MVE550 2020 Lecture 8.1
Compendium Chapter 4
Bayesian inference for Branching processes

Petter Mostad

Chalmers University

November 22, 2020

Bayesian inference for Branching processes

- Say you have observed some data, and you want to find a branching process (of the type discussed in Dobrow) that appropriately models the data, to then make predictions. How?
- A branching process is characterized by the probability vector $a = (a_0, a_1, a_2, \dots,)$ where a_i is the probabilty for i offspring in the offspring process.
- Let $y_1, y_2, \dots, y_n$ be the counts of offspring in n observations of the offspring process. If a is given we have the likelihood

$$\pi(y_1, \dots, y_n \mid a) = \prod_{i=1}^{n} a_{y_i}$$

- To complete the model, we need a prior on a. However, it is tricky to define a prior on the infinite-dimensional set of possible vectors a.
- So, instead we will consider two situations where we make additional assumptions on the offspring process.

Using a Binomial likelihood

- Assume the offspring process is Binomial(N, p) for some parameter p and a fixed known N. We get the likelihood

$$\pi(y_1, \ldots, y_n \mid p) = \prod_{i=1}^{n} \text{Binomial}(y_i; N, p).$$

- A possibility is to use a prior $p \sim \text{Beta}(\alpha, \beta)$. Writing $S = \sum_{i=1}^{n} y_i$ we get the posterior

$$p \mid \text{data} \sim \text{Beta}(\alpha + S, \beta + nN - S).$$

- More generally, if $\pi(p) = f(p)$ for any positive function integrating to 1 on $[0, 1]$, we get

$$\pi(p \mid \text{data}) \propto_p \text{Beta}(p; 1 + S, 1 + nN - S) f(p)$$

- We can then for example compute numerically the posterior probability that the branching process is supercritical, i.e., that $\Pr(p > 1/N \mid \text{data})$, with (see R computations)

$$\int_{1/N}^{1} \pi(p \mid \text{data})\, dp = \frac{\int_{1/N}^{1} \text{Beta}(1 + S, 1 + nN - S) f(p)\, dp}{\int_{0}^{1} \text{Beta}(1 + S, 1 + nN - S) f(p)\, dp}$$

Using a Multinomial likelihood

- Assume there is a maximum of N offspring and that $p = (p_0, p_1, \ldots, p_N)$ is an unknown probability vector so that p_i is the probability of i offspring. We get the likelihood

$$\pi(y_1, \ldots, y_n \mid p) = \text{Multinomial}(c; p)$$

  where $c = (c_0, \ldots, c_N)$ is the vector of counts in the data of cases with $0, \ldots, N$ offspring, respectively.
- If we use the prior $p \sim \text{Dirichlet}(\alpha)$ where $\alpha = (\alpha_0, \ldots, \alpha_N)$ is a vector of pseudocounts, we get

$$p \mid \text{data} \sim \text{Dirichlet}(\alpha + c).$$

- Note that $\text{Dirichlet}(1, \ldots, 1)$ corresponds to the uniform distribution. Using this prior, we get the posterior expectation for p

$$\mathrm{E}\,(p \mid \text{data}) = \frac{c + (1, 1, \ldots, 1)}{n + N + 1}.$$

- We can simulate from the posterior to investigate for example the probability of being supercritical.

MVE550 2020 Lecture 8.2
Compendium Chapter 5
MCMC for Bayesian inference

Petter Mostad

Chalmers University

November 24, 2020

Continuous variable Markov chains

- A discrete time continuous state space Markov chain is a sequence

$$X_0, X_1, \ldots$$

of continuous random variables with the property that, for all $n > 0$,

$$\pi(X_{n+1} \mid X_0, X_1, \ldots, X_n) = \pi(X_{n+1} \mid X_n)$$

- We work with time-homogeneous Markov chains, so that the *density* $\pi(X_{n+1} \mid X_n)$ is the same for all n.
- *Ergodicity* is defined in a similar way as for discrete state space chains: The chain needs to be irreducible, aperiodic, and positive recurrent.
- The fundamental limit theorem for ergodic Markov chains holds: In the limit as $n \to \infty$, the chain approaches a unique positive stationary distribution.

Markov chain Monte Carlo (MCMC)

- The Metropolis Hastings algorithm is defined as before, except that the proposal distribution $q(\theta_{\text{new}} \mid \theta)$ is now a probability density, not a probability mass function.
- Exactly as before, the limiting distribution of the Metropolis Hastings Markov chain is the target distribution, as long as the Markov chain is ergodic.
- The strong law of large numbers also holds in this situation, as previously noted.
- *Markov chain Monte Carlo (MCMC)* is making the approximation

$$\mathrm{E}_\pi\left(r(\theta)\right) \approx \frac{1}{N} \sum_{i=1}^{N} r(\theta_i)$$

where $\theta_1, \ldots, \theta_N$ is a realization of steps from the Metropolis Hastings Markov chain with the distribution π as its target.

Bayesian inference with MCMC

- We have some data $y_1, \ldots, y_n$ and we want to make a probability prediction for y_{new}.
- We define (in this course) a parameter θ, and a probabilistic model so that

$$\pi(y_1, \ldots, y_n, y_{new}, \theta) = \left[\prod_{i=1}^{n} \pi(y_i \mid \theta)\right] \pi(y_{new} \mid \theta)\pi(\theta)$$

- Thus $\pi(y_{new} \mid y_1, \ldots, y_n) = \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y_1, \ldots, y_n)\, d\theta$.
- Often when the dimension of θ is reasonably high: We use Metropolis Hastings to generate a sample $\theta_1, \ldots, \theta_N$ from $\pi(\theta \mid y_1, \ldots, y_n)$ and approximate

$$\pi(y_{new} \mid y_1, \ldots, y_n) \approx \frac{1}{N} \sum_{i=1}^{N} \pi(y_{new} \mid \theta_i)$$

- The acceptance probabiliby in MH may in case above be written

$$a = \min\left(1, \frac{\pi(y_1, \ldots, y_n \mid \theta^*)\pi(\theta^*)q(\theta \mid \theta^*)}{\pi(y_1, \ldots, y_n \mid \theta)\pi(\theta)q(\theta^* \mid \theta)}\right).$$

Example

- Old example from Chapter 1:

$$\begin{aligned} y \mid p &\sim \text{Binomial}(17, p) \\ p &\sim \text{Beta}(2.3, 4.1) \\ y_{new} \mid p &\sim \text{Binomial}(3, p) \end{aligned}$$

- We would like to compute $\Pr(y_{new} = 1 \mid y = 4)$.
- In this toy example we can do so
  - directly, using conjugacy
  - using discretization
  - using numerical integration
- As an illustration see R for use of Metropolis Hastings with
  - a random walk proposal
  - an independent proposal

The Normal Normal conjugacy

- Assume $y \sim \text{Normal}\left(\theta, \frac{1}{\tau_y}\right)$ where θ is unknown and the *precision* τ_y is known and fixed. Then the normal family is a conjugate family for θ.
- In fact, if $\theta \sim \text{Normal}\left(\mu, \frac{1}{\tau_\mu}\right)$ then

$$\theta \mid y \sim \text{Normal}\left(\frac{\tau_y y + \tau_\mu \mu}{\tau_y + \tau_\mu}, \frac{1}{\tau_y + \tau_\mu}\right)$$

.

- The predictive distribution is also normal. In fact,

$$y \sim \text{Normal}\left(\mu, \frac{1}{\tau_y} + \frac{1}{\tau_\mu}\right).$$

- More conjugacies in Compendium appendix, and on Wikipedia...

Example using Gibbs sampling

- Consider the model with data $y_1, \ldots, y_n$ and parameter $\theta = (\mu, \tau)$:

$$\begin{aligned} y_1, y_2, \ldots, y_n \mid \mu, \tau &\sim \text{Normal}(\mu, \tau^{-1}) \\ \mu &\sim \text{Normal}(\mu_0, \tau_0^{-1}) \\ \tau &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

  $\mu_0, \tau_0, \alpha, \beta$ are fixed known numbers.
- Recall: Gibbs sampling uses proposals where, for each component of θ: fix all other components and simulate from the conditional distribution.
- In our case we may compute

$$\mu \mid y_1, \ldots, y_n, \tau \sim \text{Normal}\left(\frac{\tau_0 \mu_0 + \tau \sum_{i=1}^n y_i}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

  and

$$\tau \mid y_1, \ldots, y_n, \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right).$$

- See R code for an implementation, computing the probability that a new y is less than y_1.

MCMC for Bayesian inference

- Works for almost all models (when the dimension of θ is less than a few thousand).
- Easy to program.
- Difficult to find proposal functions that give an acceptable rate of convergence!
- Difficult to assess the convergence in a particular case: How accurate are your results?