

Likelihood-based Inference for Stochastic Epidemic Models via Data Augmentation

Jason Xu

Department of Statistical Science, Duke University

Design and Analysis of Infectious Disease Studies

Mathematisches Forschungsinstitut Oberwolfach

Feb 21, 2023

Stochastic SIR model

For any small interval of time h ,

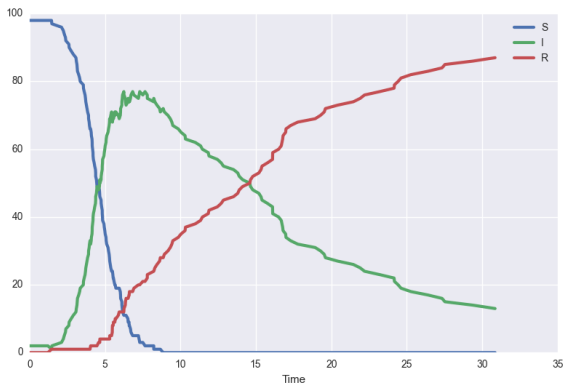
$$\Pr(\text{an infection occurs by } t + h) = \beta S(t)I(t)h + o(h)$$

$$\Pr(\text{a recovery occurs by } t + h) = \gamma I(t)h + o(h)$$

$$\Pr(\text{no event occurs by } t + h) = 1 - [\beta S(t)I(t) + \gamma I(t)]h + o(h)$$

- Replaces rates of change with jump probabilities (Kermack and McKendrick 1927); agree for large N [Kurtz 1978]
- A continuous-time Markov chain
- Enough to track S, I if $S(t) + I(t) + R(t) = N$, i.e. closed population
- Assumes random mixing: everyone is equally likely to come in contact with one another

Stochastic version of the SIR model



- Basic reproductive number $R_0 = \beta N / \gamma$: “**average** number of secondary infections” caused by an infectious individual.

Even the simple stochastic SIR is difficult to fit to data

Under the baseline model, it's already difficult to infer the parameters from data without additional simplifying assumptions

- Common to approximate the model (discrete-time, branching (early), diffusion (late), locally constant, TSIR)
- Simulation methods (SMC, particle filters, ABC) are powerful and popular, often suffer poor mixing/degeneracy, nontrivial

Instead, we pursue **likelihood-based** methods for fitting **continuous-time**, **stochastic** models in the **partially observed** setting

- In many ways, complementary to filtering methodology

Toward inference: the complete-data likelihood

A continuous-time Markov chain jumps between states according to exponential rates based on the generator \mathbf{Q} .

For instance, q_{xy} in \mathbf{Q} means the time T until the process jumps between two states $x, y \in \Omega$ is distributed $T \sim \exp(q_{xy})$.

- Diagonals $-q_j := \sum_{k \in \Omega} q_{jk}$ give rate of jumping out of j .
- Likelihood is simply a product of exponentials:

$$L_c(\boldsymbol{\tau}, \mathbf{x}; \mathbf{Q}) = q_{x_0} e^{-q_{x_0} \tau_1} q_{x_1} e^{-q_{x_1} (\tau_2 - \tau_1)} \dots \\ \cdot q_{x_{N_T-1}} e^{-q_{x_{N_T-1}} (\tau_{N_T} - \tau_{N_T-1})} e^{-q_{x_{N_T}} (T - \tau_{N_T})} \prod_{i=1}^{N_T} \frac{q_{x_{i-1} x_i}}{q_{x_{i-1}}}.$$

The complete-data likelihood of general CTMC

CTMC with rate matrix \mathbf{Q} : likelihood factorizes nicely

$$L_c(\theta; \mathbf{X}) = \prod_{\mathbf{x} \neq \mathbf{y}} \left(q_{\mathbf{xy}}^{N(\mathbf{x}, \mathbf{y})} \right) e^{\sum_{\mathbf{x}} \tau(\mathbf{x}) q_{\mathbf{xx}}}$$

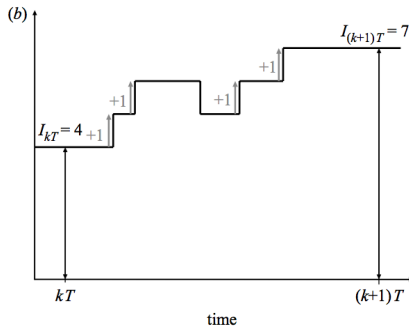
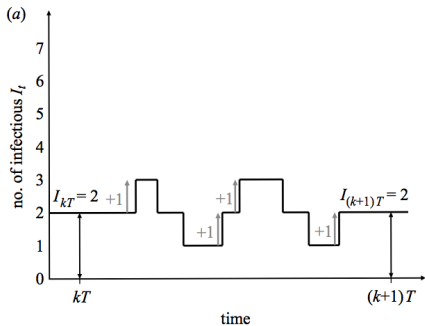
$$\ell_c(\theta; \mathbf{X}) = \sum_{\mathbf{x}} \sum_{\mathbf{x} \neq \mathbf{y}} N(\mathbf{x}, \mathbf{y}) \ln q_{\mathbf{xy}} - \sum_{\mathbf{x}} \sum_{\mathbf{x} \neq \mathbf{y}} \tau(\mathbf{x}) q_{\mathbf{xy}}$$

$N(\mathbf{x}, \mathbf{y})$ = total $\mathbf{x} \rightarrow \mathbf{y}$ transitions, $\tau(\mathbf{x})$ = total time spent in \mathbf{x}

- MLEs are easy: $\hat{q}_{\mathbf{xy}} = N(\mathbf{x}, \mathbf{y}) / \tau(\mathbf{x})$
- Posteriors with conjugate $\text{Gamma}(\alpha, \beta)$ priors also easy:

$$q_{\mathbf{xy}} | \mathbf{N}, \boldsymbol{\tau} \sim \text{Gamma}(N(\mathbf{x}, \mathbf{y}) + \alpha, \tau(\mathbf{x}) + \beta)$$

Challenge: partial observations



The discretely observed data log-likelihood

$$\ell_o(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{p=1}^m \sum_{i=0}^{n(p)-1} \log p_{\mathbf{x}^p(t_{p,i}), \mathbf{x}^p(t_{p,i+1})}(t_{p,i+1} - t_{p,i} | \boldsymbol{\theta})$$

In particular, composed of finite-time **transition probabilities**:

$$p_{\mathbf{x}, \mathbf{y}}(s) = \Pr(\mathbf{X}(t+s) = \mathbf{y} | \mathbf{X}(t) = \mathbf{x})$$

- Marginalized over infinitely many endpoint-conditioned paths (i.e. evaluating a hard integral). We've worked a lot on this!
- Classical matrix exponentiation for CTMCs is $\mathcal{O}(|\Omega|^3)$

$$\mathbf{P}(t) := \{p_{\mathbf{x}, \mathbf{y}}(t)\}_{\mathbf{x}, \mathbf{y} \in \Omega} = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{(\mathbf{Q}t)^k}{k!}.$$

Data augmentation: integration via latent sampling

The partial data likelihood is obtained by marginalizing over missing data denoted \mathbf{Z} :

$$\mathcal{L}_o(\mathbf{X}|\theta) = \int \mathcal{L}_c(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z}$$

When difficult to compute this integral directly, an alternative view entails proposing possible values of \mathbf{Z} , and construct a Markov chain to explore the *joint posterior* distribution $\pi(\mathbf{Z}, \theta|\mathbf{X})$.

By targeting the *joint posterior*, computations (MH acceptance ratios, etc) require the tractable joint likelihood \mathcal{L}_c above.

If we use MCMC to draw $\{(\mathbf{Z}_1, \theta_1), \dots, (\mathbf{Z}_m, \theta_m)\} \sim \pi(\mathbf{Z}, \theta|\mathbf{X})$, marginalizing out \mathbf{Z} in the posterior is trivial.

Data augmentation: integration via latent sampling

To answer original questions regarding model parameters θ given observations \mathbf{X} , we simply *ignore the samples of \mathbf{Z}_i* . That is, the remaining $\{\theta_1, \dots, \theta_m\}$ form a sample from the posterior $\pi(\theta|\mathbf{X})$

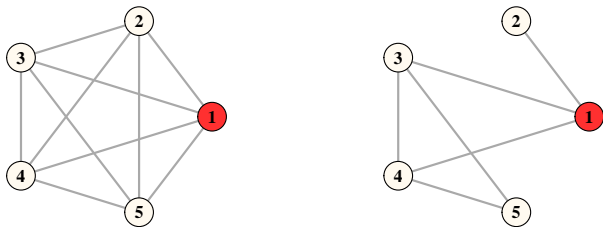
I will attempt to make the case that classical MCMC with data augmentation works well for the challenges in SIR models, and is extensible when relaxing various assumptions.

Example: relax random mixing with contact networks [Bu et al 2022]:

1. Formulate a model for co-evolution of SIR and contact network
2. Derive its complete data likelihood \mathcal{L}_c
3. Construct a sampler to augment the data to make use of \mathcal{L}_c

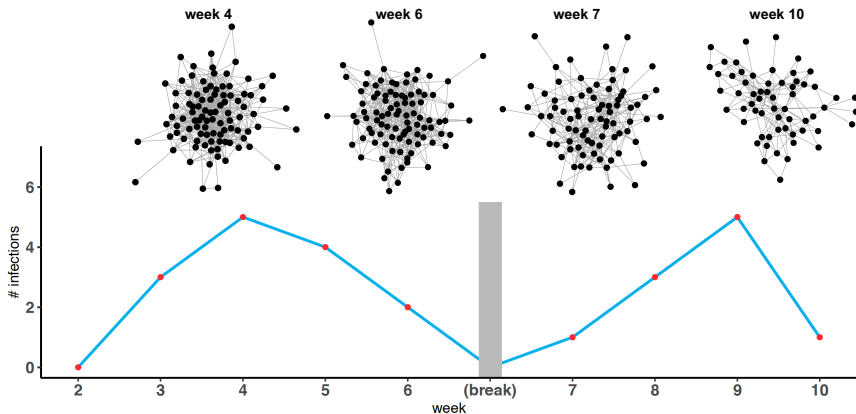
Beyond the well-mixed case

Naturally, one may be interested in transmission through **links** in a contact network (right panel).



- Recent growing interest in network setting, largely due to advances for gathering data at this resolution

SIR over a contact network: “mobile health” data



Motivating example: we will analyze eX-FLU data (Aiello et al, 2016) study; collected social contacts of 103 individuals at 5-min intervals using a Bluetooth app to study flu symptoms

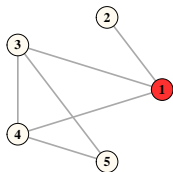
Goals

In this project, we sought to develop a framework that

1. models how the contact network impacts disease spread,
2. describes the influence of disease spread on network evolution,
3. enables inference based on **exact** likelihood of the stochastic SIR and network processes
4. accommodates **partial observations** (missing epidemic information but complete network information).

Notation and a very simple network process

Each individual's contacts can be represented via an adjacency matrix \mathbf{A} .

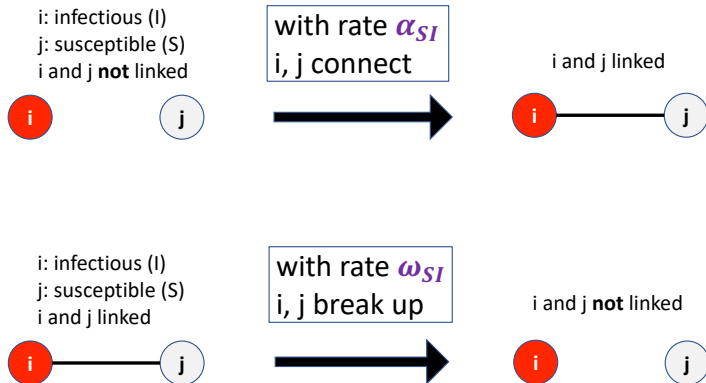


$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

To allow network to adapt, we will have each $\mathbf{A}_{ij}(t)$ evolve according to a Markov process taking values of 0 or 1.

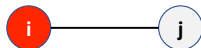
In particular, the rates of edge creation α_{AB} and deletion ω_{AB} should depend on disease status A, B of the two individuals.

Individual-level network events



Individual-level SIR dynamics

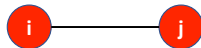
i: infectious
j: susceptible
i and j in contact



with rate β
i infects j

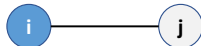


i and j both infected



with rate γ
i recovers

i recovered, j susceptible



Joint dynamics of the model

At the individual level, four competing Poisson processes:

1. **Infection:** an I individual infects an S neighbor with rate β ;
2. **Recovery:** an I individual recovers with rate γ independently;
3. **Link activation:** a link is formed at rate α_{AB} between a status A individual and a status B individual;
4. **Link termination:** a link is removed at rate ω_{AB} analogously
5. Assume recovered and susceptible individuals behave identically in terms of the network

Dynamics of the overall generative model

At the population level, the entire process sums over all individual-level rates (superposition property):

1. **Infection** occurs with rate $\beta SI(t)$, where $SI(t)$ = number of S - I links at time t ;
2. **Recovery** occurs with rate $\gamma I(t)$, $I(t)$ = number of I individuals at time t ;
3. **Link activation** for A - B pairs occurs with rate $\alpha_{AB} M_{AB}^d(t)$, $M_{AB}^d(t)$ = number of disconnected A - B pairs at time t ;
4. **Link termination** for A - B pairs is dissolved with rate $\omega_{AB} M_{AB}(t)$, $M_{AB}(t)$ = number of connected A - B pairs at time t .

Deriving the complete-data likelihood

Given the initial network structure \mathcal{G}_0 initial infective at time 0,

$$\begin{aligned} & \mathcal{L}(\beta, \gamma, \tilde{\alpha}, \tilde{\omega} | \mathcal{G}_0) \\ &= \gamma^{n_R} \beta^{n_E - 1} \alpha_{SS}^{C_{HH}} \alpha_{SI}^{C_{HI}} \alpha_{II}^{C_{II}} \omega_{SS}^{D_{HH}} \omega_{SI}^{D_{HI}} \omega_{II}^{D_{II}} \prod_{j=2}^n \left[\tilde{M}(t_j) (I_{p_{j1}}(t_j))^{F_j} \right] \\ & \times \exp \left(- \int_0^{T_{\max}} [\beta SI(t) + \gamma I(t) + \tilde{\alpha}^T \mathbf{M}_{\max}(t) + (\tilde{\omega} - \tilde{\alpha})^T \mathbf{M}(t)] dt \right). \end{aligned}$$

with parameters $\Theta = \{\beta, \gamma, \alpha_{SS}, \alpha_{SI}, \alpha_{II}, \omega_{SS}, \omega_{SI}, \omega_{II}\}$

- This takes same simple form presented in beginning of talk, just with more notation/bookkeeping (see paper for full notation!)
- Difficult to write the likelihood for very similar models, i.e. preventative rewiring [Ball & Britton 2022]

Closed form inference

Analogously to the introduction slides: when completely observed,

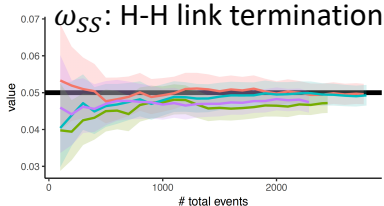
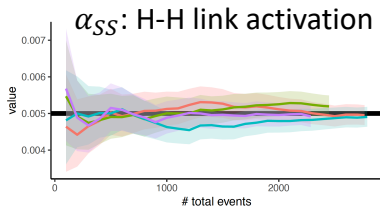
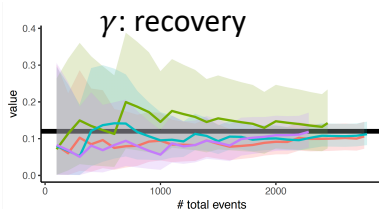
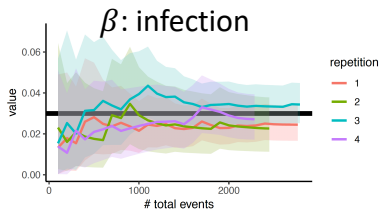
$$\begin{aligned}\hat{\beta} &= \frac{n_E - 1}{\sum_{j=1}^n SI(t_j)(t_j - t_{j-1})}, & \hat{\gamma} &= \frac{n_R}{\sum_{j=1}^n I(t_j)(t_j - t_{j-1})}, \\ \hat{\alpha}_{SS} &= \frac{C_{HH}}{\sum_{j=1}^n \left[\frac{H(t_j)(H(t_j)-1)}{2} - M_{HH}(t_j) \right] (t_j - t_{j-1})}, & \hat{\omega}_{SS} &= \frac{D_{HH}}{\sum_{j=1}^n M_{HH}(t_j)(t_j - t_{j-1})}, \\ \hat{\alpha}_{SI} &= \frac{C_{HI}}{\sum_{j=1}^n [H(t_j)I(t_j) - M_{HI}(t_j)] (t_j - t_{j-1})}, & \hat{\omega}_{SI} &= \frac{D_{HI}}{\sum_{j=1}^n M_{HI}(t_j)(t_j - t_{j-1})}, \\ \hat{\alpha}_{II} &= \frac{C_{II}}{\sum_{j=1}^n \left[\frac{I(t_j)(I(t_j)-1)}{2} - M_{II}(t_j) \right] (t_j - t_{j-1})}, & \hat{\omega}_{II} &= \frac{D_{II}}{\sum_{j=1}^n M_{II}(t_j)(t_j - t_{j-1})}.\end{aligned}$$

For Bayesian approach, Gamma priors again offer conjugacy:

$$\begin{aligned}\beta &\sim Ga(a_\beta, b_\beta), & \gamma &\sim Ga(a_\gamma, b_\gamma), \\ \alpha_{..} &\sim Ga(a_\alpha, b_\alpha), & \omega_{..} &\sim Ga(a_\omega, b_\omega).\end{aligned}$$

Simulation study: inference with complete data

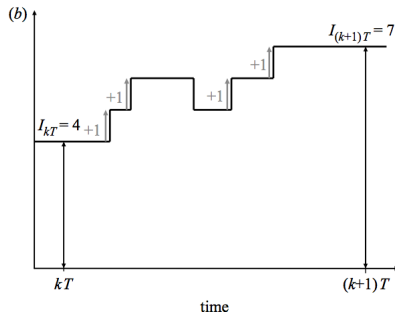
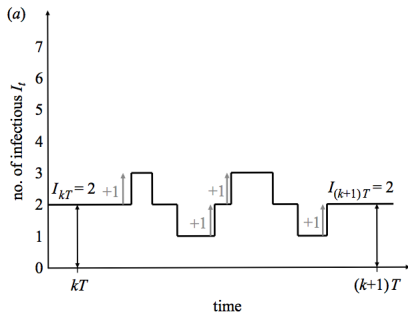
- Population size $N = 100$, 4 repetitions;
- Posterior samples concentrate around the truth (**bold line**)



Inference for partially observed case

Having validated the likelihood, how do we use it toward inference in the partially observed case?

Recall we need a way to draw valid samples of the missing events \mathbf{Z} , conditional on the model and observed data. In the case of the real data, \mathbf{Z} is the set of **unknown recovery times**.



Data setting: eX-FLU

Spread of influenza on campus (Aiello et al. 2016) with weekly health status surveys and high-resolution contact network info via Bluetooth app

- $\{(u_\ell, u_{\ell+1}]\}_{\ell=1}^L$: intervals (weeks) during which “health status reports” are collected
- Times of social link creation/deletion are **known continuously**
- In the ℓ th interval, there are R_ℓ new recoveries (summary data), but their event times are unknown
- Latent variables: $\mathbf{Z} = \{r_{\ell,1}, \dots, r_{\ell,R_\ell}\}$.
- Observed data: $\mathbf{x} = \text{events} + \text{endpoint statuses}$.

A simple algorithmic primitive

Proposed inference method: data-augmented MCMC.

for $s = 1 : \text{maxIters}$:

1. **Data augmentation.** For observation times $\ell = 1 : L$, draw recovery times $\{r_{\ell,i}^{(s)}\}_{i=1:R_\ell} = \mathbf{Z}$ from their joint **conditional**

$$p\left(\{r_{\ell,i}\}_{i=1:R_\ell} \mid \Theta^{(s-1)}, \mathbf{x}, \{r_{\ell',i}\}_{i=1:R_{\ell'}, \ell' \neq \ell}\right). \quad (1)$$

2. **Update parameter values.** Now that \mathbf{x} augmented by \mathbf{Z} comprise the **complete** data, **Gibbs sample** parameters $\Theta^{(s)}$ from their (conditional) posteriors.

Conditional sampling/simulation is nontrivial

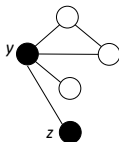
1. Simulating CTMC with fixed endpoints **notoriously difficult**
2. Must respect constraints imposed by the contact network structure.
3. Previous work in well-mixed case proposes individual disease histories one by one, and require intensive Metropolis-Hastings steps based on marginal likelihood

Fortunately,

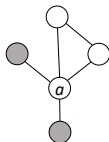
- Network structure “reduces dimension” of the latent space to be explored (**mechanistic information**)
- Smart PhD students can think hard and derive a Gibbs step!

Data Augmentation Regulated by Contact Info (DARCI)

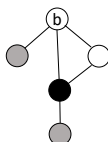
At time u , both y and z are infected and the other three are susceptible.



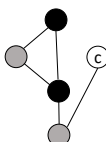
At time i_a , at least one of y and z has to be infected in order to infect a .



At time i_b , b can get infected by a , so both y and z can recover by then.



At time i_c , z has to be infected in order to infect c .



At time v , both y and z are recovered and the other three are infected.

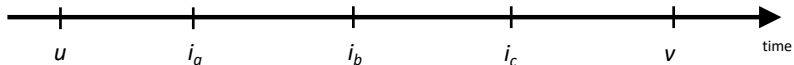
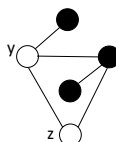


Figure: How DARCI “imputes” y and z ’s unknown recovery times in $(u, v]$. ● = known infected, ● = possibly infected, ○ = healthy.

Gibbs sampling from conditionals for parameters

$$\beta | \text{Data} \sim \text{Ga} \left(a_\beta + n_E - 1, b_\beta + \sum_{j=1}^n SI(t_j)(t_j - t_{j-1}) \right),$$

$$\gamma | \text{Data} \sim \text{Ga} \left(a_\gamma + n_R, b_\gamma + \sum_{j=1}^n I(t_j)(t_j - t_{j-1}) \right),$$

$$\alpha_{SS} | \text{Data} \sim \text{Ga} \left(a_\alpha + C_{HH}, b_\alpha + \sum_{j=1}^n [H(t_j)(H(t_j) - 1)/2 - M_{HH}(t_j)] (t_j - t_{j-1}) \right),$$

$$\alpha_{SI} | \text{Data} \sim \text{Ga} \left(a_\alpha + C_{HI}, b_\alpha + \sum_{j=1}^n [H(t_j)I(t_j) - M_{HI}(t_j)] (t_j - t_{j-1}) \right),$$

$$\alpha_{II} | \text{Data} \sim \text{Ga} \left(a_\alpha + C_{II}, b_\alpha + \sum_{j=1}^n (I(t_j)(I(t_j) - 1)/2 - M_{II}(t_j)) (t_j - t_{j-1}) \right),$$

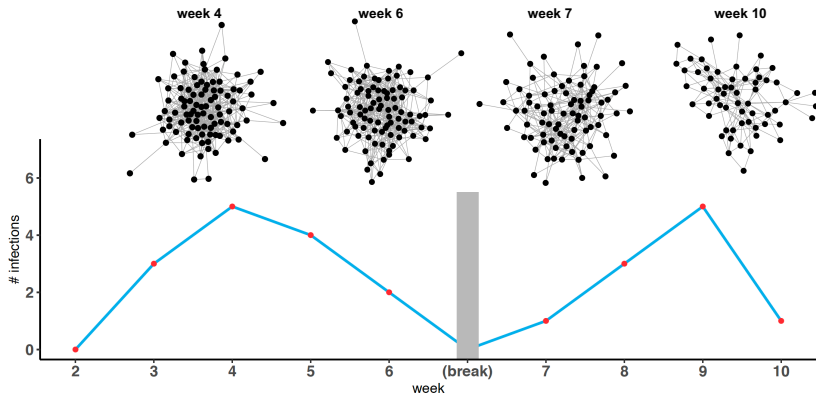
$$\omega_{SS} | \text{Data} \sim \text{Ga} \left(a_\omega + D_{HH}, b_\omega + \sum_{j=1}^n M_{HH}(t_j)(t_j - t_{j-1}) \right),$$

$$\omega_{SI} | \text{Data} \sim \text{Ga} \left(a_\omega + D_{HI}, b_\omega + \sum_{j=1}^n M_{HI}(t_j)(t_j - t_{j-1}) \right),$$

$$\omega_{II} | \text{Data} \sim \text{Ga} \left(a_\omega + D_{II}, b_\omega + \sum_{j=1}^n M_{II}(t_j)(t_j - t_{j-1}) \right).$$

eX-FLU data

- 2013 study on influenza transmission among 590 college students over 10 weeks
- 45,760 network events with low density



Results and analysis

- The data-augmented inference scheme is employed based on the exact model likelihood (posterior)
- Modification allows for external infection rate ξ
- Posterior inference gives us summaries of parameters as well as uncertainty: strong internal force of infection with slow transmission, quick recovery, short contacts.

Parameter	Mean	2.5%	97.5%
β (internal infection)	0.0695	0.0247	0.1500
ξ (external infection)	0.00331	0.00208	0.00494
γ (recovery)	0.294	0.186	0.428
α_{SS} (H - H link activation)	0.0514	0.0499	0.0529
ω_{SS} (H - H link termination)	38.26	33.55	40.62
α_{SI} (H - I link activation)	0.130	0.0785	0.194
ω_{SI} (H - I link termination)	53.5	22.5	231.7

Building upon the foundation

- Use an inhomogeneous bivariate branching process as faithful proposal density when *no event times are known* (for instance, incidence data) [Morsomme and Xu 2022+]
- Initialized at low density regions, appears to reach stationarity after 100 iterations and yields 100,000 posterior samples in a couple of minutes on a simulation with $S_0 = 1,000$

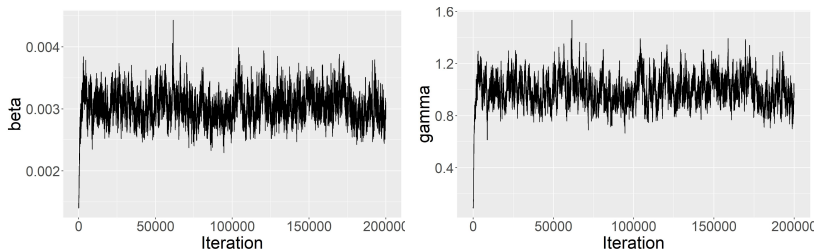


Figure: Traceplots of data-augmented MCMC with $\beta_0 = 0.003, \gamma_0 = 1$.

The piecewise-decoupled process

Under the SIR, the *population* infection rate $\mu_{pop}(t) = \beta I(t)S(t)$ changes after every event. Consider the approximation:

Let $\tilde{\mu}_{pop}$ be decoupled from only $I(t)$ dynamics (PD-SIR).

$$\tilde{\mu}_{pop}(t) := \beta S(t)I(t_{k-1}), \quad t \in [t_{k-1}, t_k).$$

- The corresponding *individual* infection rate is

$$\tilde{\mu}(t) := \frac{\tilde{\mu}_{pop}(t)}{S(t)} = \beta I(t_{k-1}) = \mu_k, \quad \text{piece-wise constant!}$$

- As a result, $S(t)$ compartment follows a linear death process (LDP)— here infections correspond to “deaths”.
- Proposal is equivalent to a **bivariate** branching process on each interval, deteriorating only on that interval before “resetting”

A Useful Theorem

Theorem

Consider a LDP with death rate μ and let $\tau_{1:N} \in (t_l, t_u]$ be the times of N deaths occurring between times t_l and t_u . Then

$$\tau_j \stackrel{d}{=} X_{(j)}, \quad j = 1, \dots, N$$

where $X_{(j)}$ is the j^{th} order statistic of N IID truncated exponential variables with rate μ , lower bound t_l and upper bound t_u .

This will allow us to Gibbs sample *all infection times* in an outbreak under the PD-SIR process in **one vectorized draw!**

Simulating the PD-SIR – pseudo-algorithm

Algorithm 1: Simulating a PD-SIR process conditionally on \mathbf{Y}

Output: $\mathbf{Z}^* = \{(z_i^I, z_i^R)\}_i$ compatible with $\mathbf{Y} = I_{1:K}$

for *interval* $k = 1, \dots, K$ **do**

 Compute the infection rate:

$$\mu_k \leftarrow \beta I(t_{k-1})$$

 Jointly generate the infection times:

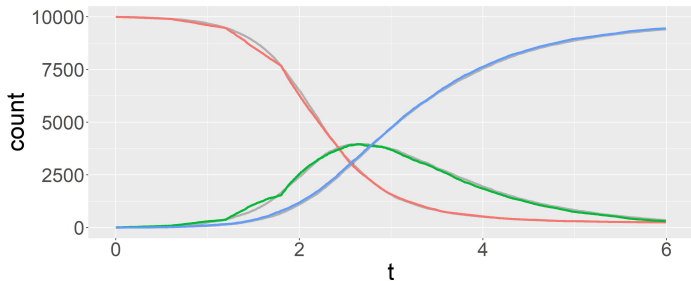
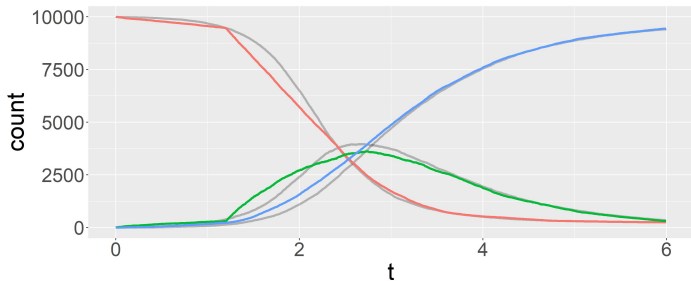
$$z_j^I \stackrel{iid}{\sim} \text{TruncExp}(\mu_k; t_{k-1}, t_k), \text{ for } j \in \mathcal{J}_k$$

 Generate the removal times:

$$z_j^R \mid z_j^I \stackrel{indep.}{\sim} z_j^I + \text{Exponential}(\gamma), \text{ for } j \in \mathcal{J}_k$$

end

A faithful approximation



A faithful approximation

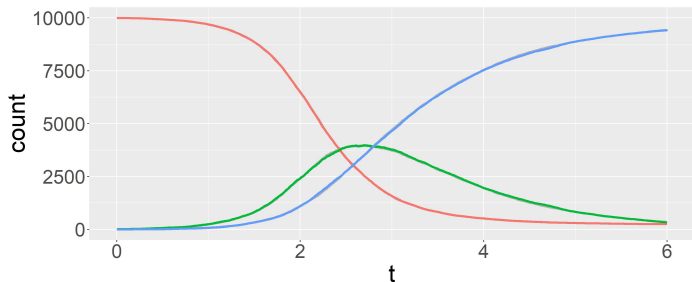


Figure: SIR (with S in red, I in green and R in blue) and a PD-SIR (in grey) trajectories under the same infection incidence data $I_{1:K}$ with $K = 5, 10, 50$ total observations.

2014-2016 Ebola Pandemic in Western Africa

- Weekly incidence data in Gueckedou, $n \approx 300,000$; $\rho = 0.1$
- Can sample 1,000,000 draws from posterior in half an hour on a Macbook, about 20% acceptance rate
- Overcomes computational limitations while enabling classical MCMC ideas [O'Neill and Roberts 1999, Gibson & Renshaw 2001, Neal & Roberts 2005]

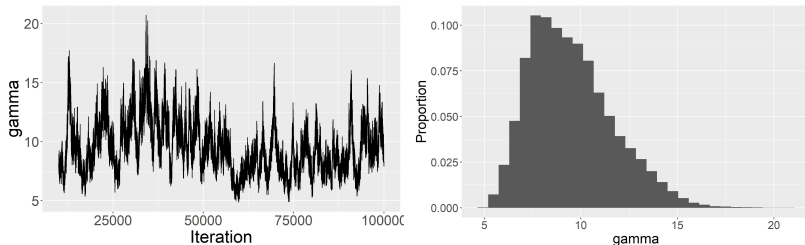


Figure: Posterior draws of $1/\gamma$, expected infection period, Ebola data

Extensions: Arbitrarily Distributed Infection Periods

- Infection times are the hard part— each removal is *conditional* on its corresponding infection event
- Weibull distribution, for instance, may be more plausible for infectious periods.
- Dropping exponentially distributed infectious period assumption \Rightarrow **non-Markovian**.
- Yet, the likelihood still has a nice, closed form! Inference carries through straightforwardly

Time-varying Infection Rate $\beta(t)$

- A good idea: $\beta \sim GP(.)$ (Kypraios, 2018)
 - expensive: invert a matrix of order $n \times n$ each iteration
- Locally adaptive: $\Delta(\beta)_k$ follows a Laplace or HS distribution (Faulkner & Minin, 2018)
 - accommodates sudden variations in β .
- Our first extension: piece-wise constant β_k per interval, promoting sparsity in $(\beta_{k+1} - \beta_k)$
 - Makes use of the nice Gibbs steps from this method while performing change point detection on where breaks occur
 - Very simple HMM prior favoring local constancy

Closing thoughts

Fully stochastic modeling: interpretability, uncertainty quantification, and [eventually] reliable forecasting/decision-making

CTMC framework is extensible in many open directions and yields tractable computation and workhorse inferential procedures + statistical guarantees

Many open directions and extensions toward model realism

- Covariates, heterogeneous networks, human behavior?
- Multiple strains, compartments, covariates, reporting rates
- Network missingness/better random graph models
- Hybrid differential equation models, sequential MCMC/ABC

Thank you!

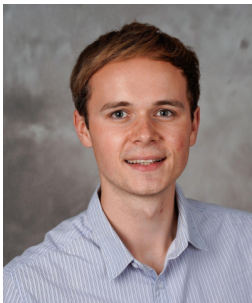


Figure: Students: Fan Bu, Raphael Morsomme, Jenny Huang

Incomplete references and collaborators

Likelihood-based Inference for Partially Observed Epidemics on Dynamic Networks. Bu, Aiello, Xu*, Volfovsky* 2022, Journal of the American Statistical Association.

Exact Inference for Stochastic Epidemic Models via Uniformly Ergodic Block Sampling. Morsomme and Xu, 2022, arXiv preprint.

Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity. Bu, Aiello, Volfovsky*, Xu*, 2021+, arXiv preprint.

Detecting Changes in the Transmission Rate of a Stochastic Epidemic Model. Huang, Morsomme, Dunson, Xu, 2023, arXiv preprint..

Birth/birth-death processes and their computable transition probabilities with biological applications. Ho, Xu, Crawford, Minin, Suchard 2018, Journal of Mathematical Biology.

Slide appendix

We've taken approaches via generating functions ϕ

$$\begin{aligned}\phi_{jk}(t, s_1, s_2; \theta) &= E_{\theta} \left(s_1^{X_1(t)} s_2^{X_2(t)} | X_1(0) = j, X_2(0) = k \right) \\ &= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} p_{(jk),(lm)}(t; \theta) s_1^l s_2^m; \quad |s_i| \leq 1\end{aligned}$$

- We derive **differential equations** governing ϕ_{jk} using the Kolmogorov forward/backward equations
- ϕ_{jk} may have closed solutions or cheap numerical solutions

Transition probabilities are related to the PGF via differentiation:

$$p_{(jk),(lm)}(t) = \left. \frac{1}{l!m!} \frac{\partial^l}{\partial s_1} \frac{\partial^m}{\partial s_2} \phi_{jk}(t) \right|_{s_1=s_2=0}.$$

Repeated differentiation impractical, numerically unstable

From differentiation to integration: series inversion

- Let $s_1 = e^{2\pi i w_1}$, $s_2 = e^{2\pi i w_2} \Rightarrow \phi$ becomes a Fourier series:

$$\phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) = \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} p_{(jk), (lm)}(t) e^{2\pi i l w_1} e^{2\pi i m w_2}$$

$$p_{(jk), (lm)}(t) = \int_0^1 \int_0^1 \phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) e^{-2\pi i l w_1} e^{-2\pi i m w_2} dw_1 dw_2$$

(Fourier inversion + Riemann approximation)

$$\approx \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) e^{-2\pi i l u/N} e^{-2\pi i m v/N}.$$

- simultaneously* compute set of probabilities $\{p_{(jk), (lm)}(t)\}$ for all $l, m = 0, \dots, N$ using **Fast Fourier Transform (FFT)** [Xu, Guttorp, Kato-Maeda, Minin 2015]

Nonlinear transition probabilities for marginal likelihood

Much harder to apply these ideas to **non-linear processes**—recall in SIR, interaction term $\beta S(t)I(t)$

Very briefly, working in the Laplace domain,

$$\phi_{ab}(s) := \mathcal{L}[p_{(jk),(ab)}(t)](s) = \int_0^\infty e^{-st} p_{(jk),(ab)}(t) dt$$

satisfies a recursion with continued fraction representation

$$\phi_{ab}^{(0)}(s) = \prod_{i=1}^b x_{ai} \frac{x_{a,b+1}}{y_{a,b+1} + \frac{x_{a,b+2} y_{ab}}{y_{a,b+2} + \frac{x_{a,b+3}}{y_{a,b+3} + \frac{x_{a,b+4}}{y_{a,b+4} + \cdots}}}}$$

- Evaluate to finite depth with bound on error

A brief outline: direct computation of SIR likelihood

“The associated mathematical manipulations required to generate solutions can only be described as heroic.”

- Eric Renshaw on computing the SIR marginal likelihood, Stochastic Population Processes.
- Compute continued fraction by recursion in Laplace domain
- Numerically invert Laplace transform to recover transition probabilities [Abate and Whitt]
- The marginal likelihood is simply a product of these quantities [Ho et al 2018, J. Math. Bio.]
- We’ve also used bivariate branching process approximations that allow for more efficient series inversion approaches via their probability generating functions

Simulation study: partially observed study

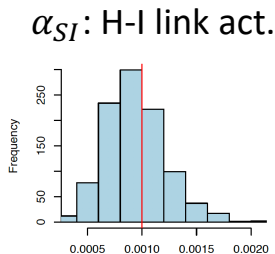
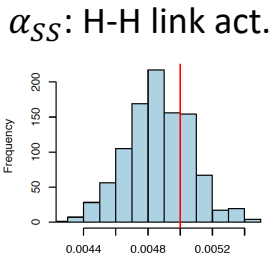
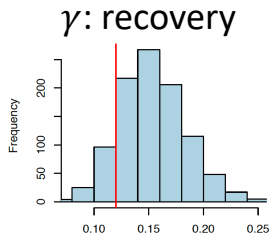
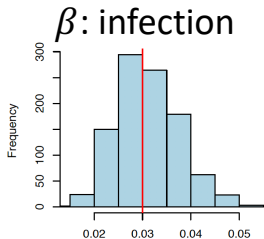
Forward simulate from SIR model and record the complete data

Time	Event Type	Individual(s) Involved
.....
95.8352	Link activation	100 & 5
95.8357	Link termination	148 & 157
95.8361	Recovery	125
95.8432	Link activation	62 & 147
95.8473	Recovery	16
95.8509	Infection	124
.....

Table: A sample of simulated complete data.

Hide all recovery times to compile “weekly health reports” to mimic discretely observed nature in real data

Accurate inference via data augmented MCMC



Posterior for infection and recovery rates

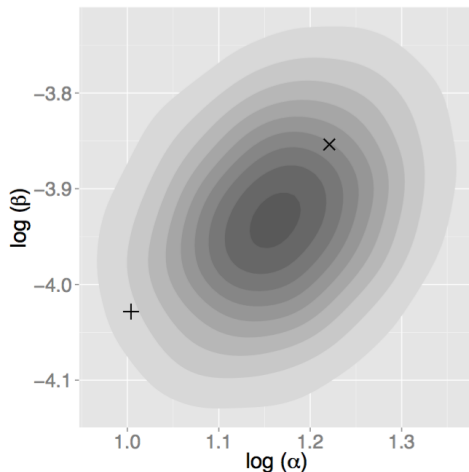


Figure: Posterior (log scale) of recovery rate α and infection rate β . The “+” and “x” symbol are previous point estimates from deterministic and approximate models [Brauer 2008; Raggett 1982].

Relaxing the closed population assumption

- So far, closed population of size N : one can only get infected from someone **inside** the community;
- What if the population under study is a sub-population?
- ξ (“external infection rate”): constant force of infection on every susceptible.
- Likelihood:

$$\begin{aligned} & \mathcal{L}(\beta, \xi, \gamma, \tilde{\alpha}, \tilde{\omega} | \mathcal{G}_0) \\ &= \gamma^{n_R} \alpha_{SS}^{C_{HH}} \alpha_{SI}^{C_{HI}} \alpha_{II}^{C_{II}} \omega_{SS}^{D_{HH}} \omega_{SI}^{D_{HI}} \omega_{II}^{D_{II}} \prod_{j=2}^n \left[\tilde{M}(t_j) (\beta I_{p_{j1}}(t_j) + \xi)^{F_j} \right] \\ & \times \exp \left(- \int_0^{T_{\max}} \left[\beta SI(t) + \xi S(t) + \gamma I(t) + \tilde{\alpha}^T \mathbf{M}_{\max}(t) + (\tilde{\omega} - \tilde{\alpha})^T \mathbf{M}(t) \right] dt \right). \end{aligned} \tag{2}$$

- No closed-form MLEs for β and ξ , but can be numerically solved.

Slide appendix: relaxing closed population

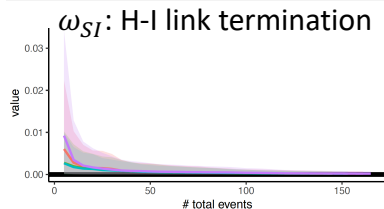
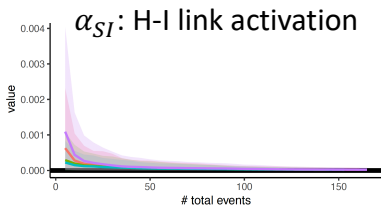
- The $N = 103$ population is in fact a sub-population, so the closed population assumption has to be relaxed;
- ξ : external infection rate.
- Slight modification; for an infection event e_j :
 - if there are possible internal infection sources, regard it as an **internal** case (labelled as $\text{Int}_j = 1, \rightarrow \beta$);
 - if there are no possible internal infection sources, regard it as an **external** case (labelled as $\text{Int}_j = 0, \rightarrow \xi$).
- Complete data likelihood:

$$\begin{aligned}
 & \mathcal{L}(\beta, \xi, \gamma, \tilde{\alpha}, \tilde{\omega} | \mathcal{G}_0) \\
 = & \beta^{(n_E^{\text{int}} - \text{Int}_1)} \xi^{(n_E^{\text{ext}} - 1 + \text{Int}_1)} \gamma^{n_R} \alpha_{SS}^{C_{HH}} \alpha_{SI}^{C_{HI}} \alpha_{II}^{C_{II}} \omega_{SS}^{D_{HH}} \omega_{SI}^{D_{HI}} \omega_{II}^{D_{II}} \\
 & \times \prod_{j=2}^n \left[\tilde{M}(t_j) I_{p_{j1}}(t_j)^{F_j \text{Int}_j} \right] \\
 & \times \exp \left(- \int_0^{T_{\max}} \left[\beta SI(t) + \xi S(t) + \gamma I(t) + \tilde{\alpha}^T \mathbf{M}_{\max}(t) + (\tilde{\omega} - \tilde{\alpha})^T \mathbf{M}(t) \right] dt \right).
 \end{aligned} \tag{3}$$

$$\tag{4}$$

Model flexibility

- The framework generalizes:
 - static network epidemic process ($\alpha_{..} = \omega_{..} \equiv 0$),
 - “decoupled” network epidemic process ($\alpha_{..} \equiv \alpha, \omega_{..} \equiv \omega$).
- It can recognize those cases (e.g., static, link rates = 0).



Efficiency of DARCI

Compared DARCI with two other data augmentation methods:

1. **Rejection sampling (“Reject”)**: For $\ell = 1 : L$, keep proposing recovery times $\{r_{\ell,i}^*\}_{i=1:R_\ell} \stackrel{iid}{\sim} \text{TEXP}(\gamma^{(s-1)}, u_\ell, v_\ell)$ until the proposed recovery times are compatible with the observed events.
2. **Metropolis-Hastings (“MH”)**: For $\ell = 1 : L$, propose recovery times $\{r_{\ell,i}^*\}_{i=1:R_\ell} \stackrel{iid}{\sim} \text{TEXP}(\gamma^{(s-1)}, u_\ell, v_\ell)$, and accept them as $\{r_{\ell,i}^{(s)}\}_{i=1:R_\ell}$ with probability

$$\min \left(1, \frac{p \left(\mathbf{x}, \{r_{\ell,i}^*\}_{i=1:R_\ell}, \{r_{\ell',i}^{(s-1)}\}_{i=1:R_{\ell'}, \ell' \neq \ell} \mid \Theta^{(s-1)} \right) \text{pTEXP} \left(\{r_{\ell,i}^{(s-1)}\}_{i=1:R_\ell}; \gamma^{(s-1)}, u_\ell, v_\ell \right)}{p \left(\mathbf{x}, \{r_{\ell,i}^{(s-1)}\}_{i=1:R_\ell, \ell=1:L} \mid \Theta^{(s-1)} \right) \text{pTEXP} \left(\{r_{\ell,i}^*\}_{i=1:R_\ell}; \gamma^{(s-1)}, u_\ell, v_\ell \right)} \right).$$

Efficiency of DARCI

Take 1000 consecutive samples using each method, and evaluate

- effective sample size (ESS),
- Geweke Z-score (Geweke et al., 1991),
- two-sided p-value for the Z-score.

Statistic	β	γ	α_{II}	ω_{SI}	Method
ESS	1000.00	1000.00	1000.00	1000.00	DARCI
Z-score	-0.90	-0.20	-0.22	-0.02	
$Pr(> Z)$	0.37	0.84	0.82	0.99	
ESS	1000.00	1160.17	1000.00	926.63	Reject
Z-score	0.48	-1.01	1.08	-0.16	
$Pr(> Z)$	0.63	0.31	0.28	0.87	
ESS	566.43	1000.00	538.12	729.33	MH
Z-score	-1.25	-1.83	-2.09	-1.52	
$Pr(> Z)$	0.21	0.07	0.04	0.57	

Efficiency of DARCI

- Also compared **DARCI** and **Reject** in running times.
- On a dataset with 5 intervals containing various numbers of missing recovery times.

Interval	#(To recover)	Min Time		Median Time	
		Reject	DARCI	Reject	DARCI
1	1	227 μ s	224 μ s	484 μ s	245 μ s
2	8	285 μ s	287 μ s	563 μ s	319 μ s
3	15	163 μ s	161 μ s	279 μ s	181 μ s
4	2	138 μ s	138 μ s	153 μ s	156 μ s
5	1	133 μ s	133 μ s	146 μ s	147 μ s