulation models that only employ the underlying code as a "black box" to generate simulations have been called **equation-free** (Kevrekidis et al., 2003, 2004).
  - This is the same concept as plug-and-play, but we prefer our label!
  - A typical goal is to determine the relationship between macroscopic phenomena (e.g. phase transitions) and microscopic properties (e.g. molecular interactions).

The cost of plug-and-play

- Approximate Bayesian methods and simulated moment methods lead to a loss of statistical efficiency.
- In contrast, iterated filtering enables (almost) exact likelihood-based inference.
- Improvements in numerical efficiency may be possible when analytic properties are available (at the expense of plug-and-play). But many interesting dynamic models are analytically intractable—for example, it is standard to investigate systems of ordinary differential equations numerically.

Summary of plug-and-play inference via iterated filtering

- **Filtering** is the extensively-studied problem of calculating the conditional distribution of the unobserved state vector x_t given the observations up to that time, $y_1, y_2, \ldots, y_t$.
- **Iterated filtering** is a recently developed algorithm which uses a sequence of solutions to the filtering problem to maximize the likelihood function over unknown model parameters.
  (Ionides, Bretó & King. *PNAS*, 2006)
- If the filter is plug-and-play (e.g. using standard sequential Monte Carlo methods) this is inherited by iterated filtering.

Key idea of iterated filtering

- The conditional distribution of time-varying parameters is a (relatively) tractable filtering problem. Set $\theta = \theta_t$ to be a random walk with

$$E[\theta_t|\theta_{t-1}] = \theta_{t-1} \qquad \mathrm{Var}(\theta_t|\theta_{t-1}) = \sigma^2$$

- The limit $\sigma \to 0$ can be used to maximize the likelihood for fixed parameters.

Theorem. (Ionides, Bretó & King, *PNAS*, 2006)

Suppose $\hat{\theta}_0$, C *and* $y_{1:T}$ *are fixed and define*

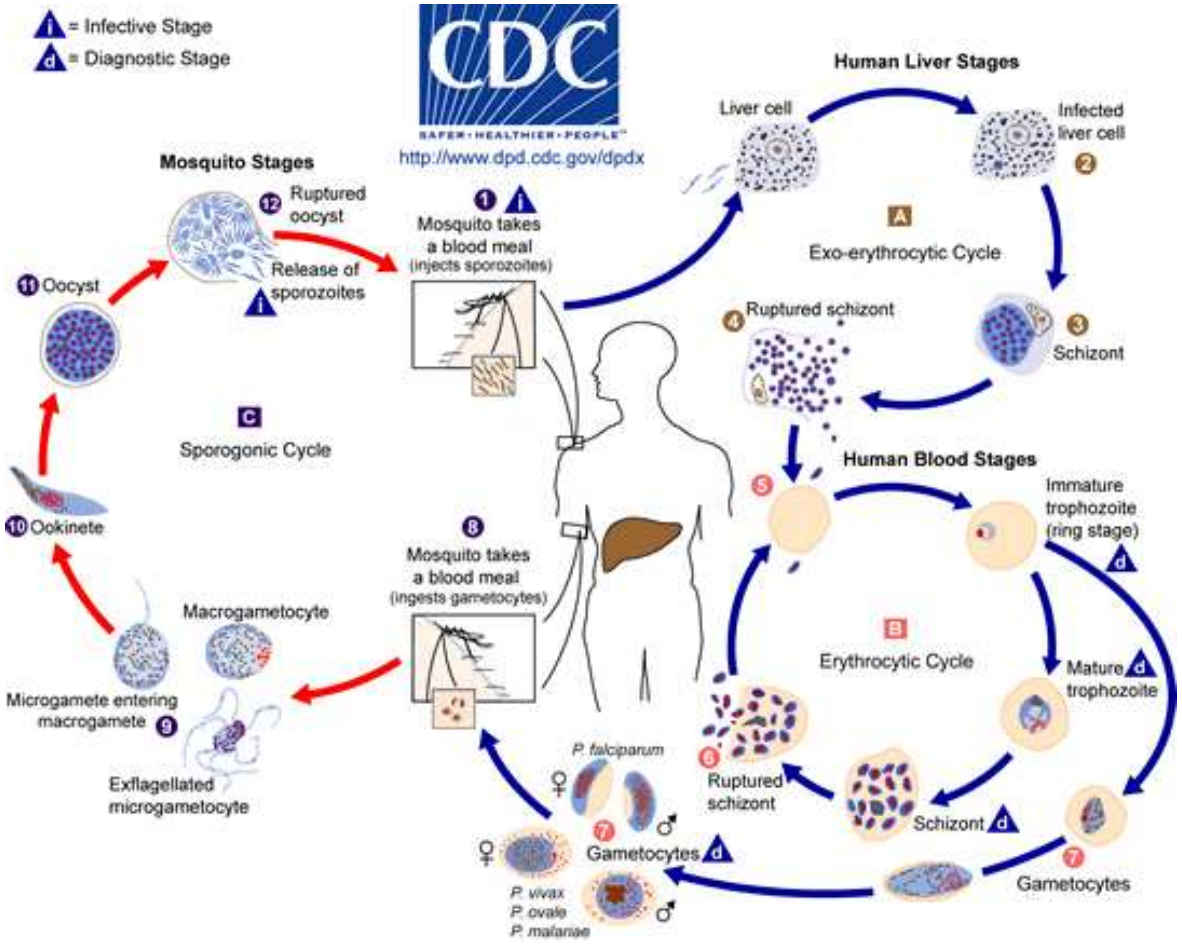
$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_t(\sigma) = E[\theta_t | y_{1:t}] \\ V_t &= V_t(\sigma) = \mathrm{Var}(\theta_t | y_{1:t-1}) \end{aligned}$$

Assuming sufficient regularity conditions for a Taylor series expansion,

$$\lim_{\sigma \to 0} \sum_{t=1}^{T} V_t^{-1} (\hat{\theta}_t - \hat{\theta}_{t-1}) = (\partial/\partial\theta) \log f(y_{1:T} | \theta, \sigma{=}0) \Big|_{\theta=\hat{\theta}_0}$$

The limit of an appropriately weighted average of local filtered parameter estimates is the derivative of the log likelihood.

Example: malaria (mosquito-transmitted *Plasmodium* infection)



Despite extensive study of the disease system (mosquito, *Plasmodium* & human immunology) ecological dynamics of malaria remain hotly debated.

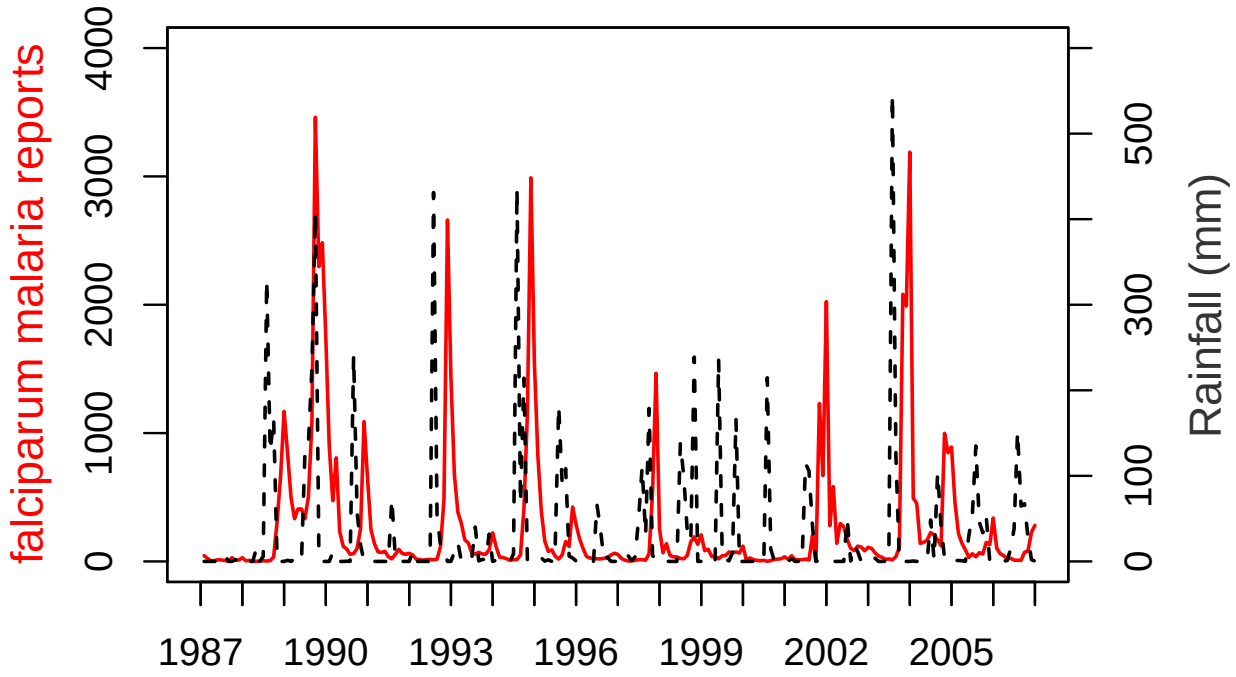
Malaria: A global challenge

- Bill Gates would like to eradicate it, but others have tried before...
- There has been extensive debate on whether/how global climate change will affect malaria burden—a model validated by data is required.

From the perspective of statistical methodology

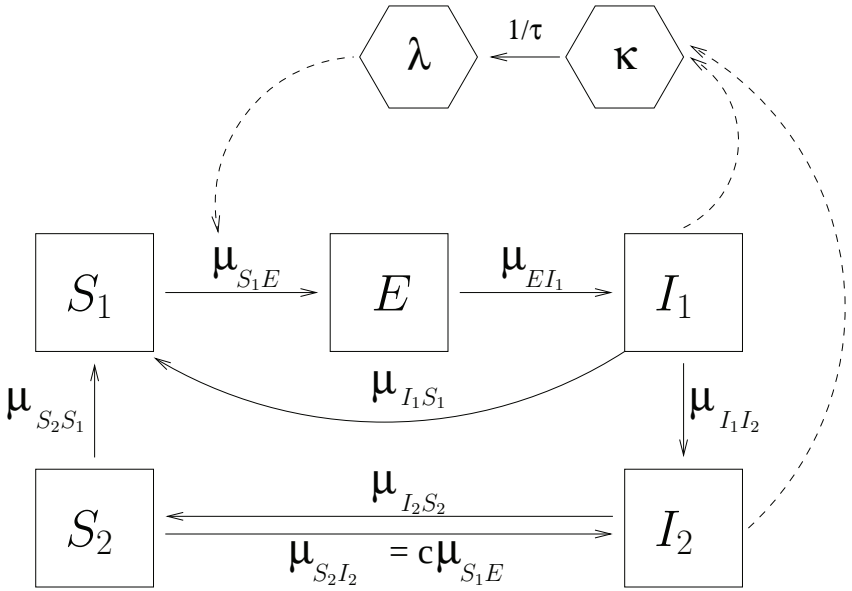
- Despite the huge literature, no dynamic model of malaria transmission has previously been fitted directly to time series data.
- Difficulties include: Incomplete and complex immunity; dynamics in both mosquito and human stages; diagnostic difficulties (the immediate symptom is non-specific fever).
- Malaria is beyond the scope of methods developed for simpler diseases.

Malaria and rainfall in Kutch (an arid region of NW India)



- To what extent are cycles driven by immunity rising and falling? To what extent are they driven by rainfall?

A dynamic model (Laneri et al, *PLoS Comp. Biol.*, 2010).



λ, force of infection; κ, latent force of infection; S_1, fully susceptible humans; S_2 clinically protected (partially immune); I_1, clinically infected; I_2, asymptomatically infected.

Minimal complexity acceptable to scientists

$\approx$ **Maximal complexity acceptable to available data**

Model representation: coupled SDEs driven by Lévy noise

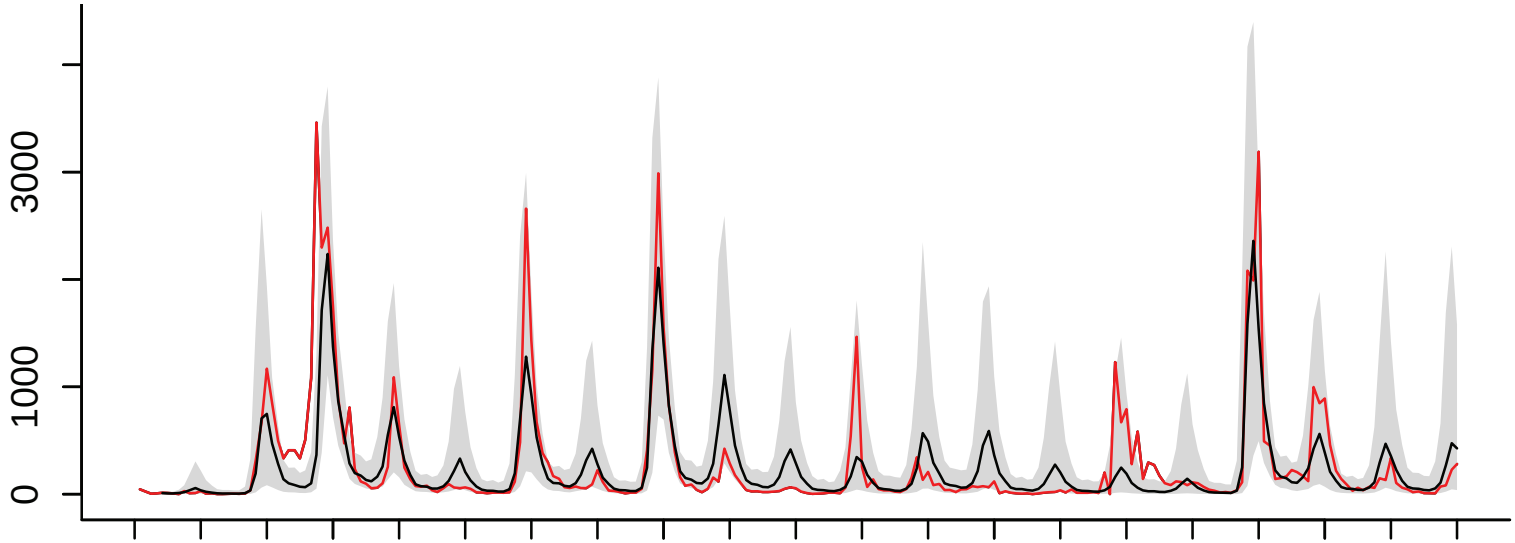
$$
\begin{aligned}
dS_1/dt &= \mu_{BS_1} P - \mu_{S_1E} S_1 + \mu_{I_1S_1} I_1 + \mu_{S_2S_1} S_2 - \mu_{S_1D} S_1 \\
dS_2/dt &= \mu_{I_2S_2} I_2 - \mu_{S_2S_1} S_2 - \mu_{S_2I_2} S_2 - \mu_{S_2D} S_2 \\
dE/dt &= \mu_{S_1E} S_1 - \mu_{EI_1} E - \mu_{ED} E \\
dI_1/dt &= \mu_{EI_1} E - \mu_{I_1S_1} I_1 - \mu_{I_1I_2} I_1 - \mu_{I_1D} I_1 \\
dI_2/dt &= \mu_{I_1I_2} I_1 + \mu_{S_2I_2} S_2 - \mu_{I_2S_2} I_2 - \mu_{I_2D} I_2 \\
d\kappa/dt &= d\lambda_0/dt \;=\; (f(t) - \kappa)\, \ell\, \tau^{-1} \\
d\lambda_i/dt &= (\lambda_{i-1} - \lambda_i)\, \ell\, \tau^{-1} \qquad \text{for } i = 1, \ldots, \ell - 1 \\
d\lambda/dt &= d\lambda_\ell/dt \;=\; (\lambda_{\ell-1} - \lambda)\, \ell\, \tau^{-1} \\
f(t) &= \frac{I_1(t) + q I_2(t)}{N(t)}\, \bar{\beta}\, \exp\Big\{ \sum_{i=1}^{n_s} \beta_i s_i(t) + Z_t \beta \Big\} \frac{d\Gamma}{dt}.
\end{aligned}
$$

Z_t is a vector of climate covariates (here, rainfall).

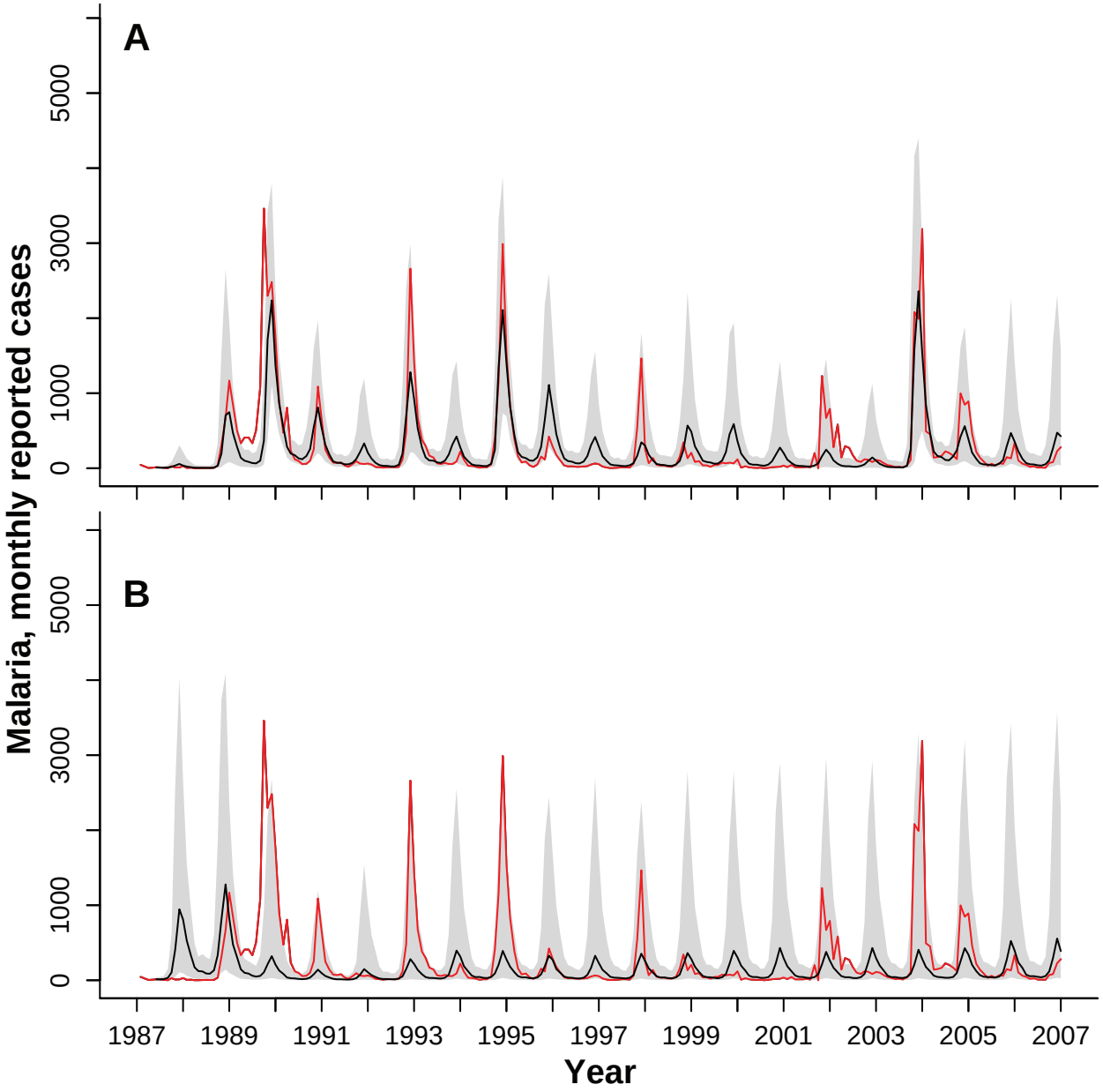
$\sum_{i=1}^{n_s} \beta_i s_i(t)$ is a spline representation of seasonality.

Conclusions from malaria data analysis

- Rainfall (with an appropriate delay and threshold) a critical role in determining interannual cycles.
- Immunity plays a role at faster timescales (controlling annual peaks)



Simulations forward from 1987 to 2007, from the MLE, with prescribed rainfall. Showing monthly case reports (red), simulation median (black) and 10th to 90th percentiles (grey). Without rainfall, the model cannot come close to this.



Simulations forward from 1987 to 2007 from fitted models (A) with rainfall; (B) without rainfall. Showing monthly case reports (red), simulation median (black) and 10th to 90th percentiles (grey).

Stochastic differential equations (SDEs) vs. Markov chains

- SDEs are a simple way to add stochasticity to widely used ordinary differential equation models for population dynamics.
- When some species have low abundance (e.g. fade-outs and re-introductions of diseases within a population) discreteness can become important.
- This motivates the consideration of discrete population, continuous time POMP models (Markov chains).

Over-dispersion in Markov chain models of populations

- **Remarkably, in the vast literatures on continuous-time individual-based Markov chains for population dynamics (e.g. applied to ecology and chemical reactions) no-one has previously proposed models capable of over-dispersion.**
- It turns out that the usual assumption that no events occur simultaneously creates fundamental limitations in the statistical properties of the resulting class of models.
- Over-dispersion is the rule, not the exception, in data.
- Perhaps this discrepancy went un-noticed before statistical techniques became available to fit these models to data.

Implicit models for plug-and-play inference

- Adding "white noise" to the transition rates of existing Markov chain population models would be a way to introduce an infinitesimal variance parameter, by analogy with the theory of SDEs.

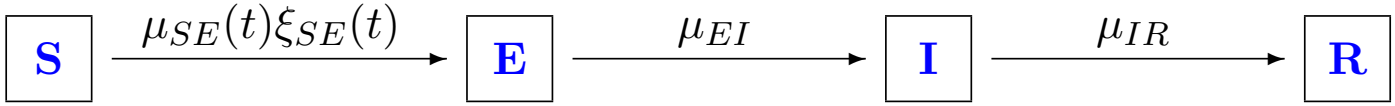
- **We do this by defining our model as a limit of discrete-time models. We call such models *implicit*.** This is backwards to the usual approach of checking that a numerical scheme (i.e. a discretization) converges to the desired model.

- Implicit models are convenient for numerical solution, by definition, and therefore fit in well with plug-and-play methodology.

- Details in Bretó et al (2009, *AoAS*); Bretó & Ionides (2011, *Stoc. Proc. Appl.*).

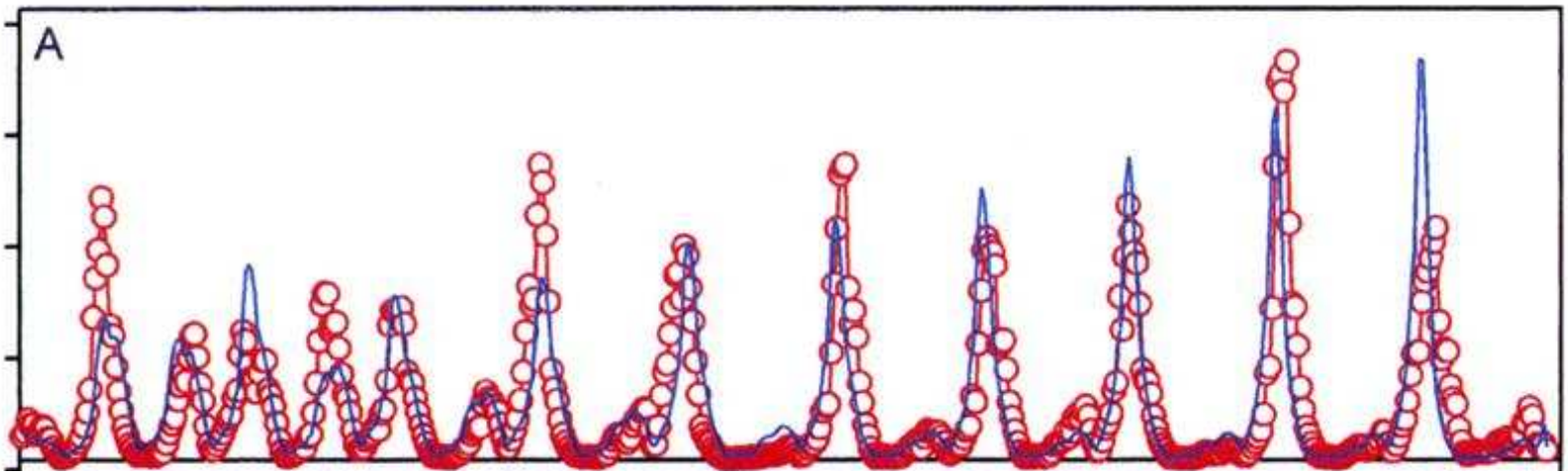
Measles: an exhaustively studied system

- Measles is simple: direct infection of susceptibles by infecteds; characteristic symptoms leading to accurate clinical diagnosis; life-long immunity following infection.



Susceptible $\longrightarrow$ Exposed (latent) $\longrightarrow$ Infected $\longrightarrow$ Recovered,
with noise intensity σ_{SE} on the force of infection.

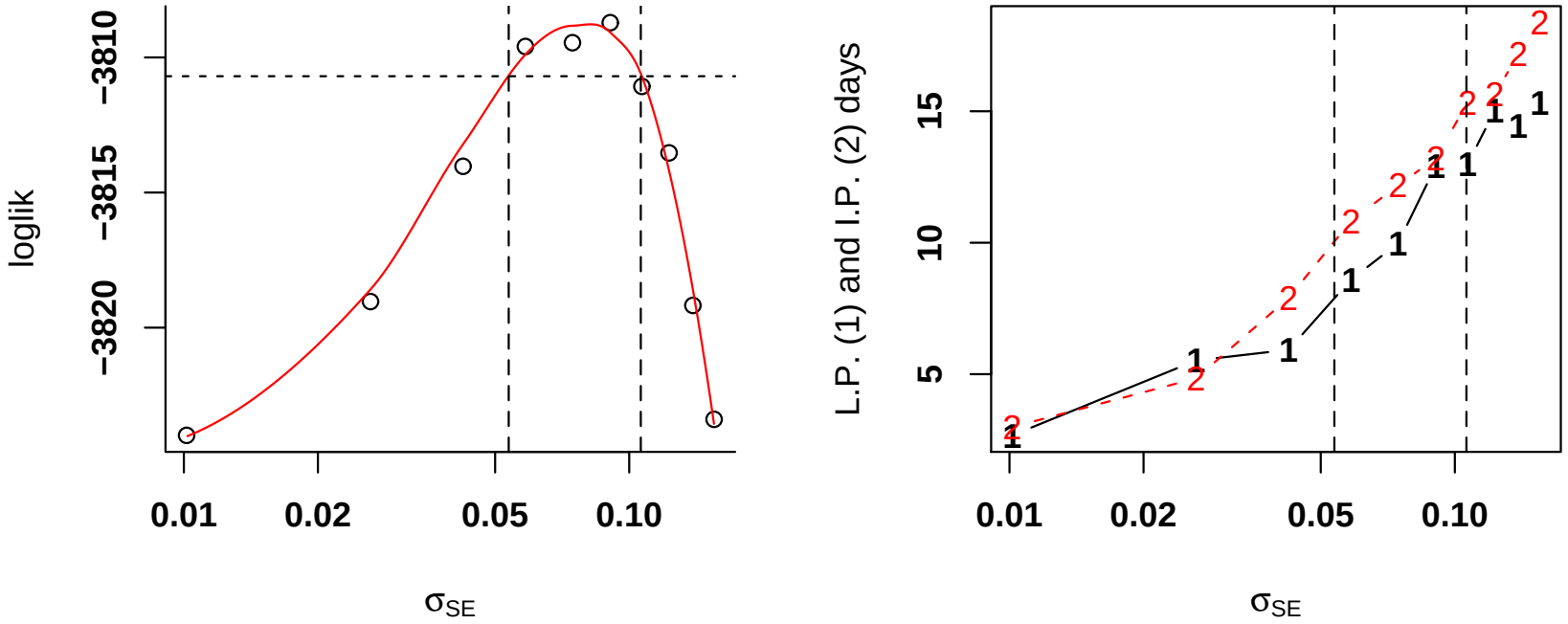
- Measles is still a substantial health issue in sub-Saharan Africa.
- A global eradication program is under debate.
- Comprehensive doctor reports in western Europe and America before vaccination (≈ 1968) are textbook data.



- Measles cases in London 1944–1965 (circles and red lines) and a deterministic SEIR fit (blue line) (from Grenfell *et al*, 2002).
- A deterministic fit, specified by the initial values in January 1944, captures remarkably many features.

Is demographic stochasticity ($\sigma_{SE} = 0$) plausible?

- Profile likelihood for σ_{SE} and effect on estimated latent period (L.P.) and infectious period (I.P.) for London, 1950–1964.
- Variability of $\approx 5\%$ per year on the infection rate substantially improves the fit, and affects scientific conclusions (He et al, *JRSI*, 2010).



Interpretation of over-dispersion

- Social and environmental events (e.g., football matches, weather) lead to stochastic variation in rates: **environmental stochasticity**.
- A catch-all for other model mis-specification? It is common practice in linear regression to bear in mind that the "error" terms contain un-modeled processes as well as truly stochastic effects. This reasoning can be applied to dynamic models as well.

Conclusions and outstanding challenges

- Plug-and-play statistical methodology permits likelihood-based analysis of flexible classes of stochastic dynamic models.

- **It is increasingly possible to carry out data analysis via nonlinear mechanistic stochastic dynamic models**. This can build links between the mathematical modeling community (within which models are typically conceptual and qualitative) and quantitative applications (testing hypotheses about mechanisms, forecasting, evaluating the consequences of interventions). Increasingly many long time series are available. **Much work remains to be done!**

- Many models of interest are beyond current algorithms & computational resources. New data types (e.g., genetic markers for some or all reported individuals) both enable and require the fitting of more complex models.

Thank you!

These slides (including references for the citations) are available at

`www.stat.lsa.umich.edu/~ionides/pubs`

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72:269–342.

Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3:319–348.

Bretó, C. and Ionides, E. L. (2011). Compound markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, (to appear).

Grenfell, B. T., Bjornstad, O. N., and Finkenstädt, B. F. (2002). Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72(2):185–202.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface*, 7:271–283.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA*, 103:18438–18443.

Kendall, B. E., Briggs, C. J., Murdoch, W. W., Turchin, P., Ellner, S. P., McCauley, E., Nisbet, R. M., and Wood, S. N. (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80:1789–1805.

Kevrekidis, I. G., Gear, C. W., and Hummer, G. (2004). Equation-free: The computer-assisted analysis of complex, multiscale systems. *American Institute of Chemical Engineers Journal*, 50:1346–1354.

Kevrekidis, I. G., Gear, C. W., Hyman, J. M., Kevrekidis, P. G., Runborg, O., and Theodoropoulos, C. (2003). Equation-free coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. *Communications in the Mathematical Sciences*, 1:715–762.

Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Yadav, R., Dhiman, R., and Pascual, M. (2010). Forcing versus feedback: Epidemic malaria and monsoon rains in NW India. *PLoS Computational Biology*, 6:e1000898.

Liu, J. and West, M. (2001). Combining parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 197–224. Springer, New York.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the USA*, 104(6):1760–1765.