MSA101/MVE187 2020 Lecture 8.1
Missing data / augmented data

Petter Mostad

Chalmers University

September 18, 2020

Missing data / augmented data

- Assume some data values are *censored*: You don't know them exactly, only that they are (for example) above some threshold. How to deal with this?
- Example application: Survival analysis. You want to know how long people live after some event. But some people are still alive at the end of the study (or they died from other causes).
- We want to learn about density $f(\cdot \mid \theta)$ from sample where $x_1, \ldots, x_k$ are observed values and $c_1, \ldots, c_n$ are observations that the corresponding x_i is greater than some a_i. The likelihood becomes

$$\pi(x_1, \ldots, x_k, c_1, \ldots, c_n \mid \theta) = \prod_{i=1}^{k} f(x_i \mid \theta) \prod_{i=1}^{n} (1 - F(a_i \mid \theta))$$

  where $F(\cdot \mid \theta)$ is the cumulative density.
- You may simulate from the posterior for θ using for example random walk MH.
- ALTERNATIVELY: You may *add to the model* variables representing the censored values, and simulate these together with the unknown θ. This may often be a simpler and better solution! See R example.

Example: Augmented data

- Example (7.7. in RC): In a genetics problem, one wants to know how close two genes are on the chromosome, measured by a parameter θ. Given n individuals, the number of individuals x_1, x_2, x_3, x_4 in each of 4 categories will be multinomially distributed accoring to

$$(x_1, x_2, x_3, x_4) \mid \theta \sim \text{Multinomial}\left(n, \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

  Given a prior on θ, how do you simulate from the posterior?
- The likelihood for θ makes necessary approximate or numerical simulation:

$$\pi(x_1, \ldots, x_4 \mid \theta) \propto_\theta \left(\frac{1}{2} + \frac{\theta}{4}\right)^{x_1} \left(\frac{1}{4}(1-\theta)\right)^{x_2} \left(\frac{1}{4}(1-\theta)\right)^{x_3} \left(\frac{\theta}{4}\right)^{x_4}.$$

- We extend the data (x_1, x_2, x_3, x_4) with a latent variable z, so that

$$(z, x_1 - z, x_2, x_3, x_4) \mid \theta \sim \text{Multinomial}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

- The likelihood becomes

$$\pi(z, x_1, \ldots, x_4 \mid \theta) \propto_\theta \theta^{x_1 - z + x_4}(1-\theta)^{x_2 + x_3}.$$

Example continued

- Note that, with the augmented data (z, x_1, x_2, x_3, x_4), the likelihood has the Beta family of densities as conjugate priors! Assume, for example, $\theta \sim \text{Beta}(\alpha, \beta)$.
- You can now use Gibbs sampling to sample from the distribution $\pi(z, \theta \mid x_1, \ldots, x_4)$:
  - $\theta \mid z, x_1, x_2, x_3, x_4 \sim \text{Beta}(\alpha + x_1 - z + x_4, \beta + x_2 + x_3)$.
  - $z \mid \theta, x_1, x_2, x_3, x_4 \sim \text{Binomial}\left(x_1, \frac{\frac{1}{2}}{\frac{1}{2} + \frac{\theta}{4}}\right)$.
- Exercise: Derive the Binomal distribution for z above.

Handling missing data

- In many classical statistical methods, missing data may present a problem.
- The standard Bayesian answer in such cases: Add to the model random variables representing the unobserved values, and simulate them together with parameters and other variables of interest.
- This solves the problem in theory, but may of course sometimes be difficult in practice.

MSA101/MVE187 2020 Lecture 8.2
Hamiltonian Monte Carlo

Petter Mostad

Chalmers University

September 24, 2020

Using the target density in the proposal

- We have looked at several ideas for constructing good proposal densities. Somehow, they take into account the properties of the target density.
- Can one construct general methods that "automatically" learns about the target density and makes good proposals based on that?
- Several methods exist that do this; they have varying degrees of success with good convergence.
- We will look at one quite popular and clever method: *Hamiltonian Monte Carlo*.

Hamiltonian Monte Carlo: Motivation

We assume given a posterior density up to a constant:
$\pi(q) \propto_\theta \exp(-U(q))$ for vectors $q = (q_1, \ldots, q_d)$.

- Idea: Look at $U(q)$ as some kind of "potential energy" for a particle at position q.
- Then: Simulate not only q at each step, but also $p = (p_1, \ldots, p_d)$ representing a "momentum" for the particle. Move particle!
- In fact we will simulate from the joint distribution

$$\pi(p, q) \propto_{p,q} \exp(-U(q)) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{p_i^2}{m_i}\right)$$

  where $m_1, \ldots, m_d$ are positive weights, so that p and q are independent, and p is normally distributed.
- Proposals will propose points (p, q) based on their "total energy"

$$H(p, q) = U(q) + \frac{1}{2}\sum_{i=1}^{d}\frac{p_i^2}{m_i}$$

  so that points with lower potential energy (i.e., higher probability in the target density) will be proposed more often.

Hamiltonian Monte Carlo: Simulation

We would like to simulate from the density

$$\pi(p,q) \propto_{p,q} \exp\left(-H(p,q)\right) = \exp\left(-U(q)\right) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{p_i^2}{m_i}\right)$$

- We will define in later overheads a transformation T_s sending the set of pairs (p,q) to itself, satisfying
  - $H(T_s(p,q)) = H(p,q)$.
  - T_s preserves volumes.
  - T_s is symmetric: $T_s : (p,q) \mapsto (p^*,q^*)$ implies $T_s(p^*,q^*) = (p,q)$. (Fixed 20/09/24)
- When the Markov chain is at (p,q) a proposal (p^*,q^*) is generated as follows:
  1. Simulate a p_0 from the marginal normal distribution for p.
  2. Set $(p^*,q^*) = T_s(p_0,q)$.
- We show below that the acceptance probability will be one.
- As the marginal distribution for the generated q's is the right one, we can in the end throw away the generated p's.

Why is the acceptance probability 1?

- A possible way to see this is to split the M.H. updates into two parts.
- In the first part p is replaced with a new p simulated from the marginal. The acceptance probability is clearly 1.
- In the second *deterministic* step, the transformation T_s is applied. As:
  - It sends points (p, q) on points (p^*, q^*) with the same density: $\pi(p, q) = \pi(p^*, q^*)$.
  - The transformation is volume preserving.
  - The transformation is symmetric.

  it follows that the transformed density is unchanged, and the acceptance rate is 1.

Hamiltonian dynamics

Given a function $H(p, q)$.

- A particle that has "position" q and "momentum" p at time t is said to follow Hamiltonian dynamics if, for $i = 1, \ldots, d$,

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \text{and} \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}.$$

- Assume for example $H(p, q) = U(q) + \frac{1}{2m} p^T p$. Then the first equation above says that the velocity of the particle is given by its momentum divided by its mass, and the second equation says that the change in the momentum is given by the negative change in its potential energy $U(q)$.
- After a specific time s, a particle with position q and momentum p will have position q^* and momentum p^*. This defines a mapping sending the set of all pairs (p, q) to itself. T_s is this mapping followed by changing the sign of p^*. (Fixed 20/09/24)
- We need to show that T_s has the required properties.

Outline of proofs

- We have $H(T_s(p,q)) = H(p,q)$ because

$$\frac{dH}{dt} = \sum_{i=1}^{d} \left(\frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i} \right) = \sum_{i=1}^{d} \left(\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} \right) = 0$$

  and because $H(-p,q) = H(p,q)$. (Fixed 20/09/24)

- We show below that the divergence of the vector field defining T_s is zero. It then follows from theorems in analysis that T_s preserves volumes:

$$\sum_{i=1}^{d} \left(\frac{\partial}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial}{\partial p_i}\frac{dp_i}{dt} \right) = \sum_{i=1}^{d} \left(\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i} \right)$$

$$= \sum_{i=1}^{d} \left(\frac{\partial^2 H}{\partial q_i \partial p_i} - \frac{\partial^2 H}{\partial p_i}\partial q_i \right) = 0.$$

Leapfrog algorithm: A numerical approximation of T_s

- The second part of our proposal function requires us to compute $(p^*, q^*) = T_s(p_0, q)$, i.e., the position q^* and momentum $-p^*$ after time s of a particle following Hamiltonian dynamics and starting with position q and momentum p_0 at time zero.
- This requires solving differential equations, which we do with *the Leapfrog method*.
- The Leapfrog method actually preserves volume exactly, although it only approximately preserves the total energy $H(p, q)$.
- We will not look more on the theory here, but rather show an example in R. For more information see for example Neal (2011) "MCMC Using Hamiltonian Dynamics".

MSA101/MVE187 2020 Lecture 8.3
Summary / overview of MCMC so far

Petter Mostad

Chalmers University

September 20, 2020

Review of most important MCMC concepts so far

- The general Metropolis Hastings algorithm
- Types of proposals: Random walk, symmetric, independent, ...
- Detecting non-convergence; improving the approximations using burn-in, thinning, ...
- Laplace approximation
- Gibbs sampling
- Slice sampling
- Handling missing data / using augmented data
- Hamiltonian Monte Carlo
- We will return to some simulation examples, with some related concepts (e.g., transformations), towards the end of the course.

Summary on MCMC so far

- Metropolis Hastings is an extremely versatile algorithm for approximate simulation from a density.
- It requires knowing the density only up to a constant.
- The main drawbacks are that it converge may be slow, and that we generally don't know how slow it is.
- Remember: You may combine different proposal functions in one algorithm, as long as each represents a valid Metropolis Hastings step.
- It seems difficult to find universal "black box" implementations of Metropolis Hastings that always work perfectly: Instead, users need to understand the method and check that it works in their case.
- Several black box (-ish) implementations exist: A popular one is Jags ("Just Another Gibbs Sampler") for hierarchical models.
- Some tips about coding Metropolis Hastings from my colleague Umberto (who will lecture later in this course): https://umbertopicchini.wordpress.com/2017/12/18/tips-for-coding-a-metropolis-hastings-sampler/
- Lots of work is still going on developing ideas related to MCMC. One simple idea, tempered MCMC, is briefly discussed below.

Tempered MCMC

- Problem: The MCMC too easily gets stuck, and then does not reach the areas of high posterior density.
- Idea: Start with a period of "improved searching" before approaching the acutal MCMC formulas.
- The posterior $\exp(h(x))$ is replaced with $\exp\left(\frac{h(x)}{T}\right)$ for some positive "temperature" T: For large T this "evens out" the posterior.
- Making T monotonically sink towards 1 gives an MCMC chain that can jump more easily in the start while simulating from the correct posterior in the end.
- Making T monotonically sink towards 0 gives an MCMC chain that finds a maximum! If T sinks sufficiently slowly, one can prove it finds the *global* optimum with probability 1. *Simulated annealing*.