MVE550 2023 Lecture 8
Compendium Chapter 3
Inference for Branching processes.
MCMC for Bayesian inference

Petter Mostad

Chalmers University

November 21, 2023

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.
- As a has infinite length and we have a finite number of observations, we need to put information from the context into the prior, to get a sensible posterior.
- Some alternatives:
  - You assume the offspring distribution has a particular parametric form, and you learn about the parameters.
  - You assume that $a_i = 0$ for $i \geq m$ for some m.

Example: 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

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

- 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 the posterior

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

- 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)

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

Example: Using a Multinomial likelihood

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

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

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

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

  with expectation

  $$\mathrm{E}\left(p_i \mid \text{data}\right) = \frac{\alpha_i + c_i}{\sum_{j=0}^{N}(\alpha_j + c_j)}$$

- Note that $\text{Dirichlet}(1, \dots, 1)$ corresponds to the uniform distribution.
- We can simulate from the posterior to investigate for example the probability of being supercritical.

Part 2: Using MCMC for Bayesian inference

We have some data $y_1, \ldots, y_n$ and we want to make a probability prediction for y_{new}.

- We (often) define a parameter θ, and a probabilistic model so that $y_1, \ldots, y_n, y_{new}$ are all conditionally independent given θ:

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

- Then

$$\begin{aligned} \pi(y_{new} \mid y_1, \ldots, y_n) &= \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y_1, \ldots, y_n)\, d\theta \\ &= \mathrm{E}_{\theta \mid y_1, \ldots, y_n}\left(\pi(y_{new} \mid \theta)\right) \end{aligned}$$

- Upshot (using "law of large numbers"): We can make predictions by
  - Simulating $\theta_1, \ldots, \theta_k$ from the posterior $\pi(\theta \mid y_1, \ldots, y_n)$.
  - Averaging

$$\mathrm{E}_{\theta \mid y_1, \ldots, y_n}\left(\pi(y_{new} \mid \theta)\right) \approx \frac{1}{k}\sum_{j=1}^{k} \pi(y_{new} \mid \theta_j).$$

Finding a sample from the posterior

- So far, we have mostly used conjugacy to be able to find and simulate from the posterior.
- Alternative, we have used numerical computations of integrals.
- What if you cannot use conjugacy, and your integral is too high-dimensional to compute well numerically?
- **Markov Chain Monte Carlo** (MCMC) comes to the rescue!
- Idea of MCMC:
  - Start with a function $f(\theta)$ that is proportinal to the posterior, e.g., $f(\theta) = \pi(\text{data} \mid \theta)\pi(\theta)$.
  - Define an *ergodic Markov chain* so that its **limiting distribution** is the distribution with density or probability mass function proportional to f.
  - Use the values of the Markov chain as an *approximate sample* in a computation like above.
  - It turns out that, in the limit as the length of the chain increases towards ∞, the approximation goes to the expected value above.

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.

The Metropolis-Hastings algorithm

Given a function $f(\theta)$, how can we define an ergodic Markov chain with limiting distribution with density (or pmf) proportional to $f(\theta)$?

- Define a *proposal distribution* $q(\theta^* \mid \theta)$ so that, for any given θ, it is possible to simulate a θ^*.
- Run the *Metropolis-Hastings algorithm*:
  - Choose or simulate some (reasonable) $\theta^{(0)}$.
  - For $i = 0, 1, 2 \ldots$:
    - Simulate a *proposal* θ^* using $q(\theta^* \mid \theta^{(i)})$.
    - Compute the acceptance probability

$$\rho = \min\left(1, \frac{f(\theta^*)q(\theta^{(i)} \mid \theta^*)}{f(\theta^{(i)})q(\theta^* \mid \theta^{(i)})}\right).$$

    - With probability ρ, set $\theta^{(i+1)} = \theta^*$, otherwise set $\theta^{(i+1)} = \theta^{(i)}$.
- The MH algorithm defines a Markov chain $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$
- IF this Markov chain is ergodic, its limiting distribution will have density proportional to $f(\theta)$.

Toy example

- Old example from compendium 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) we may also use MCMC.

Second example

- We have observed the data (x_i, y_i):

$$(2, 0.32), (3, 0.57), (4, 0.61), (6, 0.83), (9, 0.91)$$

- The context gives us the following model
  - We expect the data to follow $y = f(x, \theta_1) = \frac{\exp(\theta_1 x) - 1}{\exp(\theta_1 x) + 1}$ where θ_1 is an unknown positive parameter.
  - We have observed the data with added noise $\text{Normal}(0, \theta_2^2)$ where θ_2 is an unknown positive parameter.
  - We assume a flat prior on $\theta_1 > 0$ and $\theta_2 > 0$.
- We get the posterior

$$\pi(\theta \mid \text{data}) \propto_\theta \prod_{i=1}^{5} \text{Normal}(y_i; f(x_i, \theta_1), \theta_2^2).$$

- Use MCMC to simulate from the value of y when $x = 10$ (see R).