SIS models for recurrent infections

SISMID/July 17–19, 2023

Instructors: Volodymyr Minin, Kari Auranen, Elizabeth Halloran

Outline

- Recurrent infections
- A simple Susceptible–Infected–Susceptible (SIS) model without transmission
  - Complete-data likelihood
- Modeling transmission
- Incomplete observations
  - Continuous-time Markov processes with Bayesian data augmentation and reversible jump MCMC
- A computer class exercise of an SIS model without transmission and with completely observed data

Background

- Many infections can be considered recurrent, i.e., occurring as an alternating series of presence and absence of infection
  - Nasopharyngeal carriage of *Streptococcus pneumoniae* (Auranen et al.; Cauchemez et al.; Melegaro et al.)
  - Nasopharyngeal carriage of *Neisseria meningitidis*
  - multi-resistant *Staphylococcus aureus* (Cooper et al.)
  - HPV (human papilloma virus) infection
  - some parasitic infections (e.g. Nagelkerke et al.)
- Many of the above infections are asymptomatic, which means that observation requires active sampling to record the current epidemiological states of the study subjects
- Exact acquisition and clearance times of infection often remain unobserved $\Rightarrow$ incompletely observed data

A binary Markov process

A simple model for a recurrent infection is the binary Markov process:

- The state of the individual alternates between "susceptible" (state 0) and "infected" (state 1)
- For a susceptible individual, the rate of acquiring infection is β:

  $$\mathrm{P}(\text{acquisition in } [t, t + dt[\,|\text{ susceptible at time } t-) \simeq \beta dt$$

- For an infected individual, the rate of clearing infection is μ:

  $$\mathrm{P}(\text{clearance in } [t, t + dt[\,|\text{infected at time } t-) \simeq \mu dt$$

Complete data

- For each individual i, the complete data include the times of acquisition and clearance during the observation period $[0, T]$:
    - Denote the ordered acquisition times of individual i during $]0, T[$ by $\boldsymbol{t}^{(i)} = (t_{i1}, \ldots, t_{iN_{01}^{(i)}})$
    - Denote the ordered clearance times of individual i during $]0, T[$ by $\boldsymbol{r}^{(i)} = (r_{i1}, \ldots, r_{iN_{10}^{(i)}})$
    - Denote the ordered sequence of all acquisition and clearance times of individual i as $u_{i1} = 0, u_{i2}, u_{i3}, \ldots, u_{i,N^{(i)}} = T$
        - Note: these include times 0 and T, so that the total number of observation times for individual i is $N^{(i)} = N_{01}^{(i)} + N_{10}^{(i)} + 2$

Keeping track who is susceptible

- The binary (“yes/no”) indicators for individual i to be susceptible or infected at time t are denoted by $Y_0^{(i)}(t)$ and $Y_1^{(i)}(t)$, respectively
    - For the simple binary model, $Y_1^{(i)}(t) = 1 - Y_0^{(i)}(t)$ for all times $t \geq 0$, i.e. the individual is always either susceptible or infected
    - Both indicators are taken to be *predictable*, i.e., their values at time t are determined by their initial values and the complete data observed up to time $t-$ (i.e. time just before t)
    - In practice, this means that the values of $Y_0^{(i)}(t)$ and $Y_1^{(i)}(t)$ can be calculated from the observed data and these indicators can be easily used as shorthand when writing the likelihood function
    - Note also that the indicators $Y_0^{(i)}(t)$ and $Y_1^{(i)}(t)$ are defined such that they denote the state of susceptibility or infection *just before* time t (e.g. for someone who gets infected at time t, the indicator $Y_0^{(i)}(t) = 1$)

Process of acquisitions

- For each individual i, acquisitions (i.e. new infections) occur with rate $\beta Y_0^{(i)}(t)$
  - The rate is β when the individual is in state 0 (susceptible) and 0 when the individual is in state 1 (infected)
- The probability density of the acquisition events of individual i is

$$\prod_{k=1}^{N^{(i)}} \left[\beta^{1(u_k \text{ is a time of acq.})} e^{-\beta Y_0^{(i)}(u_k)(u_k - u_{k-1})}\right]$$

$$\propto \beta^{N_{01}^{(i)}} \times \exp\{-\beta \times \overbrace{\sum_{k=1}^{N^{(i)}} Y_0^{(i)}(u_k)(u_k - u_{k-1})}^{\text{total time spent susceptible for ind. } i}\} \qquad (1)$$

- $N_{01}^{(i)}$ is the total number of infections that occur for individual i during the study period

Process of clearances

- For each individual i, clearances of infection occur with rate $\mu Y_1^{(i)}(t)$
  - The rate is μ when the individual is in state 1 (infected) and 0 when then individual is in state 0 (susceptible)
- The probability density of the clearance events of individual i is

$$\prod_{k=1}^{N^{(i)}} \left[\mu^{1(u_k \text{ is a time of clearance})} e^{-\mu Y_1^{(i)}(u_k)(u_k - u_{k-1})} \right]$$

$$= \mu^{N_{10}^{(i)}} \times \exp\{-\mu \times \overbrace{\sum_{k=1}^{N^{(i)}} Y_1^{(i)}(u_k)(u_k - u_{k-1})}^{\text{total time infected of ind. } i}\} \qquad (2)$$

- $N_{10}^{(i)}$ is the total number of clearances that occur for individual i during the study period

Complete data likelihood

- The contribution to the likelihood function of parameters β and μ, based on the complete data from individual i, is obtained by multiplying the likehood expressions (1) and (2):

$$\begin{aligned}
&\overbrace{L_i(\beta, \mu; \boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)})}^{f(\boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)}|\beta, \mu)} \\
&= \beta^{N_{01}^{(i)}} \mu^{N_{10}^{(i)}} \times e^{-\sum_{k=1}^{N^{(i)}} (\beta Y_0^{(i)}(u_k) + \mu Y_1^{(i)}(u_k))(u_k - u_{k-1})} \\
&= \beta^{N_{01}^{(i)}} \mu^{N_{10}^{(i)}} \times \exp\left(- \int_0^T \{\beta Y_0^{(i)}(u) + \mu Y_1^{(i)}(u)\} du \right)
\end{aligned}$$

- The likelihood based on *all* M individuals is a product over individual likelihood contributions: $\prod_{i=1}^{M} L_i(\beta, \mu; \boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)})$

Modeling transmission

- The rate of infection may depend on the presence of infected individuals in the family, day care group, school class etc.
    - The statistical unit is then determined by the relevant mixing group
- Let $H_t^{(i,fam)}$ denote the joint infection status of all members in the mixing group (e.g. family) of individual i at time $t-$
- For a single-type pathogen, the rate of infections can now be modeled as follows:

$$\mathrm{P}(\text{infection for } i \text{ in } [t, t+dt[|H_{t-}^{(i,\text{fam})}) \simeq \alpha_{01}^{(i)}(t) Y_0^{(i)}(t)dt \equiv \frac{\beta C^{(i)}(t)}{M_{\text{fam}}^{(i)} - 1} Y_0^{(i)}(t)dt$$

where $C^{(i)}(t)$ is the number of infected individuals in i's family (of size $M_{fam}^{(i)}$) at time $t-$; note that $C^{(i)}(t)$ can be calculated from the state-indicator variables of the family members

Complete data likelihood: the general expression

- For M individuals followed from time 0 to time T, the *complete data* record all transitions between states 0 and 1:

$$x_{\text{complete}} = \{T_{sr}^{(ik)};\ s, r = 0, 1\ (s \neq r),\ k = 1, \dots, N_{sr}^{(i)}(T),\ i = 1, \dots, M\}$$

- The likelihood of the rate parameters $\theta = (\beta, \gamma)$, based on the complete data, is

$$\overbrace{L(\theta; x_{\text{complete}})}^{f(x_{\text{complete}}|\theta)} = \prod_{i}^{N} \prod_{r \neq s} \prod_{k}^{N_{sr}^{(i)}(T)} \left[\alpha_{sr}^{(i)}(T_{sr}^{(ik)}) \times \exp\left(-\int_0^T \alpha_{sr}^{(i)}(u) Y_s^{(i)}(u) du \right) \right]$$

Remarks

- Although the likelihood expressions above were constructed as a product of individual likelihood contributions, they are valid even when the individual processes are dependent on the infection outcomes of *other* individuals (as when modeling transmission)
- The likelihood is correctly normalized with respect to any number of events occurring between times 0 and T
  - This is crucial when performing MCMC computations through data augmentation with an unknown number of events
- These results are somewhat non-trivial and require the theory of counting processes (Andersen et al.)

Incomplete observations

- Usually we do not observe complete data (= all infection and clearance times for each study subject)
- Instead, the status (infection stage) $X_j^{(i)}$ of each individual is observed only at some pre-defined times $t_j^{(i)}$
    - This creates *incomplete data*: the process is only observed at discrete times (panel data)
    - The observed data likelihood is now a complicated function of the model parameters
- How to estimate the rate parameters of the underlying continuous process from discrete observations?
    - We can apply a similar approach that we already saw in the SIR model with incomplete observations
    - This means that the unknown event times are treated as additional model unknowns (data augmentation)
    - Another option would be to discretize the model (see e.g. Melegaro et al.)

Bayesian data augmentation

- If we retain the continuous-time model formulation, unobserved event times of acquisition and clearance can be taken as additional model unknowns (parameters)
- Statistical inference is performed on all model unknowns (parameters θ and event times x_{complete}), based on the joint probability model of all model quantities:

$$\overbrace{f(x_{\text{observed}}|x_{\text{complete}})}^{\text{observation model}} \; \overbrace{f(x_{\text{complete}}|\theta)}^{\text{complete data likelihood}} \; \overbrace{f(\theta)}^{\text{prior}}$$

- The observed data $x_{observed}$ contain only the current status of infection in each study subject at the predefined observation times
- The model of the observations model ensures agreement with the observed data; in practice, this part of the model is based on simple indicator functions for the agreement
- A specific computational problem:
  how to sample from $f(x_{\text{complete}}|x_{\text{observed}}, \theta)$?

Sampling algorithm

- Initialize the model parameters and the latent processes (i.e. the unobserved event times)
- For each individual, update the unobserved event times
    - Update the current iterates of the event times using standard MH
    - Add/delete episodes of infection and non-infection using reversible jump MH
        - with 0.5 probability propose to add a new episode
        - with 0.5 probability propose to delete an existing episode
- Update the model parameters using single-step MH (or when taking Gamma priors for the rate parameters, a Gibbs step can often be applied due to the Poisson-type complete-data likelihood)
- Iterate the above updating steps for a given number of MCMC iterations

Adding/deleting episodes

- Choose one interval at random from among the K sampling intervals (see page+2)
- Choose to add an episode (delete an existing episode) within the chosen interval with probability $\pi_{\text{add}} = 0.5$ ($\pi_{\text{delete}} = 0.5$)
  - If 'add', choose random event times $\bar{t}_1 < \bar{t}_2$ uniformly from Δ ($=$ the length of the sampling interval). These define the new episode.
  - If 'delete', delete the two event times
- The 'add' move is accepted with probability (Metropolis-Hastings acceptance ratio)

$$\min\left(\frac{f(x_{\text{observed}}|x^*_{\text{complete}})f(x^*_{\text{complete}}|\theta)q(x_{\text{complete}}|x^*_{\text{complete}})}{f(x_{\text{observed}}|x_{\text{complete}})f(x_{\text{complete}}|\theta)q(x^*_{\text{complete}}|x_{\text{complete}})}, 1\right)$$

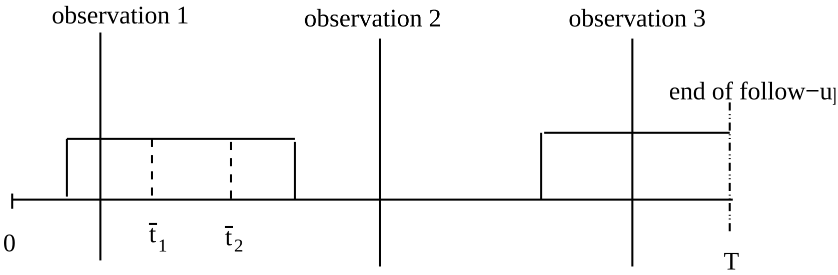
Adding/deleting episodes cont.

- The ratio of the proposal densities is

$$\frac{q(x_{\text{complete}}|x^*_{\text{complete}})}{q(x^*_{\text{complete}}|x_{\text{complete}})} = \frac{\pi_{\text{delete}}\frac{1}{K}\frac{1}{L}}{\pi_{\text{add}}\frac{1}{K}\frac{1}{L}\frac{2}{\Delta^2}} = \frac{\Delta^2}{2}$$

- The ratio of the proposal densities in the 'delete' move is the inverse of the expression above
- Technically, the add/delete step relies on so called reversible jump MCMC (see page+2)
- Reversible jump types should be devised to assure irreducibility of the Markov chain
- For a more complex example, see e.g. Hoti et al.

Adding/deleting latent processes cont.



The number of sampling intervals K= 4
The number of ’sub−episodes’ within the second interval L = 2

Reversible jump MCMC

- “When the number of things you don’t know is one of the things you don’t know”
- For example, under incomplete observation of the previous (Markov) processes, the exact number of events is not observed
- This requires a joint model over ’sub-spaces’ of different dimensions
- And a method to do numerical integration (MCMC sampling) in the joint state space

References

[1] Andersen et al. "Statistical models based on counting processes", Springer, 1993
[2] Auranen et al. "Transmission of pneumococcal carriage in families – a latent Markov process model for binary data. J Am Stat Assoc 2000; 95:1044-1053.
[3] Melegaro et al. Estimating the transmission parameters of pneumococcal carriage in families. Epidemiol Infect 2004; 132:433-441.
[4] Cauchemez et al. Streptococcus pneumoniae transmission according to inclusion in cojugate vaccines: Bayesian analysis of a longitudinal follow-up in schools. BMC Infectious Diseases 2006, 6:14.
[5] Nakelkerke et al. Estimation of parasitic infection dynamics when detectability is imperfect. Stat Med 1990; 9:1211-1219.
[6] Cooper et al. "An augmented data method for the analysis of nosocomial infection data. Am J Epidemiol 2004; 168:548-557.
[7] Bladt et al. "Statistical inference for disceretly observed Markov jump processes. J R Statist Soc B 2005; 67:395-410.
[8] Andersen et al. Multi-state models for event history analysis. Stat Meth Med Res 2002; 11:91-115.
[9] Hoti et al. Outbreaks of Streptococcus pneumoniae carriage in day care cohorts in Finland – implications to elimination of carriage. BMC Infectious Diseases, 2009 (in press)
[10] Green P. Reversible jump Markov chain Monte Carlo computation and Bayesianmodel determination. Biometrika 1995; 82:711-732.