Fitting Stochastic Epidemic Models to Gene Genealogies Using Linear Noise Approximation

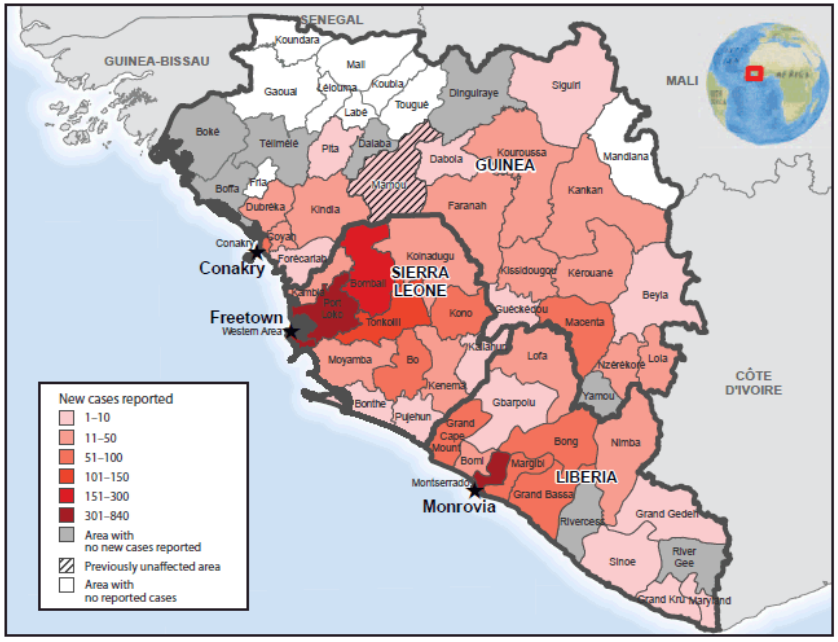
Mingwei Tang

University of Washington
STATISTICS

Joint work with: Gytis Dudas (GGBC),
Trevor Bedford (Fred Hutch),
Vladimir Minin (UCI)

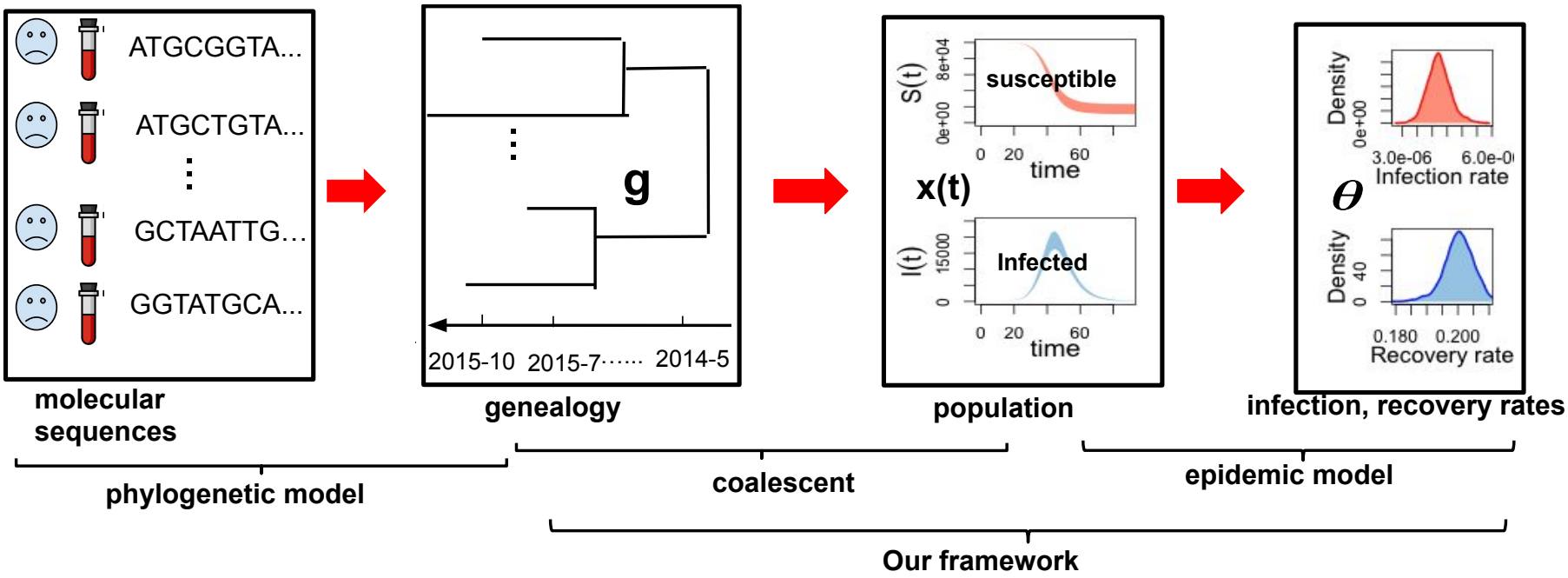
July 29, 2019

Background: Infectious Disease



- Objectives:
  - Estimating disease dynamics (population, rates)
  - Intervention (assessment, decision)
  - Forecast future infections

Background: Molecular Epidemiology/ Phylodynamics



- Phylogenetic model: construct genealogy from sequence data
- Coalescent: estimate population from genealogy
- Goal: Infer **host population trajectory X** + **rate parameters** θ from fixed **genealogy g**

$$\Pr(\mathbf{X}, \theta|\mathbf{g}) \propto \underbrace{\Pr(\mathbf{g}|\mathbf{X}, \theta)}_{\text{Structured coalescent}} \times \underbrace{\Pr(\mathbf{X}|\theta)}_{\text{stochastic epi}} \times \underbrace{\pi(\theta)}_{\text{prior}}.$$

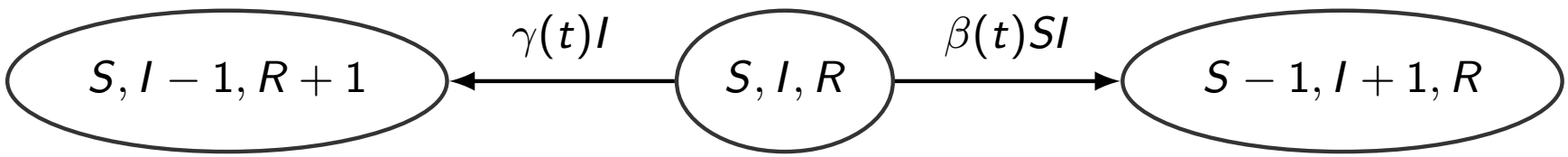
Structured Population: SIR dynamics

- Divide population into Susceptible (S), Infected (I), Recovered (R).
- Interactions between compartments:

$$\text{Susceptible} + \text{Infected} \xrightarrow{\beta(t)} 2\ \text{Infected}, \tag{1}$$

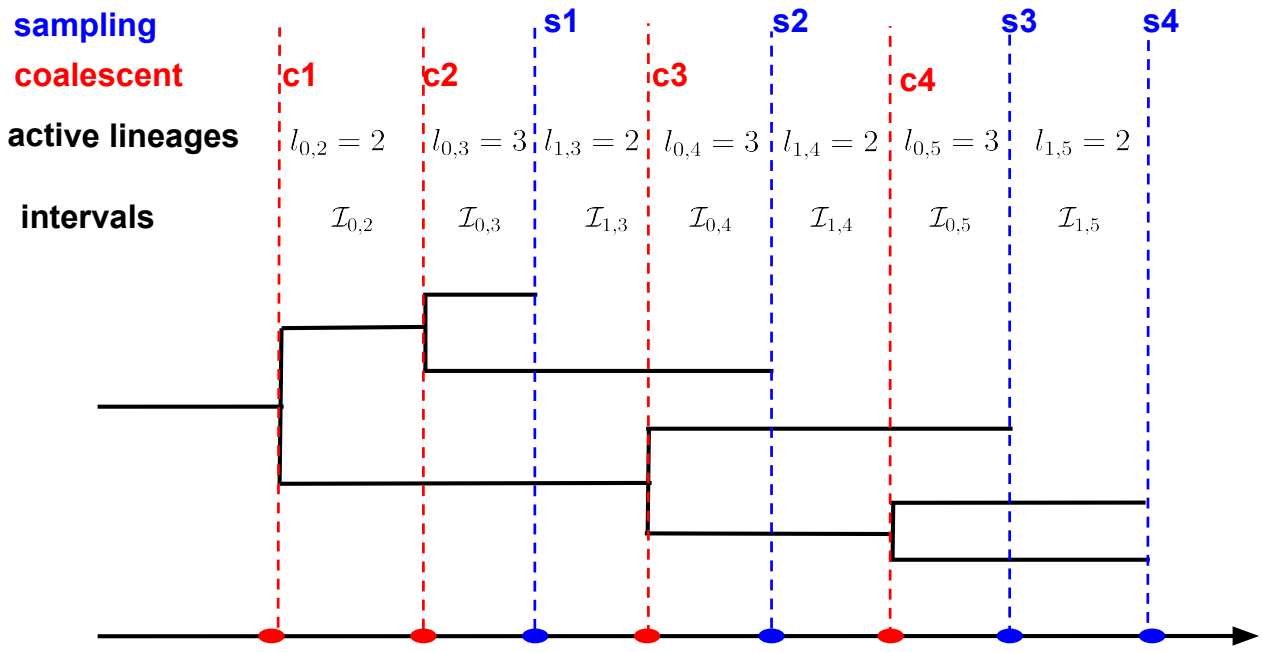
$$\text{Infected} \xrightarrow{\gamma(t)} \text{Recovered}, \tag{2}$$

$\beta(t)$: per capita infection rate, $\gamma(t)$: recovery rate



- Population vector $\mathbf{X}(t) = (S(t), I(t))$, rate vector $\boldsymbol{\theta}(t) = (\beta(t), \gamma(t))$

Coalescent Likelihood Notation



- Sufficient statistics: sampling times $s_1, \ldots, s_m$, coalescent times $c_1, \ldots, c_n$, active lineages $l_{i,j}$ at interval $\mathcal{I}_{i,j}$
- Coalescent as a inhomogenous Markov point process

$$\Pr(\mathbf{g}|\mathbf{X}(\cdot), \boldsymbol{\theta}(\cdot)) \propto \prod_{k=2}^{n} \frac{2\beta(c_{k-1})S(c_{k-1})\binom{l_{0,k}}{2}}{I(c_{k-1})} \exp\left(-\sum_{i=0}^{i_k-1}\binom{l_{i,k}}{2}\int_{\mathcal{I}_{i,k}} \frac{2\beta(\tau)S(\tau)}{I(\tau)}\mathrm{d}\tau\right).$$

Coalescent Likelihood

$$\Pr(\mathbf{g}|\mathbf{X}(\cdot),\boldsymbol{\theta}(\cdot)) \propto \prod_{k=2}^{n} \frac{2\beta(c_{k-1})S(c_{k-1})\binom{l_{0,k}}{2}}{I(c_{k-1})} \exp\left(-\sum_{i=0}^{i_k-1}\binom{l_{i,k}}{2}\int_{\mathcal{I}_{i,k}} \frac{2\beta(\tau)S(\tau)}{I(\tau)}\mathrm{d}\tau\right).$$

- $\int \frac{\beta(\tau)S(\tau)}{I(\tau)}\mathrm{d}\tau$ is non-tractable
- Construct piecewise constant approximation on a regular grid $t_0, \ldots, t_T$

$$\mathbf{X}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1},t_i)}(t)\mathbf{X}_{i-1} \quad \boldsymbol{\theta}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1},t_i)}(t)\boldsymbol{\theta}_{i-1}.$$

- Substitute $\Pr(\mathbf{g}|\mathbf{X}(\cdot),\boldsymbol{\theta}(\cdot))$ with $\Pr(\mathbf{g}|\mathbf{X}_{0:T},\boldsymbol{\theta}_{0:T})$
- Infer $\mathbf{X}_{0:T},\boldsymbol{\theta}_{0:T}$

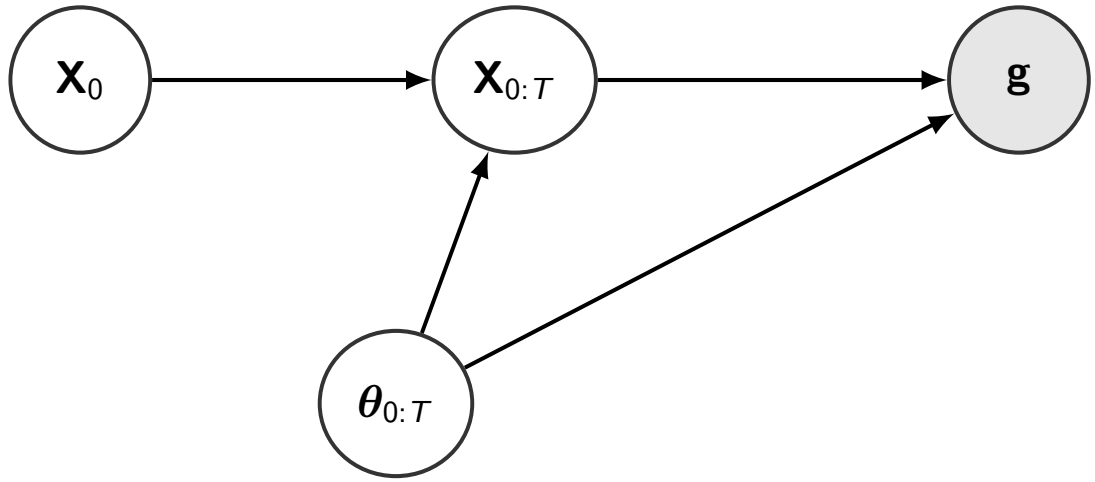
Background: State of the Art Phylodynamic

- Nonparametric curve-fitting method based on Kingman Coalescent (Drummond et al. [1], Minin et al. [5] etc)
  - Hard to interpret
- Deterministic ODE method with structured coalescent model (Volz [9], Volz et al. [10])
  - No stochasticity in the population trajectories, overconfident
- SDE method with structured coalescent model (Rasmussen et al. [7])
  - particle MCMC, heavy computation

Our goal:

- Fit stochastic epidemic model to genealogy
- Computationally efficient algorithm
- Fit realistic models to data (time-varying rates)

Data Generating Procedure



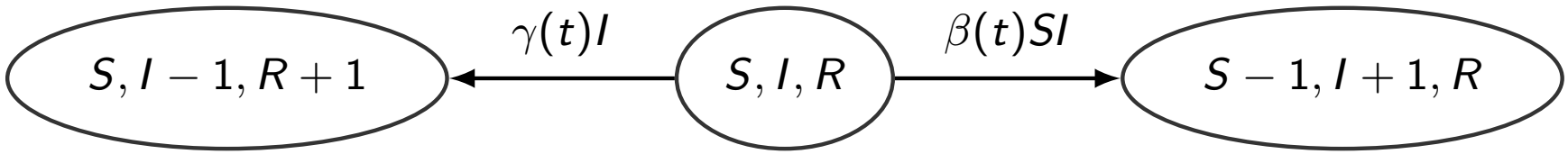
- Bayesian inference

$$\Pr(\mathbf{X}, \boldsymbol{\theta}|\mathbf{g}) \propto \underbrace{\Pr(\mathbf{g}|\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T})}_{\text{coalescent likelihood}} \times \underbrace{\Pr(\mathbf{X}_{1:T}|\mathbf{X}_0, \boldsymbol{\theta}_{0:T})}_{\text{stochastic epi model}} \times \underbrace{\pi(\boldsymbol{\theta}_{0:T})\pi(\mathbf{X}_0)}_{\text{prior}}$$

$$\propto \underbrace{\Pr(\mathbf{g}|\mathbf{X}_{1:T}, \boldsymbol{\theta}_{0:T})|\mathbf{X}_0}_{\text{coalescent likelihood}} \times \prod_{i=1}^{T} \underbrace{\Pr(\mathbf{X}_i|\mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})}_{\text{transition density}} \times \underbrace{\pi(\boldsymbol{\theta}_{0:T})\pi(\mathbf{X}_0)}_{\text{prior}}$$

Stochastic Epidemic Models: SIR Model

- Notation: $\mathbf{X}(t) = (S(t), I(t))^T$, $\boldsymbol{\theta}(t) = (\beta(t), \gamma(t))$



$$\mathbf{A} = \begin{array}{c} \begin{array}{cc} \text{susceptible} & \text{infected} \end{array} \\ \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \end{array} \begin{array}{l} (1) \\ (2) \end{array} \qquad \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}(t)) = (\beta(t)S(t)I(t), \gamma(t)I(t))^T$$

- ODE Formula using matrix algebra

$$\mathrm{d}\mathbf{X} = \mathbf{A}^T \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\mathrm{d}t$$

- MJP: Discrete space, continuous time Markov chain
- SDE formulation:

$$\mathrm{d}\mathbf{X}(t) = \mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}(t))\mathrm{d}t + \sqrt{\mathbf{A}^T \mathbf{H}(\mathbf{X}(t), \boldsymbol{\theta}(t))\mathbf{A}}\mathrm{d}\mathbf{W}_t$$

- Problem: No closed-form transition probability $\Pr(\mathbf{X}_i | \mathbf{X}_{i-1}, \boldsymbol{\theta})$ for large N

Linear Noise Approximation

- Solution: Approximate $\Pr(\mathbf{X}_i|\mathbf{X}_{i-1}, \boldsymbol{\theta})$ with normal distribution
- Linear noise approximation (LNA) (Kurtz [3, 4])

$$\mathbf{X}_i = \boldsymbol{\eta}_i + \mathbf{m}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}_{i-1}, t_i - t_{i-1}, \boldsymbol{\theta}\right) + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}_i)$$

  - $\boldsymbol{\eta}_i$: deterministic ODE solution
  - $\mathbf{m}(\mathbf{X}_{i-1} - \boldsymbol{\eta}_{i-1}, t_i - t_{i-1}, \boldsymbol{\theta})$: residual depend on previous state
  - $\boldsymbol{\epsilon}_i$: stochastic noise
- $\boldsymbol{\eta}_i, \mathbf{m}()$ and $\boldsymbol{\Phi}_i$ calculated by solve ODEs — fast!
- Non-centered parameterization: $\mathbf{X}_{1:T} \Rightarrow \boldsymbol{\xi}_{1:T}$, where $\boldsymbol{\xi}_i \sim_{iid} \mathcal{N}(0, \mathbf{I})$

$$\mathbf{X}_i = \boldsymbol{\eta}(t_i) + \mathbf{m}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), t_i - t_{i-1}, \boldsymbol{\theta}\right) + \boldsymbol{\Phi}_i^{1/2}\boldsymbol{\xi}_i$$

Assumption and Reparameterization

- Initial state: $S_0 \simeq N$, $\mathbf{X}_0 = (N, I_0)$. N — total population size **(known)**
- Time-varying $\beta_{1:T}$ and constant γ
- Parameterize using basic reproduction number

$$R_0(t) := \frac{\beta(t)N}{\gamma}$$

  - $R_0 > 1$: the infection will be able to spread in a population
  - $R_0 < 1$: the infection will be die out
- Assume piecewise constant $R_0(t)$ trajectory

$$R_0(t) = \sum_{i=1}^{T} R_{0_i} \mathbf{1}_{[t_{i-1}, t_i)}(t)$$

- GMRF, prior for log-increments: $\boldsymbol{\delta}_{1:T}$

$$\delta_i := \log\left(\frac{R_{0_i}}{R_{0_{i-1}}}\right) \sim_{iid} \mathcal{N}(0, \sigma^2) \quad i = 1, \ldots, T$$

- lognormal prior on I_0, R_0, γ, σ

MCMC Strategies

$$\Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma | \mathbf{g}) \propto \Pr(\mathbf{g} | I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma) \cdot \Pr(I_0) \Pr(R_0) \Pr(\gamma) \Pr(\boldsymbol{\delta}_{1:T}) \Pr(\boldsymbol{\xi}_{1:T}) \Pr(\sigma)$$

Difficulty/ Bottleneck in sampling latent variable $\boldsymbol{\xi}_{1:T}$, $\boldsymbol{\delta}_{1:T}$

- High dimensional: $\mathcal{O}(T)$. Highly correlated

Elliptical Slice Sampler (Murray et al. [6]):

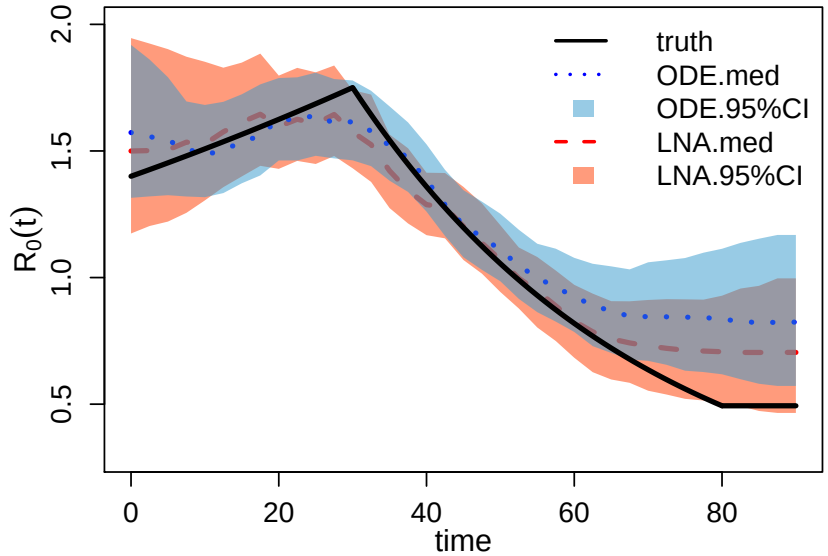
- For state space model with Gaussian distribution latent variables
- Jointly update, tuning free

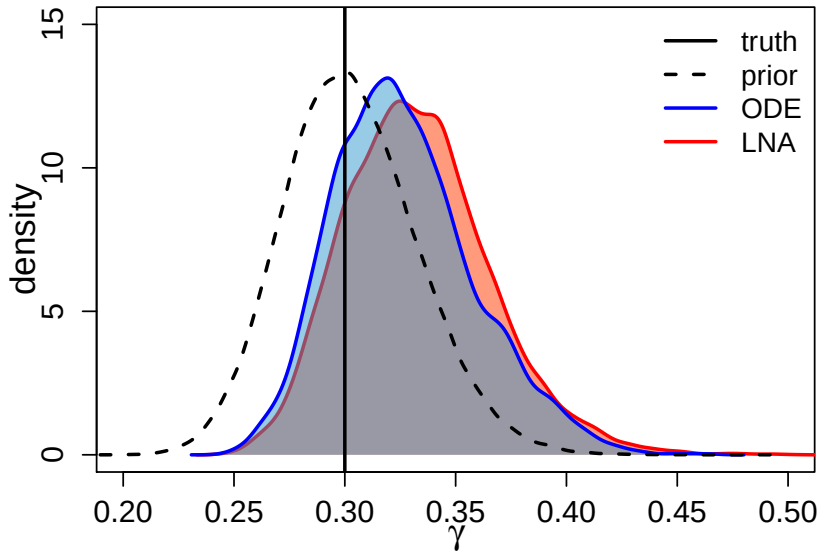
MCMC strategies:

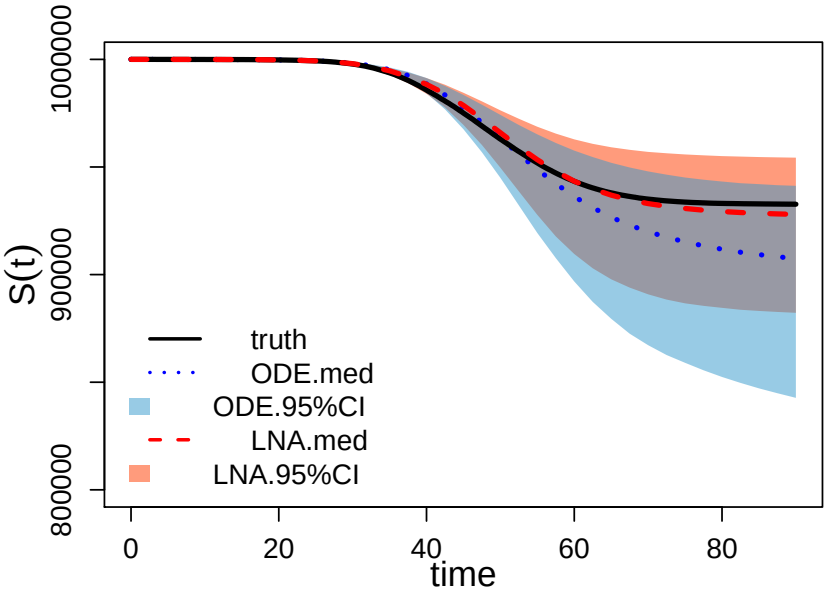
- Use elliptical slice sampler to update $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$
- Use elliptical slice sampler to update $\boldsymbol{\xi}_{1:T}$.
- Update I_0, γ using univariate MH algorithm

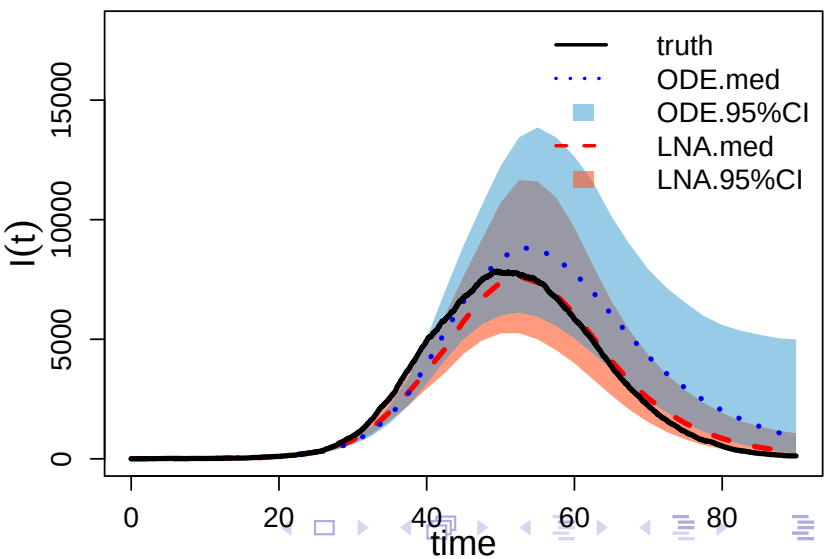
Simulation Study

- Simulate one realization of trajectory, then genealogy
- Fit both LNA-based and ODE-based model, informative prior on γ







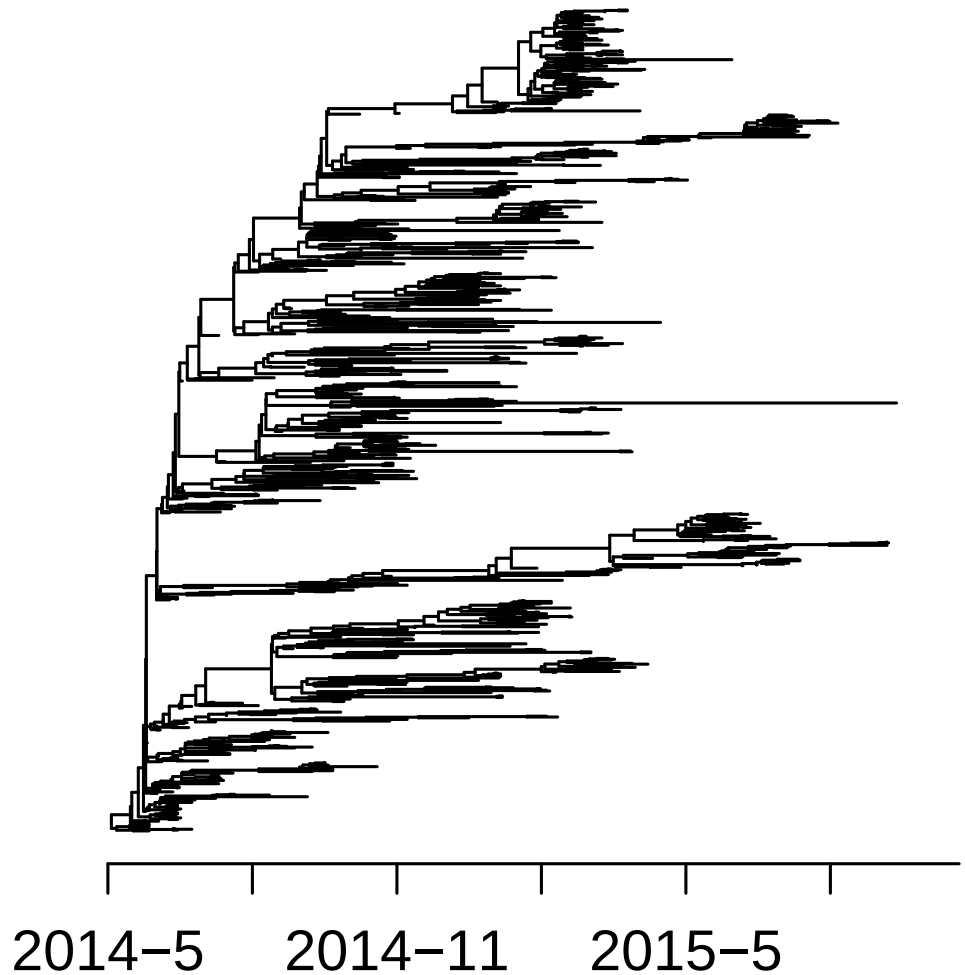


Results of Repeated Simulations

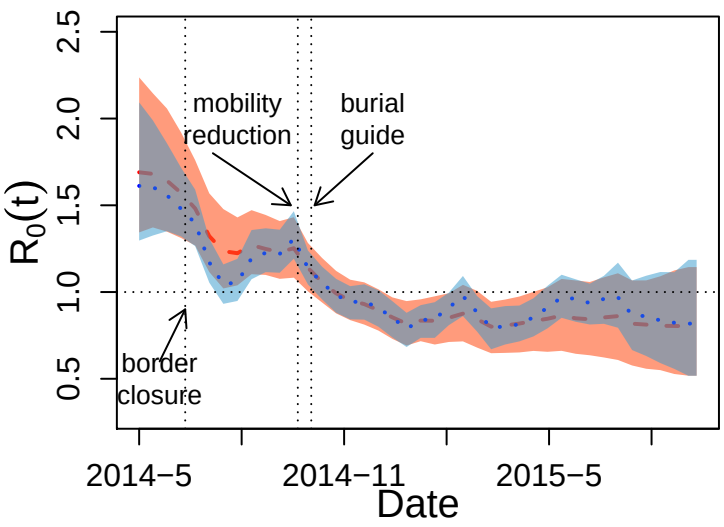
- Repeat simulation for 100 times
- Evaluate estimation of $R_0(t)$, $I(t)$: **bias, precision, BCI coverage**
- Results: LNA-based method (vs ODE-based) has
  - lower bias
  - wider BCI
  - better coverage

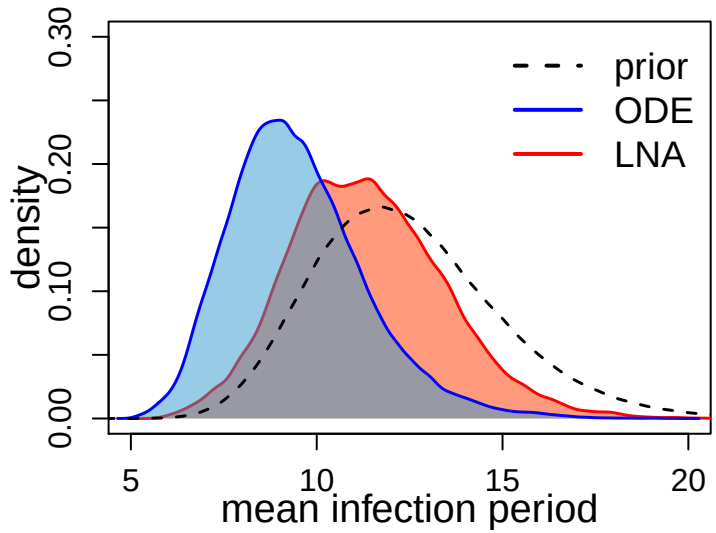
Ebola Genealogies in 2014 West Africa Outbreak

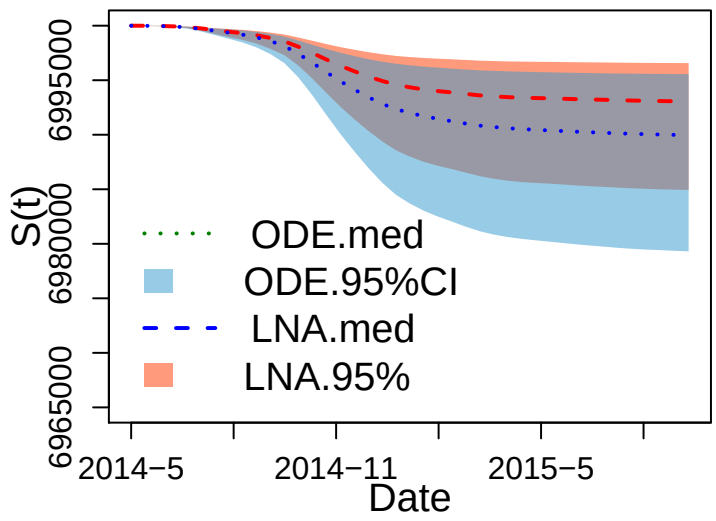
- Ebola genealogies from Dudas et al. [2]
- Sierra Leone, 1010 sequences from 2014-05-25 to 2015-09-12

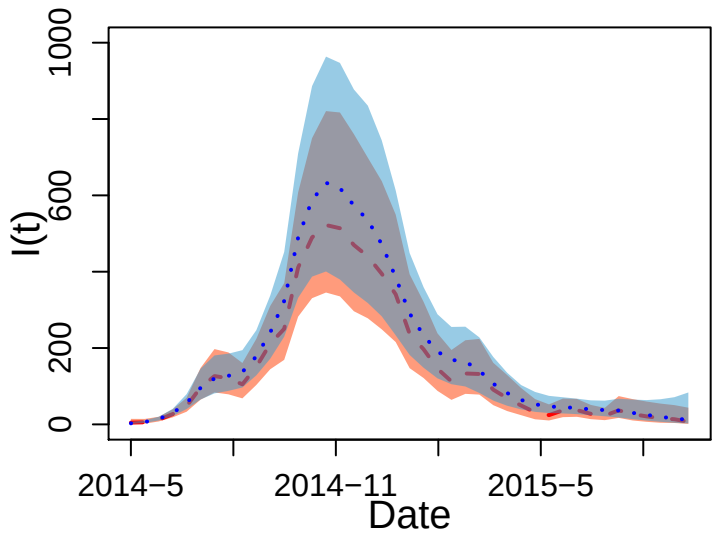


Real Data: Sierra Leone



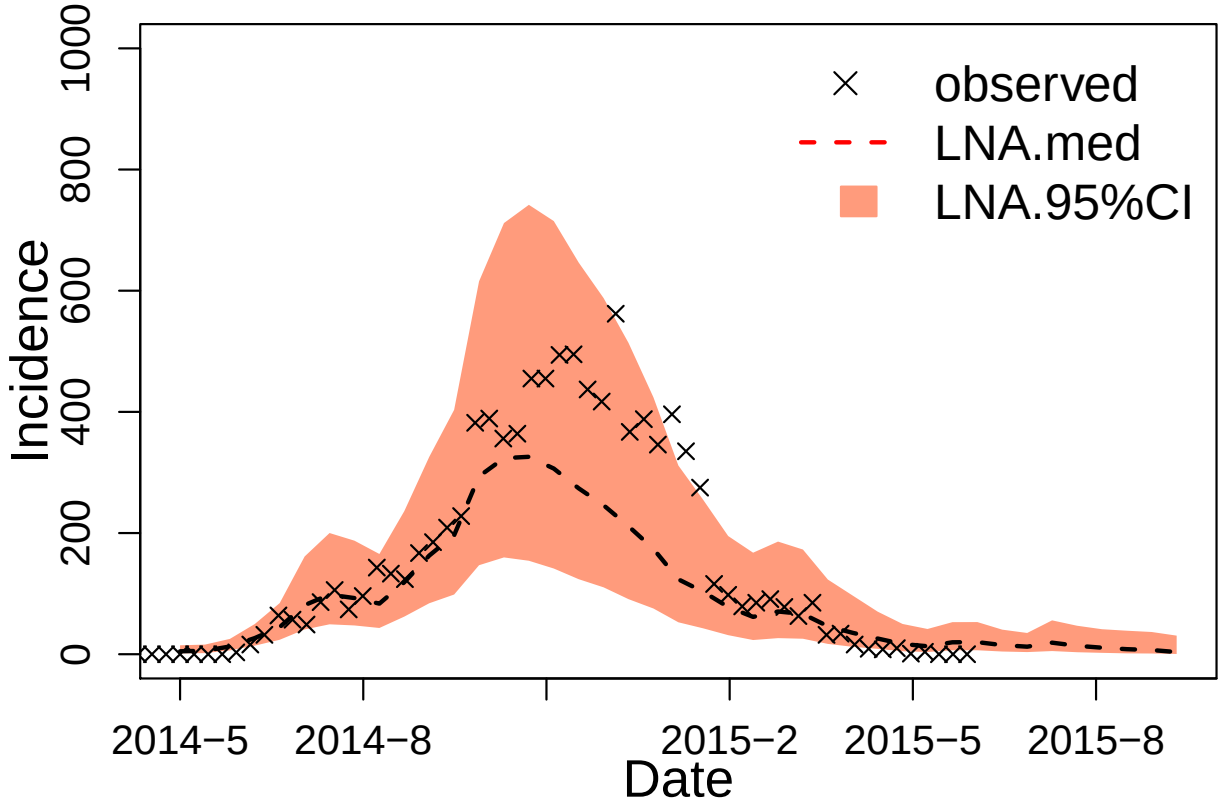






- Final epidemic size: $(3397, 14870)$. CDC: 8706 confirmed

Real Data: Out-of-Sample Validation



- Use WHO incidence data $\times$ as out-of-sample validation
- Posterior estimated incidence

Conclusions and future directions

- Contributions
  - Bayesian semi-parametric framework to estimate epidemic via genealogy
  - Apply LNA to approximate population dynamics
  - Propose efficient sampling algorithm for approximating the posterior
- Issues and concerns
  - Identifiability (prior sensitivity issues) on recovery rate γ
- Future directions
  - Combining genealogy data with other data source, e.g incidence data
  - More complicated stochastic epidemic models, SEIR
  - Sampling genealogies from sequences

Paper and Software

- Tang, M, Dudas, G, Bedford, T and Minin, VN. Fitting Stochastic Epidemic Model to Gene Genealogies using Linear Noise Approximation. arXiv preprint arXiv:1902.08877, 2019. (Tang et al.[8])
- Package: LNAPhyloDyn
  `https://github.com/MingweiWilliamTang/LNAphyloDyn`

References I

[1] AJ Drummond, A Rambaut, B Shapiro, and OG Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005.

[2] G Dudas, LM Carvalho, T Bedford, AJ Tatem, G Baele, NR Faria, DJ Park, JT Ladner, A Arias, D Asogun, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544(7650): 309–315, 2017.

[3] TG Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[4] TG Kurtz. Limit theorems for sequences of jump Markov processes. *Journal of Applied Probability*, 8(2):344–356, 1971.

[5] VN Minin, EW Bloomquist, and MA Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008.

[6] I Murray, RP Adams, and DJC MacKay. Elliptical slice sampling. In *AISTATS*, volume 13, pages 541–548, 2010.

[7] DA Rasmussen, O Ratmann, and K Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, 7(8):e1002136, 2011.

References II

[8] M Tang, G Dudas, T Bedford, and VN Minin. Fitting stochastic epidemic models to gene genealogies using linear noise approximation. *arXiv preprint arXiv:1902.08877*, 2019.

[9] EM Volz. Complex population dynamics and the coalescent under neutrality. *Genetics*, 190(1):187–201, 2012.

[10] EM Volz, SLK Pond, MJ Ward, AJL Brown, and SDW Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4): 1421–1430, 2009.