

Data augmentation in the continuous-time SIR model

July 2014

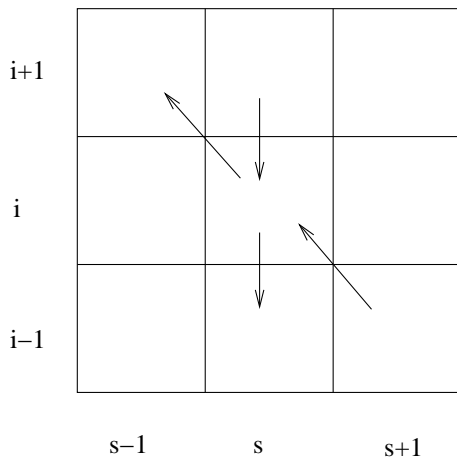
Outline

- ▶ The general epidemic model
 - ▶ A simple Susceptible–Infected–Removed (SIR) model of an outbreak of infection in a closed population
- ▶ Poisson likelihood for infection and removal rates
 - ▶ Complete data: both infection and removal times are observed
 - ▶ Under Gamma priors for the infection and removal rates, their full conditionals are also Gamma, so Gibbs updating steps can be used
- ▶ Incomplete data: only removal times are observed
 - ▶ Augment the unknown infection times
 - ▶ Additional Metropolis-Hastings steps for sampling infection times, requiring explicit computation of the complete data likelihood

The SIR model

- ▶ Consider a closed population of M individuals
- ▶ One introductory case (infective) introduces the infection into a population of initially susceptible individuals, starting an outbreak
- ▶ Once the outbreak has started, the hazard of infection for a still susceptible individual depends on the number of infectives in the population: $(\beta/M)I(t)$
- ▶ If an individual becomes infected, the hazard of clearing infection (and stopping being infective) is γ , i.e., he/she remains infective for an exponentially distributed period of time. He/she then becomes *removed* and does not contribute to the outbreak any more
- ▶ There is no latency

Transitions in the state space



The complete data

- ▶ Assume one introductory case whose infection takes place at time $t = 0$ (i.e. this fixes the time origin)
- ▶ For M individuals followed from time 0 until the end of the outbreak at time T (after which time the number of infectives $I(t) = 0$), the *complete data* record all event times
- ▶ This is equivalent to observing $n - 1$ infection times and n removal times, and the fact the $M - n$ individuals escaped infection throughout the outbreak

$$\overbrace{\{0 = i_1 < i_2 < \dots < i_n\}}^{\text{infection times}} \quad \text{and} \quad \overbrace{\{r_1 < \dots < r_{n-1} < r_n = T\}}^{\text{removal times}}$$

- ▶ N.B. Here, the i_k and r_k need not correspond to the same individual

Counting infectives and susceptibles

- ▶ Denote the ordered event times i_1, \dots, i_n and r_1, \dots, r_n jointly as $0 = u_1 < u_2 < \dots < u_{2n} = T$
- ▶ Denote the indicators of time u_k being an infection or removal time by D_k and R_k , respectively
- ▶ Denote the number of infectives at time t by $I(t)$
 - ▶ it is a piecewise constant (left-continuous) function, assuming values in the set $\{0, 1, \dots, M\}$
 - ▶ it jumps at times $u_2 < \dots < u_{2n}$
- ▶ Denote the number of susceptibles at time t by $S(t)$
 - ▶ it is a piecewise constant (left-continuous) function, jumping at times $i_2 < \dots < i_n$
- ▶ Both $I(t)$ and $S(t)$ are determined by the complete data

The process of infections

- ▶ The model of new infections is a non-homogeneous Poisson process with rate $\beta I(t)S(t)/M$
 - ▶ the rate is a piecewise constant (left-continuous) function
 - ▶ it jumps at times $u_2 < \dots < u_{2n}$, with levels $\beta I(u_2)S(u_2)/M, \beta I(u_3)S(u_3)/M, \dots, \beta I(u_{2n})S(u_{2n})/M$
- ▶ The probability density of the infection events is thus proportional to

$$\prod_{k=2}^{2n} \left[((\beta/M)I(u_k)S(u_k))^{D_k} \exp^{-(\beta/M)I(u_k)S(u_k)(u_k - u_{k-1})} \right]$$

total time for "infectious pressure"

$$\propto \prod_{k=2}^{2n} (\beta I(u_k)S(u_k))^{D_k} \times \exp^{-\underbrace{(\beta/M) \sum_{k=2}^{2n} I(u_k)S(u_k)(u_k - u_{k-1})}_{\text{total time for "infectious pressure"}}$$

The process of removals

- ▶ The model of removals is a non-homogeneous Poisson process with rate $\gamma I(t)$
 - ▶ the rate is a piecewise constant (left-continuous) function
 - ▶ it jumps at times $u_2 < \dots < u_{2n}$, with levels $\gamma I(u_2), \gamma I(u_3), \dots, \gamma I(u_{2n})$
- ▶ The probability density of the removal events is thus proportional to

$$\begin{aligned} & \prod_{k=2}^{2n} \left[(\gamma I(u_k))^{R_k} \exp^{-\gamma I(u_k)(u_k - u_{k-1})} \right] \\ & = \prod_{k=2}^{2n} (\gamma I(u_k))^{R_k} \times \exp \underbrace{-\gamma \sum_{k=2}^{2n} I(u_k)(u_k - u_{k-1})}_{\text{total time spent infective}} \end{aligned}$$

The complete data likelihood

- ▶ The joint likelihood of parameters β and γ , based on the complete data:

$$\begin{aligned} \overbrace{L(\beta, \gamma; \mathbf{i}, \mathbf{r})}^{f(\mathbf{i}, \mathbf{r} | \beta, \gamma)} &= \prod_{k=2}^{2n} (\beta I(u_k) S(u_k))^{D_k} \prod_{k=2}^{2n} (\gamma I(u_k))^{R_k} \\ &\times \exp^{-\sum_{k=2}^{2n} ((\beta/M)I(u_k)S(u_k) + \gamma I(u_k))(u_k - u_{k-1})} \\ &= \prod_{k=2}^n \{\beta I(i_k) S(i_k)\} \prod_{k=1}^n \{\gamma I(r_k)\} \\ &\times \exp^{-\sum_{k=2}^{2n} ((\beta/M)I(u_k)S(u_k) + \gamma I(u_k))(u_k - u_{k-1})} \end{aligned}$$

Simplifying the notation

- ▶ Note that $\sum_k I(u_k)S(u_k)(u_k - u_{k-1}) = \int_0^T I(u)S(u)du$
- ▶ Similarly $\sum_k I(u_k)(u_k - u_{k-1}) = \int_0^T I(u)du$
- ▶ The likelihood function can thus be written as

$$\prod_{k=2}^n \{\beta I(i_k)S(i_k)\} \prod_{k=1}^n \{\gamma I(r_k)\} \\ \times \exp\left(-\int_0^T \{(\beta/M)I(u)S(u) + \gamma I(u)\}du\right)$$

Poisson likelihood and Gamma priors

- ▶ This above likelihood is the so called Poisson likelihood for parameters β and γ
- ▶ In particular, Gamma distributions can be used as conjugate priors for β and γ
- ▶ It follows that the full conditional distributions of β and γ are also Gamma and can be updated by Gibbs steps

Gamma prior distributions

- ▶ Rate parameters β and γ are given independent Gamma priors

$$f(\beta) \propto \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta)$$

$$f(\gamma) \propto \gamma^{\nu_\gamma - 1} \exp(-\lambda_\gamma \gamma)$$

- ▶ This allows easy updating of these parameters using Gibbs sampling (the next two pages)

The full conditional of β

- ▶ Parameter β can be updated through a Gibbs step

$$f(\beta|\mathbf{i}, \mathbf{r}, \gamma) \propto f(\beta, \gamma, \mathbf{i}, \mathbf{r}) \propto f(\mathbf{i}, \mathbf{r}|\beta, \gamma)f(\beta)$$

$$\propto \beta^{n-1} \exp\left(-(\beta/M) \int_0^T I(u)S(u)du\right) \beta^{\nu_\beta-1} \exp(-\lambda_\beta\beta)$$

- ▶ This means that

$$\beta|\mathbf{i}, \mathbf{r}, \gamma \sim \Gamma\left(n-1 + \nu_\beta, (1/M) \int_0^T I(u)S(u)du + \lambda_\beta\right)$$

The full conditional of γ

- ▶ Parameter γ can be updated through a Gibbs step:

$$f(\gamma|\mathbf{i}, \mathbf{r}, \beta) \propto f(\beta, \gamma, \mathbf{i}, \mathbf{r}) \propto f(\mathbf{i}, \mathbf{r}|\beta, \gamma)f(\gamma)$$

$$\propto \gamma^n \exp\left(-\gamma \int_0^T l(u)du\right) \gamma^{\nu_\gamma-1} \exp(-\lambda_\gamma \gamma)$$

- ▶ This means that

$$\gamma|(\mathbf{i}, \mathbf{r}, \beta) \sim \Gamma\left(n + \nu_\gamma, \int_0^T l(u)du + \lambda_\gamma\right)$$

Computation of the integral terms

- ▶ In practice, the integral terms can be calculated as follows:

total time spent infective

$$\int_0^T I(u) du = \sum_{k=1}^n (r_k - i_k)$$

total time for "infectious pressure"

$$\int_0^T I(u) S(u) du = \sum_{k=1}^n \sum_{j=1}^M (\min(r_k, i_j) - \min(i_k, i_j))$$

where $i_j = \infty$ for $j > n$, i.e., for those never infected

- ▶ These expressions are invariant to choice of which r_k corresponds to which i_k

Incomplete data

- ▶ Assume that only the removal times $\mathbf{r} = (r_1, \dots, r_n)$ have been observed
- ▶ Augment the set of unknowns (β and γ) with infection times $\mathbf{i} = (i_2, \dots, i_n)$
- ▶ The aim is to do statistical inference about rates β and γ (and times \mathbf{i}), based on their posterior distribution $f(\beta, \gamma, \mathbf{i} | \mathbf{r})$
- ▶ The posterior distribution is proportional to the joint distribution of all model quantities:

$$f(\beta, \gamma, \mathbf{i} | \mathbf{r}) \propto f(\beta, \gamma, \mathbf{i}, \mathbf{r}) = \overbrace{f(\mathbf{i}, \mathbf{r} | \beta, \gamma)}^{\text{complete data likelihood}} \overbrace{f(\beta)f(\gamma)}^{\text{prior}},$$

Updating infection times

- ▶ The full conditional distributions of β and γ are as above
- ▶ The unknown infection times require a Metropolis–Hastings step, including explicit evaluations of the Poisson likelihood
- ▶ If the current iterate of i_k is $i_k^{(j)}$, a new value \tilde{i}_k is first proposed (e.g.) from a uniform distribution on $[0, T]$
- ▶ The proposal is then accepted, i.e., $i_k^{(j+1)} := \tilde{i}_k$, with probability

$$\min\left\{1, \frac{f(\tilde{\mathbf{i}}, \mathbf{r} | \beta, \gamma)}{f(\mathbf{i}, \mathbf{r} | \beta, \gamma)}\right\}$$

- ▶ Here $\tilde{\mathbf{i}}$ is \mathbf{i} except for the k th entry which is \tilde{i}_k (instead of $i_k^{(j)}$)

Augmenting individual histories

- ▶ The likelihood above was constructed for the aggregate processes, i.e., to count the total numbers of susceptibles and infectives
- ▶ In such case, the corresponding augmentation model must not consider individuals
 - ▶ In particular, times i_2, \dots, i_n must not be tied to particular removal times, i.e., individual event histories must not be reconstructed
- ▶ If one considers individual event histories as pairs of times (i_k, r_k) for individuals $k = 1, \dots, M$, the appropriate complete data likelihood is (cf. above)

$$\gamma^n \prod_{k=2}^n \{\beta I(i_k)\} \exp\left(-\int_0^T (\gamma I(u) + (\beta/M)I(u)S(u))du\right)$$

Example: a smallpox outbreak

- ▶ The Abakaliki smallpox outbreak
 - ▶ A village of $M = 120$ inhabitants
 - ▶ One introductory case
 - ▶ 29 subsequent cases; this means that $n = 1 + 29 = 30$
- ▶ The observations are given as time intervals between detection of cases (removals) (0 means that symptoms occurred at the same day):

13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5

- ▶ The problem: to estimate rates β and γ from these outbreak data
- ▶ See the computer class exercise

References

- [1] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, 1999; 162: 121–129.

- [2] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.

- [3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

- [4] Andersen et al. Statistical models based on counting processes. Springer Verlag, New York, 1993.