Sequential Monte Carlo methods for inferring transmission dynamics from pathogen genetic sequences

Mathematisches Forschungsinstitut Oberwolfach

workshop on

Design and analysis of infectious disease studies

November 13, 2013

Edward Ionides

The University of Michigan, Department of Statistics

Outline

- What is sequential Monte Carlo (SMC)?
- Inferring unknown parameters via SMC.
- The plug-and-play property and its relationship to SMC.
- SMC for transmission dynamics: lessons from case studies.
- Sequence data to inform disease transmission dynamics.
    i. SMC for transmission dynamics conditional on a phylogeny.
    ii. SMC to estimate a phylogeny.
    iii. SMC to jointly estimate a phylogeny and transmission dynamics.
- Exorcising the curse of dimensionality for SMC.

To talk about SMC, we first have to introduce the models on which SMC operates... **partially observed Markov process** (POMP) models.

Partially observed Markov process (POMP) models

- A Markov process is a time-indexed stochastic process for which the past and future are conditionally independent given the present.
- We allow discrete-time, continuous-time, discrete-valued, continuous-valued, vector-valued, function-valued, etc.

  For phylodynamics, the state of the Markov process could be a tree, with branches and leaves being added through time.
- If any variable that affects the future evolution of a system is modeled in the current state, then the Markov property holds tautologously.
- Delays cannot usually be modeled in a finite dimensional Markov process, except in specific cases (e.g., gamma-distributed delays).
- Partial observations are noisy functions of the process, observed at a discrete set of times.

Motivations for the POMP framework

- POMP models have repeatedly been proposed as a general framework for modeling biological systems.
- A reasonable tradeoff between generality and tractability.
- Computationally practical algorithms exist for reconstructing unobserved variables from data (filtering) and for evaluating the likelihood function.
- Difficulties arise for large state spaces.
- Theoretical properties of POMPs are well studied.

Some history of sequential Monte Carlo (SMC)

- SMC grew in the 1990s, simultaneously called particle filtering, bootstrap filtering, Monte Carlo filtering, sequential importance sampling, and the condensation algorithm.
- Used in physics and chemistry since the 1950s: marketed as "Poor man's Monte Carlo" (Hammersley & Morton, 1954).
- In modern theory, SMC and MCMC have similar asymptotic guarantees.
- SMC provides an alternative to MCMC for many computations.
- For dynamic systems, SMC is preferred.

An evolutionary description of SMC algorithms

- An SMC algorithm consists of a **swarm** of **particles** evolves according to a stochastic dynamic system.
- Particles consistent with data are propagated; those inconsistent with data are pruned. Ancestry within the swarm can be represented as a tree.

  When elements of the swarm are trees, we have a tree of trees!
- The propagation and pruning are done in such a way that SMC approximates an ideal nonlinear filter.
- SMC gives unbiased likelihood estimates, with more particles giving reduced variance.

Basic SMC for a Markov process $\{X(t)\}$ with data $\{y_1, y_2, \dots\}$ observed at times $\{t_1, t_2, \dots\}$

- The conditional distribution of $X(t_n)$ given data up to time t_n is represented by a swarm of J particles, $\{X^F_{n,j}, j = 1, \dots, J\}$.
- Each particle $X^F_{n,j}$ is simulated forward from time t_n to time t_{n+1}. The resulting swarm, denoted $\{X^P_{n+1,j}, j = 1, \dots, J\}$, represents the distribution of $X(t_{n+1})$ given data up to time t_n.

  Stochasticity in the update adds phenotypic variation to the swarm

- Each particle in the prediction swarm is resampled with probability proportional to its likelihood given the observation at time t_{n+1}. The resulting swarm represents the filtering distribution at time t_{n+1}.

  This natural selection reduces phenotypic variation in the swarm.

When and why does SMC fail?

- We expect evolution to be good at finding fitness improvements locally in genetic space.
- We do not expect evolution (or evolutionary optimization algorithms, or any other known methods) to be good at finding globally optimal solutions to complex problems.

What is the globally "optimal" animal?

- The theory for SMC (and more generally, for MCMC, simulated annealing, etc) typically assures global convergence given sufficient computer time.
- Practical interpretation of this theory requires care! Practitioners should usually interpret global convergence results as guarantees of good local behavior.

Particle depletion

- Evolution since the most recent common ancestor (MRCA) defines the 'local' neighborhoods which an evolutionary search can hope to explore.
- When selection is strong (in other words, the offspring distribution is highly skewed) the MRCA can be only a few generations back even for a large number of particles (say, $J = 10^5$).
- The unpleasant phenomenon of the proximity of the MRCA and its consequences for global search is known as **particle depletion**.

Inference for POMPs

- SMC estimates latent dynamic variables, and integrates them out to approximate the likelihood, for a given model.
- Using these noisy and computationally expensive likelihood evaluations in Bayesian or frequentist inference is tricky. The issue has inspired much research over the past decade.
- **Iterated filtering** aims to maximize the likelihood by adding a random walk in parameter space.
  This adds diversity to the genotype of the swarm, so that selection on the phenotype improves its fitness (i.e., the likelihood).
- **Particle MCMC** aims to simulate a posterior distribution by plugging an SMC estiamte of the likelihood into an MCMC procedure in the parameter space.

The plug-and-play property

- Iterated filtering and particle MCMC share an important property: they require simualation from the dynamic model but they do NOT require explicit computation of transition probabilities.
- This is called the **plug-and-play** property.
- Plug-and-play facilitates model development (we can simualate from many models of interest for which transition probabilities are hard to compute).
- Plug-and-play facilitates software development: inference software simply takes as its argument code to generate simulations.
- Other plug-and-play methods, such as synthetic likelihood and approximate Bayesian computation (ABC) have also proved popular.

Categorizing some POMP inference methods

Plug-and-play

	Frequentist	Bayesian
Full-information	iterated filtering	particle MCMC
Feature-based	simulated moments	ABC

Not plug-and-play

	Frequentist	Bayesian
Full-information	EM algorithm	MCMC
Feature-based	Yule-Walker*	???

*Yule-Walker is the method of moments for ARMA, a linear Gaussian POMP.

Six problems of Bjørnstad and Grenfell (Science, 2001)

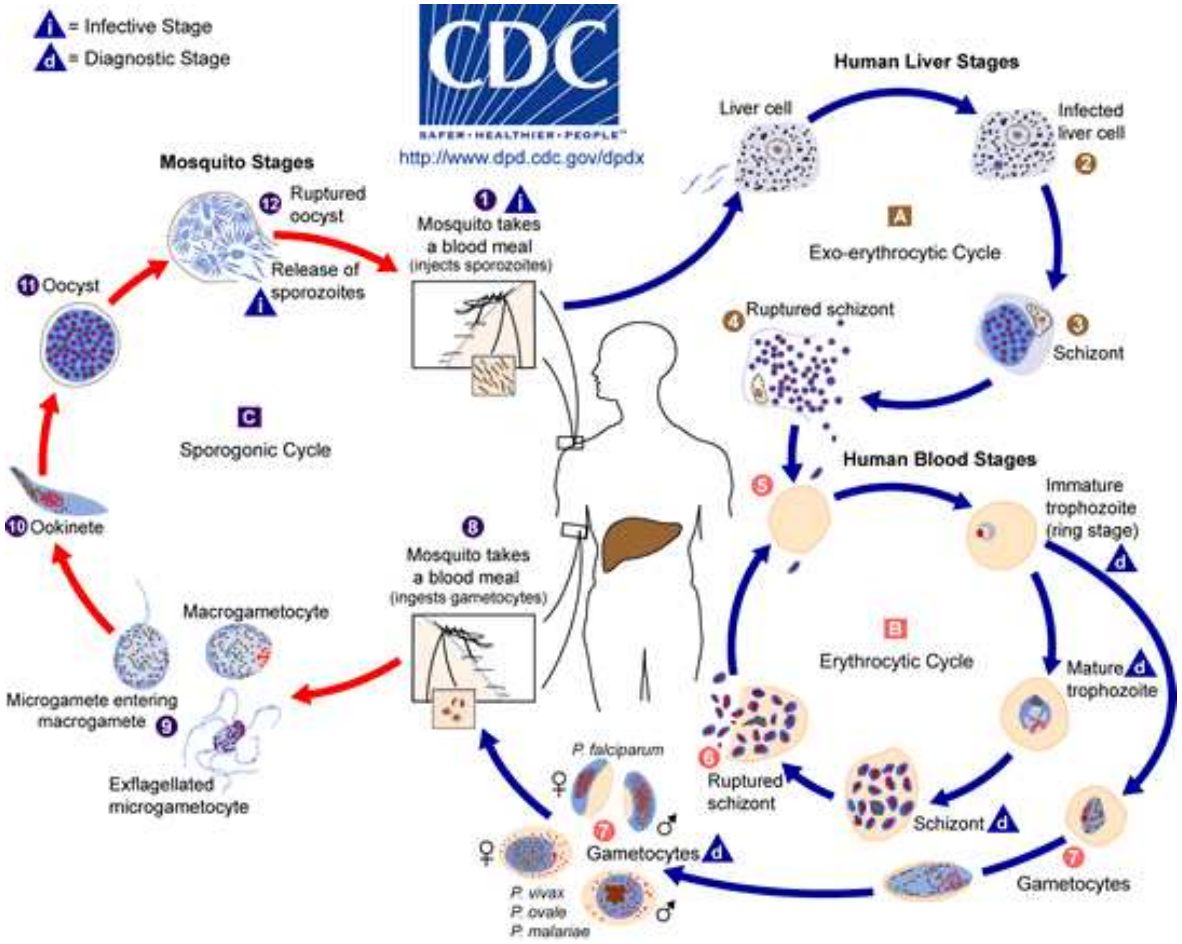
Obstacles for *ecological* inference via nonlinear mechanistic models:

1. Combining measurement noise and process noise.
2. Including covariates in mechanistically plausible ways.
3. Continuous time models.
4. Modeling and estimating interactions in coupled systems.
5. Dealing with unobserved variables.
6. Modeling spatial-temporal dynamics.

Spatial-temporal and high dimensional systems remain a challenge.

Genetic data is a new frontier!

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



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

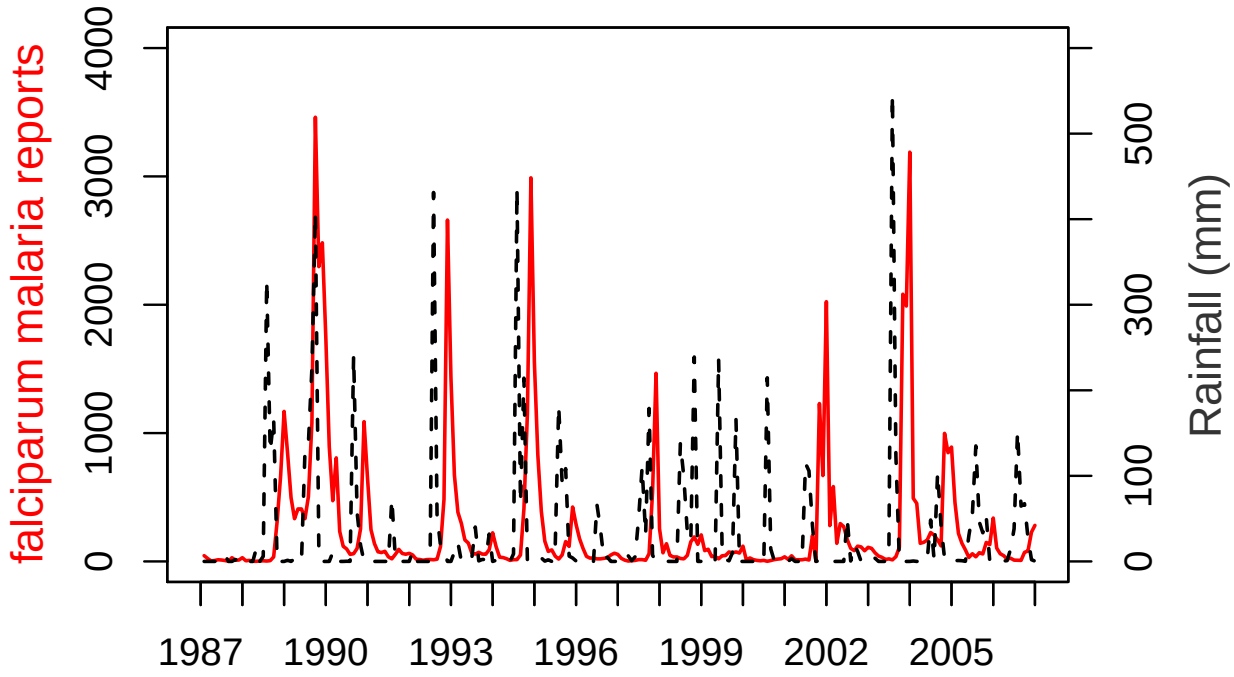
Malaria transmission: Modeling and inference

- The Gates Foundation targets eradication. The previous Global Malaria Eradication Program (1955-1969) ultimately failed, though with some lasting local successes.
- Malaria transmission dynamics have much local variation (vectors and their ecology; human behaviors).

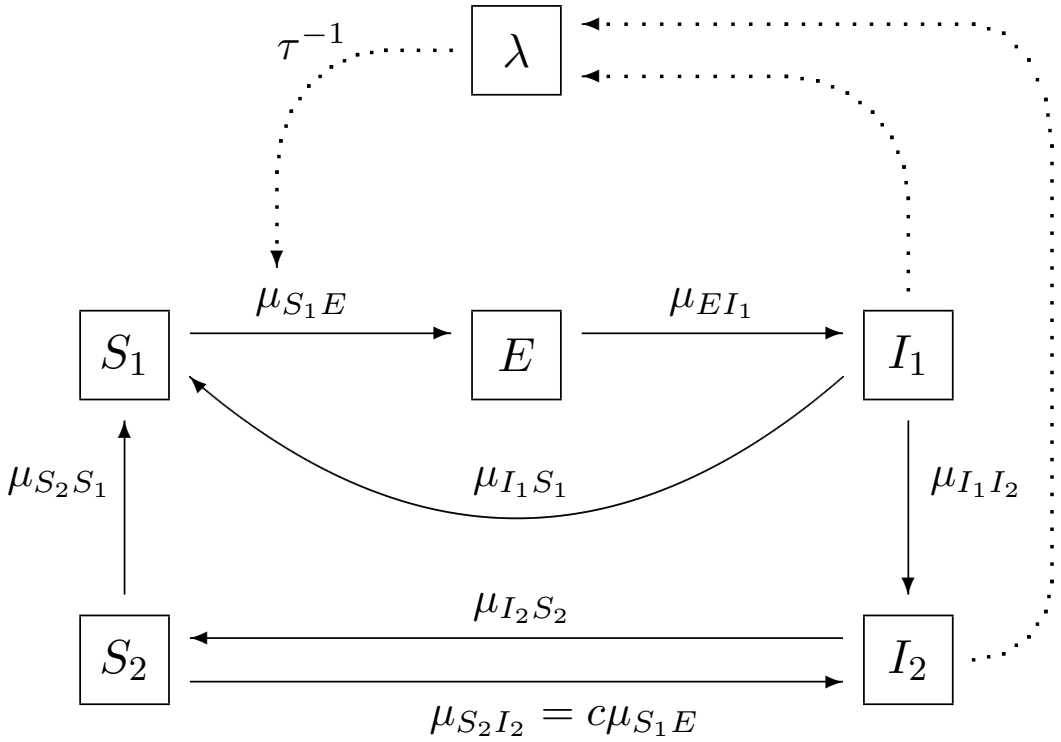
From the perspective of statistical methodology

- Despite the huge literature, no dynamic model of malaria transmission has previously been fitted directly to population-level time series data.
- Difficulties include: Incomplete and complex immunity; dynamics in both mosquito and human stages; non-specific diagnosis via 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?



Latent force of infection

Disease status of the human population

(Laneri et al, *PLoS Comp. Biol.*, 2010; Bhadra et al, *JASA*, 2011)

μ_{S_1E}, force of infection; λ, 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

$$
\begin{aligned}
dS_1/dt &= \mu_{BS_1} P - \mu_{S_1E} S_1 + \mu_{I_1S_1} I_1 + \mu_{S_2S_1} S_2 - \mu_{S_1D} S_1 \\
dS_2/dt &= \mu_{I_2S_2} I_2 - \mu_{S_2S_1} S_2 - \mu_{S_2I_2} S_2 - \mu_{S_2D} S_2 \\
dE/dt &= \mu_{S_1E} S_1 - \mu_{EI_1} E - \mu_{ED} E \\
dI_1/dt &= \mu_{EI_1} E - \mu_{I_1S_1} I_1 - \mu_{I_1I_2} I_1 - \mu_{I_1D} I_1 \\
dI_2/dt &= \mu_{I_1I_2} I_1 + \mu_{S_2I_2} S_2 - \mu_{I_2S_2} I_2 - \mu_{I_2D} I_2 \\
d\lambda_i/dt &= (\lambda_{i-1} - \lambda_i)\, k\, \tau^{-1} \qquad \text{for } i = 1, \ldots, k \\
\mu_{S_1E}(t) &= \lambda_k(t) \\
\lambda(t) &= \lambda_0(t) = \frac{I_1(t) + qI_2(t)}{N(t)} \exp\left\{ \sum_{i=1}^{n_s} \beta_i s_i(t) + Z_t\beta \right\} \frac{d\Gamma}{dt}.
\end{aligned}
$$

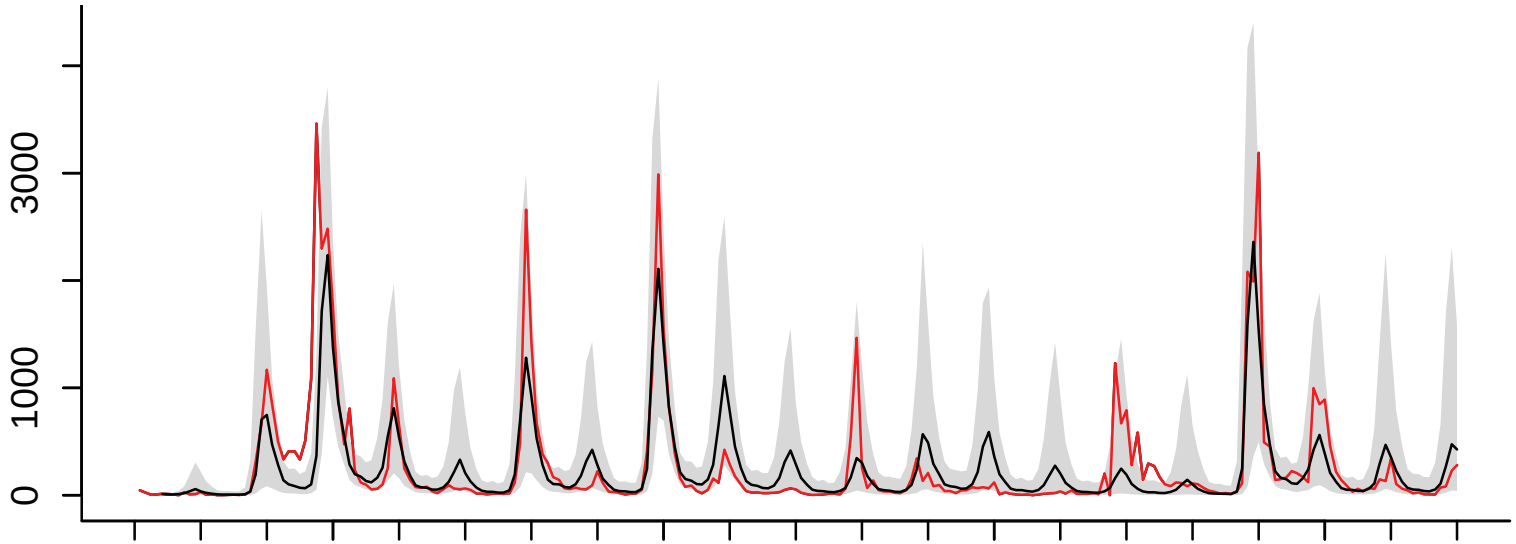
Z_t is a vector of climate covariates (here, rainfall).

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

Parasite latency within the vector has mean τ and shape parameter k.

Conclusions from malaria data analysis

- Rainfall (with an appropriate delay and threshold) has a critical role in determining interannual cycles.
- Immunity has a minor role, at a fast timescale (limiting 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.

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

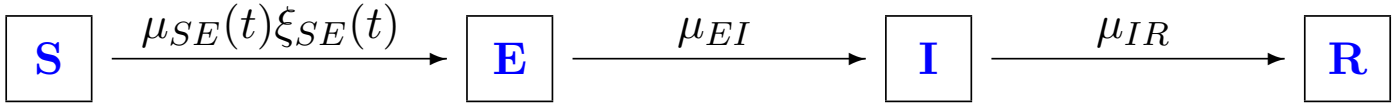
- **Remarkably, in the vast literatures on continuous-time individual-based Markov chains for population dynamics (e.g. applied to ecology and chemical reactions) no-one has previously proposed models capable of over-dispersion.**
- It turns out that the usual assumption that no events occur simultaneously 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 & Ionides (2011, *Stoc. Proc. Appl.*).

Measles: an exhaustively studied system

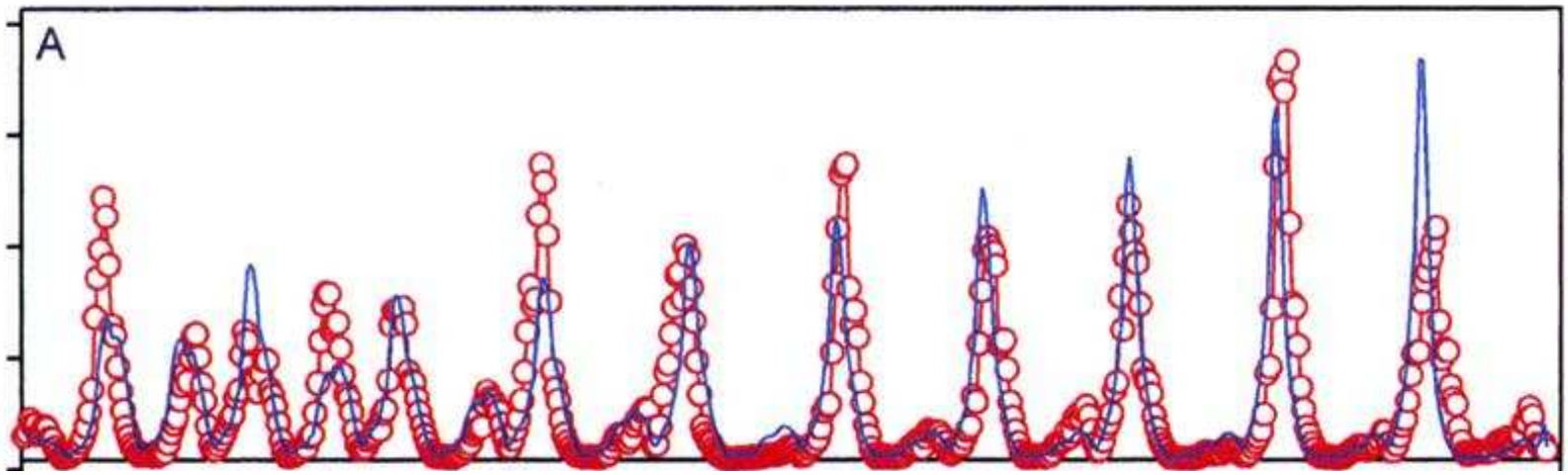
- Measles is simple: direct infection of susceptibles by infecteds; characteristic symptoms leading to accurate clinical diagnosis; life-long immunity following infection.



Susceptible $\longrightarrow$ Exposed (latent) $\longrightarrow$ Infected $\longrightarrow$ Recovered,

with noise intensity σ_{SE} on the force of infection.

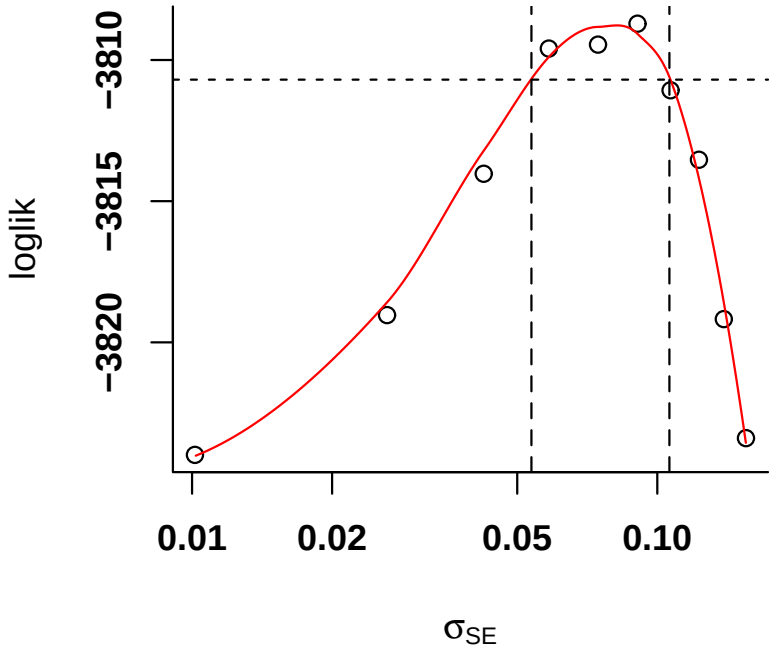
- Measles is still a substantial health issue in sub-Saharan Africa.
- A global eradication program is under debate.
- Comprehensive doctor reports in western Europe and America before vaccination (≈ 1968) are textbook data.

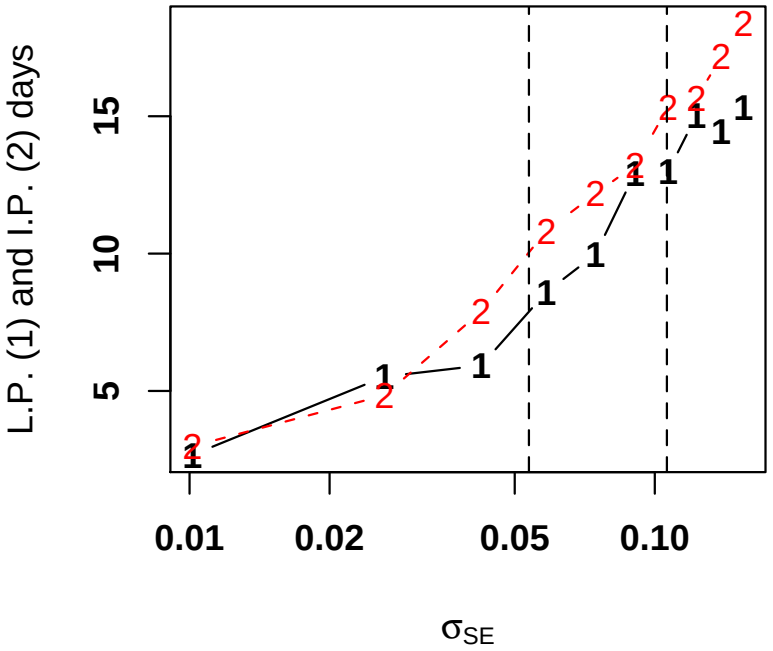


- Measles cases in London 1944–1965 (circles and red lines) and a deterministic SEIR fit (blue line) (from Grenfell *et al*, 2002).
- A deterministic fit, specified by the initial values in January 1944, captures remarkably many features.

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

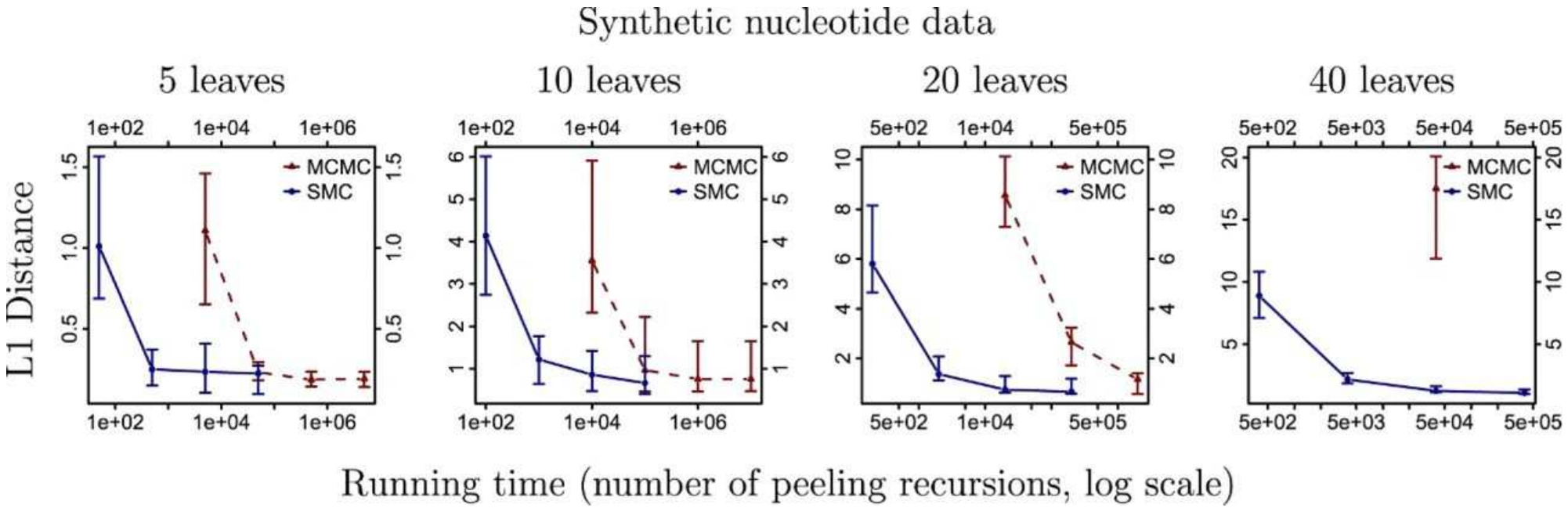
- Social and environmental events (e.g., football matches, weather) lead to stochastic variation in rates: **environmental stochasticity**.
- A catch-all for **model misspecification**? It is common practice in linear regression to bear in mind that the "error" terms contain un-modeled processes as well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

SMC conditional on a phylogeny (Rasmussen et al, 2011)

- If uncertainty in the phylogeny is negligible, then a coalescent process on this phylogeny can be used as a measurement model for applying SMC techniques to stochastic dynamic transmission models.
- This reduces the problem to nonlinear time series analysis, where the data are a time series of the number of coalescent times in small, discrete time intervals.
- Some uncertainty in the estimated phylogeny can be accounted for, but mutually consistent estimation of the phylogeny and the transmission model is currently unresolved.

SMC to estimate a phylogeny (Bouchard-Côté et al, 2012)

- Build a phylogeny backwards in time, so a “particle” is a forest of trees that combine as the filtering proceeds.
- Combine a coalescent prior on the forest with usual microevolutionary models for the sequences.



SMC for joint estimation of transmission dynamics and phylogeny

(Joint work with Alex Smith and Aaron King)

- For concreteness, focus on a high sequence fraction situation (HIV).
- Each particle is a transmission tree of all infected individuals in a population. Tree growth follows the forward-time dynamic model.
- Observations are assignments of sequences to branches on the tree.
- Currently, we can filter simulated data from simple models with, say, 100 observed sequences.
- Some improvements are expected through refining the code.
- **Is there hope for fundamental algorithmic developments to enable, say, 1000 sequences and 20-parameter models?**

Exorcising the curse of dimensionality for SMC

- Many things we'd like to do become exponentially harder with increasing dimension of the data and/or model. This is **the curse**.
- In general, SMC could require a number of particles exponential in the length of the time series. This is infeasible. SMC is numerically stable only when the Markov process has **temporal mixing** properties.
- Sadly, SMC requires a number of particles exponential in the state dimension. How can one take advantage of weak spatial coupling (here, space is a space of trees). A current research area (Rebschini and van Handel, 2013) that I'm currently investigating with Joon Ha Park.
- Weak coupling arises when lineages interact only through competition for susceptibles. It has been said that genetic data 'decorrelate' the lineages. But ecological competition still exists and can be critical.

Thank you!

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

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

References

[1] Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *Journal of the American Statistical Association*, 106:440–451.

[2] Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Systematic Biology*, 61(4):579–593.

[3] 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.

[4] 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*, 121:2571–2591.

[5] 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.

[6] Hammersley, J. M. and Morton, K. W. (1954). Poor man's Monte Carlo. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 16:23–38.

[7] 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.

[8] Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. A. (2011). Iterated filtering. *Annals of Statistics*, 39:1776–1802.

[9] 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.

[10] 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.

[11] Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear

epidemiological models using genealogies and time series. *PLoS Computational Biology*, 7(8):e1002136.

[12] Rebeschini, P. and van Handel, R. (2013). Can local particle filters beat the curse of dimensionality? *Arxiv*, page 1301.6585.