Module 11: Lesson 3 Lecture Notes

Callum Arnold

20 July, 2021

Contents

Non-Markov SIR epidemic model

- Each infectious individual remains so for length of time T_I drawn from $f(x|\vec{\theta})$
  - Markov model draws from an exponential distribution, so is a special case
  - Common choices are Gamma and Weibull distribution for non-Markov model
    - Both flexible to allow modelling real-life infectious period i.e. can model mean and variance separately, unlike exponential
  - For non-Exponential T_I, $\{(S(t), I(t)) : t \geq 0\}$ is not a Markov process
- Infectious contact occur with each susceptible according to a Poisson process of rate β/N
  - Overall infection rate is $\frac{\beta S(t) I(t)}{N}$
- Inference problem is the same as Markov
  - Try to estimate model parameters β, θ
  - Find/sample the posterior density $\pi(\beta, \theta | r_1, r_2, ..., r_n)$ where r_i refers to the removal times observed
- Generally the likelihood $\pi(r_1, r_2, ..., r_n | \beta, \theta)$ is hard to compute
  - Tractable in the special case of constant infectious period as if you know r_k, you know i_k (given you know the infectious period)

  - Introduce the infection times as extra variables to use augmented likelihood

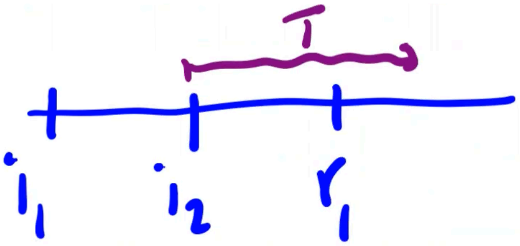
- Define:
  - $\vec{r} = r_1, ..., r_n$
  - $\vec{i} = i_1, ..., i_{a-1}, i_{a+1}, ..., i_n$
    - Note that it does not include i_a, which is the time of infection for the first infected individual a
- Let $f(x|\theta)$ denote the probability density function of the infectious period distribution

Augmented likelihood

- The augmented likelihood is

$$\pi(\vec{i}, \vec{r}|\beta, \theta, i_a, a) = \underbrace{\prod_{j \neq a} \frac{\beta}{N} I(i_j -) \times \exp\left(\frac{\beta}{N} \int S(t) I(t) dt\right)}_{\text{how likely that individuals get infected at inf times observed}} \times \underbrace{\prod_{1 \leq j \leq n} f(r_j - i_j|\theta)}_{\text{how likely to observe removals}}$$

- Can derive this intuitively using simulation
  - For an infectious individual (say the 1st), generate the infectious time $(r_1 - i_1) \sim T_i, f(x|\theta)$ by sampling from the distribution
    - This gives you the removal time
  - Now, what's the likelihood of observing this time?
    - $f(r_1 - i_1|\theta)$
  - Generate the next infection, which is a Poisson process, and therefore the waiting times follow an exponential distribution ($\text{Exp}(\lambda) = \lambda e^{-\lambda x}$)
    - Our infection rate is $\frac{\beta S(t) I(t)}{N}$
    - $\frac{\beta S(t) I(t)}{N} e^{-\frac{\beta S(t) I(t)}{N}(i_2 - i_1)}$
  - We only keep the first part of the exponential term for individuals who infect others before they are removed $(i_2 - i_1) < (r_1 - i_1)$
    - Results in $\prod_{j \neq a} \frac{\beta}{N} I(i_j -)$
  - If the next infection generated is after the next removal scheduled, this happens with probability



    - $e^{-\frac{\beta}{N} SI(r_1 - i_2)}$, which is $P(T > r_1 - i_2)$
  - The exponential terms result in an integral when we extend to all individuals

Target posterior

- The target posterior density
  - $\pi(\beta, \theta, \vec{i}, i_a, a|\vec{r}) \propto \pi(\vec{i}, \vec{r}|\beta, \theta, i_a, a)\pi(\beta, \theta, i_a, a)$
- Set independent priors as:
  - $\beta \sim \Gamma(m_\beta, \lambda_\beta)$
  - $a \sim \mathrm{U}[1, n]$
  - $i_a \sim \mathrm{U}[\infty, r_1]$
  - $\theta \sim$???
    - Depends on what θ is!
- $\beta, \vec{i}, i_a$ can be updated as for the Markov model (Gibbs step for β, M-H otherwise)
- Updates for θ depends on what θ is

Debugging tips

- Test each piece of code separately
  - Most MCMC algorithms involve various components e.g. Gibbs update, M-H update
  - Check that impossible situations get 0 likelihood!
- Validate output using simulations
  - E.g. Simulate model M times,
    - Run MCMC on each output to infer parameters
    - Average parameter estimates from MCMC should be close to known true values
  - If MCMC code time-consuming, e.g. spatial where likelihood has to incorporate distance from each infected to each susceptibles
    - Use one large outbreak that should give reasonable information about the model parameters
- Beware 0s
  - E.g. 0/0 without reporting error
    - Can happen in M-H acceptance ratio with the likelihoods
- Try a very small data set
  - Can work out required inference by hand and check against MCMC output
- Use log likelihood
  - Likelihood often require calculation of products which can lead to numerical instabilities and run-time errors
  - R has built in functions `lgamma(k) = log(gamma(k))`

What to do with MCMC output

- Marginal summaries
  - Look at 1-D aspects of a parameter
  - Useful to plot the marginal posterior density/histogram of each parameter

- Joint summaries
    - Assess the extent to which β and γ can be estimated separately
    - Scatter plots and contour plots common way to visualize
    - Can compute correlation statistics
    - Can help design MCMC algorithm
        - When very strong correlation, indication data doesn't allow separate estimation
- Functions of model parameters
    - R_0 is a common function of parameters
        - $R_0 = \beta E(T_I) = \frac{\beta}{\gamma}$
        - Can use MCMC output to create a new file containing $(\beta_1/\gamma_1), (\beta_2/\gamma_2), ..., (\beta_M/\gamma_M)$
            - Samples from the posterior density of R_0
    - Translate inference for **rates** into inference for **probabilities**
        - e.g. $1 - \exp(-\beta/N)$ is the probability that one infective individual infects a given susceptible in one time unit

What can be estimated

- Think about how informative the data are about the model parameters of interest
- Can I feasibly estimate the parameters I care about?
    - Sometimes not obvious

Latent periods

- Think about Markov SEIR with fixed latent period c days
- Introduce exposure times $e_k = i_k - c$
- Define:
    - $\vec{e} = e_1, e_{a-1}, e_{a+1}, ..., e_n$
    - $\vec{r} = r_1, ..., r_n$
    - $\vec{i} = i_1, ..., i_n$
        - Note that the exposed time vector now misses the original infected individual, and the infection time vector contains all individuals as we're now more concerned when the first person was exposed
- The augmented likelihood:
    - Same as the usual Markov SIR augmented likelihood with the addition of the indicator function
        - Given we only need to one of i_k and e_k to know the other (given a constant c), the indicator function just insists we have a constant c

$$\pi(\vec{e}, \vec{i}, \vec{r} | \beta, \gamma, i_a, a) = \prod_{j \neq a} \frac{\beta}{N} I(i_j-) \times \exp\left(-\frac{\beta}{N} \int S(t) I(t) dt\right) \times \gamma^n \exp\left(-\gamma \sum (r_j - i_j)\right) \times 1_{\{e_k - i_k = c, k=1,\ldots,n\}}$$

- We can update the MCMC algorithm to include c as an extra parameter (M-H updates for c), but it would be uninformative about c given removal data alone
  - For one data point (r_k) you're trying to estimate 2 parameters (either e_k or i_k, and c)
  - End up with strong positive posterior correlation between c and β
    - Nothing in data to say that you could have very long infection periods with very small amount of infection rate (or visa versa)
- Instead, we fix c to reasonable values then perform estimation of β and γ

Gamma infectious periods

- Instead of exponentially distributed infectious periods, have gamma distributed
- Each infective remains so for a period of time T_I, where $T_I \sim \Gamma(c, d)$
  - $E(T_I) = c/d$ where c is the shape, and d is the rate
- The augmented likelihood is:

$$\pi(\vec{i}, \vec{r} | \beta, c, d, i_a, a) = \prod_{j \neq a} \frac{\beta}{N} I(i_j -) \times \exp\left(\frac{\beta}{N} \int S(t) I(t) dt\right) \times \prod_{1 \leq j \leq n} f(r_j - i_j | c, d)$$

$$\text{where: } f(x|c, d) = x^{c-1} d^c \frac{e^{-dx}}{\Gamma(c)}$$

- Not immediately obvious if it is possible to esimate both parameters separately from the removal data
- Might expect that the mean infectious period ($E(T_I)$) be estimated with reasonable precision
  - Doesn't tell you anything about c and d as it's a ratio
- Might want to use a Gamma distribution parameterised by mean and variance rather than shape and rate
  - Equations are less pleasant, particularly $f(x|c, d)$

Data for Markov SIR model

- Estimation depends on the detail of the data
- For example, if $n = 0$, no inference to be made as no one becomes infected
- What value of n is the most informative?

Discussion

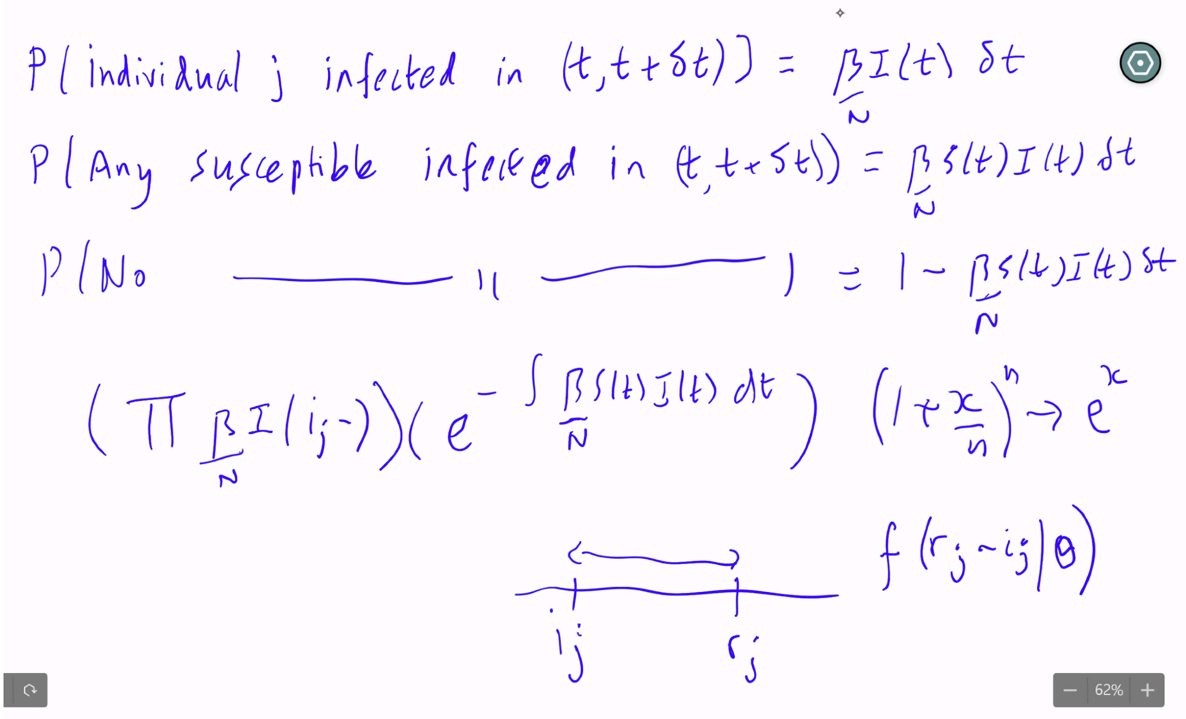
When do you decide to use mean and variance vs shape and rate parameters (in Gamma infectious periods)

- Not always obvious what's easy to estimate
- Worth writing down the math/think about what's going to happen, **before coding**

Deriving the augmented likelihood

- In video think about simulating and drawing from exponential distributions
  - Put together and you get the likelihood

- exponential integral based on Taylor expansion in the limit as $\delta t \to 0$



$$P(\text{individual } j \text{ infected in } (t, t+\delta t)) = \frac{\beta}{N} I(t)\, \delta t$$

$$P(\text{Any susceptible infected in } (t, t+\delta t)) = \frac{\beta}{N} S(t) I(t)\, \delta t$$

$$P(\text{No} \;\rule{3em}{0.4pt}\; '' \;\rule{3em}{0.4pt}\;) = 1 - \frac{\beta}{N} S(t) I(t)\, \delta t$$

$$\left(\prod \frac{\beta}{N} I(i_j -) \right) \left(e^{-\int \frac{\beta}{N} S(t) I(t)\, dt} \right) \qquad \left(1 + \frac{x}{n}\right)^n \to e^x$$

$$f(r_j - i_j \mid \theta)$$

-