PhD Preliminary Exam Report on
"Pair-based likelihood approximations for stochastic epidemic models"

Seth D Temple[1]

Abstract

This report examines the approximation methods of Stockdale et al. (2019) for efficient evaluation of likelihoods in stochastic epidemic models with partially observed data. I introduce the stochastic epidemic model as a continuous-time Markov chain, and I motivate likelihood approximations by identifying computational challenges faced by Bayesian data-augmented MCMC procedures. Next, I derive various likelihood approximations and discuss the practical implications of their independence assumptions. Simulation studies demonstrate when PBLAs are reasonable approximations, and three data analyses showcase their flexibility in modeling complex transmission dynamics.

1 Introduction

Epidemiology affords vital insights into the incidence, spread, and control of contagions that threaten human, animal, and plant welfare. Fitting exact statistical models for such pursuits is computationally demanding when disease transmission is partially observed. Stockdale et al. (2019) propose fast likelihood approximations that exploit the underlying structure of an epidemiological model. The *stochastic epidemic model (SEM)* is a specific example of a bivariate continuous-time Markov chain with discrete state space. For some index set T, we say that the collection of random variables $\{X(t) : t \in T\}$ is a stochastic process, and we label these continuous-time processes if T is a continuous interval. A stochastic process satisfies the Markov property if its future conditional on its history only depends on its most recent observation; namely, for any times $0 = t_0 \leq t_1 \leq \cdots \leq t_n < t$ and $s > 0$,

$$P(X(t+s)\,|\,X(t), X(t_n), \ldots, X(t_1), X(t_0)) = P(X(t+s)\,|\,X(t))$$

Together, these two conditions define a continuous-time Markov chain (CTMC).

[1] Department of Statistics, University of Washington Seattle, WA, 98195, USA

For a given epidemic, let N be the size of a closed population with an initial infected. Each new infection is due to an infected individual from within this population. At time t, let $S(t)$, $I(t)$, and $R(t)$ denote counts of susceptible, infected, and removed individuals, where a removal means that an infected may no longer transmit the disease. The epidemic starts at $S(0) = N - 1$ and $I(0) = 1$, and it concludes when $I(t^*) = 0$. Because $N = S(t) + I(t) + R(t)$ for all t, we study the bivariate CTMC $\{(S(t), I(t)) : t \geq 0\}$. Transitions between states in this CTMC are based on independent and adversarial Poisson processes (PPs), with some processes governing infections and other processes governing removals. For instance, at t, there are $S(t) \cdot I(t)$ Poisson processes with rate β and $I(t)$ Poisson process with rate γ. β-PPs correspond to a susceptible being infected and γ-PPs correspond to an infected being removed. Figure 6 (Appendix A) illustrates these dynamics of individuals moving about an enclosed space spreading and coalescing from pathogens. Following the arguments of Durrett (1999), Allen (2008), and Andersson and Britton (2012), I derive transition rates for infections and removals:

$$P\{\,(S(t), I(t)) \to (S(t) - 1, I(t) + 1)\,\} = \beta \cdot S(t) \cdot I(t) \tag{1}$$

$$P\{\,(S(t), I(t)) \to (S(t), I(t) - 1)\,\} = \gamma \cdot I(t) \tag{2}$$

I interpret these transitions as the byproduct of $S(t) \cdot I(t)$ β-PPs and $I(t)$ γ-PPs racing to the first renewal. Since renewals for PPs are exponentially distributed, these races are the minima of independent exponential random variables. From this observation, I developed Algorithm 1, an event-driven simulator for epidemics. It illustrates *exponential racing*, a concept that I apply in Section 2.2 to evaluate various expressions.

Algorithm 1 Simulating a general stochastic epidemic

```
Initiate $S(0) = N - 1$, $I(0) = 1$, $R(0) = 0$, and $t = 0$.
while $I(t) > 0$ do
    Compute $\eta_1 = \beta \cdot I(t) \cdot S(t)$ and $\eta_2 = \beta \cdot I(t)$.
    Draw $s \sim$ Exponential($\eta_1 + \eta_2$). Update $t = t + s$.
    if $q \sim$ Bernoulli($\frac{\eta_1}{\eta_1 + \eta_2}$) then Decrement $S(t)$ and increment $I(t)$ by 1.
    else Decrement $I(t)$ and increment $R(t)$ by 1.
end while
```

This CTMC formulation may be generalized to individual-based infection rates β_{kj} and removal rates γ_j. Moreover, we may require that a removal occur after m_j marginal renewals of a rate γ_j exponential random variable, facilitating Gamma(m_j, γ_j)-distributed infectious periods. We refer to the SEM with $\beta_{kj} = \beta/N$ for all pairs (k, j) and $\theta_j := (m_j, \gamma_j) = (1, \gamma)$ for all j as the *general stochastic epidemic model*.

Based on the aforementioned SEM as a CTMC, I derive a likelihood conditional on an initial infected and infection times $i_1, \ldots, i_n$ and removal times $r_1, \ldots, r_n$ from $n \leq N$ infected individuals. Without loss of generality, I assume infecteds are sorted according to their removal times, and, for simplicity, I assume the first removed is the initial infected. Let β_{kj} be the infection rate that k applies to j, $\theta_j := (m_j, \gamma_j)$ parameterize the infectious period $r_j - i_j \sim$ Gamma(m_j, γ_j), and τ_{kj} be the total time that k tries to infect j. Next, I define three terms ψ_j, χ_j, and ϕ_j that serve as the probabilities for j evading infection until time i_j, j being infected at time i_j, and j failing to infect the $N - n$ never-infecteds. With respect to the CTMC, ψ_j and ϕ_j come as survival probabilities and χ_j is a transition rate:

$$
\begin{aligned}
\psi_j &= \exp\Big(-\sum_{k \neq j} \beta_{kj}\tau_{kj}\Big) = \prod_{k \neq j} \exp(-\beta_{kj}\tau_{kj}) = \prod_{k \neq j} \psi_{kj} \\
\chi_j &= \sum_{k \neq j} \beta_{kj} 1_{\{i_k < i_j < r_k\}} \\
\phi_j &= \exp\Big(-\sum_{k=n+1}^{N} \beta_{jk}\tau_{jk}\Big) = \exp\Big(-\sum_{k=n+1}^{N} \beta_{jk}(r_j - i_j)\Big) = \exp\Big(-B_j(r_j - i_j)\Big)
\end{aligned}
$$

Marginal term $\psi_{kj} = \exp(-\beta_{kj}\tau_{kj})$ is the probability that j avoids infection from k. These ψ_{kj} reappear in the main method (Section 2.1.1), as Stockdale et al. (2019) approximate $\mathbb{E}[\psi_j]$ as the product of $\mathbb{E}[\psi_{kj}]$ terms. Given this notation, the model likelihood is

$$
\pi(\mathbf{i}_{-1}, \mathbf{r} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta}, i_1) = \left\{ \prod_{j=2}^{n} \psi_j \chi_j \phi_j f_j(r_j - i_j|\theta_j) \right\} \phi_1 f_1(r_1 - i_1|\theta_1), \tag{3}
$$

where f denotes densities for independent infectious periods. In practice, infection times $i_1, \ldots, i_n$ are not known, leaving us with partially observed data $r_1, \ldots, r_n$. We refer to (3) as the *augmented model likelihood*. Standard model fitting procedures either draw infection times within some sampling scheme or integrate over infection times.

Marginalizing over infection times involves numerical approximation, which is resource-intensive for large n. As a result, the gold standard approach from the current literature is to sample $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\mathbf{i}$ from some posterior distribution via a data-augmented Monte Carlo Markov Chain (DAMCMC) routine. However, this method is not without flaws. Suppose we have a general SEM, and we assume conjugate priors $\beta \sim \text{Gamma}(m_\beta, \nu_\beta)$ and $\gamma \sim \text{Gamma}(m_\gamma, \nu_\gamma)$. Algorithm 2 provides details for a Metropolis-within-Gibbs sampling procedure. This scheme may mix poorly or diverge if the augmented infection times $\mathbf{i}$ are strongly correlated with β and γ (Stockdale, 2019). Terms $A_\mathbf{i}$ and $C_\mathbf{i}$ characterize how posterior samples of β and γ depend on $\mathbf{i}$. To ameliorate these posterior correlations, Kypraios (2007) and Neal and Roberts (2005) suggest reparameterization techniques that simultaneously update rate γ and infection times $\mathbf{i}$. (See Section 5.5 for an example algorithm.) While these methods address concerns about mixing and convergence, the computation time involved in exploring the parameter space increases considerably for large n. These shortcomings motivate the development of computationally efficient likelihood approximations that do not depend on unknown infection times.

Algorithm 2 Metropolis-within-Gibbs sampling for general SEM

```
Draw compatible[2] infection times $i_1^{(0)}, \ldots, i_n^{(0)}$.
for $l$ in $L$ iterations do
    Compute $A_\mathbf{i}^{(l)} = \sum_{j=1}^n \sum_{k=1}^N \tau_{jk}^{(l-1)}$.
    Draw $\beta^{(l)} \mid \gamma^{(l-1)}, \mathbf{i}^{(l-1)}, \mathbf{r} \sim \text{Gamma}(m_\beta + n - 1, \nu_\beta + A_\mathbf{i}^{(l)})$.
    Compute $C_\mathbf{i}^{(l)} = \sum_{j=1}^n (r_j - i_j^{(l-1)})$.
    Draw $\gamma^{(l)} \mid \beta^{(l)}, \mathbf{i}^{(l-1)}, \mathbf{r} \sim \text{Gamma}(m_\gamma + n, \nu_\gamma + C_\mathbf{i}^{(l)})$.
    for $M$ updates do
        Draw $j$ from $\{1, \ldots, n\}$.
        Draw a compatible $i'_j$ based on some proposal density $f$.
        Update $i_j^{(l)}$ based on the Hastings ratio.
    end for
end for
```

[2] An infection time is incompatible if it implies $I(t^*) = 0$ for $t^* < r_n$, a premature end to the epidemic.

2 Methods

2.1 Pair-based Likelihood Approximations

Rather than DAMCMC, Stockdale et al. (2019) propose approximate likelihoods for the observed removal times $\mathbf{r}$. They start with the augmented model likelihood and marginalize over unknown infection times $\mathbf{i}$. Instead of numerically approximating the integral, they take *bona fide* independence assumptions that simplify expectations of products. In general, they approximate expectations of products with products of expectations. Later, I will demonstrate in a simulation study that these independence assumptions break down when the infected proportion n/N exceeds fifty percent.

I now derive *pair-based likelihood approximations (PBLAs)*, state their independence assumptions, and offer formulas relevant to their evaluation. A key observation by the authors is that the integral over infection times is simplified via a change of variable. They absorb information about ϕ_j terms into densities by setting $a(\theta_j, -B_j)g_j(r_j - i_j|\theta_j) = \phi_j f_j(r_j - i_j|\theta_j)$, where moment generating function $a(\theta_j, \cdot)$ of $r_j - i_j$ and g_j are defined below.

$$
\begin{aligned}
f_j(r_j - i_j|\theta_j) &= \frac{\gamma_j^{m_j}}{\Gamma(m_j)}(r_j - i_j)^{m_j - 1}\exp(-\gamma_j(r_j - i_j)) \\
a(\theta_j, -B_j) &= \left(\frac{\gamma_j}{\gamma_j + B_j}\right)^{m_j} = \left(\frac{\gamma_j}{\delta_j}\right)^{m_j} \\
g_j(r_j - i_j|\theta_j) &= \frac{\exp(-B_j(r_j - i_j))f_j(r_j - i_j|\theta_j)}{a(\theta_j - B_j)} = \frac{\delta_j^{m_j}}{\Gamma(m_j)}(r_j - i_j)^{m_j - 1}\exp(-\delta_j(r_j - i_j))
\end{aligned}
$$

Above g_j is a Gamma density, but now with rate $\delta_j = \gamma_j + B_j$. Given that our theoretical developments rely on Poisson processes, this permanence of the Gamma family is crucial. The resulting partial likelihood is

$$
\begin{aligned}
\pi(\mathbf{r} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \pi(\mathbf{i}_{-1}, \mathbf{r}|\boldsymbol{\beta}, \boldsymbol{\theta}, i_1)\pi(i_1)\, d\mathbf{i}_{-1}\, di_1 \\
&= \left\{\prod_{j=1}^{n} a(\theta_j, -B_j)\right\}\mathbb{E}_{\mathbf{g}}[\pi(i_1)]\,\mathbb{E}_{\mathbf{g}}\left[\left\{\prod_{j=2}^{n}\psi_j\chi_j\right\}\right] \qquad (4)
\end{aligned}
$$

Next, we approximate the expected product in (4) by products of expectations.

$$\begin{aligned}\mathbb{E}_{\mathbf{g}}\left[\left\{\prod_{j=2}^{n}\psi_j\chi_j\right\}\right] &\approx \prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\psi_j\chi_j] \\ &\approx \prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\psi_j]\cdot\mathbb{E}_{\mathbf{g}}[\chi_j] \qquad (5) \\ \mathbb{E}_{\mathbf{g}}\left[\left\{\prod_{j=2}^{n}\psi_j\chi_j\right\}\right] &\approx \left\{\prod_{j=2}^{n}\mathbb{E}_{\mathbf{g}}[\chi_j]\right\}\left\{\mathbb{E}_{\mathbf{g}}\left[\prod_{j=2}^{n}\psi_j\right]\right\} \qquad (6)\end{aligned}$$

Approximation (5) is the basis for three general approximations, whereas approximation (6) is the basis for two approximations with distributional assumptions. For those two PBLAs, we require $r_j - i_j \sim \text{Exponential}(\delta)$ for all j. Both paradigms involve $\mathbb{E}_{\mathbf{g}}[\chi_j]$ terms which can be evaluated directly or via exponential racing. They differ in how they study ψ_j terms.

2.1.1 Standard PBLA

The *standard PBLA* involves further approximation for the product terms in (5); namely,

$$\begin{aligned}\mathbb{E}_{\mathbf{g}}[\psi_j]\cdot\mathbb{E}_{\mathbf{g}}[\chi_j] &\approx \left\{\prod_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}_{g_l,g_j}[\psi_{lj}]\right\}\left\{\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{g_k,g_j}\left[1_{\{i_k<i_j<r_k\}}\frac{\psi_{kj}}{\psi_{kj}}\right]\right\} \\ &\approx \left\{\prod_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}_{g_l,g_j}[\psi_{lj}]\right\}\left\{\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{g_k,g_j}\left[1_{\{i_k<i_j<r_k\}}\psi_{kj}\right](\mathbb{E}_{g_k,g_j}[\psi_{kj}])^{-1}\right\} \qquad (7)\end{aligned}$$

This approximation is appealing in that the marginal ψ_{kj} terms are jointly studied with indicators $1_{\{i_k<i_j<r_k\}}$ from χ_j. Stockdale et al. (2019) call it the standard method because it is the most general and it performs well in extensive simulation studies (Appendix B).

2.1.2 Product PBLAs

PBLAs based on (6) do not assume independent ψ_j terms and do not approximate $\mathbb{E}[\psi_j]$ as the product of marginal expectations. Instead, we require $r_j - i_j \sim \text{Exponential}(\delta)$ for all j. This distributional assumption means that $\theta_j = (m_j, \gamma_j) = (1, \gamma)$ for all j and $\beta_{kj} = \beta/N$ for all (k, j), which severely restricts the model. In Section 2.3, I state two theoretical results to evaluate the expected product, begetting two additional likelihood approximations. These results characterize the distribution of W, the cumulative time infective pressure is exerted.

$$W = \sum_{j=2}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \tau_{kj} = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (\tau_{kj} + \tau_{jk}) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \omega_{jk}$$

These methods then calculate the moment generating function at $-\beta/N$

$$\mathbb{E}_{\mathbf{g}}\Big[\prod_{j=2}^{n} \psi_j\Big] = \mathbb{E}_{\mathbf{g}}\Big[\exp\Big(- \beta/N \sum_{j=2}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \tau_{kj}\Big)\Big] = \mathbb{E}_{\mathbf{g}}[\exp(-(\beta/N)W)]$$

2.2 Pair-based Expectations

Our first two theoretical results concern terms $\mathbb{E}_{g_k,g_j}[\psi_{kj}]$ and $\mathbb{E}_{g_k,g_j}[1_{\{i_k < i_j < r_k\}}\psi_{kj}]$ in the standard PBLA. Formulas for these terms are exactly what is required to compute the likelihood approximation. I present such for $m_j, m_k = 1$ and $m_j, m_k \in \mathbb{Z}^+$.

Lemma 2.1. Let $r_j - i_j$ and $r_k - i_k$ be distinct independent exponentially distributed infectious periods with rates δ_j and δ_k.

$$\begin{aligned} &\mathbb{E}_{g_k,g_j}[\psi_{kj}] \\ &= \begin{cases} 1 - \beta_{kj}\delta_j\{(\delta_j + \delta_k)(\beta_{jk} + \delta_k)\}^{-1} \exp(-\delta_k(r_k - r_j)), & r_j < r_k \\ \delta_k(\beta_{kj} + \delta_k)^{-1} + \beta_{kj}\delta_k\{(\delta_j + \delta_k)(\beta_{kj} + \delta_k)\}^{-1} \exp(-\delta_j(r_j - r_k)), & r_j > r_k \end{cases} \end{aligned} \tag{8}$$

$$\begin{aligned} &\mathbb{E}_{g_k,g_j}[1_{\{i_k < i_j < r_k\}}\psi_{kj}] \\ &= \begin{cases} \delta_j\delta_k\{(\delta_j + \delta_k)(\beta_{kj} + \delta_k)\}^{-1} \exp(-\delta_k(r_k - r_j)), & r_j < r_k \\ \delta_j\delta_k\{(\delta_j + \delta_k)(\beta_{kj} + \delta_k)\}^{-1} \exp(-\delta_j(r_j - r_k)), & r_j > r_k \end{cases} \end{aligned} \tag{9}$$

Proof. I provide a partial proof as a sketch of the general strategy for these arguments. First, I decompose $\mathbb{E}_{g_k,g_j}[\psi_{kj}]$ into three terms based on τ_{kj}. To simplify notation, I drop the subscript ${}_{g_k,g_j}$ and define $\beta := \beta_{kj}$.

$$\tau_{kj} := r_k \wedge i_j - i_k \wedge i_j = \begin{cases} 0, & i_j < i_k \\ i_j - i_k, & i_k < i_j < r_k \\ r_k - i_k, & i_j > r_k \end{cases} \tag{10}$$

$$\begin{aligned} \mathbb{E}[\psi_{jk}] &= \mathbb{E}[\exp(-\beta\tau_{kj})] \\ &= \mathbb{E}[1_{\{i_j<i_k\}}] + \mathbb{E}[1_{\{i_j>r_k\}} \exp(-\beta(r_k - i_k))] + \mathbb{E}[1_{\{i_k<i_j<r_k\}} \exp(-\beta(i_j - i_k))] \end{aligned} \tag{11}$$

I consider these three terms in (11) individually. Although solving double integrals is possible, I put forward a more illuminating probabilistic argument based on Poisson processes.

I evaluate the third term, which is (9). Assume $r_j > r_k$. I traverse the stochastic process backwards from the latest removal time r_j. Between (r_k, r_j), I have a δ_j-PP, with the probability of no renewal in (r_k, r_j) being $\exp(-\delta_j(r_j - r_k))$. At r_k, I transition to a superposition of the δ_j-PP with a δ_k-PP, with probability $\delta_j(\delta_j + \delta_k)^{-1}$ of the first renewal being i_j. At i_j, I transition to a superposition of the δ_k-PP with a β-PP, with probability $\delta_k(\delta_k + \beta)^{-1}$ of the first renewal being i_k. Because these events are over independent increments, I can multiply the event probabilities.

$$\begin{aligned} \mathbb{E}[1_{\{i_k<i_j<r_k\}} \exp(-\beta(i_j - i_k))] &= \mathbb{E}[\exp(-\beta(i_j - i_k)) \,|\, i_k < i_j < r_k] \cdot P(i_k < i_j < r_k) \\ &= \frac{\delta_k}{\delta_k + \beta} \cdot \frac{\delta_j}{\delta_j + \delta_k} \cdot \exp(-\delta_j(r_j - r_k)) \end{aligned}$$

I take this general strategy of traversing the process backwards and setting up exponential races between independent PPs to determine the remaining terms and to handle the terms under the case $r_j < r_k$. I relegate these arguments to (Appendix A, Section 5.3). □

If $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{Exponential}(\delta)$, then $X = \sum_{j=1}^m X_j \sim \text{Gamma}(m, \delta)$. This sum property of Gamma random variables can be exploited to extend Lemma 2.1 to the broader class of Erlang infectious periods. Since $\mathbb{E}[X] = m/\delta$ and $\text{Var}(X) = m/\delta^2$, this enables modeling with more *a priori* belief that infectious periods are long.

Lemma 2.2. Let $r_j - i_j$ and $r_k - i_k$ be distinct independent Gamma-distributed infectious periods with positive integer shapes m_j and m_k and rates δ_j and δ_k.

$$\begin{aligned}
&\mathbb{E}_{g_k,g_j}[\psi_{kj}] \qquad (12)\\
&= \begin{cases}
1 + \exp(-\delta_k(r_k - r_j))\delta_j^{m_j} \sum_{l=0}^{m_k-1} \delta_k^l[(\delta_k\{\delta_k + \beta_{kj}\}^{-1})^{m_k-l} - 1] & \\
\times \sum_{p=0}^{l} \frac{1}{(l-p)!}\binom{m_j+p+1}{p}(r_k - r_j)^{l-p}(\delta_j + \delta_k)^{-(m_j+p)}, & r_j < r_k\\
1 - G_j(r_j - r_k)[1 - (\delta_k\{\delta_k + \beta_{kj}\}^{-1})^{m_k}] & \\
+ \exp(-\delta_j(r_j - r_k))\delta_j^{m_j} \sum_{l=0}^{m_k-1} \delta_k^l[(\delta_k\{\delta_k + \beta_{kj}\}^{-1})^{m_k-l} - 1] & \\
\times \sum_{p=0}^{m_j-1} \frac{1}{(m_j-p-1)!}\binom{l+p}{p}(r_k - r_j)^{m_j-p-1}(\delta_j + \delta_k)^{-(l+p+1)}, & r_j > r_k
\end{cases}\\
&\mathbb{E}_{g_k,g_j}[1_{\{i_k<i_j<r_k\}}\psi_{kj}] \qquad (13)\\
&= \begin{cases}
\exp(-\delta_k(r_k - r_j))\delta_j^{m_j} \sum_{l=0}^{m_k-1}(\{\delta_k + \beta_{kj}\}^{-1})^{m_k-l} & \\
\times \sum_{p=0}^{l} \frac{1}{(l-p)!}\binom{m_j+p+1}{p}(r_k - r_j)^{l-p}(\delta_j + \delta_k)^{-(m_j+p)}, & r_j < r_k\\
\exp(-\delta_j(r_j - r_k))\delta_j^{m_j} \sum_{l=0}^{m_k-1}(\{\delta_k + \beta_{kj}\}^{-1})^{m_k-l} & \\
\times \sum_{p=0}^{m_j-1} \frac{1}{(m_j-p-1)!}\binom{l+p}{p}(r_k - r_j)^{m_j-p-1}(\delta_j + \delta_k)^{-(l+p+1)}, & r_j > r_k
\end{cases}
\end{aligned}$$

where G_j is the cumulative distribution function corresponding to density g_j.

Proof. I approach these arguments with the same strategy as in Lemma 2.1, but I must handle some combinatorics in that a removal occurs after m marginal renewals. Figure 1 offers a helpful schematic for (13) under the case $r_j > r_k$. The δ_j-PP is renewed $m_j - p - 1$

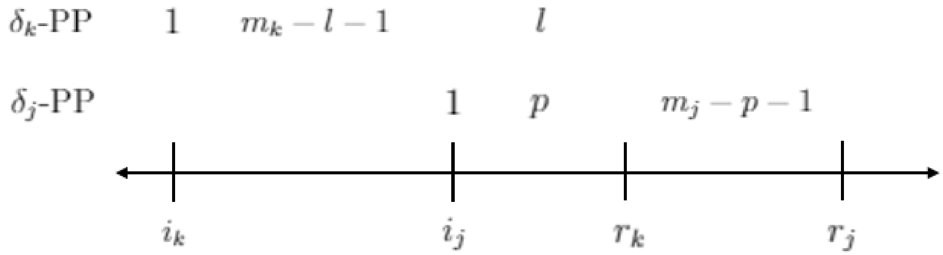

Figure 1: Marginal renewals of PPs.

times between (r_k, r_j). Then, there are $l+p$ exponential races between the δ_j- and δ_k-PPs, with δ_j-PP winning p times and δ_k-PP winning l times. Next, there is a single race where δ_j-PP wins at time i_j, achieving its m_j^{th} renewal. Finally, the δ_k-PP races against and beats a β-PP $m_k - l$ times, the last of which happens at time i_k. Summing over indices l and p,

$$\begin{aligned}\mathbb{E}[1_{\{i_k<i_j<r_k\}}\exp(-\beta(i_j-i_k))] &= \mathbb{E}[\exp(-\beta(i_j-i_k))\,|\,i_k<i_j<r_k]\cdot P(i_k<i_j<r_k)\\ &= \sum_{l=0}^{m_k-1}\sum_{p=0}^{m_j-1}\Bigg[e^{-\delta_j(r_j-r_k)}\frac{(\delta_j(r_j-r_k))^{m_j-p-1}}{(m_j-p-1)!}\\ &\times\binom{l+p}{p}\left(\frac{\delta_j}{\delta_j+\delta_k}\right)^{p+1}\times\left(\frac{\delta_k}{\delta_k+\beta}\right)^{m_k-l}\Bigg]\end{aligned}$$

Similar probabilistic arguments based on Poisson processes may be made for the other terms and for the case $r_j < r_k$. □

2.3 Product Expectations

For a general SEM with exponential infectious periods, we may evaluate the expectation of the product of ψ_j terms in (6) by establishing distributional results for W, the total time during which infective pressure is exerted. First, we study a subtotal time V for a subset of infectious periods. This general result describes cumulative times that individuals try to infect each other. It suffices to characterize W when all infecteds are considered.

Lemma 2.3. Let $\mathcal{K} \subseteq \{1,\ldots,n\}$ with $K := |\mathcal{K}| \geq 2$. If $\{r_k - i_k : k \in \mathcal{K}\} \overset{\text{ind}}{\sim} \text{Exponential}(\delta)$,

$$V = \sum_{j\in\mathcal{K}}\sum_{\substack{k\in\mathcal{K}\\ k\neq j}}\tau_{kj} \sim \sum_{j=1}^{K-1} j\cdot Y_j$$

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

Proof. Without loss of generality, I resort and relabel $\mathcal{K}$ with indices $\{1,\ldots,K\}$ according to removal times. I traverse the CTMC $\{(S(t),I(t)) : t < r_K\}$ in reverse. This involves infections $(S(t),I(t)) \to (S(t){+}1, I(t){-}1)$ at rate $\delta I(t)$ and removals $(S(t),I(t)) \to (S(t), I(t){+}1)$. I express V as a finite summation and make changes of variable $t' = t\cdot(I(t))^{-1}$ over piecewise intervals. See (Appendix A, Section 5.4.1) for a complete argument. □

For large n, conditional on another independence assumption, we have a normal approximation for W. This asymptotic result is based on a U-statistic for multiple comparisons from Barbour and Eagleson (1985). Let $D_n := \{(j,k) : 1 \leq j < k \leq n\}$ denote 2-element subsets of $\{1, \ldots, n\}$, and let $\{X_{jk} : (j,k) \in D_n)\}$ be a collection of zero-mean random variables. This collection is *dissociated* if for all pairs X_{jk} and X_{lp} the condition $|(j,k) \cap (l,p)| = 0$ implies independence, where $|\cdot|$ denotes the cardinality of shared indices.

Lemma 2.4. Let infectious periods be $r_1 - i_1, \ldots, r_n - i_n \overset{\text{iid}}{\sim} \text{Exponential}(\delta)$. Moreover, given a dissociated collection of random variables $\{\omega_{jk} : (j,k) \in D_n\}$, define

$$\begin{aligned}
s_n^2 &:= \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \sum_{l=1}^{n-1} \sum_{m=l+1}^{n} \left(\mathbb{E}[\omega_{jk}\omega_{lm}] - \mathbb{E}[\omega_{jk}]\mathbb{E}[\omega_{lm}]\right) \\
&= \frac{(4n-5)}{3\delta^2}\binom{n}{2} \\
Z_n &:= \frac{\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(\omega_{jk} - \mathbb{E}[\omega_{jk}])}{s_n} \\
&= \frac{\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(\omega_{jk} - \delta^{-1})}{s_n} \\
&= \frac{W - \frac{1}{\delta}\binom{n}{2}}{s_n}
\end{aligned}$$

Then, $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$. (Proof in Appendix A, Section 5.4.2.)

Given Lemmas 2.3 and 2.4, we derive expressions for the second term in (6) as moment generating functions at $-\beta/N$.

$$\mathbb{E}_{\mathbf{g}}\left[\prod_{j=2}^{n}\psi_j\right] = \prod_{j=1}^{n-1}\left(\frac{\delta}{\delta + j(\beta/N)}\right) \tag{14}$$

$$\mathbb{E}_{\mathbf{g}}\left[\prod_{j=2}^{n}\psi_j\right] \approx \exp\left(-\frac{\beta}{N\delta}\binom{n}{2} + \frac{\beta^2}{6N^2\delta^2}(4n-5)\binom{n}{2}\right) \tag{15}$$

I refer to the likelihood approximations based on 14 and 15 as *product* and *weak PBLAs*. (14) is a more attractive as an exact result whereas (15) assumes dissociation. Albeit (14) is an $O(n)$ task and (15) is an $O(1)$ task, both PBLAs involve the expected product of χ_j terms, which is an $O(n^2)$ task.

3 Results

3.1 Simulation Studies

In extensive simulation studies, Stockdale et al. (2019) found that standard PBLA equaled or outperformed competitor methods in inferential accuracy and/or computational runtime. Namely, they demonstrated (i) that MLEs $\hat{\beta}$ and $\hat{\gamma}$ from standard PBLA center about the truth even when varying infection rate β, removal rate γ, population size N, and shape m, (ii) that an MCMC sampler with random walk proposal densities and Hastings ratios based on standard PBLA mixes as well as DAMCMC, and (iii) that the various PBLAs provide nearly identical inference. Assuming PBLAs are good approximations, these results are unsurprising in that we expect MLEs to be good estimators and we know DAMCMC mixes slowly. I have reproduced these findings in Appendix B using my `R` package sdtemple/pblas. Additionally, I conducted simulation studies to (iv) measure compute times for PBLAs, (v) study the performance of PBLAs under increasing infected proportion n/N, and (vi) assess inference in the presence of disease case underreporting.

3.1.1 Computational Runtime

PBLAs apply to individual-based models where interactions between pairs of individuals are of interest. As a result, these methods scale quadratically in n, the number of infected individuals. Using Algorithm 1, I simulated candidate epidemics with $n/N \approx 0.5$ and timed how long it took to calculate likelihoods. Table 1 presents runtimes in seconds for likelihood approximations as n increases. For maximum likelihood estimation, optimizers evaluate the likelihood approximation tens to hundreds of times, so individual compute times less than 1 minute are preferred. As a comparison, I report runtimes for a method requiring numerical approximation (Eichner and Dietz, 2003). Even for medium-sized epidemics, this method may take hours to days to find an MLE. Pair-based likelihood approximations evaluatec considerably faster. Given that product and weak PBLAs scaled similarly, I favor product PBLA because its expected product is an exact result, not based on a normal approximation. Standard and product PBLAs differed by a constant multiplier of about 5, both experiencing $O(n^2)$ behavior. Parallel computing may accelerate these runtimes (Appendix B, Table 3).

Table 1: Time in seconds to compute likelihood for standard, product, and weak PBLAs, and Eichner-Dietz approximation. Based on simulated epidemics with increasing n and infected proportion $n/N \approx 0.5$. Superscript * denotes very small, nonzero times.

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	

3.1.2 Infected Proportion

Stockdale et al. (2019) suggest that PBLAs degrade in performance when the infected proportion n/N exceeds seventy percent. I study this claim by simulating general stochastic epidemics, subsetting by the infected proportion, and comparing inferences between PBLAs. Figure 2 summarizes such a study with $N = 200$, $(\beta, \gamma) = (1.5, 1)$, and enough simulations so that each decile has at least 50 observations for density estimation. (Plots for γ show similar patterns and are available in Appendix B.) For increasing infected proportion n/N, inferences on (β, γ) start to deviate between standard, product, and weak PBLAs. One explanation for this behavior is that the relative likelihood contributions of the pairwise approximated $\mathbb{E}[\psi_j]$ terms increases with n/N (see Appendix B). Another explanation is that each PBLA takes on different independence assumptions. For highly contagious epidemics, independent transmission is doubtful. I interpret divergence between standard PBLA and product PBLA as an indicator for when the *bona fide* independence of marginal ψ_{kj} terms begins to break. Based on Figure 2, I caution against PBLAs if n/N exceeds fifty percent. Weak PBLA offers different inferences for finite n and ought to ignored in light of the more exact product PBLA.

Andersson and Britton (2012, Theorems 4.1-2) indicate an asymptotic relationship between the basic reproduction number $R_0 = \beta/\gamma$ and the infected proportion n/N. PBLA inference on R_0 generally follows this rule. Figure 3 reports kernel densities for R_0 as n/N increases. Again, for n/N exceeding fifty percent, standard and product PBLAs disagree.

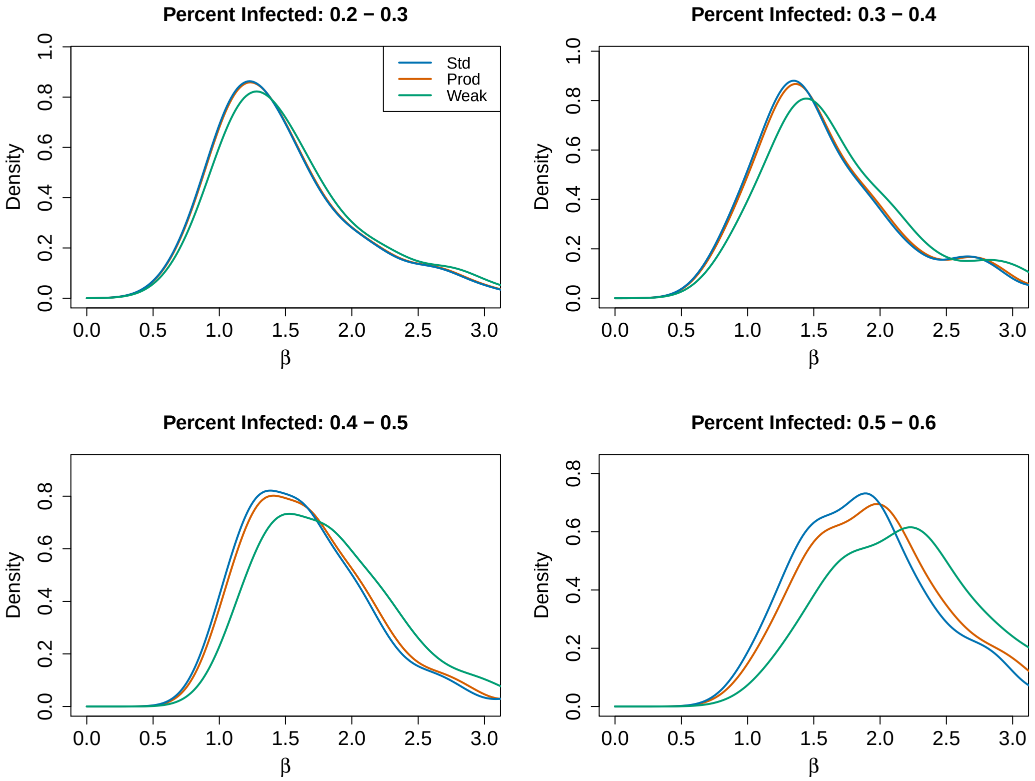

Figure 2: Inferences on (β, γ) for increasing infection proportion n/N. Titles above graphs denote ranges for infected proportions in simulation study. Legend in top left defines colors for standard, product, and weak PBLAs.

Standard PBLA seldom estimates $R_0 > 2$ even when n/N is large, and inferences from product PBLA become less certain as n/N increases. Most importantly, Figure 3 highlights that the partial data only provide intel on either β or γ, with the other implied by $R_0 = \beta/\gamma$. This deficiency makes inference on the partial data challenging. Given partial and complete simulated data from a general stochastic epidemic, I estimated MLEs for $(\beta, \gamma) = (1.5, 1)$ with the augmented model likelihood and product PBLA. Density estimation in Figure 4 points out that inference with complete data is more certain and more accurate. Moreover, the flat and bimodal curve for product PBLA stresses how inadequate PBLAs can be for inference on R_0. Since PBLAs base their R_0 guesses on a single number, n/N, they are unable to detect the adversarial dynamics of the infection and removal processes. This stresses the need to collect multiple dates and early dates for epidemic monitoring.

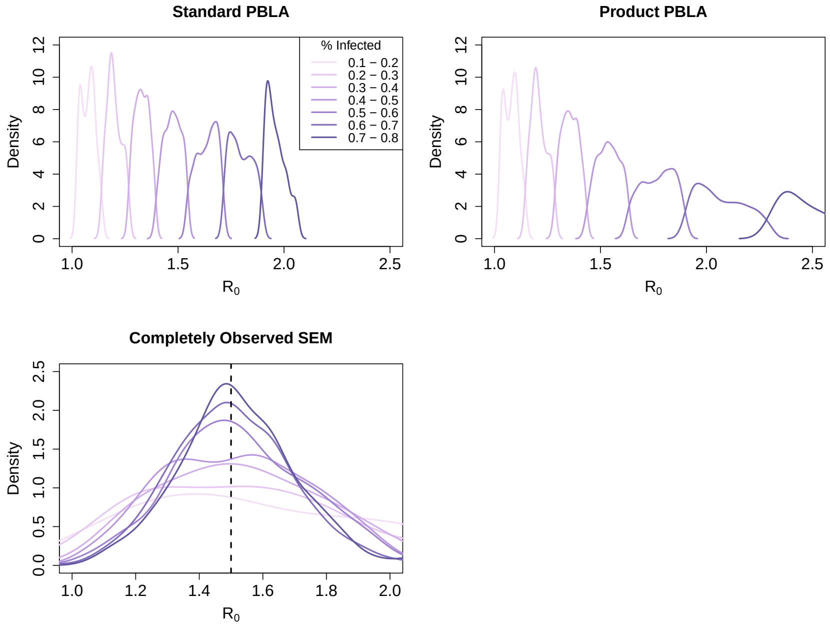

Figure 3: Inferences on $R_0 = \beta/\gamma$ for increasing infection proportion n/N. Legend in top right defines gradient colors for increasing n/N. Partial data methods estimate R_0 conditional on the infected proportion, which impacts simultaneous inference of (β, γ).

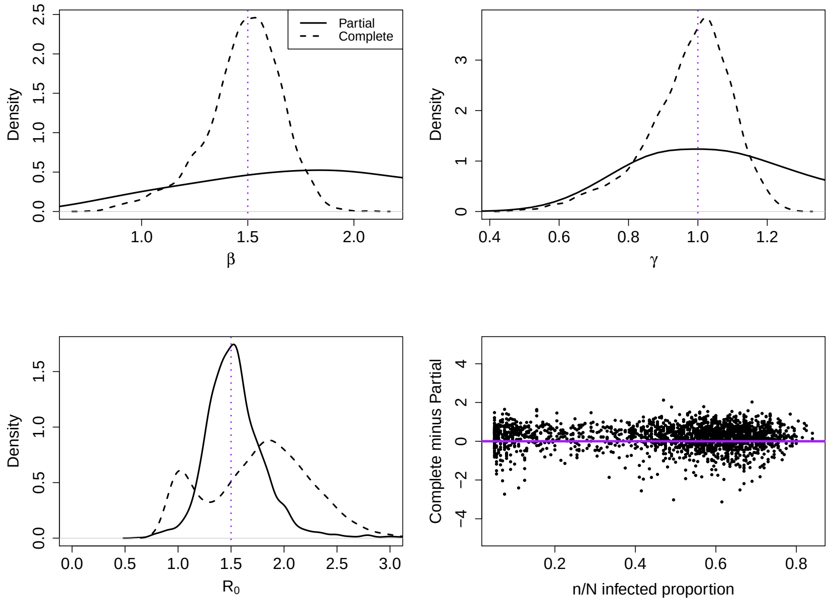

Figure 4: Inferences on β (top left), γ (top right), R_0 (bottom left) with partial (bold) and complete (dashed) data. $R_0^{\text{complete}} - R_0^{\text{partial}}$ as n/N increases in the bottom left pane. Dotted purple line is the simulated true value.

3.1.3 Case Underreporting

Underreporting of disease cases in localities with limited public health infrastructure and epidemiological surveillance may bias statistical inferences. If the reported epidemic size n is an undercount, R_0 will be calibrated lower (Section 3.1.2). I suggest two ad hoc adjustments for underreporting: sampling from a kernel density estimator to acquire pseudo-removal times $\tilde{\mathbf{r}}$, or scaling down N. Given an estimate $\hat{\eta}$ for the reporting rate η, I draw $\lceil \frac{n}{\hat{\eta}} \rceil - n$ pseudo-removal times from a kernel density estimator based on the reported removal times $\mathbf{r}$. This procedure assumes (i) that the reported removal times $\mathbf{r}$ are a large and representative sample from the population of removal times and (ii) that $\hat{\eta}$ is a good (consistent) estimator for the reporting rate. In principle, sampling pseudo-removal times is flexible enough to handle piecewise shifts in the reporting rate. On the other hand, sampling pseudo-removal times increases runtime and would involve sampling other covariates for complex SEMs. Scaling $\tilde{N} := N\hat{\eta}$ does not affect runtime, but assumes a constant reporting rate.

I conducted a simulation study to test the performance of my underreporting adjustments. For simulated epidemics, I first estimated $\hat{\beta}_0$ and $\hat{\gamma}_0$ with full reporting. Then, I randomly excluded some proportion of the removal times and estimated β and γ with underreporting and with pseudo-removals and scaling adjustments. Finally, I considered these in ratio relative to numerators $\hat{\beta}_0$ and $\hat{\gamma}_0$ and I assessed their implied R_0 values. For low reporting rates, $(\hat{\beta}_{\text{scaled}}, \hat{\gamma}_{\text{scaled}})$ and $(\hat{\beta}_{\text{pseudo}}, \hat{\gamma}_{\text{pseudo}})$ are not recovered, but, together recover $\hat{R}_0 = \hat{\beta}_0/\hat{\gamma}_0$ (Appendix B, Section 6.6). This is due to the link between n/N and $\hat{R}_0$ (Section 3.1.2).

In the case of underreporting, the general SEM views transition rates (1) and (2) for infections and removals as

$$\beta(\eta S(t)I(t) + \eta(1-\eta)I(t)^2) = \beta\left(\eta\left(1 + (1-\eta)\frac{I(t)}{S(t)}\right)\right)S(t)I(t) = \beta_{\eta,t}S(t)I(t) \tag{16}$$

$$\gamma(\eta I(t)) = \gamma_\eta I(t) \tag{17}$$

For many real epidemics, $N - n \gg n$, in which case $\frac{I(t)}{S(t)} \approx 0$ and $\beta_{\eta,t} \approx \beta_\eta := \beta\eta$. This heuristic presents η as a nonidentifiable parameter, since η scales β and γ in the same way. In short, an estimator $\hat{\eta}$ must be based on data ancillary to the observed removal times.

3.2 Data Analyses

PBLAs based on pairwise expectations may be employed to study complex epidemics where infection and removal rates vary between individuals. For example, Stockdale et al. (2019) apply PBLA methods to investigate three epidemics: (i) a small epidemic where infection rates depend on age groups, (ii) a large epidemic where infection rates change over time, and (iii) a large epidemic where infection rates depend on a Euclidean distance. In these examples, it suffices to specify formulas for β_{jk} and γ_j. (iii) uses confidential data from the United Kingdom that I could not access, so I have replaced this dataset with a small and underreported rabies epidemic in Bangui, Central African Republic. My analyses of (i) and the rabies epidemic and some commentary on (iii) are available in Appendix C.

3.2.1 Ebola Virus in West Africa

Outbreaks of ebola hemorrhagic fever rose to some prominence in Guinea, Sierra Leone, and Liberia during 2014. I apply standard PBLA to death records from the Centers for Disease Control and Prevention (CDC). Althaus (2014) researched this epidemic with a mechanistic Susceptible-Exposed-Infective-Removed (SEIR) model, where removals were recoveries or deaths. This approach solves ordinary differential equations and then takes the output as means for a Poisson likelihood. Additionally, Althaus (2014) postulated infection rates that decrease over time in light of control efforts. Stockdale et al. (2019) restricted that analysis to deaths data only in comparing to PBLA inference. Since PBLA operates under fixed infection rates between individuals j and k, they suggested the formula $\beta_{jk} = \beta_0 \exp(-k_0(T_{jk}))$ where β_0 is a null infection rate, k_0 is a decay rate, and T_{jk} is the expected midpoint time in which j can infect k. For the exposed and infectious periods, they assumed a fixed lag of $c = 5.3$ days, modifying pairwise expectations accordingly, and a fixed removal rate $\gamma = 5.61^{-1}$. (Details on the expected midpoint time T_{jk} and the fixed lag can be found in Appendix C.) Proceeding with this model specification, I determined maximum likelihood estimates for β_0, k_0, and $R_0 := \beta_0/\gamma$.

Table 2 juxtaposes my analysis against Stockdale et al. (2019) and Althaus (2014). Despite initializing at different values and employing a different optimizer, I arrived at the exact same MLEs as Stockdale et al. (2019). This observation establishes my implementation as a faithful reproduction of PBLA. I also constructed elliptical level sets for (β_0, k_0) with the output Hessian from my Newton-type optimizer. These are visualized on a log likelihood surface for the Guinea epidemic in Figure 5 and implied R_0 ranges are reported in Table 2. (Contour plots for the Sierra Leone and Liberia epidemics and other visuals are available in Section 7.2.) Log likelihood contours suggest that these MLEs are unique. PBLA and the Poisson-based SEIR model largely agree. As a whole, this case study exhibits the flexibility of PBLA to study more elaborate transmission dynamics and time-varying infection rates. Finally, I benchmarked compute times for these analyses. On a standard laptop with an Intel i7 core, it took 12, 31, and 51 minutes for the Guinea, Sierra Leone, and Liberia datasets. I found my implementation to be marginally faster than the authors', yet these times are slower than reported in Stockdale et al. (2019) (less than 1 minute).

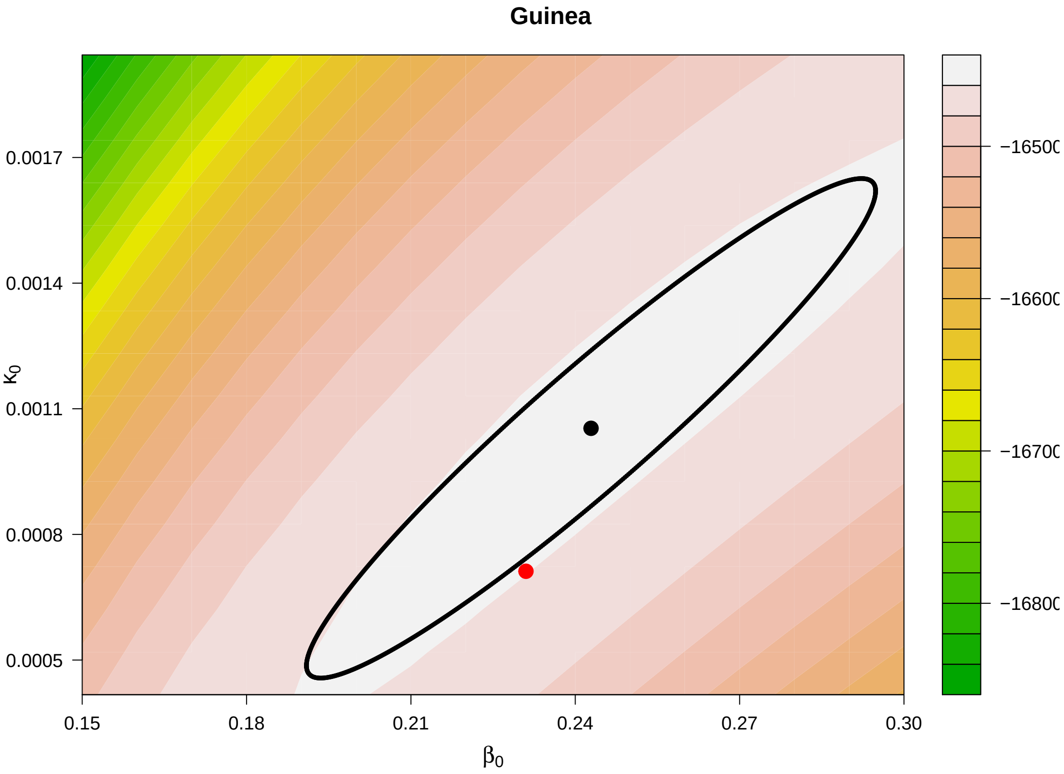

Figure 5: Log likelihood contours for Ebola virus epidemic in Guinea. Ellipses denote level set perimeters and dots denote MLEs from PBLA (black) and Althaus (2014) (red).

Table 2: 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	Author	β_0	k_0	R_0	Interval R_0
Guinea	Temple, Stockdale	0.243	0.00105	1.36	(1.07, 1.65)
	Althaus	0.231	0.00071	1.30	
Sierra Leone	Temple, Stockdale	0.335	0.00289	1.88	(1.55, 2.20)
	Althaus	0.277	0.00180	1.55	
Liberia	Temple, Stockdale	0.266	0.00180	1.49	(1.30, 1.69)
	Althaus	0.303	0.00251	1.70	

4 Discussion

Pair-based likelihood approximations for partially observed epidemics offer a fast frequentist alternative to Bayesian DAMCMC when datasets are large or model parameters are correlated *a posteriori*. Despite scaling quadratically with n infecteds, these methods compute quick enough for analyses of large epidemics ($n > 10000$) on a standard laptop within some hours. I conducted a runtime comparison to set user expectations for the time burden, and I found that weak PBLA ran as fast as the more exact product PBLA. In general, I recommend standard PBLA over other PBLAs for its flexibility. Parallel computing and/or modifying the methods so as to only consider some pairwise terms may expedite analyses. In another simulation study, I investigated the behavior of PBLAs as the infected proportion n/N increases. I found that PBLAs arrive at different likelihood approximations when n/N exceeds fifty percent, casting doubt about the pairwise independence assumption. Moreover, I diagnosed that the partial SEM approach may not have enough information to simultaneously infer β and γ. Simulation studies in Stockdale et al. (2019) obfuscate in density plots that $\hat{\beta}$ and $\hat{\gamma}$ are moderated by $R_0 := \beta/\gamma$, with R_0 closely linked to the epidemic size. On the other hand, these methods matched other epidemiological analyses on real data, alleviating this concern for more plausible transmission dynamics in which covariates help to distinguish between β and γ effects. PBLAs appear broadly applicable for most large epidemics in which their speed outperforms competitor methods.

I offer six likelihood approximations in my `R` package sdtemple/pblas. Whereas Stockdale et al. (2019) use `C` programming and depend on nonstandard libraries, I opted for the scriptability of R programming. I also wrote efficient code to contend with the authors' (compiled language) runtime performance. Using my simple implementations as a basis, I adapted PBLAs to study disease transmission dynamics that depend on covariate effects, reporting practices, and control strategies over time. Copious examples of such are available as R scripts in my repository. Practitioners may experiment with other dynamics like vaccinations and fatal/non-fatal removals by specifying formulas for β_{kj} and γ_j.

Relaxing various model assumptions, say allowing some immigration and emigration or making inferences during a live epidemic, would extend the usefulness of these methods. Unfortunately, such extensions are not immediately obvious. The partial SEM likelihood framework explicitly considers interactions between known pairs over the entire course of the epidemic. A model with demography would have individuals present only part of the time. For an epidemic in progress, the main challenge is that the final epidemic size n is unknown, and PBLA inference depends strongly on n/N to calibrate R_0. These limitations confine PBLA methods to retrospective analyses of past epidemics.

Another reality in infectious disease monitoring is underreporting. Assuming missingness completely at random, I proposed two corrections to de-bias PBLA estimates. Missingness in epidemic data usually depends on covariates, so this assumption is likely unreasonable. DAMCMC for partial SEMs addresses undercounts again by reversible-jump steps (O'Neill and Roberts, 1999). However, augmenting individuals becomes more elaborate for covariate-dependent SEMs. Assessing the utility of individual-based models relative to count-based models would be a valuable contribution to epidemiology, especially since most data come as counts (Althaus, 2014; Fintzi et al., 2021; Irons and Raftery, 2021) and individual-based methods scale poorly for pandemics.

Pair-based likelihood approximations build on Poisson process theory to establish pairwise general moments for Gamma random variables. I connected these ideas to develop an event-driven epidemic simulator, unifying pairwise exponential racing with the generative Markov jump process. Approximating expected products as products of expectations is akin to, but distinct from, composite likelihood methods (Besag, 1975; Varin et al., 2011). Both (mis)specify models so as to make involved calculations more tractable. Like composite likelihoods, PBLAs are useful and sensible when model misspecification is robust to inputs.

References

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

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

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

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

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

J. C. Blackwood and L. M. Childs. An introduction to compartmental modeling for the budding infectious disease modeler. *Letters in Biomathematics*, 5(1):195–221, 2018.

H. Bourhy, E. Nakouné, M. Hall, P. Nouvellet, A. Lepelletier, C. Talbi, L. Watier, E. C. Holmes, S. Cauchemez, P. Lemey, et al. Revealing the micro-scale signature of endemic zoonotic disease transmission in an african urban setting. *PLoS pathogens*, 12(4):e1005525, 2016.

A. Cori, P. Nouvellet, T. Garske, H. Bourhy, E. Nakouné, and T. Jombart. A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLoS computational biology*, 14(12):e1006554, 2018.

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

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.

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

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

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

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.

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

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.

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

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

P. D. O'Neill and G. O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1): 121–129, 1999.

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.

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

5 Appendix A

Here I provide additional details on methodology. Figure 6 visualizes the stochastic epidemic model. Sections 5.2.1, 5.2.2, and 5.2.3 derive three inferior pair-based likelihood approximations. Complete theoretical arguments are given for Lemmas 2.1, 2.3, and 2.4 in Sections 5.3, 5.4.1, and 5.4.2. Finally, I discuss non-centered data-augmentation in Section 5.5.

5.1 Stochastic Epidemic

The stochastic model postulates that individuals move about in an enclosed space, spread pathogens based on between-individual infection rates, and are removed (no longer infectious) based on individual-specific removal rates. Epidemiologists propose formulas β_{kj} and γ_j to elucidate key dynamics governing these processes, say space, time, group effects, mitigation/control strategies, etc.

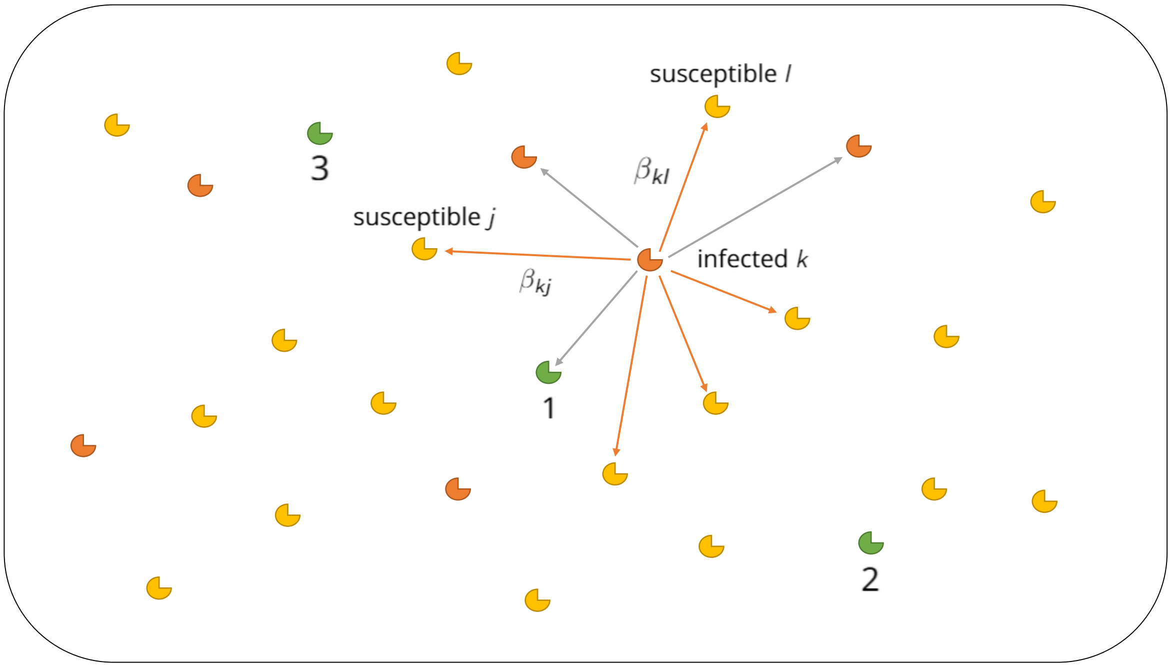

Figure 6: Susceptible (yellow), infected (orange), and removed (green) pacmen interact in a closed population world. Orange and grey arrows correspond to possible and null infections, where an infection is null because an already-infected cannot be reinfected. Parameters β_{kj} and β_{lj} denote infection rates between individuals. Removeds (green) are labeled according to their chronological order.

5.2 Other Approximations

5.2.1 Separated PBLA

Like standard PBLA, the *separated PBLA* estimates the expected product of marginal ψ_{kj} terms with the product of $\mathbb{E}[\psi_{kj}]$.

$$\begin{aligned}\mathbb{E}_{\mathbf{g}}[\psi_j\chi_j] &\approx \mathbb{E}_{\mathbf{g}}[\psi_j]\mathbb{E}_{\mathbf{g}}[\chi_j] \\ &= \mathbb{E}_{\mathbf{g}}\left[\prod_{\substack{l=1\\l\neq j}}^{n}\psi_{jl}\right]\mathbb{E}_{\mathbf{g}}\left[\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}1_{\{i_k<i_j<r_k\}}\right] \\ &\approx \left\{\prod_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}_{g_j,g_l}[\psi_{jl}]\right\}\left\{\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{g_j,g_k}[1_{\{i_k<i_j<r_k\}}]\right\} \end{aligned} \tag{18}$$

$\mathbb{E}[\psi_{kj}]$ and $\mathbb{E}[\chi_j]$ terms contribute independently to the likelihood, hence the nomenclature. This PBLA provides nearly identical computation and inference to those of standard PBLA.

5.2.2 f-based PBLA

The f-based PBLA involves densities $\mathbf{f}$ instead of $\mathbf{g}$, i.e. we do not perform the change of variable in (4). Otherwise, its derivation is similar to standard PBLA.

$$\mathbb{E}_{\mathbf{f}}[\psi_j\chi_j\phi_j] \approx \mathbb{E}_{\mathbf{f}}[\chi_j\phi_j]\mathbb{E}_{\mathbf{f}}[\psi_j] \approx \mathbb{E}_{\mathbf{f}}[\chi_j\phi_j]\prod_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}_{f_j,f_l}[\psi_{jl}]$$

Stockdale (2019) uses similar probabilistic arguments (Section 2.2) to determine formulas for this method. In Section 6.1 of Appendix B, I demonstrate that f-based PBLA is worse at estimating (β,γ) than other PBLAs. I suspect this is due to the unattractive specification that ψ_j and ϕ_j are independent.

5.2.3 Eichner-Dietz Approximation

Stockdale et al. (2019) present the likelihood expression of Eichner and Dietz (2003) for a smallpox transmission model as an approximation. They derive the E+D likelihood approximation by assuming (i) independence between infecteds and never-infecteds and (ii) independence of ψ_j and χ_j terms conditional on infection times.

$$\pi_{\mathrm{E+D}}(\mathbf{r} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \left(\prod_{j=1}^{n} \mathbb{E}_{f_j} \big[\, \mathbb{E}_{\mathbf{f}}[\psi_j \,|\, i_j] \, \mathbb{E}_{\mathbf{f}}[\chi_j \,|\, i_j] \, \big] \right) \left(\prod_{j=1}^{n} \mathbb{E}_{\mathbf{f}}[\phi_j] \right)$$

Eichner and Dietz (2003) do not numerically integrate in their original work. They backcalculate infection times $\mathbf{i}$ given a fixed mean infectious period length. Thus, the E+D approximation in Stockdale et al. (2019) is more an illustrative example for numerical approximation techniques than it is a faithful reproduction of the original method. Unsurprisingly, numerically integrating over infection times is slow (Table 1). Besides this practical difficulty, I show in Appendix B that the E+D approximation is worse at inferring infection rates β and removal rates γ than standard PBLA. Like f-based PBLA, I suspect this is due to the unattractive specification that ψ_j and ϕ_j are independent.

5.3 Pair-based Expectations

Here I provide complete arguments for Lemma 2.1. Additionally, I discuss how to generalize these arguments when shape $m > 1$.

Proof. Recall that $\mathbb{E}[\psi_{kj}]$ decomposes into three terms, and that we evaluate these three terms under cases $r_j < r_k$ and $r_j > r_k$.

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

Assume $r_j > r_k$. I traverse the stochastic process backwards from r_j. Between (r_k, r_j), I have a δ_j-PP, with the probability of no renewal in (r_k, r_j) being $\exp(-\delta_j(r_j - r_k))$. At r_k, I transition to a superposition of the δ_j-PP with a δ_k-PP, with probability $\delta_j(\delta_j + \delta_k)^{-1}$ of the first renewal being i_j. At i_j, I transition to a superposition of the δ_k-PP with a β-PP, with probability $\delta_k(\delta_k + \beta)^{-1}$ of the first renewal being i_k. Because these events are over independent increments, I can multiply the event probabilities. The first and third terms involve no renewal in (r_k, r_j). The first term has δ_k-PP renewing before the δ_j-PP.

$$\mathbb{E}[1_{\{i_k > i_j\}}] = P(i_k > i_j) = P(i_j < i_k < r_k < r_j) = \frac{\delta_k}{\delta_j + \delta_k} \exp(-\delta_j(r_j - r_k))$$

The third term describes δ_j-PP renewing before the δ_k-PP and then the δ_k-PP renewing before the β-PP.

$$\begin{aligned} \mathbb{E}[1_{\{i_k < i_j < r_k\}} \exp(-\beta(i_j - i_k))] &= \mathbb{E}[\exp(-\beta(i_j - i_k)) \,|\, i_k < i_j < r_k] \cdot P(i_k < i_j < r_k) \\ &= \frac{\delta_k}{\delta_k + \beta} \cdot \frac{\delta_j}{\delta_j + \delta_k} \cdot \exp(-\delta_j(r_j - r_k)) \end{aligned}$$

The second term involves the renewal i_j in (r_k, r_j), which I take a complement to evaluate. Finally, the δ_k-PP renews before the β-PP.

$$\begin{aligned} \mathbb{E}[\exp(-\beta_{kj}(r_k - i_k))1_{\{i_j > r_k\}}] &= \mathbb{E}[\exp(-\beta_{kj}(r_k - i_k))|i_k < r_k < i_j < r_j] \cdot P(i_k < r_k < i_j < r_j) \\ &= \delta_k/(\delta_k + \beta_{kj}) \cdot [1 - \exp(-\delta_j(r_j - r_k))] \end{aligned}$$

Now, assume $r_j < r_k$. I traverse the stochastic process backwards from r_k. Between (r_j, r_k), I have a δ_k-PP, with the probability of no renewal in (r_j, r_k) being $\exp(-\delta_k(r_k - r_j))$. At r_j, I transition to a superposition of the δ_k-PP with a δ_j-PP, with probability $\delta_k(\delta_j + \delta_k)^{-1}$ of the first renewal being i_k. At i_k, I transition to a superposition of the δ_j-PP with a β-PP, with probability $\delta_j(\delta_j + \beta)^{-1}$ of the first renewal being i_j. Because these events are over independent increments, I can multiply the event probabilities. The first and third terms involve no renewal in (r_j, r_k). For the first term, I take the complement of no renewal in (r_j, r_k) and the δ_j-PP renewing before the δ_k-PP.

$$\mathbb{E}[1_{\{i_k > i_j\}}] = P(i_k > i_j) = 1 - P(i_k < i_j < r_j < r_k) = 1 - \frac{\delta_j}{\delta_j + \delta_k} \exp(-\delta_k(r_k - r_j))$$

The third term describes δ_j-PP renewing before the δ_k-PP and then the δ_k-PP renewing before the β-PP.

$$\begin{aligned}\mathbb{E}[1_{\{i_k<i_j<r_k\}}\exp(-\beta(i_j-i_k))] &= \mathbb{E}[\exp(-\beta(i_j-i_k))\,|\,i_k<i_j<r_k]\cdot P(i_k<i_j<r_k)\\ &= \frac{\delta_k}{\delta_k+\beta}\cdot\frac{\delta_j}{\delta_j+\delta_k}\cdot\exp(-\delta_j(r_j-r_k))\end{aligned}$$

The second term is zero because $i_j > r_k$ is impossible under $r_j < r_k$.

Combining these terms, I arrive at the desired results under $r_j > r_k$ and $r_j < r_k$. Note that the third terms are expressions (9) in Lemma 2.1, so I derived these en route. To generalize this result for $m > 1$, I draw schematics like Figure 1. These facilitate combinatorial arguments when an infection or removal occurs after m marginal renewals. Whereas the survival probabilities in Lemma 2.1 can be succinctly written as $\exp(-\delta_j(r_j - r_k))$ and $\exp(-\delta_k(r_k - r_j))$, I write $1 - G_j(r_j - r_k)$ and $1 - G_k(r_k - r_j)$ for the survival probabilities in Lemma 2.2, with G denoting the cumulative distribution function of a Gamma random variable. Taking complements requires more bookkeeping as well when $m > 1$. With these changes, proofs proceeds as before. See Stockdale (2019) for complete arguments. □

5.4 Product Expectations

5.4.1 Lemma 2.3

Proof. Without loss of generality, I assume $\mathcal{K}$ is reordered and relabeled such that r_K is the latest removal time and i_1 is the earliest infection time. I traverse the CTMC $\{(S(t), I(t)) : t < r_K\}$ backwards in time from r_K to i_1. This reverse process follows two transition rules:

$$(S(t), I(t)) \to (S(t) + 1, I(t) - 1) \tag{19}$$

$$(S(t), I(t)) \to (S(t), I(t) + 1) \tag{20}$$

(19) refers to infections that occur at rate $\delta I(t)$; (20) refers to removals. I characterize W, the total time in which infective pressure is applied. Suppose during time interval (a, b) there is constant (S, I). Each infected individual applies infective pressure to each susceptible

individual for time $(b-a)$, i.e. the aggregated time is $(b-a)IS$.

$$V = \int_{i_\alpha}^{r_K} S(t) \cdot I(t)\, dt$$

More generally, for $u < r_K$, I define

$$T(u) = \int_{u}^{r_K} S(t) \cdot I(t)\, dt$$

$T(\cdot)$ is a piecewise linear function because the integrand is a jump function. Based on this piecewise property, I partition the integral into finite sums. Let $i_1 = \tilde{t}_1 < \cdots < \tilde{t}_{2K} = r_K$ be the ordered infection and removal times. Then,

$$T(i_1) = \sum_{k=2}^{2K} S(\tilde{t}_k) \cdot I(\tilde{t}_k) \cdot (\tilde{t}_k - \tilde{t}_{k-1})$$

For each interval $(\tilde{t}_{k-1}, \tilde{t}_k)$, I apply the change of variable $t' = t \cdot I(\tilde{t}_k)^{-1}$. (Division is well-defined here because $I(t) = 0$ corresponds to the end of the epidemic.) This effectively slows the clock according to the count of infected individuals. Moreover, this clock change means that (19) now happens at rate δ and (20) no longer affects the finite summation. Let $i_1 = \bar{t}_1 < \cdots < \bar{t}_K$ be the ordered infection times. After the clock change,

$$T(i_1) = \sum_{k=2}^{K} S(\bar{t}_k)(\bar{t}_k - \bar{t}_{k-1})$$

$$Y_k := \bar{t}_k - \bar{t}_{k-1} \sim \text{Exponential}(\delta)$$

where $Y_2, \ldots, Y_K$ are independent. In reverse time, piecewise linear T starts at zero, transitions to rate 1 at infection time $\bar{t}_K$, transitions to rate 2 at infection time $\bar{t}_{K-1}$, and so on. In conclusion,

$$V = T(i_1) = \sum_{k=2}^{K} (K-k+1) \cdot Y_k$$

□

5.4.2 Central Limit Theorem for Dissociated Comparisons

Theorem 5.1. (Barbour and Eagleson, 1985) 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 $\sigma_n^2 := \sum_{(i,j),(k,l) \in D_n} \mathbb{E}[X_{ij}X_{kl}]$ and $Z_n := \sigma_n^{-1} \sum_{(i,j)\in D_n} X_{ij}$. Then, $Z_n \rightsquigarrow N(0,1)$ if

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

5.4.3 Lemma 2.4

Proof. I show that Theorem 5.1 applies in this case. Set $X_{jk} = \omega_{jk} - \mathbb{E}[\omega_{jk}]$ and $\sigma_n^2 = s_n^2$. From Lemma 2.3, if $\mathcal{K} = \{j,k\}$, then $\omega_{jk} \sim \text{Exponential}(\delta)$, and

$$\begin{aligned}
\mathbb{E}[X_{jk}] &= 0 \\
\mathbb{E}[X_{jk}^2] &= \mathbb{E}[(\omega_{jk} - \mathbb{E}[\omega_{jk}])^2] = \mathbb{E}[\omega_{jk}^2] - \mathbb{E}[\omega_{jk}]^2 = \delta^{-2} \\
0 &< \mathbb{E}[|X_{jk}|^3] \\
&= \mathbb{E}[|\omega_{jk} - \mathbb{E}[\omega_{jk}]|^3] \\
&\leq \mathbb{E}[(\omega_{jk} + \mathbb{E}[\omega_{jk}])^3] < \infty
\end{aligned}$$

Thus, for exponential infectious periods,

$$\begin{aligned}
\sum_{(j,k)\in D_n} (\mathbb{E}[|X_{jk}|^3])^{1/3} \Bigg(\sum_{(l,p):|(j,k)\cap(l,p)|=0} (\mathbb{E}[|X_{lp}|^3])^{1/3} \Bigg)^2 &= c^3 \binom{n}{2}\binom{n-2}{2} \\
&= c^3 \frac{n(n-1)(n-2)(n-3)}{2(2)} \\
&= O(n^4)
\end{aligned}$$

Next, I compute σ_n^2.

$$\begin{aligned}
\sigma_n^2 &= \sum_{(j,k),(l,p)\,\in\, D_n} \mathbb{E}[X_{jk}X_{lp}] \\
&= \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=0}} \mathbb{E}[X_{jk}X_{lp}] + \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=1}} \mathbb{E}[X_{jk}X_{lp}] + \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=2}} \mathbb{E}[X_{jk}X_{lp}] \\
&= \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=0}} \mathbb{E}[X_{jk}]\mathbb{E}[X_{lp}] + \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=1}} \mathbb{E}[X_{jk}X_{lp}] + \sum_{(j,k)\,\in\, D_n} \mathbb{E}[X_{jk}^2] \\
&= 0 + \sum_{\substack{(j,k),(l,p)\,\in\, D_n \\ |(j,k)\cap(l,p)|=1}} \mathbb{E}[X_{jk}X_{lp}] + \delta^{-2}\binom{n}{2}
\end{aligned}$$

Evaluating the middle term where there is one element shared between the 2-element subsets requires more work. For $1 \le a < b < c \le n$, define $\Omega_{abc} = \omega_{ab}\omega_{bc} + \omega_{ab}\omega_{ac} + \omega_{ac}\omega_{bc}$.

$$\begin{aligned}
\mathbb{E}[\Omega_{abc}] &= \mathbb{E}[\omega_{ab}\omega_{bc} + \omega_{ab}\omega_{ac} + \omega_{ac}\omega_{bc}] \\
&= \mathbb{E}[\omega_{ab}\omega_{bc}] + \mathbb{E}[\omega_{ab}\omega_{ac}] + \mathbb{E}[\omega_{ac}\omega_{bc}] \\
&= \mathrm{Cov}(\omega_{ab}\omega_{bc}) + \mathbb{E}[\omega_{ab}]\mathbb{E}[\omega_{bc}] + \mathrm{Cov}(\omega_{ab}\omega_{ac}) + \mathbb{E}[\omega_{ab}]\mathbb{E}[\omega_{ac}] + \mathrm{Cov}(\omega_{ac}\omega_{bc}) + \mathbb{E}[\omega_{ac}]\mathbb{E}[\omega_{bc}] \\
&= \mathrm{Cov}(\omega_{ab}\omega_{bc}) + \mathrm{Cov}(\omega_{ab}\omega_{ac}) + \mathrm{Cov}(\omega_{ac}\omega_{bc}) + \mathbb{E}[\omega_{ab}]\mathbb{E}[\omega_{bc}] + \mathbb{E}[\omega_{ab}]\mathbb{E}[\omega_{ac}] + \mathbb{E}[\omega_{ac}]\mathbb{E}[\omega_{bc}] \\
&= \big(\mathrm{Var}(\omega_{ab} + \omega_{ac} + \omega_{bc}) - \mathrm{Var}(\omega_{ab}) - \mathrm{Var}(\omega_{ac}) - \mathrm{Var}(\omega_{bc})\big)/2 + 3\delta^{-2} \\
&= (5\delta^{-2} - 3\delta^{-2})/2 + 3\delta^{-2} \\
&= 4\delta^{-2},
\end{aligned}$$

since $\mathrm{Var}(\omega_{ab} + \omega_{ac} + \omega_{bc}) = \mathrm{Var}(Y_1 + 2Y_2) = \mathrm{Var}(Y_1) + 4\mathrm{Var}(Y_2) = 5\delta^{-2}$ by Lemma 2.3.

Returning to the former computation,

$$\begin{aligned}\sigma_n^2 &= 0 + \sum_{\substack{(j,k),(l,m) \in D_n \\ |(j,k)\cap(l,m)|=1}} \mathbb{E}[X_{jk}X_{lm}] + \binom{n}{2}\sigma^{-2} \\ &= 0 + \sum_{1\le a<b<c\le n} \mathbb{E}[\Omega_{abc}] + \binom{n}{2}\sigma^{-2} \\ &= 0 + 4\delta^{-2}\binom{n}{3} + \delta^{-2}\binom{n}{2} \\ &= \delta^{-2}\binom{n}{2}\frac{4n-5}{3}\end{aligned}$$

This computation suggests that $\sigma_n^{-3} = O(n^{-9/2})$. Therefore,

$$\begin{aligned}\sigma_n^{-3}\sum_{(j,k)\in D_n}(\mathbb{E}[|X_{jk}|^3])^{1/3}\Bigg(\sum_{(l,p):|(j,k)\cap(l,p)|=0}(\mathbb{E}[|X_{lp}|^3])^{1/3}\Bigg)^2 &= O(n^{-9/2})O(n^4) \\ &= O(n^{-1/2}) \\ &= o(1),\end{aligned}$$

satisfying the sufficient condition.

□

5.5 Non-centered Metropolis-within-Gibbs Sampler

Standard DAMCMC (Algorithm 2) may mix poorly or not converge if infection rates, removal rates, and/or infection times are highly correlated. I observed this when implementing standard DAMCMC for a general SEM. Stockdale (2019, pp. 98-99) discusses this dependency, citing that γ and $B_{\mathbf{i}}$ become more correlated as n increases. For some SEM models, simulataneously updating γ and infection times $i_1, \ldots, i_n$ may remedy this issue. Kypraios (2007) focuses on such alternatives in his PhD thesis, and Neal and Roberts (2005) present some sample algorithms for these purposes. Based on these references, I developed Algorithm 3 to sample β and γ from a general SEM. I marginalize over infection times when possible to further break up the known dependency. Derivations related to Algorithm 3 are shown below. (20) and (22) are useful in Hastings ratios.

$$
\begin{aligned}
\pi(\gamma, \mathbf{i}|\mathbf{r}) &= \int \pi(\gamma, \mathbf{i}, \beta|\mathbf{r})\, d\beta \\
&= \int \frac{\pi(\mathbf{i}, \mathbf{r}|\beta, \gamma)\pi(\beta)\pi(\gamma)}{\pi(\mathbf{r})}\, d\beta \\
&= \frac{\pi(\gamma)}{\pi(\mathbf{r})}\left\{B_{\mathbf{i}} f_\gamma(\mathbf{r}-\mathbf{i})\right\} \int \frac{{\nu_\gamma}^{m_\gamma} \beta^{m_\gamma - 1} \exp(-\beta\nu_\gamma)}{\Gamma(m_\gamma)} \beta^{n-1} \exp(-\beta A_{\mathbf{i}})\, d\beta \\
&= \frac{\pi(\gamma)}{\pi(\mathbf{r})}\left\{B_{\mathbf{i}} f_\gamma(\mathbf{r}-\mathbf{i})\right\} \frac{\Gamma(m_\gamma + n - 1)\nu_\gamma^{m_\gamma}}{\Gamma(m_\gamma)(\nu_\gamma + A_{\mathbf{i}})^{m_\gamma + n - 1}} \qquad (21) \\
\frac{\pi(\gamma', \mathbf{i}'|\mathbf{r})}{\pi(\gamma, \mathbf{i}|\mathbf{r})} \frac{f_\gamma(\mathbf{r}-\mathbf{i})}{f_{\gamma'}(\mathbf{r}-\mathbf{i}')} &= \frac{\pi(\gamma')}{\pi(\gamma)} \frac{B_{\mathbf{i}'}}{B_{\mathbf{i}}} \frac{(\nu_\gamma + A_{\mathbf{i}})^{m_\gamma + n - 1}}{(\nu_\gamma + A_{\mathbf{i}'})^{m_\gamma + n - 1}} \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
\pi(i_k|\gamma, \mathbf{i}_{-k}, \mathbf{r}) &= \int \pi(i_k, \beta|\gamma, \mathbf{i}_{-k}, \mathbf{r})\, d\beta \\
&= \int \frac{\pi(\mathbf{i}, \mathbf{r}|\beta, \gamma)\pi(\beta)\pi(\gamma)}{\pi(\gamma, \mathbf{i}_{-k}, \mathbf{r})}\, d\beta \\
&= \frac{\pi(\gamma)}{\pi(\gamma, \mathbf{i}_{-k}, \mathbf{r})}\left\{B_{\mathbf{i}} f_\gamma(\mathbf{r}-\mathbf{i})\right\} \frac{\Gamma(m_\gamma + n - 1)m_\gamma^{m_\gamma}}{\Gamma(m_\gamma)(m_\gamma + A_{\mathbf{i}})^{m_\gamma + n - 1}} \qquad (23) \\
\frac{\pi(i_k'|\gamma, \mathbf{i}_{-k}, \mathbf{r}) f_\gamma(r_k - i_k)}{\pi(i_k|\gamma, \mathbf{i}_{-k}, \mathbf{r}) f_\gamma(r_k - i_k')} &= \frac{B_{\mathbf{i}'}(\nu_\gamma + A_{\mathbf{i}})^{m_\gamma + n - 1}}{B_{\mathbf{i}}(\nu_\gamma + A_{\mathbf{i}'})^{m_\gamma + n - 1}} \qquad (24)
\end{aligned}
$$

I used Algorithm 3 for a simulation study in Appendix B. However, this was not the focus of this research, as even non-centered DAMCMC methods are computationally demanding for large epidemics. PBLAs provide an efficient alternative to augmenting infection times.

Algorithm 3 Non-centered DAMCMC

```
Draw compatible infection times $i_1^{(0)}, \dots, i_n^{(0)}$.
for $l$ in $L$ iterations do
    Compute $A_{\mathbf{i}}^{(l)} = \sum_{j=1}^{n} \sum_{k=1}^{N} \tau_{kj}^{(l-1)}$.
    Draw $\beta^{(l)} \mid \gamma^{(l-1)}, \alpha^{(l-1)}, i_{\alpha}^{(l-1)}, \mathbf{i}_{-\alpha}^{(k-1)}, \mathbf{r} \sim \text{Gamma}(m_\beta + n - 1, \nu_\beta + A_{\mathbf{i}}^{(l)})$.
    Propose $\gamma'$ using a random walk. Edit $\mathbf{i}'$ using $i'_j = r'_j - \frac{1}{\gamma'}(r_j - i_j)$.
    Update $\gamma^{(l)}$ and $\mathbf{i}^{(l)}$ based on a Hastings ratio (22).
    for $M$ updates do
        Draw $j$ from $\{1, \dots, n\}$.
        Draw a compatible $i'_j$ based on some proposal density $f$.
        Update $i_j^{(l)}$ based on a Hastings ratio (24).
    end for
end for
```

6 Appendix B

I replicated the extensive simulation studies in the supplement of Stockdale et al. (2019). These studies verified that MLEs $\hat{\beta}_{\text{PBLA}}$ and $\hat{\gamma}_{\text{PBLA}}$ center about the simulated truth, except when the infected proportion n/N is large. In Section 3.1.2, I argue that these findings reflect the asymptotic behavior of the general stochastic epidemic (Andersson and Britton, 2012, Theorems 4.1-2), not any statistical guarantees of PBLA inference (consistency). For these studies, I generated 500-1000 general stochastic epidemics vis-á-vis Algorithm 1 before finding MLEs with the `nlm` optimizer. If a generated epidemic had less than 5 removal times, I threw out that epidemic and simulated another one. Epidemics with too few removal times do not contain enough information to study the infection and removal dynamics.

Section 6.1 compares five pair-based likelihood approximations in their ability to infer (β, γ). Sections 6.2 and 6.3 assess inference against the E+D approximation for exponential and Erlang infectious periods. Various runtime comparisons are made in Section 6.4, including against DAMCMC samplers and with parallel computing. Additional details and figures for the main simulation studies, infected proportion and underreporting, are available in Sections 6.5 and 6.6.

6.1 Comparison Study of PBLAs

By and large, these approximations are equivalent when the infected proportion n/N is less than fifty percent. Figure 7 displays density estimates for MLEs of various PBLAs from 1000 simulations. This figure is deceiving in that it suggests that the likelihoods differ. Rather, PBLAs differ when n/N exceeds fifty percent. f-based PBLA takes an especially questionable independence assumption, as evidenced by its poor estimation.

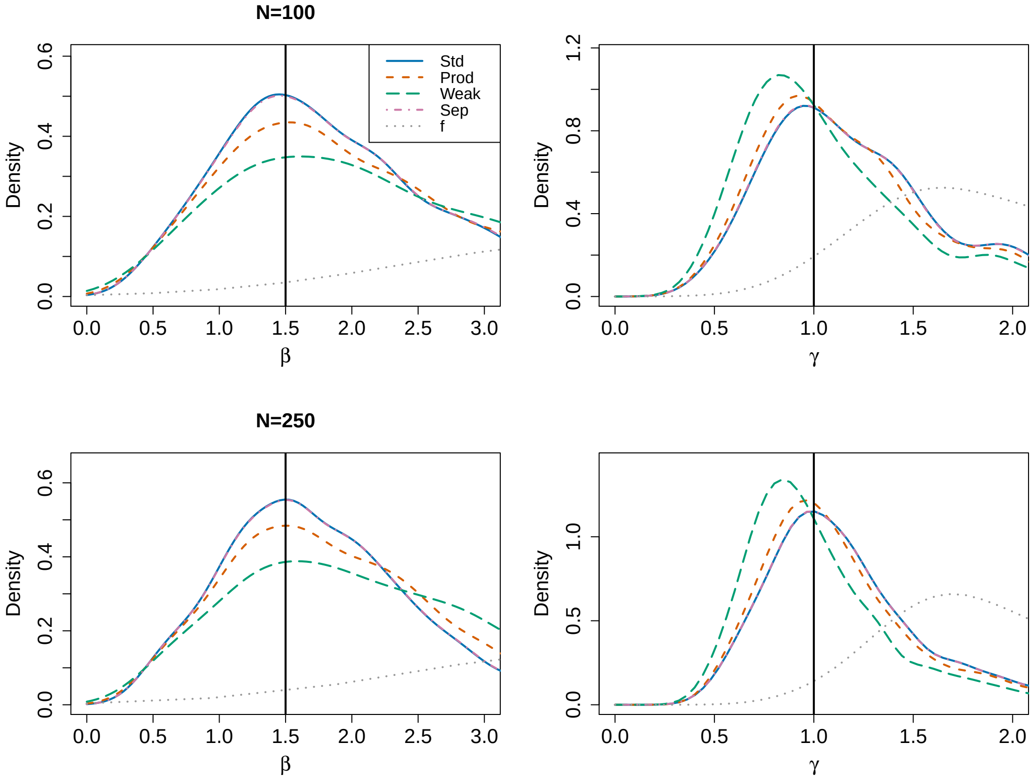

Figure 7: Comparison of pair-based likelihood approximations. MLEs from 1000 simulations with $(\beta, \gamma) = (1.5, 1)$. Top panels for $N = 100$ and bottom panels for $N = 250$. Legend in top left graph defines colors and line types for standard, product, weak, separated product, and f-based PBLAs. Note that standard and separated PBLAs provide nearly identical inference.

6.2 Exponential Infectious Periods

Under exponential infectious periods, PBLAs more accurately infer (β, γ) than the Eichner-Dietz approximation. Figures 8 and 9 indicate that these methods are robust to varying parameters (β, γ) and total population size N. Figure 10 exposes R_0 values for which PBLA and E+D methods struggle. Large R_0 values correspond to larger epidemic sizes, in which case these approximations fail. In this simulation study, PBLAs appear miscalibrated for small R_0 values because these correspond to small epidemic sizes n.

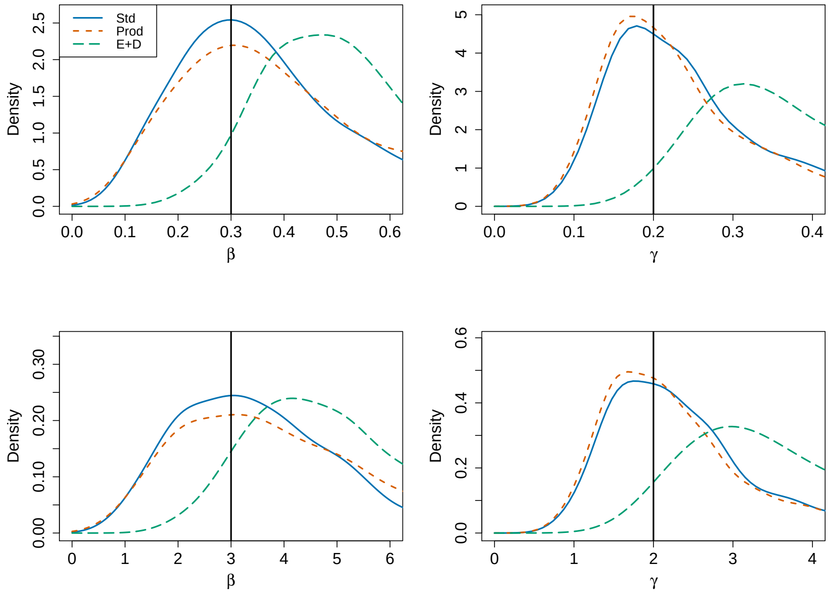

Figure 8: Varying parameters (β, γ): (0.3, 0.2) (top) and (3, 2) (bottom). MLEs from 1000 simulations with exponential infectious periods, $N = 100$, and $R_0 = \beta/\gamma = 1.5$. Legend in top left graph defines colors and line types for standard PBLA, product PBLA, and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

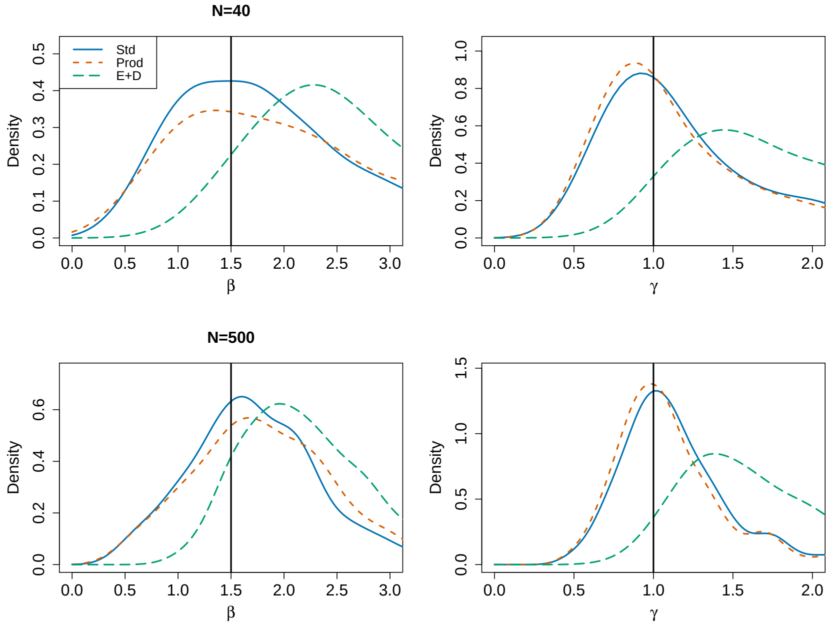

Figure 9: Varying total population size N: 40 (top) and 500 (bottom). MLEs from 1000 simulations with exponential infectious periods, and $(\beta, \gamma) = (1.5, 1)$. Legend in top left graph defines colors and line types for standard PBLA, product PBLA, and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

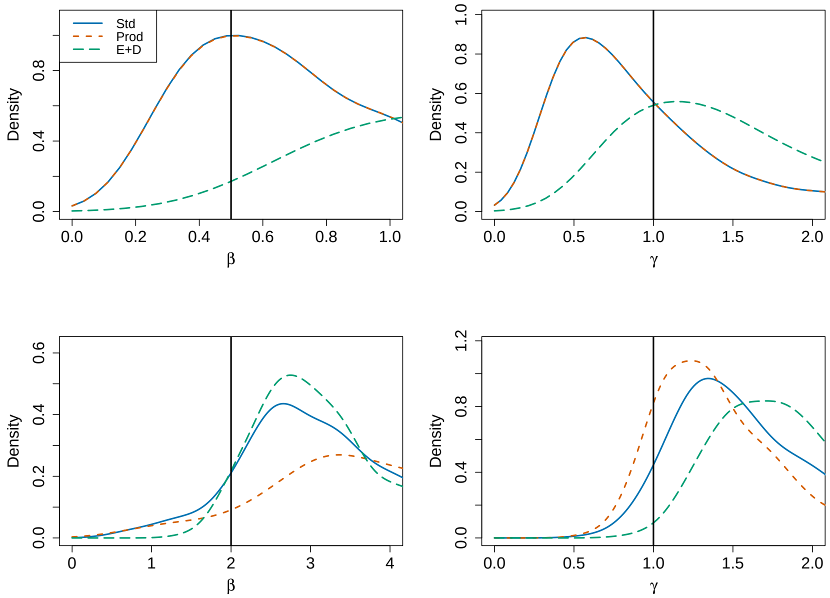

Figure 10: Varying basic reproduction number $R_0 = \beta/\gamma$: 0.5 (top) and 2 (bottom). MLEs from 1000 simulations with exponential infectious periods, $N = 100$, and $\gamma = 1$. Legend in top left graph defines colors and line types for standard PBLA, product PBLA, and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

6.3 Erlang Infectious Periods

Simulation studies highlight the same strengths and weaknesses for Gamma distributed infectious periods (Figures 11, 12, and 14). Additionally, Figure 13 tells the predictable story that the variance of the estimator decreases with more samples. In light of Section 3.1.2, this hints that either $\hat{\beta}_{\text{PBLA}}$ or $\hat{\gamma}_{\text{PBLA}}$ are individually consistent, but jointly $(\hat{\beta}_{\text{PBLA}}, \hat{\gamma}_{\text{PBLA}})$ is not consistent. Figure 15 extends robustness to the Erlang shape parameter m.

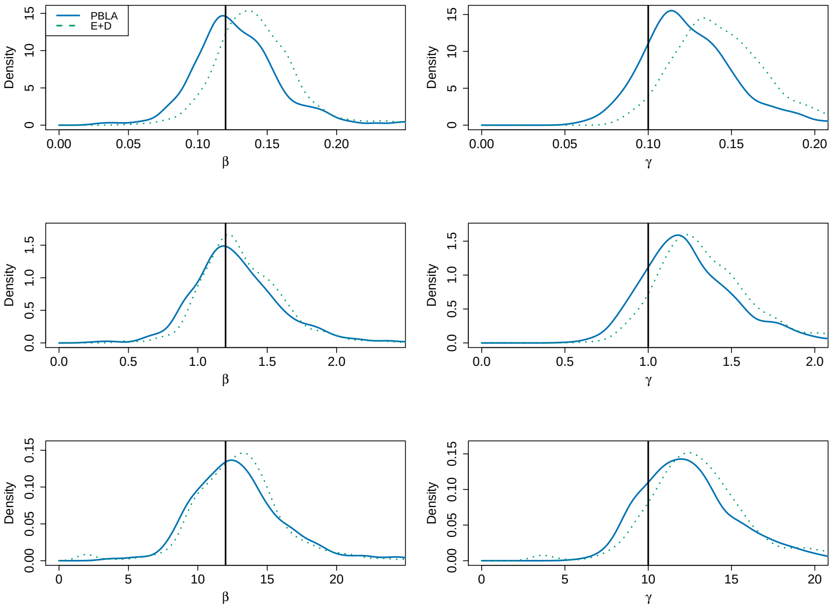

Figure 11: Varying parameters (β, γ): (0.12, 0.1) (top), (1.2, 1) (middle), and (12, 10) (bottom). MLEs from 1000 simulations with Erlang infectious periods (shape $m = 2$), $N = 100$, and $R_0 = \beta(m/\gamma) = 2.4$. Legend in top left graph defines colors and line types for standard PBLA and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

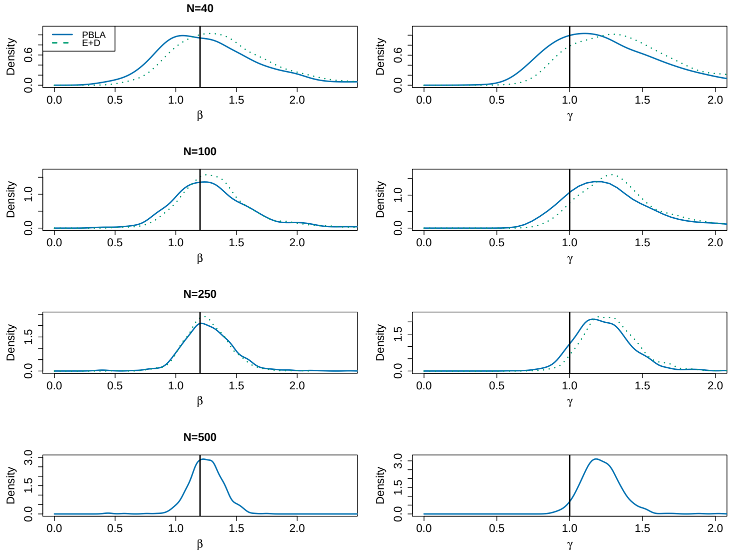

Figure 12: Varying total population size N: 40 (top), 100 (top middle), 250 (bottom middle), and 500 (bottom). MLEs from 500-1000 simulations with Erlang infectious periods ($m = 2$), $(\beta, \gamma) = (1.2, 1)$, and $R_0 = \beta(m/\gamma) = 2.4$. Legend in top left graph defines colors and line types for standard PBLA and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

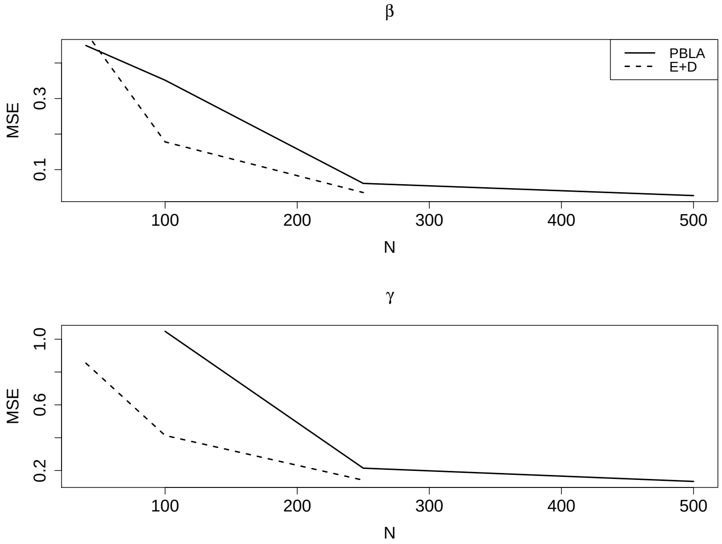

Figure 13: Mean squared error (MSE) for varying total population size N from 500-1000 simulations with with Erlang infectious periods ($m = 2$), $(\beta, \gamma) = (1.2, 1)$, and $R_0 = 2.4$. Legend in left figure defines line types for standard PBLA and the Eichner-Dietz likelihood approximation. (Same simulation study as Figure 12.)

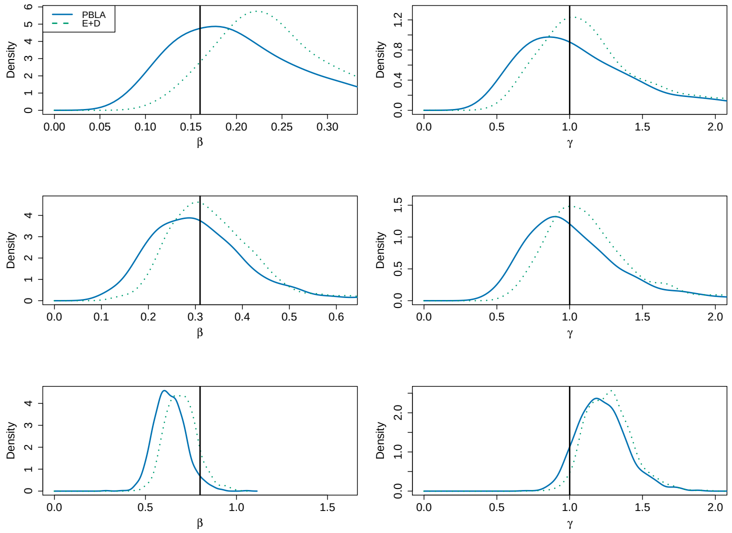

Figure 14: Varying basic reproduction number (R_0, β): (0.8, 0.16) (top), (1.55, 0.31) (middle), and (4, 0.8) (bottom). MLEs from 1000 simulations with Erlang infectious periods $(m = 5)$, $\gamma = 1$, and $N = 80$. Legend in top left graph defines colors and line types for standard PBLA and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

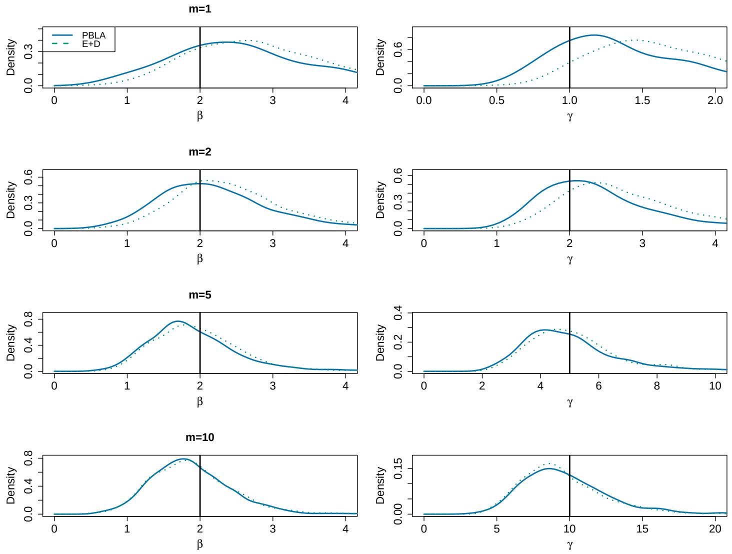

Figure 15: Varying Erlang shape and rate (m, γ): (1, 1) (top), (2, 2) (top middle), (5, 5) (bottom middle), and (10, 10) (bottom). MLEs from 1000 simulations with $(R_0, \beta) = (2, 2)$ and $N = 80$. Legend in top left graph defines colors and line types for standard PBLA and the Eichner-Dietz likelihood approximation. Vertical black line is parameter passed to simulator.

6.4 Computation Time

6.4.1 Markov Chain Monte Carlo

The main motivation for PBLAs is that data augmentation takes a long time for large models. I compared the effective sample size per second (Gelman et al., 2013, pp. 286-287) of a non-centered DAMCMC against MCMC with random walk proposal densities and Hastings ratio based on standard PBLA. For $\beta = 1.5$ and $\gamma = m = 1, 2, 5$, I simulated epidemics with population size $N = 100, 200, 500$ and infected proportion $n/N \approx 0.5$. Next, I initialized my MCMC chains at MLES for β and γ from standard PBLA, and I ran them for 500 iterations. To ensure reasonable mixing and convergence, I looked at trace plots and posterior means for each simulation. After dispensing of a 400 iteration burn-in, I computed the effective sample size using `LaplacesDemon::ESS()` before dividing by the time to generate 100 samples. Vignettes for non-centered DAMCMC and PBLA MCMC are available at sdtemple/pblas. See `scripts/ncda-mcmc-example.R` and `scripts/pbla-mcmc-example.R`.

For both samplers, I assumed independent priors Gamma$(1, 0.0001)$ for β and γ, and I set $\sigma = 0.2$ for all random walk proposals. For the DAMCMC sampler, I considered updating 2 infection times per iteration. There is a trade-off between exploring more infection times and computational performance. In an applied analysis, thoughtful consideration should be put into selecting these hyperparameters.

Figure 16 summarizes the results of this simulation study. In general, the non-centered DAMCMC achieves larger effective sample sizes, but at the cost of increased runtime. For increasing shape m, each standard PBLA takes $O(m^2n^2)$ runtime, affecting the PBLA MCMC computational performance. Non-centered DAMCMC is less affected by increasing m since it only comes in when sampling Gamma random variables. On the other hand, non-centered DAMCMC would slow down considerably if I updated more infection times per iteration. This simulation study ignored inference of β and γ, which I am critical of. Stockdale (2019, pp. 175-180) shows that DAMCMC and PBLA MCMC struggle to infer β and γ in the general SEM, despite inferring the basic reproduction number $R_0 = \beta/\gamma$ reasonably well. For these parsimonious models, β and γ can be highly correlated, making inference challenging for either Bayesian method. Finally, PBLAs may be employed in maximum likelihood estimation, in which case an off-the-shelf optimizer like `nlm()` may converge to a solution after 10-100 approximations. This use of PBLAs requires less computation. For large epidemics, maximum likelihood estimation with PBLAs may be the only viable option.

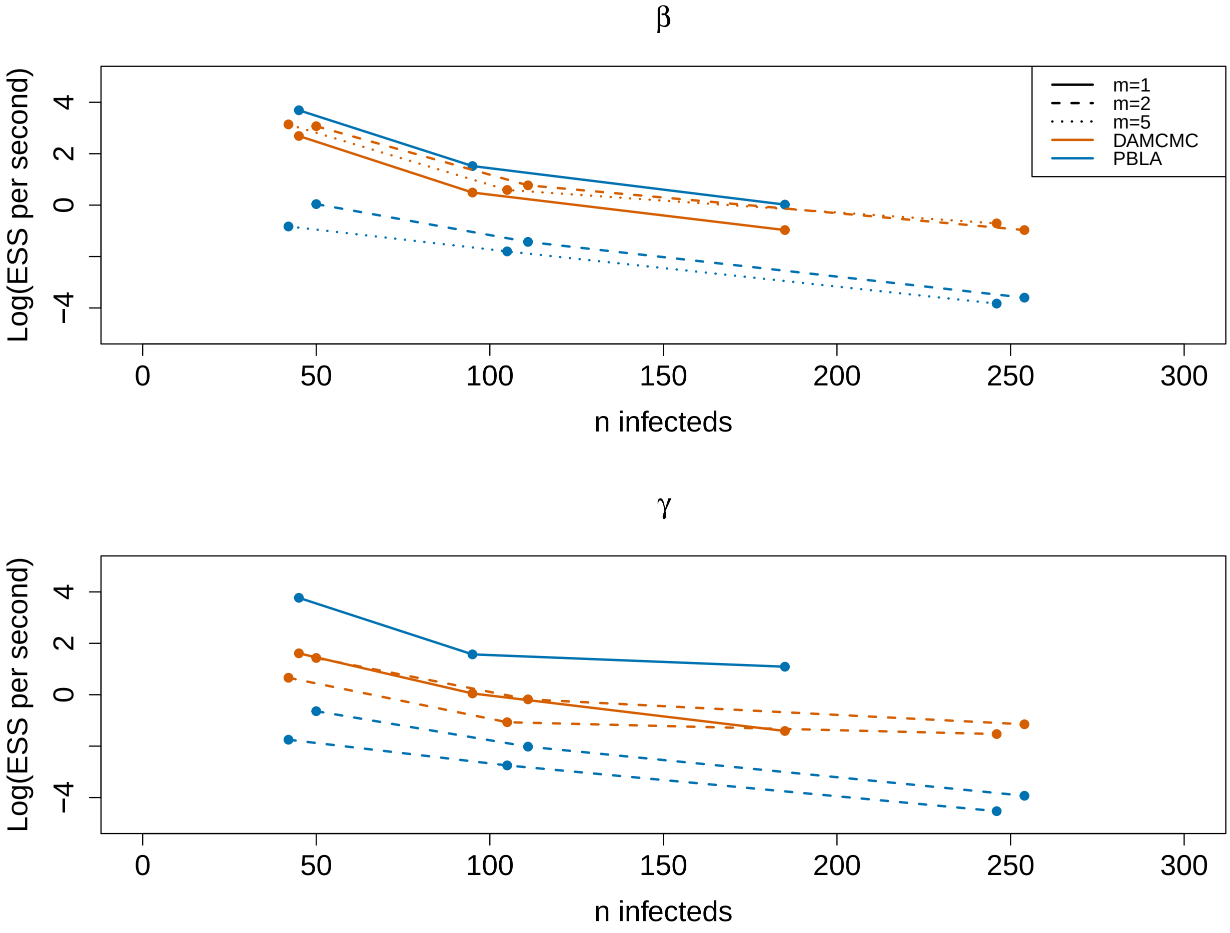

Figure 16: Log effective sample sizes per second for DAMCMC (orange) and PBLA MCMC (blue) methods when $m = 1, 2, 5$ (solid, dashes, dots).

6.4.2 Parallel Computing

I implemented parallel computing for product PBLA using the `R` package `foreach`. Similar to Table 1, I studied compute times for 7 simulated epidemics using 1 core versus 4 cores. Table 3 contains the results of this study. For $n > 5000$, the parallelized method took half as long as the serial method; for smaller epidemic sizes, the overhead associated with parallel computing was not worth it. I only parallelized the product PBLA method, as more elaborate analyses may involve careful configuration of memory access in parallel computing. See `scripts/pbla-parallel.R` and `R/pbla_prod_parallel.R` as references. Finally, Figure 17 makes the quadratic scaling in Table 1 apparent, motivating this parallelization.

Table 3: Time in seconds to compute product PBLA in serial and in parallel. Based on simulated epidemics with increasing n and infected proportion $n/N \approx 0.5$. Superscript * denotes very small, nonzero times.

n	N	Serial	Parallel
95	200	0.02	0.08
185	500	0.00^*	0.06
428	$1,000$	0.01	0.11
$1,483$	$2,500$	0.25	0.40
$2,830$	$5,000$	0.92	0.89
$5,927$	$10,000$	4.13	2.53
$11,819$	$20,000$	18.28	8.20
$29,024$	$50,000$	110.97	56.42

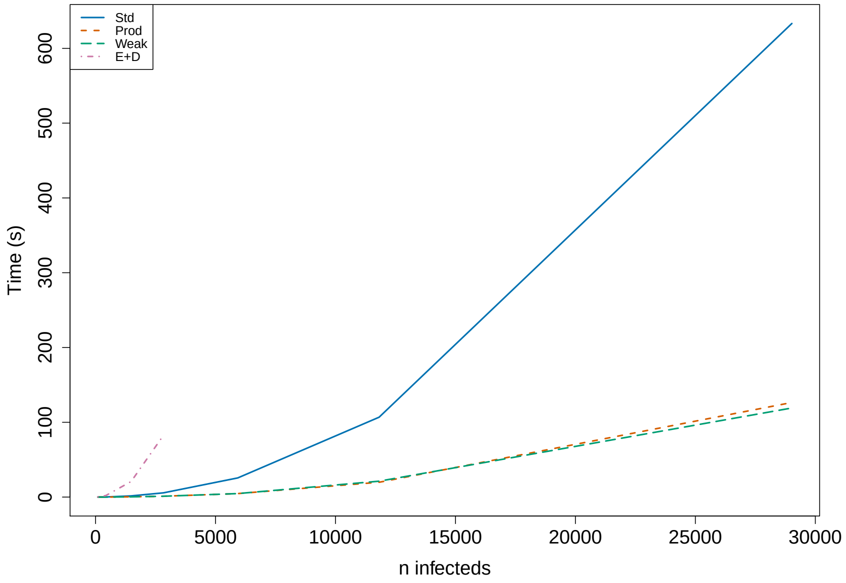

Figure 17: Time in seconds to compute likelihood approximation based on n infecteds. Legend in top left defines colors and line types for standard PBLA, product PBLA, weak PBLA, and the E+D likelihood approximation. Note that product and weak PBLAs have nearly identical computational burden. Standard, product, and weak PBLAs incur quadratic runtime, but standard PBLA has a larger constant.

6.5 Infected Proportion

How the infected proportion influences PBLA inference is discussed in Section 3.1.2. To support a claim made, I simulated epidemics and recorded the likelihood contributions of $\mathbb{E}[\psi_j]$ approximations relative to $\mathbb{E}[\chi_j]$. Since $\mathbb{E}[\psi_j]$ terms are the most aggressively approximated, high relative contributions may indicate when the likelihood approximations are fraught. Figure 18 illustrates that ψ_j terms, those involving infecteds, grow in importance as ever-infecteds make up a larger proportion of the population.

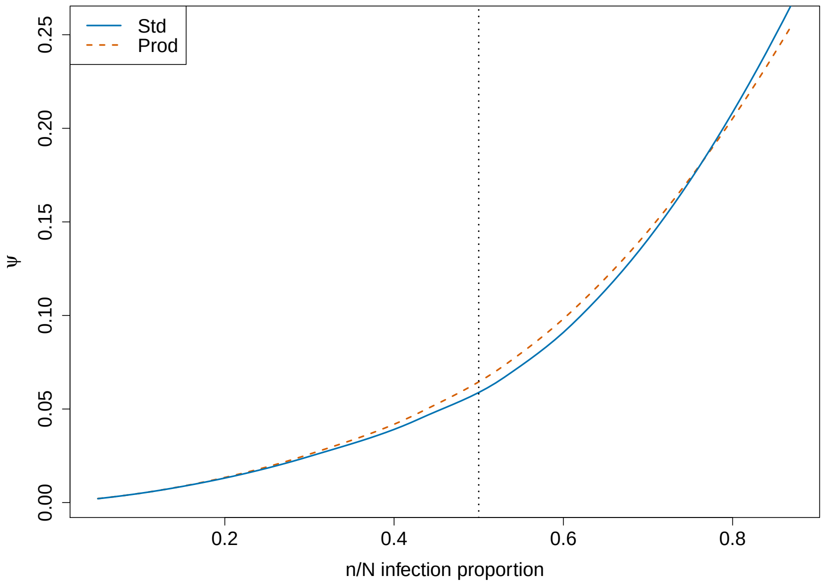

Figure 18: Likelihood contributions of $\mathbb{E}[\psi_j]$ terms relative to $\mathbb{E}[\chi_j]$ terms as n/N increases.

In Section 3.1.2, I compare to MLEs $\hat{\beta}_{\text{complete}}$ and $\hat{\gamma}_{\text{complete}}$. For simulation studies, I have removal times $\mathbf{r}$ and infection times $\mathbf{i}$, using only $\mathbf{r}$ for partial SEM methods like PBLAS. Using $\mathbf{r}$ and $\mathbf{i}$ from a general epidemic, finding $(\hat{\beta}_{\text{complete}}, \hat{\gamma}_{\text{complete}})$ and verifying that it is a global maximum involves elementary calculus.

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

$$\hat{\gamma}_{\text{complete}} = \frac{\sum_{j=1}^{n} (r_j - i_j)}{n}$$

6.6 Underreporting

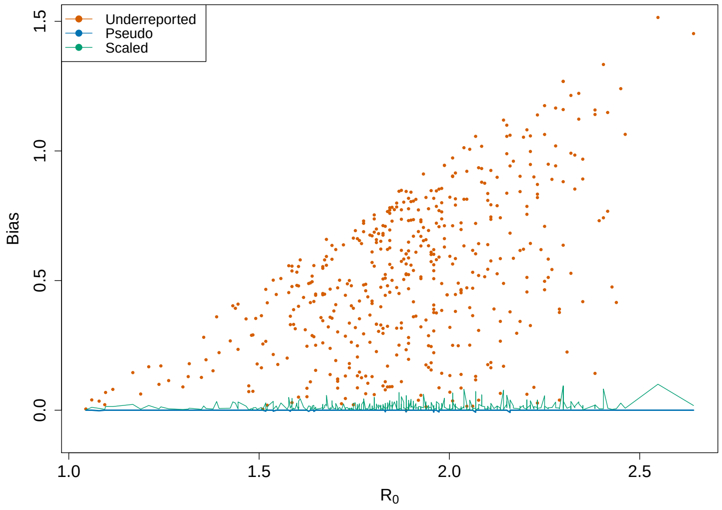

Figure 19: Differences between $\hat{R}_0$ with full partial data versus underreported partial data. Pseudo-removals (blue) and scaling (green) as adjustments recover the original PBLA inference, whereas undercounts decrease estimates.

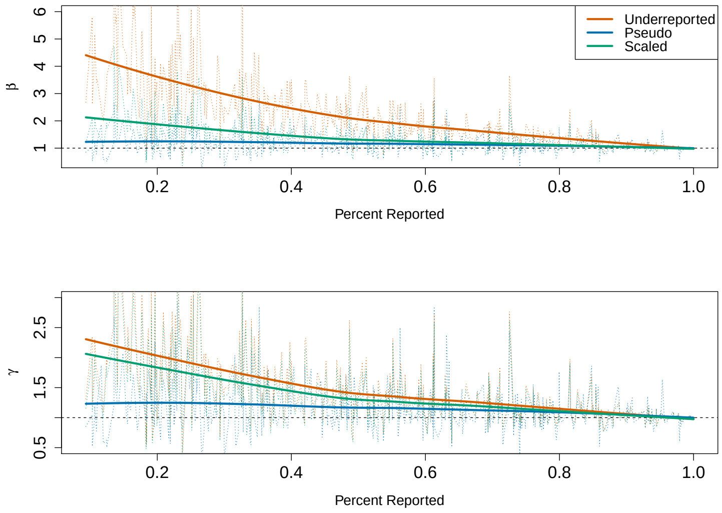

Figure 20: Ratios for β (top) and γ (bottom) estimates under varying reporting rates. Blue for PBLA with underreporting adjustment and orange for PBLA without adjustment. Solid lines for fitted LOESS curves and dotted lines for data observations. Horizontal black line for constant 1. Biases decrease as reporting rate increases.

7 Appendix C

This appendix includes details on the real datasets and their analyses. A multitype SEM for a small epidemic is studied in Section 7.1. The large Ebola virus epidemic in West Africa is discussed in Section 7.2. Commentary on the 2001 foot-and-mouth disease outbreak in the United Kingdom is available in Section 7.3. Lastly, Section 7.4 examines a dog rabies epidemic in Bangui, Central African Republic.

7.1 Common Cold in Tristan da Cunha

This dataset consists of diagnosis times for patients on a remote island in the Atlantic Ocean. I treated the diagnosis times as removal times, assuming patients quarantined after diagnosis. During October and November of 1967, 40 of 255 island residents acquired the common cold. Moreover, the islanders are classified according to three age groups: infants (1), children (2), and adults (3). Total subpopulation sizes are $N_1 = 25$, $N_2 = 36$, and $N_3 = 192$. Infecteds by group are $n_1 = 9$, $n_2 = 6$, and $n_3 = 25$.

Following Hayakawa et al. (2003), I considered a multitype SEM such that $\beta_{jk} = \beta_{G(k)}$, where $G(k)$ denotes the group k belongs to. This model assumes that the population mixes homogeneously, but that susceptibility varies by age group. I conducted a Bayesian analysis using a random walk proposal density for the four model parameters and standard PBLA in the Hastings ratios. Like in Hayakawa et al. (2003), I used independent priors $\gamma \sim \text{Gamma}(10^{-4}, 10^{-3})$ and $\beta_j \sim \text{Gamma}(10^{-8}, 10^{-5})$ for $i = 1, 2, 3$. Lastly, I used the same hyperparameters as Stockdale et al. (2019) for the random walk proposals, and I initialized at $\gamma = 0.5$ and $\beta_i = 0.001$ for $i = 1, 2, 3$.

I ran my MCMC sampler for 10,000 iterations after a 1,000 iteration burn-in. Trace plots in Figure 21 suggest that an even smaller burn-in may be reasonable. As well, I found maximum likelihood estimates for the four model parameters using the `nlm()` optimizer and standard PBLA and the E+D approximation. From the four model parameters, I calculated the basic reproduction number

$$R_0 := \frac{\beta_1 \cdot N_1 + \beta_2 \cdot N_2 + \beta_3 \cdot N_3}{\gamma}$$

I juxtapose my posterior results, the two MLE sets, and results from a Metropolis-within-Gibbs sampler (Hayakawa et al., 2003) in Table 4 and in Figure 22. These methods show near agreement on the model parameters. Most importantly, PBLA is here able to infer parameters in a multitype SEM that match that of the gold standard DAMCMC approach.

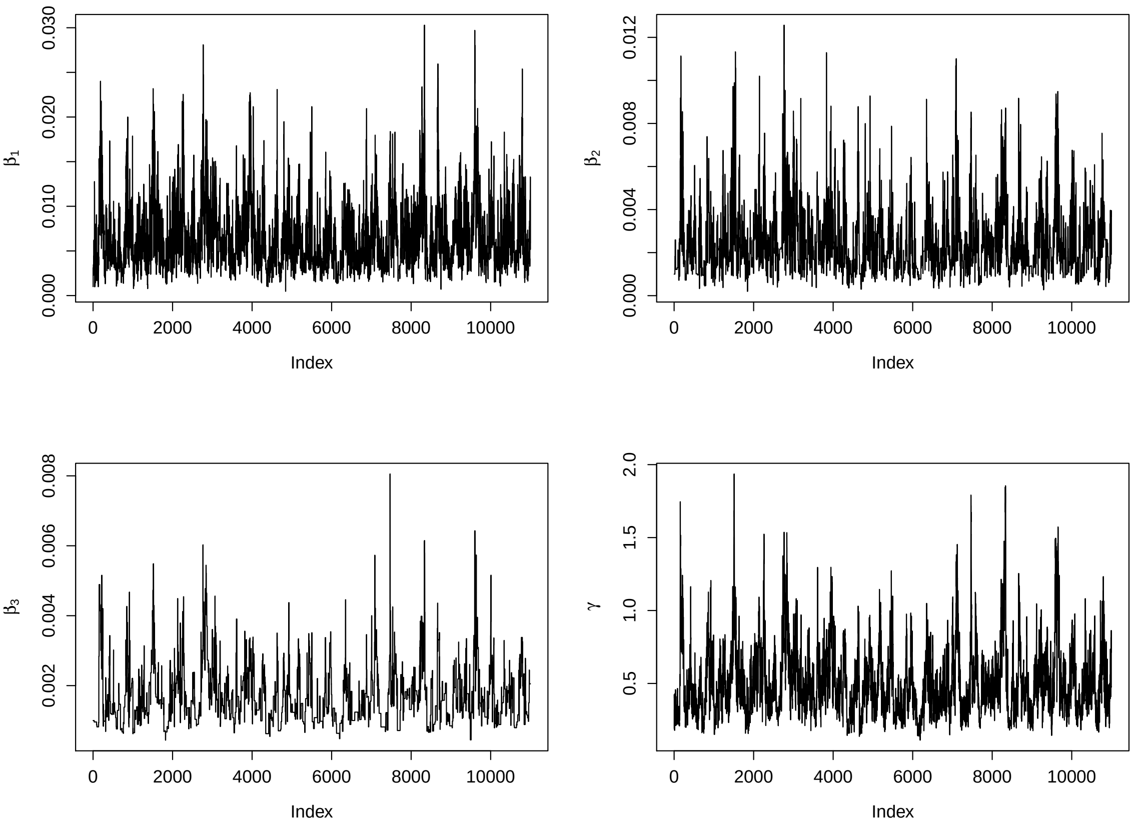

Figure 21: Trace plots of model parameters $(\beta_1, \beta_2, \beta_3, \gamma)$ for the Tristan da Cunha common cold epidemic using PBLA MCMC.

Table 4: Posterior means from PBLA MCMC and DAMCMC methods, and MLEs using the E+D approximation and standard PBLA, for the Tristan da Cunha common cold epidemic.

	PBLA MCMC	DAMCMC	E+D MLE	PBLA MLE
β_1	0.00648	0.00451	0.00568	0.00584
β_2	0.00244	0.00181	0.00224	0.00219
β_3	0.00171	0.00131	0.00166	0.00156
γ	0.50565	0.37100	0.48273	0.45562
R_0	1.17580	1.16102	1.12396	1.15301

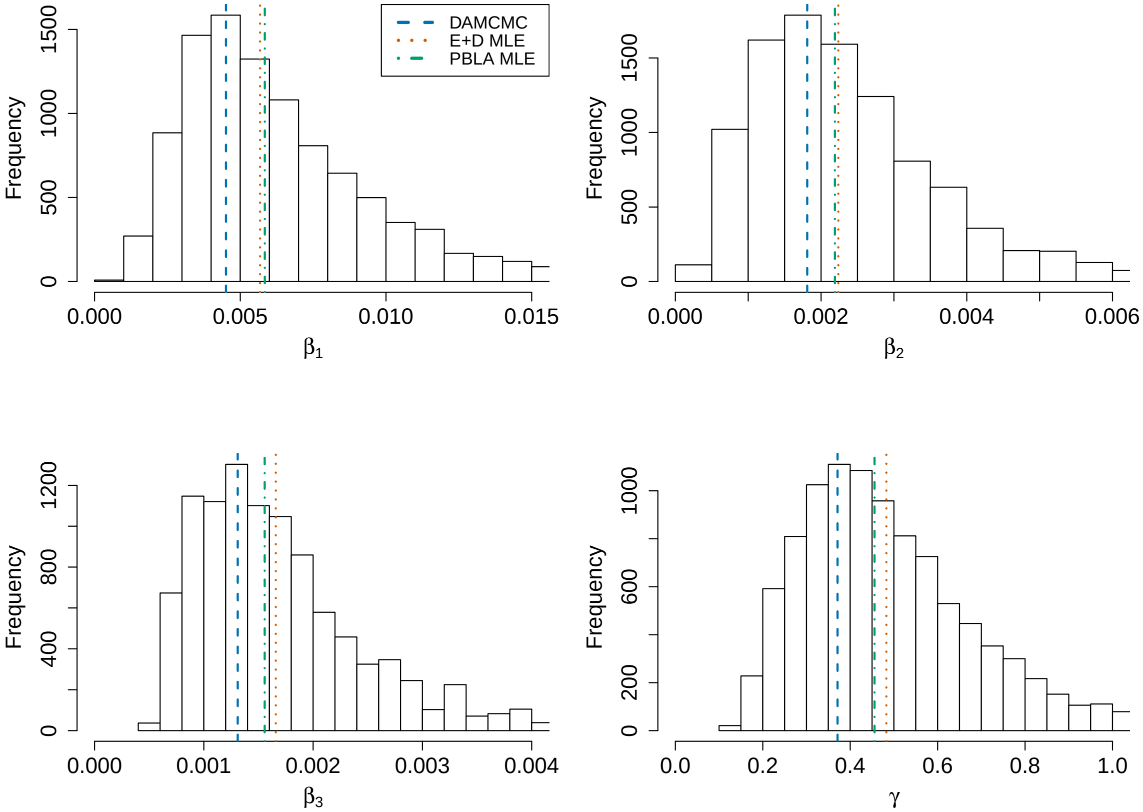

Figure 22: Histograms from $(\beta_1, \beta_2, \beta_3, \gamma)$ posterior samples for the Tristan da Cunha common cold epidemic using PBLA MCMC. For comparison, DAMCMC posterior mean in blue, E+D MLE in orange, and PBLA MLE in green.

7.2 Ebola Virus in West Africa

Death records consist of 2536, 3956, and 4809 removals in Guinea, Sierra Leone, and Liberia. For each epidemic, I assumed a total population size of $N = 1$ million and shifted removal times relative to start times τ_0 from Althaus (2014). Susceptible populations in these African countries are larger than this N, but deaths are still a tiny fraction of this N. Likelihood approximations did not change much for larger N. To accommodate a fixed exposed period, say c, the quantity τ_{kj} becomes $r_k \wedge e_j$ - $i_k \wedge e_j$ where $e_j = i_j - c$. Replacing r_j with $r_j - c$ throughout PBLA formulas makes the SEIR model possible. With this new τ_{kj}, we define the expected midpoint time in which infective pressure is applied as

$$T_{kj} = (\mathbb{E}[r_k \wedge e_j] + \mathbb{E}[i_k \wedge e_j])/2$$

I derived the expressions for T_{jk} by studying exponential periods about times $r_j - c$ or r_k.

$$T_{jk} = \begin{cases} r_k - \frac{1}{2\gamma}, & j > n \\ (r_j - c) - \frac{1}{\gamma} - \frac{1}{4\gamma}\exp(-\gamma(r_k - (r_j - c))), & j, k \leq n,\ r_k > r_j - c \\ r_k - \frac{1}{2\gamma} - \frac{3}{4\gamma}\exp(-\gamma((r_j - c) - r_k)), & j, k \leq n,\ r_k < r_j - c \end{cases}$$

Note that the third case has $-\frac{3}{4\gamma}$, whereas Stockdale et al. (2019) and Stockdale (2019) have the typo $+\frac{3}{4\gamma}$. This typo is not present in the code at jessicastockdale/PBLA and therefore did not affect their Ebola analysis. Below I illustrate the argument for the third case, considering exponential infectious periods shifted so that $r_k = 0$.

$$\begin{aligned} \mathbb{E}[r_k \wedge e_j] &= \mathbb{E}[0 \wedge e_j] = \mathbb{E}[(0 \wedge e_j)(1_{\{e_j<0\}} + 1_{\{e_j>0\}})] \\ &= \mathbb{E}[e_j 1_{\{e_j<0\}}] = \mathbb{E}[e_j | e_j < 0]P(e_j < 0) = -\frac{1}{\gamma}\exp(-\gamma((r_j - c) - r_k)) \\ \mathbb{E}[i_k \wedge e_j] &= \mathbb{E}[(i_k \wedge e_j)(1_{\{e_j<0\}} + 1_{\{e_j>0\}})] = \mathbb{E}[i_k + (i_k \wedge e_j)1_{\{e_j<0\}}] \\ &= \mathbb{E}[i_k] + \mathbb{E}[(i_k \wedge e_j)|e_j < 0]P(e_j < 0) = -\frac{1}{\gamma} - \frac{1}{2\gamma}\exp(-\gamma((r_j - c) - r_k)) \end{aligned}$$

Finally, I provide marginal profile log likelihoods in Figure 23, contour plots for the Sierra Leone and Liberia analyses in Figures 24 and 25. These plots are consistent with results in the main text.

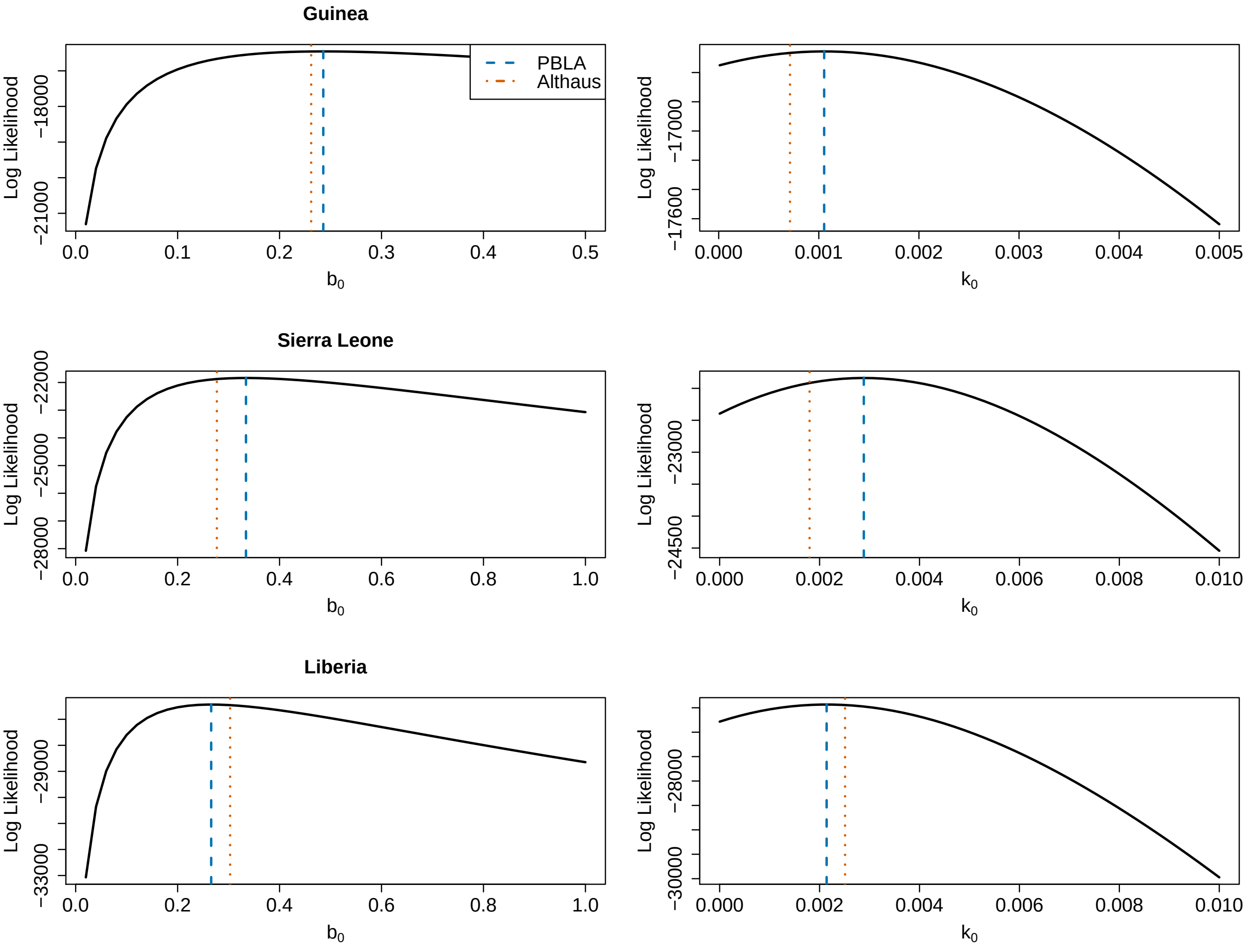

Figure 23: Profile log likelihoods for Ebola virus epidemic in Guinea (top), Sierra Leone (middle), and Liberia (bottom). Graphs on the left study varying β_0 with k_0 fixed at its MLE and vice versa for graphs on the right. Legend in top left defines colors and line types for PBLA and Althaus (2014) MLEs as vertical lines.

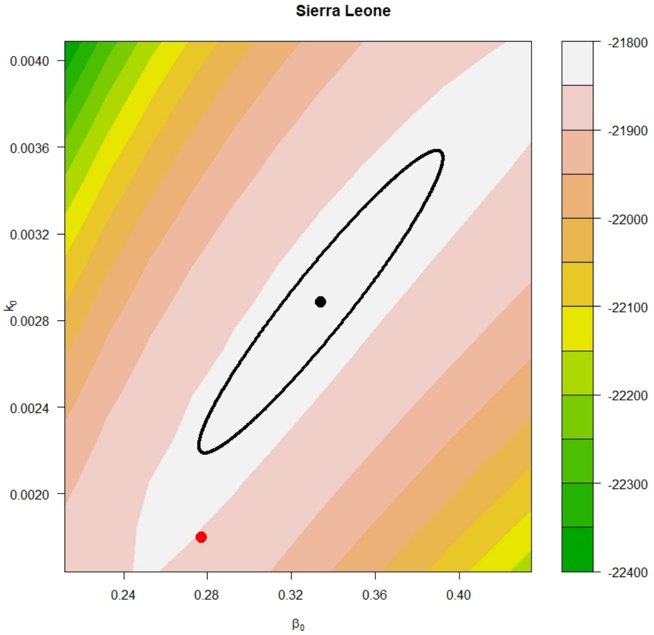

Figure 24: Log likelihood contours for Ebola virus epidemic in Guinea. Ellipses denote confidence regions and colored dots denote MLEs from Stockdale et al. (2019) (black), and Althaus (2014) (red).

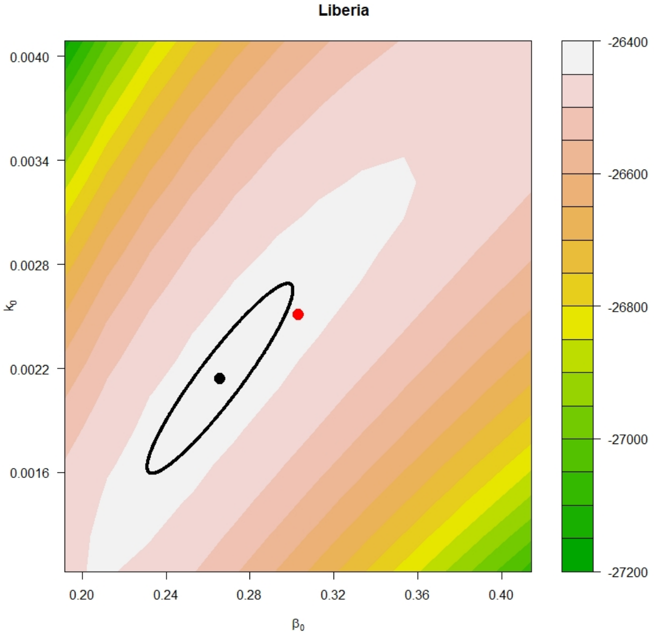

Figure 25: Log likelihood contours for Ebola virus epidemic in Guinea. Ellipses denote confidence regions and colored dots denote MLEs from Stockdale et al. (2019) (black), and Althaus (2014) (red).

7.3 Foot-and-Mouth Disease in United Kingdom

I could not acquire data from the 2001 foot-and-mouth disease (FMD) outbreak in the United Kingdom. This dataset comprises 1021 out of 5378 farms with infected livestock. Kypraios (2007) studied this epidemic with a postulated six parameter SEM where infection rates vary based on Euclidean distances and cattle and sheep counts. Even though this model is spatial and multitype, PBLA can accommodate this via a formula for infection rates. Stockdale et al. (2019) verified this by finding maximum *a posteriori* estimates with PBLA and checking against the DAMCMC results of Kypraios (2007). Since $n = 1021$, less than in the Ebola virus epidemics, this analysis ought to compute in less than an hour.

7.4 Rabies in Central African Republic

Rabies is a zoonotic viral disease endemic among dog populations in Africa where surveillance is subject to undercounts. I applied my pseudo-removals adjustment to infer SEM parameters for an epidemic in Bangui, Central African Republic. I acquired this dataset in the `R` package `outbreaks`. Originally, the data included 151 case dates, ranging from January 2002 to March 2012. There were no cases between January 2005 and May 2006, indicating two separate epidemics. I elected to analyze the larger epidemic comprising 123 cases over seven years. These observations include longitude and latitude coordinate pairs, enabling the consideration of spatially-informed infection rates. Akin to Neal and Roberts (2004), I considered the following spatial formula for infection rates:

$$\beta_{kj} = \beta_0 \exp(-\theta(\rho_{kj})),$$

where ρ_{kj} is the Euclidean distance between case locations. I found the MLE $\hat{\theta}$ to be vanishingly small, and henceforth abandoned the spatial model. Bourhy et al. (2016) likewise found little evidence for transmission dynamics mediated by the proximity of dog subpopulations. Figure 26 illustrates the spatiotemporal distribution of rabies in Bangui. There is no visibly evident spatial pattern. A follow-up study ought to examine space and time, say with $\beta_{kj} = \beta_0 \exp(-\theta(\rho_{kj})) \exp(-k_0 \cdot |r_k - r_j|)$. Moreover, dogs have an abbreviated average lifespan, so the ideal model would involve births and deaths.

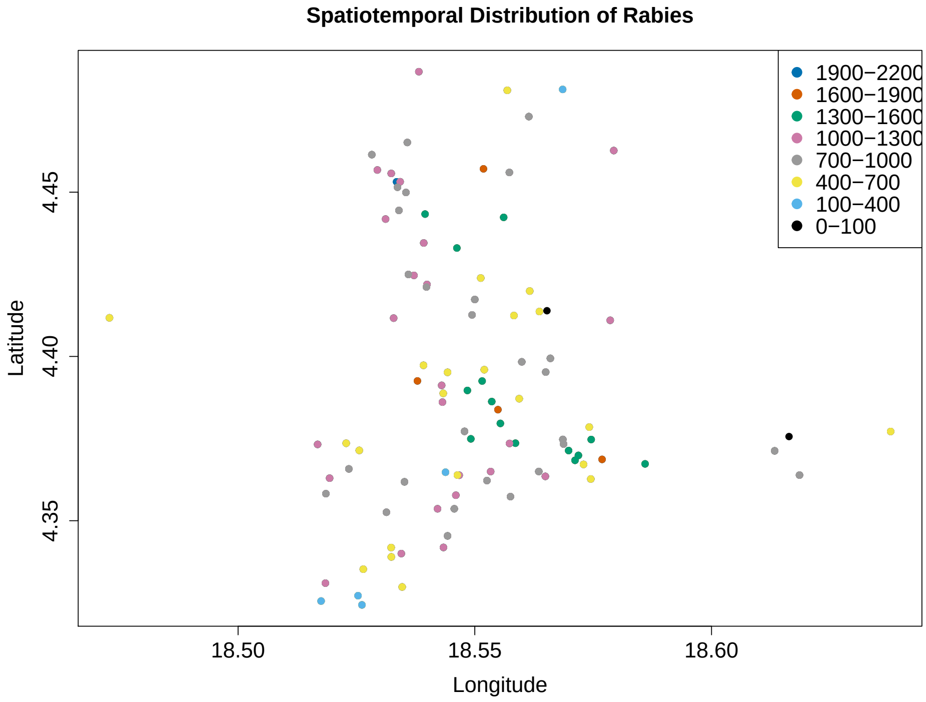

Figure 26: Spatiotemporal distribution of rabies cases in Bangui, Central African Republic. Legend in top right defines colors for time periods in days since May 2006.

Bourhy et al. (2016) and Cori et al. (2018) coupled phylogenetic analyses with mathematical models to characterize case clusters and estimate R_0 around 1. They acknowledged underreporting in the dataset with simplifying assumptions of constant reporting rates between 0.1 and 0.5. Likewise, I conducted a product PBLA analysis assuming reporting rates $\eta = 0.1, 0, 2, 0.5$, and 1 and population size $N = 10000$. Figure 27 visualizes the log likelihood surface in these scenarios. As expected, my adjustment lifted R_0 estimates in accordance with input reporting rates. I also show ranges for R_0 based on elliptical level sets for (β, γ). Since decreasing η corresponds to more pseudo removal times, these ellipses tighten and the ranges narrow. Both Bourhy et al. (2016) and Cori et al. (2018) found evidence of local outbreak die-offs and persistent importation of new infections from the surrounding area, giving more credence to R_0 about 1. Finally, I assessed the impact of the total population size N on my inference. Table 5 shares my main analysis for $N = 10000, 25000, 50000$, and 100000. For N increasing, the underreporting adjustment is less pronounced, as even 123 removal times and 1000 pseudo-removal times is a tiny fraction of N.

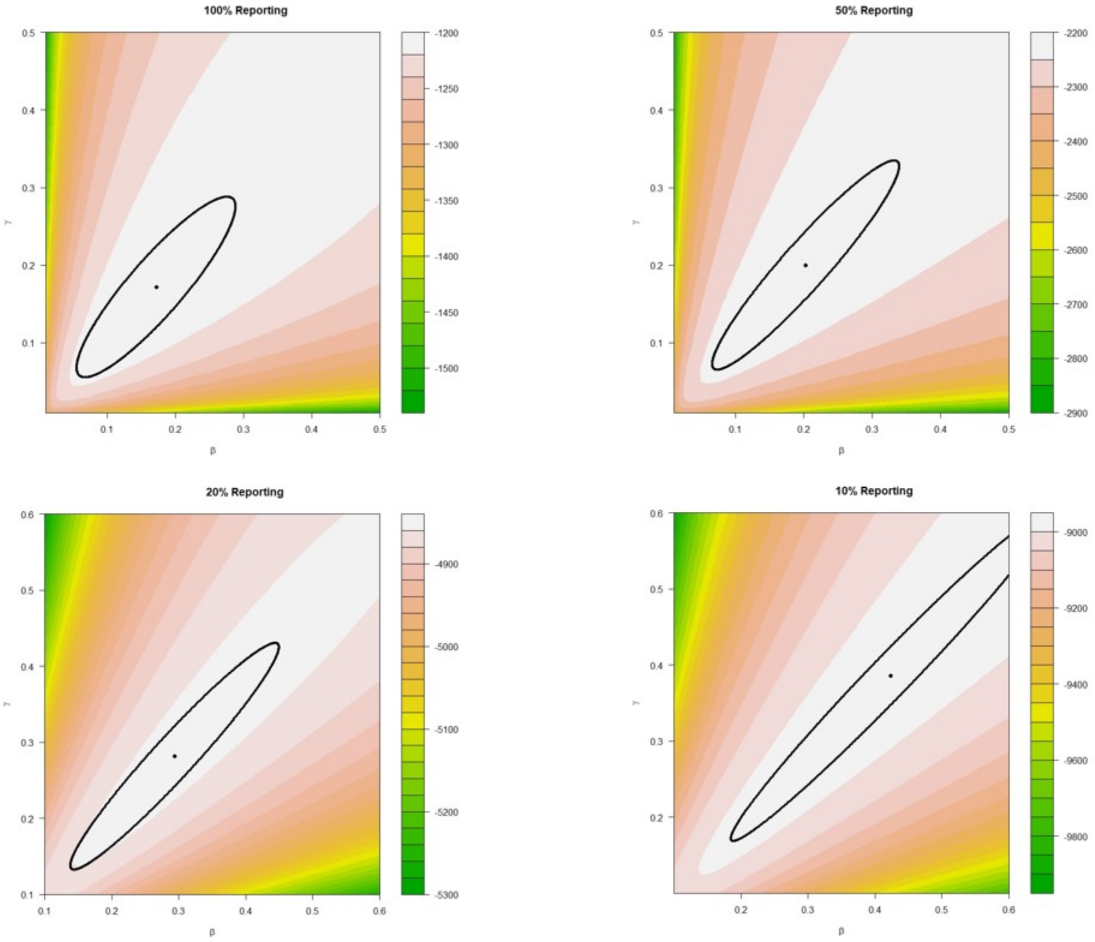

Figure 27: Log likelihood contours for rabies virus epidemic in Bangui, Central African Republic, assuming reporting rates η: 1 (top left), 0.5 (top right), 0.2 (bottom left), and 0.1 (bottom right). Ellipses denote confidence regions and dots denote MLEs. $N = 10,000$.

Table 5: Inferences on (β, γ, R_0) using product PBLA, pseudo-removals adjustment, and increasing total population size N for rabies epidemic in Bangui, Central African Republic.

N	η	β	γ	R_0	Interval R_0
10,000	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)
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)