Statistical analysis of compartmental models: epidemiology, molecular biology, and everything in between

Vladimir N. Minin

Department of Statistics
University of California, Irvine

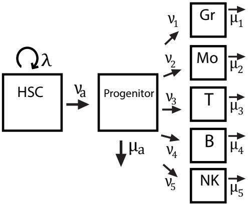
BIRS Workshop “Challenges in the Statistical Modeling of Stochastic Processes for the Natural Sciences"

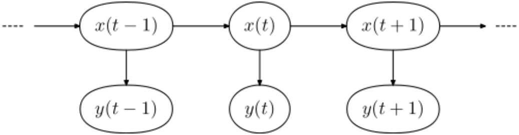
July 2017

joint work with Jason Xu (UW → UCLA), Sam Koelle (UW), Peter Guttorp (UW), Chuanfeng Wu (NIH), Cynthia Dunbar (NIH), Jan Abkowitz (UW), Amanda Koepke (NIST), Ira Longini (UF), Betz Halloran (Fred Hutch and UW), Jon Wakefield (UW), and Jon Fintzi (UW)

funding: NIH R01-AI107034 and NIH U54 GM111274

Hematopoiesis HMM driven by a branching process





- Latent process: each barcode lineage evolves as a multi-type branching process $\mathbf{X}(t)$ whose components are counts of each cell type.
- Observation process: multivariate hypergeometric distribution $\widetilde{\mathbf{Y}} \sim \text{mvhypgeo}(\mathbf{X})$ — driven by experimental design.
- Read data: read counts $\mathbf{Y}$ are proportional to $\widetilde{\mathbf{Y}}$ with unknown amplification constant.
- Likelihood is intractable due to massive size of the state space of the branching process.

Systems (molecular) biology

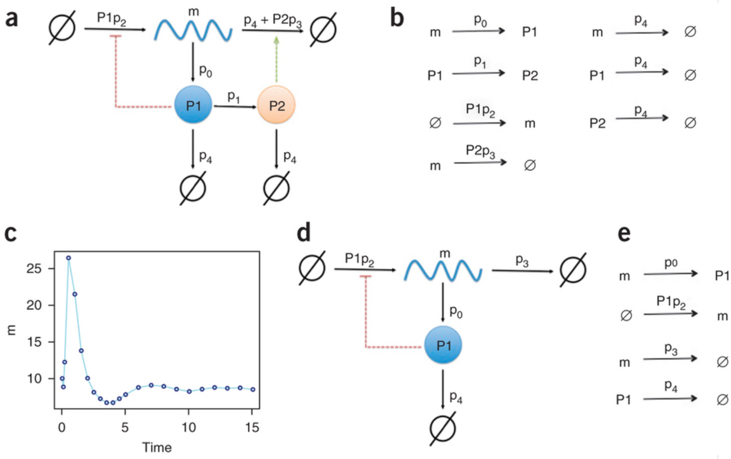
From

A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation

Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes & Michael P H Stumpf

Nature Protocols **9**, 439–456 (2014) doi:10.1038/nprot.2014.025

Published online 23 January 2014



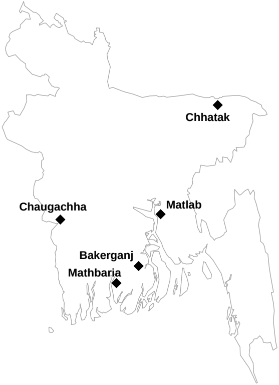
a
∅
P1p2
m
p4 + P2p3
∅
p0
P1
p1
P2
p4
p4
∅
∅
b
$m \xrightarrow{p_0} P1$
$P1 \xrightarrow{p_1} P2$
$\varnothing \xrightarrow{P1p_2} m$
$m \xrightarrow{P2p_3} \varnothing$
$m \xrightarrow{p_4} \varnothing$
$P1 \xrightarrow{p_4} \varnothing$
$P2 \xrightarrow{p_4} \varnothing$
c
m
25
20
15
10
0
5
10
15
Time
d
∅
P1p2
m
p3
∅
p0
P1
p4
∅
e
$m \xrightarrow{p0} P1$
$\varnothing \xrightarrow{P1p_2} m$
$m \xrightarrow{p_3} \varnothing$
$P1 \xrightarrow{p_4} \varnothing$

(**a**–**e**) The full mRNA self-regulation model is shown in **a**. mRNA (m) produces protein P1, which can be transformed into protein P2. P1 is required to produce mRNA, whereas P2 degrades mRNA into the empty set (–). P1 and P2 can also be degraded. The reactions that occur according to this model are shown in **b**. Fitting of the model to the data (**c**), which comprise mRNA measurements over time. The second model (**d**) is based on the first model, but it does not contain protein P2. The relevant reactions are shown in **e**.

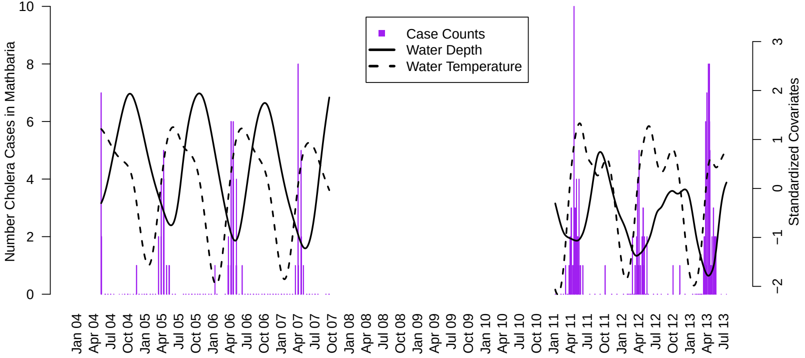
Case study: cholera in Bangladesh

Goal: To understand the dynamics of endemic cholera in Bangladesh and to develop a model that will be able to predict outbreaks several weeks in advance.

- Specifically, to understand how the disease dynamics are related to environmental covariates.



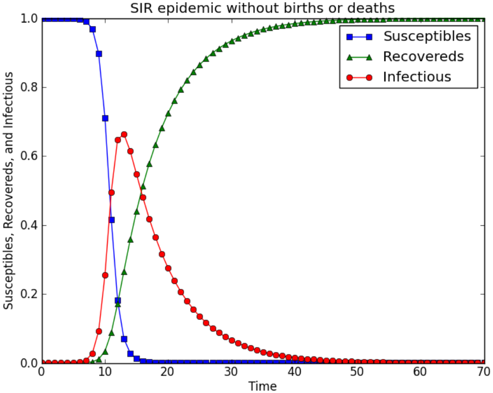
Mathbaria cholera incidence data and covariates



- Covariates are the smoothed standardized daily values $C_{WT}(i) = (WT(i) - \overline{WT})/s_{WT}$, $C_{WD}(i) = (WD(i) - \overline{WD})/s_{WD}$.
- We could use a phenomenological model (e.g., Poisson regression), but this would not tell us anything about the number of infected people, and disease transmission parameters.

Deterministic SIR-like models

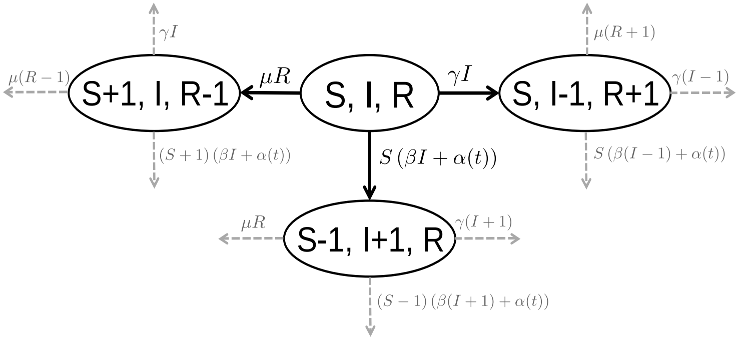
$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{SI}{N} - \nu S \\
\frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I \\
\frac{dR}{dt} &= \gamma I + \nu S \\
N &= S + I + R
\end{aligned}$$



Easy system of nonlinear differential equations. Models mean behavior of a stochastic system.

Problem: Poorly mimics stochastic dynamics when one of the compartments is low. Not clear how to model noisy data.
Solution: Use a real stochastic model (e.g., continuous-time Markov chain or SDE).

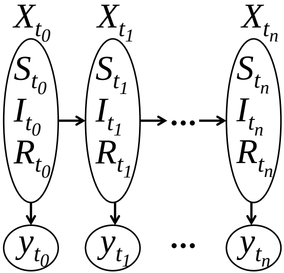
SIRS model



- S, I, and R denote the numbers of susceptible, infectious, and recovered individuals at time t; $N = S + I + R$

- β represents the infectious contact rate, $\alpha(t) = \alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 C_{WD}(i-\kappa) + \alpha_2 C_{WT}(i-\kappa)\right]$ represents the time-varying environmental force of infection, γ is the recovery rate, and μ is the rate at which immunity is lost.

Hidden SIRS model



- $\mathbf{X}_t$ only indirectly observed through $y_t \Rightarrow$ hidden Markov model (HMM)
- $y_t \sim \text{Binomial}(I_t, \rho)$
- ρ depends on the number of symptomatic infectious individuals that seek treatment

We are interested in the posterior distribution

$$\Pr(\boldsymbol{\theta}|\boldsymbol{y}) \propto \Pr(\boldsymbol{y}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta}),$$

where $\boldsymbol{y} = (y_{t_0}, \ldots, y_{t_n})$, $\boldsymbol{\theta} = (\beta, \gamma, \rho, \alpha_0, \ldots, \alpha_k)$, and

$$\Pr(\mathbf{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \left(\prod_{i=0}^{n} \Pr(y_{t_i}|I_{t_i}, \rho) \left[\Pr(\boldsymbol{X}_{t_0}|\phi_S, \phi_I) \prod_{i=1}^{n} p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_{i-1}}, \boldsymbol{\theta}) \right] \right) .$$

Problem: This likelihood is intractable, because the state space of X_t is too large even for moderately high population size N.

Estimation results

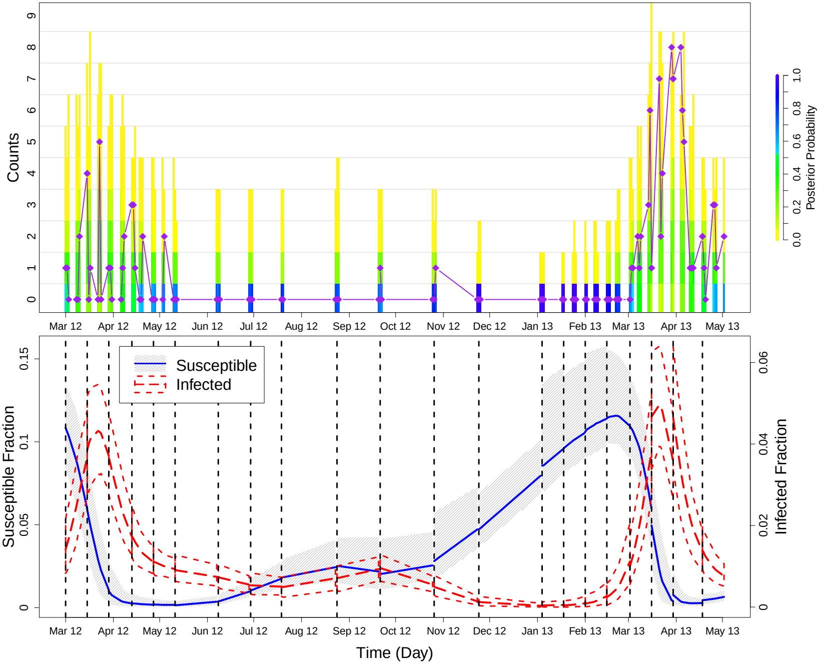
Posterior medians and 95% CIs for the parameters of the SIRS model.

Coefficient	Estimate	95% CIs
$\beta \times N$	0.491	(0.103, 0.945)
γ	0.115	(0.096, 0.142)
$(\beta \times N)/\gamma$	4.35	(0.99 , 7.15)
α_0	-5.32	(-6.63 , -4.51)
α_1	-1.37	(-1.98 , -0.98)
α_2	2.18	(1.8 , 2.62)
$\rho \times N$	55.8	(43.4 , 73.5)

Recall:

- N is the population size (set artificially to 10,000).
- β is the infectious contact rate.
- γ is the recovery rate.
- ρ is the reporting rate.
- $\alpha(t) = \alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 C_{WD}(i-\kappa) + \alpha_2 C_{WT}(i-\kappa)\right]$

Prediction results



What models cause all this trouble?

- Too broad of an answer: partially observed stochastic processes

What models cause all this trouble?

- Too broad of an answer: partially observed stochastic processes

- Too narrow of an answer: intractable HMMs

What models cause all this trouble?

- Too broad of an answer: partially observed stochastic processes
- Too narrow of an answer: intractable HMMs
- Mouthful, but good enough answer: Markov compartmental models, with a subset of compartments are observed over time with noise

Statistical inference options

- “Exact" likelihood-based methods:
  - Bayesian inference with particle MCMC (Koepke et al.)
  - Maximum likelihood inference with particle MCMC (Ionides et al.)

Statistical inference options

- “Exact" likelihood-based methods:
  - Bayesian inference with particle MCMC (Koepke et al.)
  - Maximum likelihood inference with particle MCMC (Ionides et al.)
- “Exact" summary statistics based methods:
  - Approximate Bayesian Computation (Liepe et al.)
  - Quasi- and pseudo-maximum likelihood estimators and generalized methods of moments (Chen and Hyrien, Xu et al.)

Statistical inference options

- “Exact" likelihood-based methods:
  - Bayesian inference with particle MCMC (Koepke et al.)
  - Maximum likelihood inference with particle MCMC (Ionides et al.)

- “Exact" summary statistics based methods:
  - Approximate Bayesian Computation (Liepe et al.)
  - Quasi- and pseudo-maximum likelihood estimators and generalized methods of moments (Chen and Hyrien, Xu et al.)

- Likelihood approximations:
  - Trajectory matching with iid normal error
  - Pretending some compartments change slowly (diffusion, branching processes)
  - Emulating with GPs to reduce costly likelihood evaluations (Jandarov et al.)
  - Linear noise approximation — makes Markov transition densities Gaussian, with complicated mean and covariance functions.

Open problems and glimpses of possible solutions

- Open problem:
  - Computationally stable inference procedure that can handle at least 10-100 compartments
- Current lines of attack:
  - Ho et al. use clever re-parameterization to develop a new deterministic algorithm for computing SIR (and more complex) transition densities “exactly." Doesn’t solve all the problems, because not all compartments are observed perfectly.
  - Approximate Bayesian computation and/or indirect inference. But model comparison is tough!
  - New data augmentation methods with block updates of latent variables.

References

- Xu J, Koelle S, Guttorp G, Wu C, Dunbar CE, Abkowitz JL, Minin VN. Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis, https://arxiv.org/abs/1610.07550.

- Koepke AA, Longini, IM, Halloran ME, Wakefield J, Minin VN. Predictive modeling of cholera outbreaks in Bangladesh, *Annals of Applied Statistics*, 10, 575–595, 2016, arXiv:1402.0536 [stat.AP].

- Fintzi J, Wakefield J, Minin VN. Efficient data augmentation for fitting stochastic epidemic models to prevalence data, arXiv:1606.07995 [stat.CO].

- Ho LST, Xu J, Crawford FW, Minin VN, Suchard, MA. Birth(death)/birth-death processes and their computable transition probabilities with statistical applications, arXiv:1603.03819 [stat.CO].

- http://www.stat.osu.edu/ pfc/BIRS/