# Pair-based likelihood approximations for stochastic epidemic models

Stockdale et al. (*Biostatistics, 2019*)

## Seth D Temple

Research Preliminary Exam
Seattle, WA, USA
June 2021

## Problem framing

- Epidemic models afford insight into the incidence, spread, and control of contagions that threaten human welfare
  - ▶ How do mitigation efforts affect disease spread?
  - ▶ How do diseases spread between subpopulations?

- Partially observed process
  - ▶ Removal[1] times $r_1, \ldots, r_n$ ✓
  - ▶ Infection times $i_1, \ldots, i_n$ ✗

- Analyses with partial data
  - ▶ Integrate over $i_1, \ldots, i_n$
  - ▶ Or, augment $i_1, \ldots, i_n$ in Bayesian framework

- Independence between pairs of individuals provides **likelihood approximations** as fast frequentist methods.

---

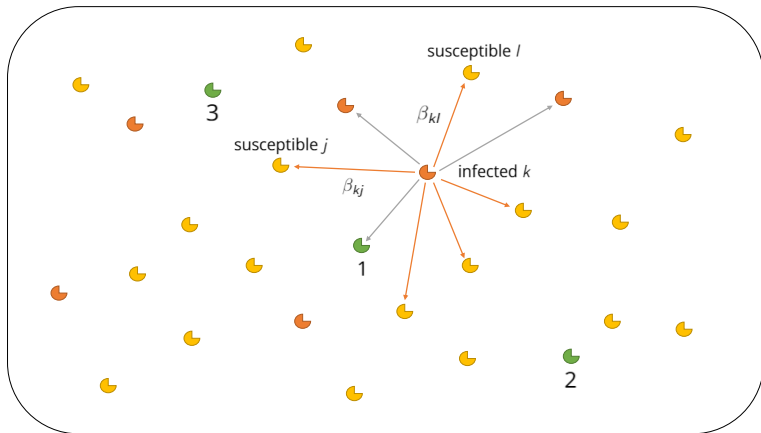[1] Removal means an individual can no longer infect others.

# Stochastic epidemic

Infection rates $\beta_{kj}$ and removals after $r_j - i_j \sim \text{Gamma}(m_j, \gamma_j)$
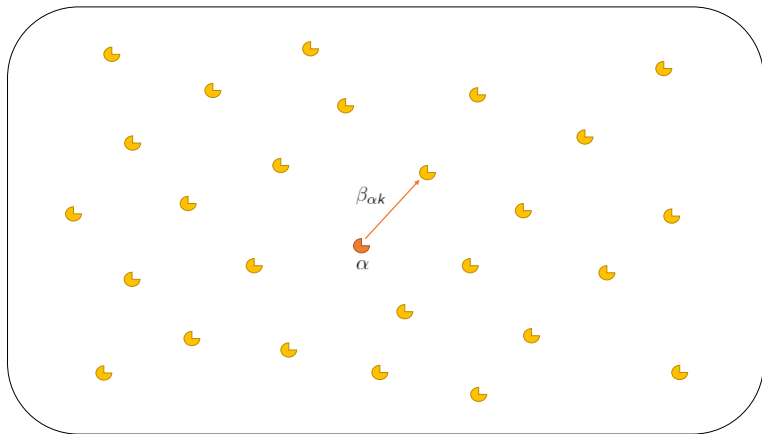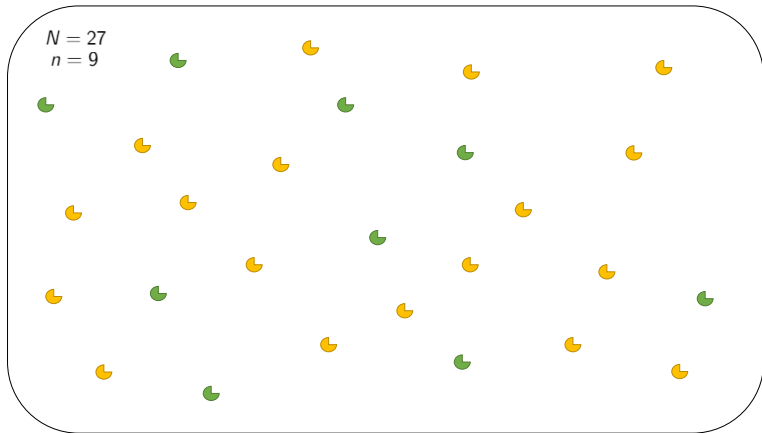
# Stochastic epidemic

At time $t$, $S(t)$ susceptibles, $I(t)$ infecteds, and $R(t)$ removeds, with $N = S(t) + I(t) + R(t)$.

# Stochastic epidemic

Epidemic ends when $I(t) = 0$.



$N = 27$
$n = 9$

# Stochastic epidemic

To simulate an epidemic, we exploit **Poisson processes (PPs)**. Define a *race* as the minimum of (exponential) rvs.

## Algorithm (Epidemic Simulator)

1. $S(0) = N - 1, I(0) = 1$
2. Until $I(t) = 0$:
   2.1 **Race $S(t)I(t)$ PPs with rate $\beta$ and $I(t)$ PPs with rate $\gamma$,** where $t_1$ is the winning race time.
   2.2 If a $\gamma$-PP wins, $I(t_1) = I(t) - 1$ and $R(t_1) = R(t) + 1$.
   2.3 If a $\beta$-PP wins, $S(t_1) = S(t) - 1$ and $I(t_1) = I(t) + 1$.
   2.4 Update $t = t_1$.

# Stochastic epidemic model

$\{S(t), I(t)\}$ is a continuous-time Markov chain (CTMC) with "jumps" based on an underlying Poisson process.

- $\tau_{kj} := r_k \wedge i_j - i_k \wedge i_j$
  - ▶ Time $k$ tries to infect $j$

- $\psi_j = \exp(-\sum_{k \neq j}^{n} \beta_{kj} \tau_{kj})$
  - ▶ Probability $j$ not infected before $i_j$
  - ▶ $\psi_{kj} = \exp(-\beta_{kj} \tau_{kj})$ is marginal term

- $\chi_j = \sum_{k \neq j}^{n} \beta_{kj} 1_{\{i_k < i_j < r_k\}}$
  - ▶ Probability $j$ infected at $i_j$

- $\phi_j = \exp(-\sum_{k=n+1}^{N} \beta_{jk}(r_j - i_j))$
  - ▶ Probability $j$ doesn't infect never-infecteds

# Stochastic epidemic model

With $i_1, \ldots, i_n$ known, the *augmented model likelihood* is

$$\pi(\mathsf{i}_{-1}, \mathsf{r} | \boldsymbol{\beta}, \boldsymbol{\theta}, i_1) = \left\{ \prod_{j=2}^{n} \psi_j \chi_j \phi_j f_j(r_j - i_j | \theta_j) \right\} \phi_1 f_1(r_1 - i_1 | \theta_1) \quad (1)$$

MLE is easy with complete data. For common $(\beta, \gamma)$ and $m = 1$:

$$\hat{\beta} = \frac{N(n-1)}{\sum_{j=2}^{n} \sum_{k \neq j} \tau_{kj} + (N - n) \sum_{j=1}^{n} r_j - i_j}$$

$$\hat{\gamma} = \frac{1}{n} \sum_{j=1}^{n} r_j - i_j$$

---

[0] We can generalize to unknown patient zero, i.e. $\alpha$ unknown.

## DAMCMC for SEM

Construct a Metropolis-within-Gibbs sampling routine.
Assume common $\beta$ and $\theta = (m, \gamma)$ for infectious periods.

$$\beta \mid \gamma, \alpha, i_\alpha, i_{-\alpha}, r \sim \text{Gamma}(m_\beta + n - 1, \nu_\beta + A_i)$$

$$\gamma \mid \beta, \alpha, i_\alpha, i_{-\alpha}, r \sim \text{Gamma}(m_\gamma + n, \nu_\gamma + C_i)$$

$$A_i = \sum_{j=1}^{n} \sum_{k=1}^{N} \tau_{jk}$$

$$C_i = \sum_{j=1}^{n} r_j - i_j$$

$$i_1, \ldots, i_n \sim f(\cdot) \text{ (Metropolis-Hastings)}$$

Scheme suffers from **high posterior correlations**. Either fix this issue (Kypraios, 2007; Neal and Roberts, 2005), or evade it.

## Pair-based likelihood approximations

First, Stockdale et al. (2019) derive *partial data likelihood*:

1. Integrate over $i_1, \ldots, i_n$
2. Change of variable $a(\theta_j, -B_j)g_j = \phi_j f_j$

   ▶ Absorbs info on never-infecteds into density

   ▶ Permanence of $g_j \sim \text{Gamma}(m_j, \delta_j)$

   ▶ New rate $\delta_j = \gamma_j + B_j = \gamma_j + \sum_{k=n+1}^{N} \beta_{jk}$

   ▶ $a(\theta_j, \cdot)$ is mgf of $r_j - i_j$

$$\pi(\mathbf{r}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int \pi(\mathbf{i}_{-1}, \mathbf{r}|\boldsymbol{\beta}, \boldsymbol{\theta}, i_1)\pi(i_1) \, d(i_1, \ldots, i_n)$$

$$= \left\{ \prod_{j=1}^{n} a(\theta_j, -B_j) \right\} \mathbb{E}_{\mathbf{g}}[\pi(i_1)] \underbrace{\mathbb{E}_{\mathbf{g}}\left[ \left\{ \prod_{j=2}^{n} \psi_j \chi_j \right\} \right]}_{\text{approximate}} \quad (2)$$

## Pair-based likelihood approximations

Second, they approximate the expected product.

$$\mathbb{E}_{\mathbf{g}}\left[\left\{\prod_{j=2}^{n}\psi_j\chi_j\right\}\right] \approx \prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\psi_j]\cdot\mathbb{E}_{\mathbf{g}}[\chi_j] \qquad (3)$$

$$\mathbb{E}_{\mathbf{g}}\left[\left\{\prod_{j=2}^{n}\psi_j\chi_j\right\}\right] \approx \left\{\prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\chi_j]\right\}\left\{\mathbb{E}_{\mathbf{g}}\left[\prod_{j=2}^{n}\psi_j\right]\right\}$$

$$= \left\{\prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\chi_j]\right\}\mathbb{E}_{\mathbf{g}}\left[\exp\left(-\sum_{j=2}^{n}\sum_{k\neq j}^{n}\beta_{kj}\tau_{kj}\right)\right]$$

$$= \left\{\prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\chi_j]\right\}\underbrace{\mathbb{E}_{\mathbf{g}}\left[\exp\left(-\frac{\beta}{N}\sum_{j=2}^{n}\sum_{k\neq j}^{n}\tau_{kj}\right)\right]}_{\text{mgf of }W\text{ at }-\beta/N}$$

$$(4)$$

___

[0] $W$ is cumulative time that infecteds try to infect susceptibles.

## Pair-based likelihood approximations

The *standard pair-based likelihood approximation (PBLA)* assumes **marginal pairwise independence**.

$$\mathbb{E}_{\mathbf{g}}[\psi_j]\mathbb{E}_{\mathbf{g}}[\chi_j] = \mathbb{E}_{\mathbf{g}}\bigg[\prod_{l\neq j}^{n}\psi_{lj}\bigg]\mathbb{E}_{\mathbf{g}}\bigg[\sum_{k\neq j}^{n}\beta_{kj}1_{\{i_k<i_j<r_k\}}\bigg]$$

$$\approx \bigg\{\prod_{l\neq j}^{n}\mathbb{E}_{g_l,g_j}[\psi_{lj}]\bigg\}\bigg\{\sum_{k\neq j}^{n}\beta_{kj}\mathbb{E}_{g_k,g_j}\bigg[1_{\{i_k<i_j<r_k\}}\frac{\psi_{kj}}{\psi_{kj}}\bigg]\bigg\}$$

$$\approx \bigg\{\prod_{l\neq j}^{n}\mathbb{E}[\psi_{lj}]\bigg\}\sum_{k\neq j}^{n}\beta_{kj}\mathbb{E}[\psi_{kj}1_{\{i_k<i_j<r_k\}}](\mathbb{E}[\psi_{kj}])^{-1}$$

$$(5)$$

* Two similarly derived PBLAs are proposed.

## Pair-based expectations

### Lemma (1)

Let $1 \leq j, k \leq n$ with $j \neq k$, and $\beta_{kj} > 0$. For each $j$, suppose $r_j - i_j \sim$ Exponential($\delta_j$). Then,

$$\mathbb{E}[\psi_{kj}] = \mathbb{E}[\exp(-\beta_{kj}\tau_{kj})] \tag{6}$$

$$= \begin{cases} 1 - \frac{\beta_{kj}\delta_j}{(\delta_j+\delta_k)(\beta_{jk}+\delta_k)}\exp(-\delta_k(r_k-r_j)), & r_j < r_k \\ \frac{\delta_k}{\beta_{kj}+\delta_k} + \frac{\beta_{kj}\delta_k}{(\delta_j+\delta_k)(\beta_{kj}+\delta_k)}\exp(-\delta_j(r_j-r_k)), & r_j > r_k \end{cases}$$

$$\mathbb{E}[1_{\{i_k<i_j<r_k\}}\exp(-\beta_{kj}\tau_{kj})] \tag{7}$$

$$= \begin{cases} \frac{\delta_j\delta_k}{(\delta_j+\delta_k)(\beta_{kj}+\delta_k)}\exp(-\delta_k(r_k-r_j)), & r_j < r_k \\ \frac{\delta_j\delta_k}{(\delta_j+\delta_k)(\beta_{kj}+\delta_k)}\exp(-\delta_j(r_j-r_k)), & r_j > r_k \end{cases}$$

## Pair-based expectations

*Proof.* Based on cases of $\tau_{kj}$, partition $\mathbb{E} := \mathbb{E}_{g_k, g_j}$.

$$\tau_{kj} := r_k \wedge i_j - i_k \wedge i_j = \begin{cases} 0, & i_j < i_k \\ i_j - i_k, & i_k < i_j < r_k \\ r_k - i_k, & i_j > r_k \end{cases}$$

$$\begin{aligned}
&\mathbb{E}[\exp(-\beta_{kj}\tau_{kj})] \\
&= \mathbb{E}[e^{-\beta_{kj}\tau_{kj}}1_{\{i_j < i_k\}}] + \mathbb{E}[e^{-\beta_{kj}\tau_{kj}}1_{\{i_j > r_k\}}] + \mathbb{E}[e^{-\beta_{kj}\tau_{kj}}1_{\{i_k < i_j < r_k\}}] \\
&= \mathbb{E}[1_{\{i_j < i_k\}}] + \mathbb{E}[e^{-\beta_{kj}(r_k - i_k))}1_{\{i_j > r_k\}}] + \mathbb{E}[e^{-\beta_{kj}(i_j - i_k))}1_{\{i_k < i_j < r_k\}}]
\end{aligned}$$

$$(8)$$

Evaluate terms in (8) separately. Direct integration is possible, but an argument using Poisson processes is more illuminating.
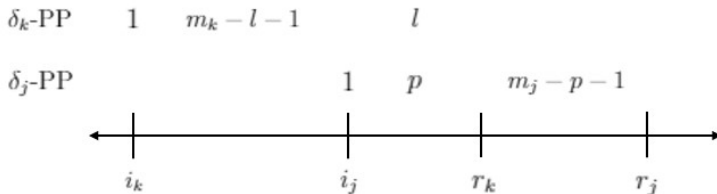
# Pair-based expectations

1. Assume $r_k < r_j$
2. Traverse process backwards from $r_j$
3. $\delta_j$-PP between $(r_k, r_j)$
   3.1 $\exp(-\delta_j(r_j - r_k))$ is probability of no renewal
   3.2 If renewed, renewal time is $i_j$
4. At $r_k$, $\delta_k$-PP begins
   4.1 Compound $(\delta_j + \delta_k)$-PP if $i_j < r_k$.
   4.2 For $(\delta_j + \delta_k)$-PP, we have an **exponential race** with probabilities for renewals $\frac{\delta_j}{\delta_j + \delta_k}$, $\frac{\delta_k}{\delta_j + \delta_k}$

Races may also be set up with a $\beta_{kj}$-PP. With such probabilistic arguments, we derive expressions for terms in (8).

## Pair-based expectations

Likewise, we can derive formulas for Erlang infectious periods. Draw line graphs to support combinatorial extension.

$\delta_k$-PP $\quad$ 1 $\quad$ $m_k - l - 1$ $\qquad\qquad$ $l$

$\delta_j$-PP $\qquad\qquad\qquad$ 1 $\quad$ $p$ $\qquad$ $m_j - p - 1$

$i_k \qquad\qquad i_j \qquad r_k \qquad\qquad r_j$

---

[0] Schematics based on Stockdale (2019)

# Product expectation

## Lemma (2)

Let $\mathcal{K}$ be any subset of $\{1, \ldots, n\}$ with $K = |\mathcal{K}| \geq 2$. Suppose $\{r_k - i_k : k \in \mathcal{K}\} \overset{\text{iid}}{\sim} \text{Exponential}(\delta)$. Then

$$V = \sum_{\substack{j,k \,\in\, \mathcal{K} \\ j < k}} (\tau_{jk} + \tau_{kj}) = \sum_{\substack{j,k \,\in\, \mathcal{K} \\ j < k}} \omega_{jk} \sim \sum_{j=1}^{K-1} j \cdot Y_j$$

where $Y_1, \ldots, Y_{K-1} \sim \text{Exponential}(\delta)$.

*Proof.* Again, traverse process in reverse and make a convenient change of variable.

---

[0] If $\mathcal{K} = \{1, \ldots, n\}$, we have $W$.
[0] Recall $W$ is cumulative time that infecteds try to infect susceptibles.
[0] We require its moment-generating function at $-\beta/N$.

# Methods recap

Stockdale et al. (2019) propose 2 (6) PBLAs.

- MLE now possible despite partial observance
- Utilize properties of Poisson processes
- Expected product as product of expecteds
- Considering all pairs is $O(n^2)$

## Simulation studies

I conducted additional[2] simulation studies[3] to ask:

- How fast are PBLAs?

- Behavior of PBLA-based MLEs
  - ▶ When is pairwise independence inappropriate?
  - ▶ Does PBLA inference offer consistent estimators?

- How does underreporting impact inference?
  - ▶ Undercounts result in lower $R_0$
  - ▶ Ad hoc adjustments assuming MCAR

---

[2] Stockdale et al. (2019) suggest that PBLAs can learn $(\beta, \gamma)$ (Appendix B).

[3] To simulate epidemics, I used algorithm on slide 6 with inputs $\beta, \gamma, N$.

# Runtime comparisons

Table: Time in seconds to compute likelihood for standard, product, and weak PBLAs, and Eichner-Dietz approximation (2003).

| n | N | Std | Prod | Weak | E+D |
|---|---|---|---|---|---|
| 95 | 200 | 0.01 | 0.00* | 0.00* | 0.19 |
| 185 | 500 | 0.02 | 0.01 | 0.01 | 0.49 |
| 428 | 1,000 | 0.13 | 0.03 | 0.01 | 2.03 |
| 1,483 | 2,500 | 1.58 | 0.33 | 0.29 | 20.83 |
| 2,830 | 5,000 | 5.66 | 1.12 | 1.12 | 82.56 |
| 5,927 | 10,000 | 25.61 | 4.67 | 4.67 | |
| 11,819 | 20,000 | 106.81 | 19.81 | 21.24 | |
| 29,024 | 50,000 | 633.27 | 126.47 | 119.12 | |

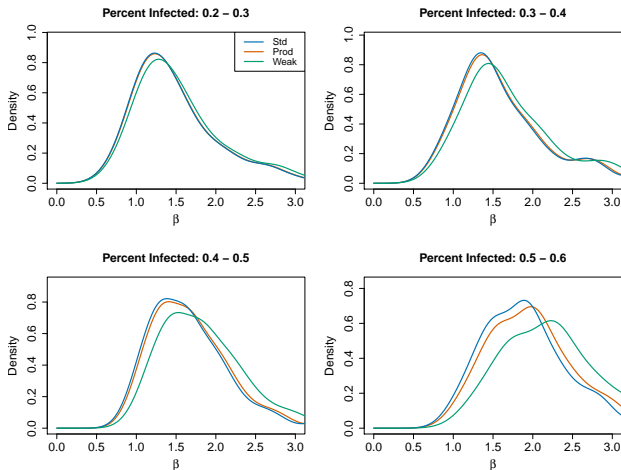* denotes very small, nonzero times.

# Infected proportion



Figure: Inferences on $(\beta, \gamma)$ for increasing infected proportion $n/N$. 2000 simulated epidemics with $(\beta, \gamma) = (1.5, 1)$. Plots for $\gamma$ similar.
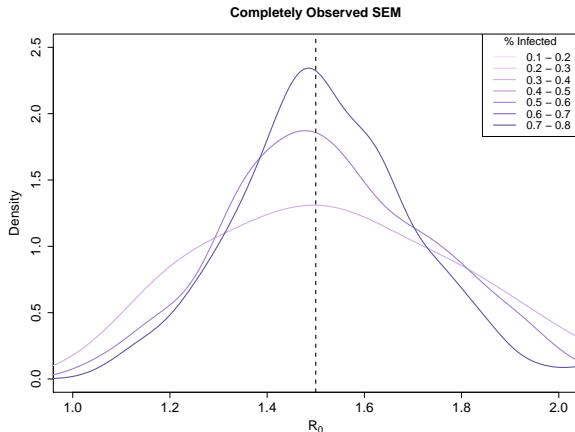
# Properties of PBLA MLEs



Figure: Inference on $R_0 := \beta/\gamma$ for increasing infected proportion. 2000 simulated epidemics with $(\beta, \gamma) = (1.5, 1)$.
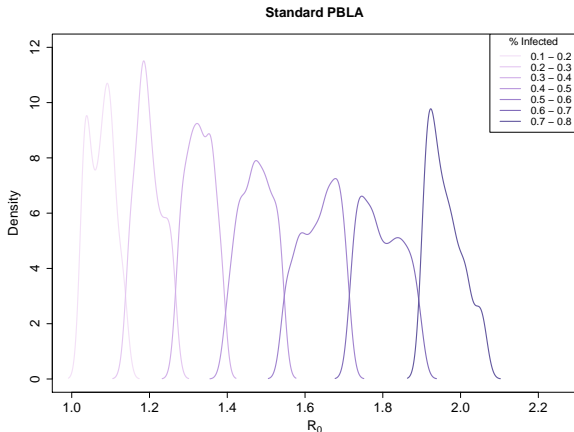
# Properties of PBLA MLEs



**Standard PBLA**

Figure: Inference on $R_0 := \beta/\gamma$ for increasing infected proportion.
2000 simulated epidemics with $(\beta, \gamma) = (1.5, 1)$.

# Properties of PBLA MLEs
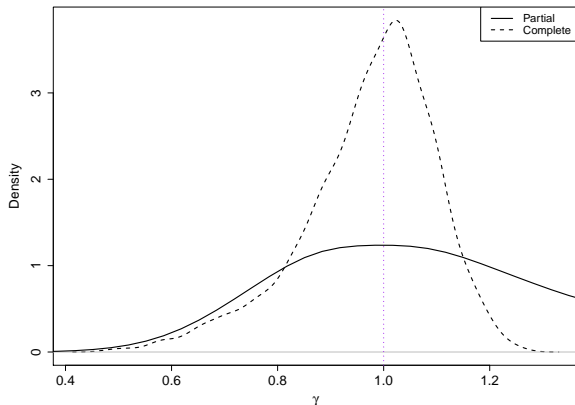


Figure: Inference on $\gamma$ based on complete versus partial data.
2000 simulated epidemics with $(\beta, \gamma) = (1.5, 1)$.

## Properties of PBLA MLEs

- Infected proportion $n/N$ calibrates $\hat{R}_{0,\text{PBLA}}$
  - Conditional on $R_0$, $n/N$ is approximately normal (Andersson and Britton, 2012, Theorems 4.1-2)

- $(\hat{\beta}_{\text{PBLA}}, \hat{\gamma}_{\text{PBLA}})$ do not estimate true $(\beta, \gamma)$
  - In fact, even with fixed true $\beta$, $\gamma$, PBLA inference cannot consistently estimate the other

- **Partial data appears inadequate** for inferring adversarial dynamics of infection and removal processes

# Real data analyses

- **Ebola virus in West Africa**
  - ▶ $n \in (2000, 5000)$
  - ▶ SEIR with fixed exposed period $c$
  - ▶ Time-varying infections

- Dog rabies in Central African Republic
  - ▶ Cases underreported

- Common cold on a remote island
  - ▶ $N = 254$, split into age groups
  - ▶ Accommodates multitype infections

- Foot-and-mouth disease in UK
  - ▶ Rich covariate set
  - ▶ $\beta$ depend on distance

# Ebola virus in West Africa

- Replace $r_j$ with $r_j - c$ for fixed exposed period $c = 5.3$ days

- Fixed $\gamma^{-1} = 5.61$ days

- $\beta_{kj} = \beta_0 \exp(-k_0(T_{kj}))$ where $T_{kj}$ is expected midpoint

- Compare Poisson model with deterministic SEIR fit (Althaus, 2014)
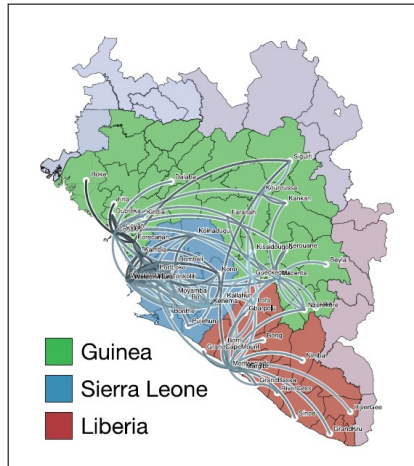


Figure: Suchard et al. (2018)

# Ebola virus in West Africa: Estimates

Table: Parameter estimates from SEIR model of Ebola virus in West Africa. Temple and Stockdale et al. (2019) use PBLA whereas Althaus (2014) uses a deterministic model.

| Country | Method | $\beta_0$ | $k_0$ | $R_0$ |
|---|---|---|---|---|
| Guinea | PBLA | 0.243 | 0.00105 | 1.36 |
| | ODE | 0.231 | 0.00071 | 1.30 |
| Sierra Leone | PBLA | 0.335 | 0.00289 | 1.88 |
| | ODE | 0.277 | 0.00180 | 1.55 |
| Liberia | PBLA | 0.266 | 0.00180 | 1.49 |
| | ODE | 0.303 | 0.00251 | 1.70 |

# Stockdale et al. (2019) contributions

1. Propose likelihood approximations
   - ▶ MLE for partially observed epidemic
   - ▶ Faster than existing methods

2. Promote flexible framework for epidemiology
   - ▶ Specify formulas for $\beta_{kj}$ and $\gamma_j$
   - ▶ Copious simulated and real examples

3. Address partial observance without data augmentation
   - ▶ Motivated by transmission dynamics

# My contributions

1. R package sdtemple/pblas
   - ▶ Highly scriptable; documented; no dependencies
   - ▶ Reproduces **all** simulation studies and data analyses

2. Additional simulation studies
   - ▶ PBLA MLEs for $(\beta, \gamma)$ are **not consistent**
   - ▶ Pairwise independence fails for $n/N > 0.5$
   - ▶ Weak PBLA 👎
   - ▶ Underreporting biases estimation

3. Some corrections
   - ▶ Methods scale with $n$, not $N$
   - ▶ Ebola analysis takes longer than reported
   - ▶ Minor typos: $\pm$, constants

# Future work

- **Consistency**
  - ▶ Are methods consistent with varying $\beta_{kj}$, $\gamma_j$ ?
  - ▶ Can a partially observed SEM achieve consistency?
  - ▶ If so, develop consistent estimators.

- Compare to count-based models
  - ▶ Usually have aggregate counts
  - ▶ (Irons and Raftery, 2021; Fintzi et al., 2021)

- Relax various model assumptions
  - ▶ Set some $\beta_{kj} = 0$ (faster computations)
  - ▶ Adjust for underreporting
  - ▶ Study epidemics in progress (online inference)
  - ▶ Models with demography

# References I

📑 J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

📑 D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

📑 A. D. Barbour and G. Eagleson. Multiple comparisons and sums of dissociated random variables. *Advances in applied probability*:147–162, 1985.

📑 R. Durrett. *Essentials of stochastic processes*. Volume 1. Springer, 1999.

# References II

M. Eichner and K. Dietz. Transmission potential of smallpox: estimates based on detailed data from an outbreak. *American Journal of epidemiology*, 158(2):110–117, 2003.

Y. Hayakawa, P. D. O'Neill, D. Upton, and P. S. Yip. Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Australian & New Zealand Journal of Statistics*, 45(4):491–502, 2003.

P. J. Neal and G. O. Roberts. Statistical inference and model selection for the 1861 hagelloch measles epidemic. *Biostatistics*, 5(2):249–261, 2004.

# References III

📄 P. Neal and G. Roberts. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15(4):315–327, 2005.

📄 T. Kypraios. *Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models*. PhD thesis, Lancaster University, 2007.

📄 L. J. Allen. An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer, 2008.

📄 C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*:5–42, 2011.

📄 H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Volume 151. Springer Science & Business Media, 2012.

📄 C. L. Althaus. Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLoS currents*, 6, 2014.

📄 M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(1), 2018.

📄 J. E. Stockdale. *Bayesian computational methods for stochastic epidemics*. PhD thesis, University of Nottingham, 2019.

📄 J. E. Stockdale, T. Kypraios, and P. D. O'Neill. Pair-based likelihood approximations for stochastic epidemic models. *Biostatistics*, 2019.

📄 J. Fintzi, J. Wakefield, and V. N. Minin. A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. 2021. arXiv: 2001.05099.

📄 N. J. Irons and A. E. Raftery. Estimating sars-cov-2 infections from deaths, confirmed cases, tests, and random surveys. *arXiv: 2102.10741*, 2021.

# Color scheme

- Orange: times
- Violet: before infection probabilities
- Blue: infection rates
- Purple: removal rates

# Glossary

- $N$ total individuals and $n$ infected individuals
- $r_j$ and $i_j$ are removal and infection times for $j$
- $\beta_{kj}$ is infection rate $k$ applies to $j$
- $\theta_j$ parameterizes infectious period $r_j - i_j \sim P_{\theta_j}$
  - $\theta_j = (m_j, \gamma_j)$ for Erlang periods
- $\tau_{kj}$ is time $k$ applies pressure to $j$
  - $\omega_{jk} = \tau_{jk} + \tau_{kj}$ is joint time
- $\psi_j$ is $P(j$ evades infection until time $i_j)$
  - $\psi_{jk}$ is $P(j$ evades infection from $k$ until time $i_j)$
- $\chi_j$ is infective pressure on $j$ at $i_j$
- $\phi_j$ is $P(j$ fails to infect the $N - n$ never-infecteds$)$

# Paper corrections

- Major
  - Methods scale with $n$, not $N$
  - Ebola virus epidemic analyses take 12, 31, and 51 minutes
    - Standard laptop with Intel i7 core

- Minor
  - $-\frac{3}{4}$ instead of $+\frac{3}{4}$ in $\mathbb{E}[T_{kj}]$ (Ebola virus epidemic)
    - Correct in code at jessicastockdale/PBLA
    - No impact on results
  - $(4n - 5)$ instead of $(2n - 1)$ (Lemma (3))
    - No impact on results
  - Subscript $_j$ instead of $_k$ for an Erlang case (Lemma (4))

# Proof of Lemma (2)

1. Define infection and removal transitions in reverse time
   - $(S(t), I(t)) \to (S(t) + 1, I(t) - 1)$ with rate $\delta \cdot I(t)$
   - $(S(t), I(t)) \to (S(t), I(t) + 1)$
2. Express $\int S(t)I(t) \, dt$ as a piecewise linear function
   - $T(i_1) = \sum_{k=2}^{2K} S(\tilde{t}_k) \cdot I(\tilde{t}_k) \cdot (\tilde{t}_k - \tilde{t}_{k-1})$
3. In each interval, make change of variable $t' = t \cdot I(t)^{-1}$
4. Consider weighted sum of renewal times
   - $T(i_1) = \sum_{k=2}^{K} S(\bar{t}_k)(\bar{t}_k - \bar{t}_{k-1})$
   - $Y_k := \bar{t}_k - \bar{t}_{k-1} \sim \text{Exponential}(\delta)$

# A weak limit for multiple comparisons

## Lemma (3)

Suppose $r_1 - i_1, \ldots, r_n - i_n \overset{\text{iid}}{\sim}$ Exponential($\delta$). If $\{\omega_{jk} : j, k \in \{1, \ldots, n\}\}$ are **dissociated**, then

$$W = \sum_{j=1}^{n} \sum_{k=j+1}^{n} \omega_{jk} \rightsquigarrow \text{Normal}\left(\frac{1}{\delta}\binom{n}{2}, \frac{4n-5}{3\delta^2}\binom{n}{2}\right)$$

- Central limit theorem for class of $U$-statistics
- Dissociation is independence assumption
- Appeal to Barbour and Eagleson (1985, Theorem 2.1)

---

[0] Recall $W$ is cumulative time that infecteds try to infect susceptibles.
[0] We require its moment-generating function at $-\beta/N$.

## Theorem 2.1 of Barbour and Eagleson (1985)

### Theorem

Let $D_n = \{(i,j) : 1 \leq i < j \leq n\}$, and consider $\{X_{ij} : (i,j) \in D_n\}$ to be a collection of mean-zero *dissociated* random variables such that $\mathbb{E}[|X_{ij}|^3] < \infty$ for all $(i,j) \in D_n$. Define $\sigma_n^2 := \sum_{(i,j),(k,l) \in D_n} \mathbb{E}[X_{ij}X_{kl}]$. Then

$$\sigma_n^{-3} \sum_{(i,j) \in D_n} (\mathbb{E}[|X_{ij}|^3])^{1/3} \bigg( \sum_{(k,l):|(i,j) \cap (k,l)|=0} (\mathbb{E}[|X_{kl}|^3])^{1/3} \bigg)^2 \to 0$$

implies $\sigma_n^{-1} \sum_{(i,j) \in D_n} X_{ij} = Z_n \rightsquigarrow N(0,1)$.

Pairwise comparisons are dependent if they share any indices.

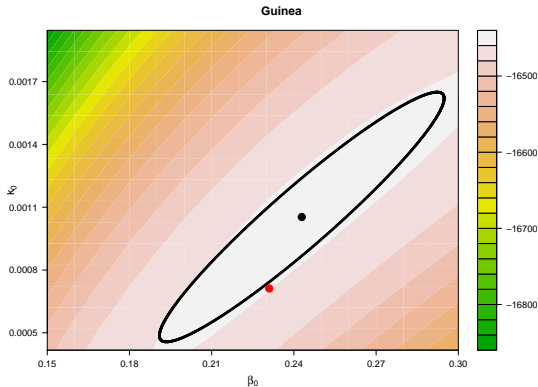# Ebola virus in West Africa: Likelihood Surface



Figure: Log likelihood contours for Ebola virus epidemic in Guinea.
Ellipses denote level set perimeter and dots denote MLEs.
PBLA in black and Althaus (2014) in red.

# Tristan da Cunha: A mutitype SEM

- $N = 254$ individuals split into three age groups
- 36% infants, 17% kids, and 13% adults
- $\beta_{kj} = \beta_{G(j)}$ depend on age group of susceptible
- Random walk MCMC vs. gold standard DAMCMC



Figure: Remote island community

---

[0] Compared against Hayakawa et al. (2003)
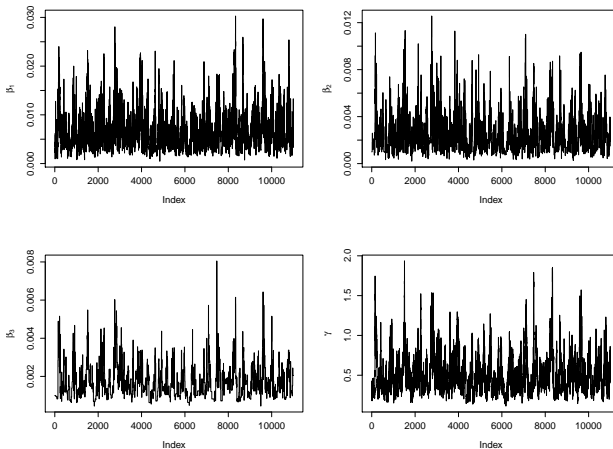
## Tristan da Cunha: Trace plots



Figure: Trace plots of $(\beta_1, \beta_2, \beta_3, \gamma)$ for Tristan da Cunha common cold epidemic using PBLA MCMC.
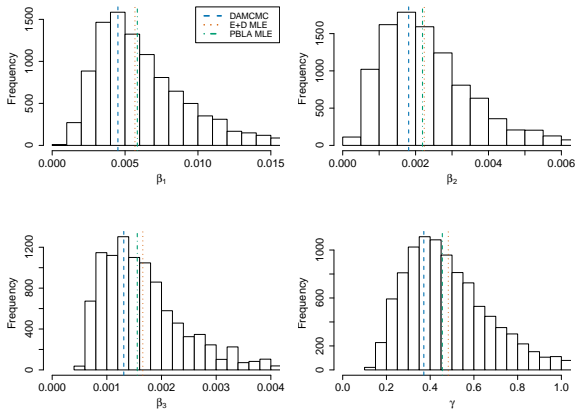
## Tristan da Cunha: Posterior Samples



Figure: Histograms of $(\beta_1, \beta_2, \beta_3, \gamma)$ posterior samples for Tristan da Cunha common cold epidemic using PBLA MCMC. DAMCMC posterior mean (blue), E+D MLE (orange), and PBLA MLE (green).

## Tristan da Cunha: Summary

Table: Posterior means from PBLA MCMC and DAMCMC methods, and MLEs using the E+D approximation and standard PBLA, for Tristan da Cunha common cold epidemic.

|  | PBLA MCMC | DAMCMC | E+D MLE | PBLA MLE |
|---|---|---|---|---|
| $\beta_1$ | 0.00648 | 0.00451 | 0.00568 | 0.00584 |
| $\beta_2$ | 0.00244 | 0.00181 | 0.00224 | 0.00219 |
| $\beta_3$ | 0.00171 | 0.00131 | 0.00166 | 0.00156 |
| $\gamma$ | 0.50565 | 0.37100 | 0.48273 | 0.45562 |
| $R_0$ | 1.17580 | 1.16102 | 1.12396 | 1.15301 |

# Dog rabies in CAR

- $n = 123$ infecteds between 2006 and 2012
- $N$ unknown
- Undercounts likely
- $\beta_{kj} = \beta_0 \exp(-\theta \cdot \rho(i, j))$ where $\rho$ is distance
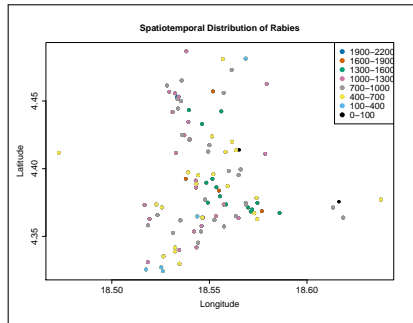  - Vanishingly small $\theta$



Figure: Spread of dog rabies

## Dog rabies in CAR: Underreporting

Table: Inferences on $(\beta, \gamma, R_0)$ using product PBLA, psuedo-removals adjustment, and total population size $N = 10,000$ for rabies epidemic in Bangui, Central African Republic.

| $\eta$ | $\beta$ | $\gamma$ | $R_0$ | Interval $R_0$ |
|--------|---------|----------|-------|----------------|
| 1.000  | 0.172   | 0.172    | 1.001 | (0.656, 1.527) |
| 0.500  | 0.203   | 0.200    | 1.015 | (0.752, 1.368) |
| 0.200  | 0.294   | 0.281    | 1.045 | (0.884, 1.235) |
| 0.100  | 0.424   | 0.385    | 1.099 | (0.974, 1.240) |

# Dog rabies in CAR: Sensitivity analysis for *N*

Table: Inferences on $(\beta, \gamma, R_0)$ with increasing total population size *N*.

| N | $\eta$ | $\beta$ | $\gamma$ | $R_0$ | Interval $R_0$ |
|---|---|---|---|---|---|
| 25,000 | 1.000 | 0.171 | 0.171 | 0.996 | (0.652, 1.518) |
| | 0.500 | 0.209 | 0.208 | 1.003 | (0.747, 1.347) |
| | 0.200 | 0.399 | 0.392 | 1.017 | (0.766, 1.348) |
| | 0.100 | 0.447 | 0.431 | 1.037 | (0.916, 1.173) |
| 50,000 | 1.000 | 0.170 | 0.171 | 0.994 | (0.651, 1.516) |
| | 0.500 | 0.312 | 0.312 | 1.000 | (0.733, 1.363) |
| | 0.200 | 0.263 | 0.261 | 1.007 | (0.841, 1.207) |
| | 0.100 | 0.384 | 0.377 | 1.017 | (0.901 1.149) |
| 100,000 | 1.000 | 0.170 | 0.171 | 0.993 | (0.650, 1.514) |
| | 0.500 | 0.189 | 0.189 | 0.998 | (0.757, 1.315) |
| | 0.200 | 0.263 | 0.262 | 1.003 | (0.847, 1.187) |
| | 0.100 | 0.331 | 0.329 | 1.008 | (0.894, 1.136) |

## Underreporting

- Undercount of size $n$ biases estimation of $R_0$
  - ▶ Theorems 4.1-2 (Andersson and Britton, 2000) say that $n/N$ is asymptotically normally distributed conditional on $R_0$

- Given reporting rate $\eta$, I propose bias corrections:
  1. Draw pseudo-removal times from KDE
  2. Scale $N^* = N \cdot \eta$

- I suggest that $\eta$ is not identifiable from removal times only.
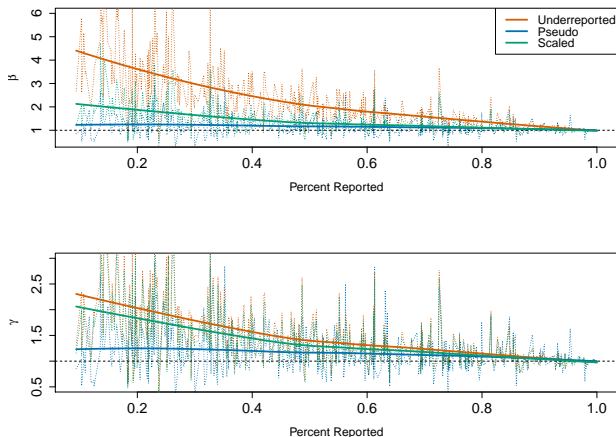
# Underreporting



Figure: Scaled ratio of PBLA MLEs with full partial data versus underreported partial data. $\beta$ (top) and $\gamma$ (bottom).
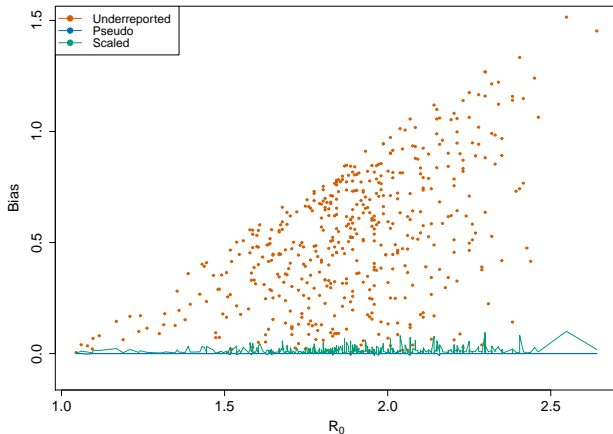
# Underreporting



Figure: Difference in $R_0$ from PBLA MLEs with full partial data versus underreport partial data.
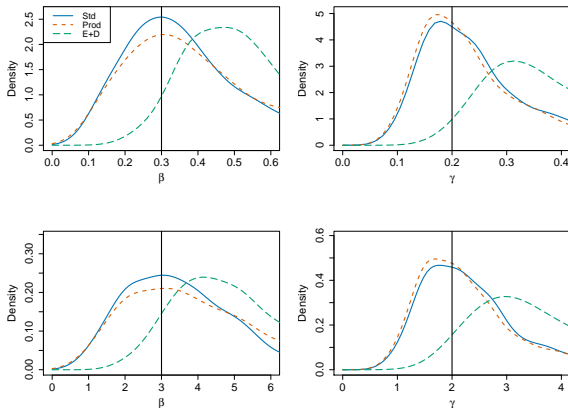
# **Varying** $(\beta, \gamma)$



Figure: Varying parameters $(\beta, \gamma)$: (0.3, 0.2) (top) and (3, 2) (bottom). MLEs from 1000 simulations with exponential infectious periods, $N = 100$, and $R_0 = \beta/\gamma = 1.5$.
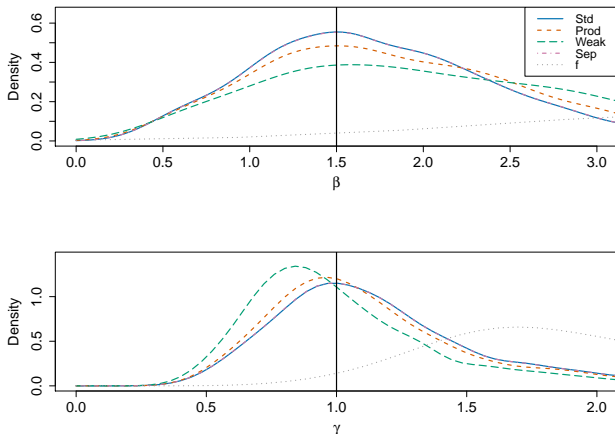
## PBLAs comparison



Figure: Comparison of pair-based likelihood approximations.
MLEs from 1000 simulations with $(\beta, \gamma) = (1.5, 1)$, $N = 250$.

## Partial data likelihood

$$\pi(\mathsf{r}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int \pi(\mathsf{i}_{-\alpha},\mathsf{r}|\boldsymbol{\beta},\boldsymbol{\theta},\alpha,i_\alpha)\pi(i_\alpha,\alpha)\,d\mathsf{i}_{-\alpha}d\,i_\alpha\,d\alpha$$

$$= \sum_{j=1}^{n} \pi(\alpha) \int \left\{\prod_{j\neq\alpha}^{n}\psi_j\chi_j\right\}\pi(i_\alpha|\alpha)\left\{\prod_{j=1}^{n}\phi_j f_j(r_j - i_j|\theta_j)\right\}d\mathsf{i}$$

$$= \left\{\prod_{j=1}^{n}a(\theta_j,-B_j)\right\}\sum_{j=1}^{n}\pi(\alpha)\,\mathbb{E}_{\mathsf{g}}\left[\pi(i_\alpha|\alpha)\left\{\prod_{j\neq\alpha}^{n}\psi_j\chi_j\right\}\right]$$

$$\approx \left\{\prod_{j=1}^{n}a(\theta_j,-B_j)\right\}\sum_{j=1}^{n}\pi(\alpha)\,\mathbb{E}_{\mathsf{g}}[\pi(i_\alpha|\alpha)]\,\mathbb{E}_{\mathsf{g}}\left[\left\{\prod_{j\neq\alpha}^{n}\psi_j\chi_j\right\}\right]$$

$$(9)$$