Perfect sampling: Review from last time

Given ergodic Markov chain with finite sample space of size k and limiting distribution π.

- When using this setup for MCMC, the goal is to get a sample from π.
- Perfect sampling: Simulating from the chain, we prove that the last simulated value actually has distribution π.
- If we start k chains from the k different states, and they all end up in the same state, we have *forgotten* the initial state, and have reached the limiting distribution.
- *Coupling*: Simulate so that if two chains have identical states at step i, they are also identical at step $i+1$ (they *coalesce*): Use function $X_{i+1}^{(j)} = g(X_i^{(j)}, U_i)$ where $U_i \sim \text{Uniform}(0, 1)$.
- Length of simulation must be decided independently of values! Simulate by extending backwards!

Monotonicity

- Do we need to keep track of *all* k chains?
- We define a *partial ordering* on a set as a relation $x \leq y$ between *some* pairs x and y in the set, such that:
  - If $x \leq y$ and $y \leq x$ then $x = y$.
  - If $x \leq y$ and $y \leq z$ then $x \leq z$ (in fact we don't use this).
- We will need that our partial ordering has a minimal element (an m such that $m \leq x$ for all x) and a maximal element (an M such that $x \leq M$ for all x).
- If we have a partial ordering on the state space of the Markov chain, and if $x \leq y$ implies $g(x, U) \leq g(y, U)$, then g is *monotone*.
- We can then prove that we only need to keep track of the chain starting at m and the chain starting at M!

Example: Perfect simulation from the Ising model

- Given an Ising model with $\beta > 0$.
- Define partial ordering on Ω (the set of all configurations) as follows

$$\sigma \leq \tau \text{ if } \sigma_v \leq \tau_v \text{ for all vertices } v$$

- We have a minimal and a maximal configuration (all -1's and +1's, respectively).
- We can arrange for g, the updating of chains, to be monotone: Assuming $\sigma \leq \tau$,

$$\Pr(\sigma_v = 1 \mid \sigma_{-v}) = \frac{1}{1 + \exp(-2\beta \sum_{v \sim w} \sigma_w)} \leq \frac{1}{1 + \exp(-2\beta \sum_{v \sim w} \tau_w)} = \Pr(\tau_v = 1 \mid \tau_{-v}).$$

- So perfect simulation from the Ising model proceeds as follows: Start one chain m at all -1's and one chain M at all +1's. Cycle through the vertices. Compute the conditional probabilities p_m and p_M of +1 at each vertex. We know that $p_m \leq p_M$. Simulate $U \sim \text{Uniform}(0,1)$. If $U < p_m$ set $\sigma_v = -1$ for both chains, and if $U > p_M$ set $\sigma_v = +1$ for both chains. Otherwise set $\sigma_v = +1$ for the M chain and $\sigma_v = -1$ for the m chain. Determine coalescence as above.

MVE550 2022 Lecture 8
Compendium chapters 4 and 5
Inference for Branching processes. MCMC for Bayesian inference

Petter Mostad

Chalmers University

November 24, 2022

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 $a = (a_0, a_1, a_2, \ldots,)$ where a_i is the probabilty for i offspring in the offspring process.
- Let $y_1, y_2, \ldots, y_n$ be the counts of offspring in n observations of the offspring process. If a is given we have the likelihood

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

- 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.
- We will look at alternatives where you either decide that $a_i = 0$ for $i \geq m$ for some m, or where the offspring distribution has a particular parametric form.

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 now $p = (p_0, p_1, \ldots, p_N)$ is an unknown probability vector so that p_i is the probability of i offspring. We get the likelihood

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

  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

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

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

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) with continuous variables

- 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 extends to this situation.
- *Markov chain Monte Carlo (MCMC)* is making the approximation

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

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 (often) define 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

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

Bayesian inference with MCMC, cont.

Often when the dimension of θ is reasonably high:

- We use Metropolis Hastings (MH) to generate an approximate sample $\theta_1, \ldots, \theta_N$ from $\pi(\theta \mid y_1, \ldots, y_n)$ and approximate

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

- We may also simulate from $\pi(y_{new} \mid y_1, \ldots, y_n)$ by simulating the $\theta_1, \ldots, \theta_N$ as above and then from $\pi(y_{new} \mid \theta_1), \ldots, \pi(y_{new} \mid \theta_N)$.
- Note that the acceptance probabiliby in MH may in our case be written

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

where θ^* is the proposed value based on θ.

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.

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 parameter.
  - We have observed the data with added noise $\text{Normal}(0, \theta_2^2)$ where θ_2 is an unknown 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).