Bayesian inference of reproduction number using epidemic and genomic data

Alicia Gill, joint work with Xavier Didelot, Richard Everitt and Jere Koskela

Department of Statistics, University of Warwick, Coventry, UK

Alicia.Gill@warwick.ac.uk

Introduction

Suppose that we wish to find the reproduction number $R(t)$ of an epidemic. Typically, only epidemic data is used to infer $R(t)$. However, epidemic data is often noisy, partially observed or biased. Genomic data is therefore increasingly being used to understand infectious disease epidemiology. The aim of this work is to incorporate both epidemic and genomic information into a joint model to infer $R(t)$. The epidemic data considered will be prevalence, which is defined as the total number of cases per day. However, we are unlikely to have access to complete prevalence data. Instead, we will have observed prevalence which may be noisy and/or incomplete. The genomic data will take the form of a dated phylogenetic tree.

Epidemic process

We model the epidemic as a non-homogeneous linear birth-death process with unknown birth rate $\beta(t)$ and known constant death rate γ.

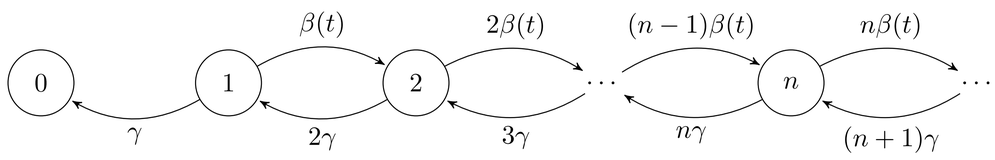

Figure 1: Birth-death process

In a birth-death model of disease outbreak, $R(t) = \beta(t)/\gamma$.

State-space model (SSM)

The epidemic is observed discretely on days $1, \ldots, M$, so we discretise $\beta(t)$, the phylogeny G and the latent epidemic E to fit into a state-space model framework.

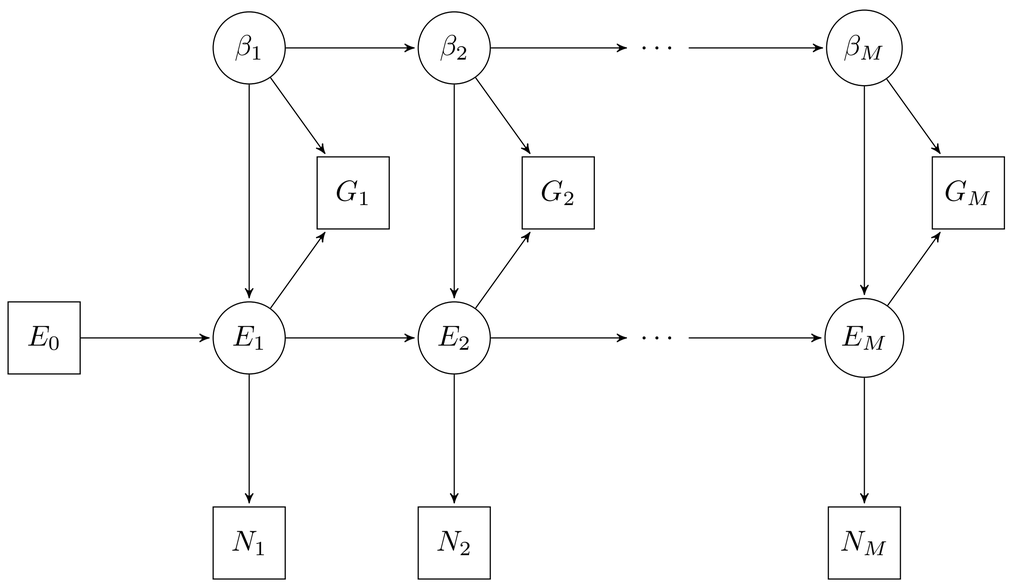

Figure 2: SSM showing relationships between birth rates $\beta_{1:M}$, the latent epidemic $E_{0:M}$, the observed epidemic $N_{1:M}$ and the dated phylogeny $G_{1:M}$.

We use a Bayesian approach to find the birth rate trajectory β and hyper-parameters θ given the known death rate γ, the observed epidemic N and the dated phylogeny G:

$$\begin{aligned} p(\beta, \theta \mid \gamma, N, G) &\propto p(\beta, \theta) p(\gamma, N, G \mid \beta, \theta) \\ &= p(\theta) p(\beta \mid \theta) \int p(E \mid \beta, \gamma) p(N \mid E, \theta) p(G \mid \beta, E)\, dE. \end{aligned} \tag{1}$$

Bayesian modelling

Model hyper-parameters are $\theta = (b, \sigma, p)$ where b is the birth rate on day 1, σ is the standard deviation of the change in birth rate between days and p is the proportion of true cases observed. Priors used are:

- $b \sim \text{Exponential}(1)$
- $\sigma \sim \text{Exponential}(10)$
- $p \sim \text{Uniform}(0, 1)$
- $\beta_1 \mid b \sim \text{Uniform}(0, b)$
- $\beta_m \mid \beta_{m-1}, \sigma \sim \text{Normal}(\beta_{m-1}, \sigma^2)$ with reflection off 0

Likelihoods used are:

- $E_m \mid \beta_m, \gamma, E_{m-1} = x_{m-1} \sim \text{Skellam}(\beta_m x_{m-1}, \gamma x_{m-1})$
- $G_m \mid \beta_m, E_m = x_m \sim \text{Binomial}\left(\binom{a_m}{2}, 1 - \exp(2\beta_m / x_m)\right)$, where a_m is the number of lineages and c_m is the number of coalescences on day m
- $N_m \mid E_m = x_m, p \sim \text{Binomial}(x_m, p)$

Particle-marginal Metropolis–Hastings (PMMH)

We have implemented a PMMH algorithm [1] to target $p(\beta, \theta \mid \gamma, N, G)$. θ^* is proposed according to a multivariate Normal distribution. We then run a sequential Monte Carlo (SMC) algorithm with adaptive resampling [2] to get an unbiased estimator for $p(\beta \mid \theta) \int p(E \mid \beta, \gamma) p(N \mid E, \theta) p(G \mid \beta, E)\, dE$. Within the SMC, we propose $\beta_m^{1:K}$ according to its prior and we propose $E_m^{1:K}$ using a negative binomial distribution if $N_m > 0$ and according to its prior otherwise. The estimate of the marginal likelihood $\hat{p}_{\theta^*}(\beta, \gamma, N, G)$ as well as one birth rate and prevalence trajectory β^*, E^* are carried into the random walk Metropolis–Hastings. We have implemented backward simulation [3] to choose the single trajectories β^* and E^* to keep in order to avoid the problem of path degeneracy. θ^*, β^* and E^* are jointly accepted with probability:

$$1 \wedge \frac{p(\theta^*) \hat{p}_{\theta^*}(\beta^*, \gamma, N, G)}{p(\theta) \hat{p}_{\theta}(\beta, \gamma, N, G)}.$$

Simulations

We have simulated a 30-day epidemic with constant $R(t) = 3$. The death rate $\gamma = 0.1$, so $\beta(t) = 0.3$. 5% of the true prevalence is observed each day and 5% of the cases on day 30 are used to generate the phylogeny. We have applied our method using only epidemic data and using both epidemic and genomic data to see whether genomic data provides any benefit. The chains were both initialised at $b_0 = 1$, $\sigma_0 = 0.01$ and $p_0 = 0.5$ and run for 100,000 iterations. The number of particles used was 100.

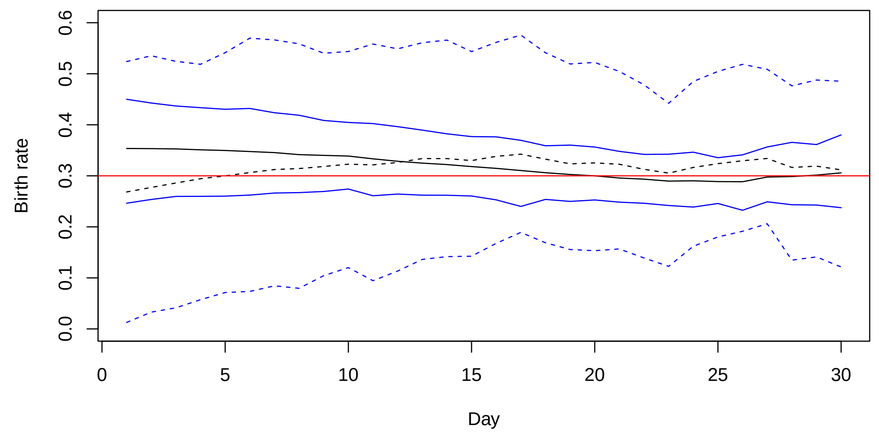

Figure 3: Posterior birth rate. Black lines denote the posterior mean. Blue lines denote the posterior 95% credible intervals. The red line denotes the true birth rate trajectory. Solid black/blue lines used epidemic and genomic data. Dashed blue/black lines used only epidemic data.

It can be seen from Figure 3 that whilst epidemic data alone provides a good estimate of the posterior mean, incorporating the genomic data has significantly reduced the uncertainty around this estimate. This is done with minimal increase to the run time; the chain using only epidemic data ran in 76 minutes and the chain using epidemic and genomic data ran in 80 minutes.

References

(1) C. Andrieu, A. Doucet et al., *J. Royal Stat. Soc. Ser. B*, 2010, **72**, 269–342.

(2) A. Doucet and A. Johansen, The Oxford Handbook of Nonlinear Filtering, ed. D. Crisan and B. Rozovskii, Oxford University Press, 2011, ch. 24.

(3) F. Lindsten and T. B. Schön, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 3845–3848.