MSA101/MVE187 2022 Lecture 11
Missing data / augmented data
Hamiltonian MCMC

Petter Mostad

Chalmers University

October 3, 2022

Review: Bayesian framework

- Prediction variable Y_{pred}, data Y_{data}, parameter θ.
- Specify a complete model by specifying prior $\pi(\theta)$, likelihood $\pi(Y_{data} \mid \theta)$, and prediction distribution $\pi(Y_{pred} \mid \theta)$.
- Derive the posterior $\pi(\theta \mid Y_{data})$.
- Make predictions using

  $$\pi(Y_{pred} \mid Y_{data}) = \int \pi(Y_{pred} \mid \theta, Y_{data}) \pi(\theta \mid Y_{data}) \, d\theta$$

  (in some cases $\pi(Y_{pred} \mid \theta, Y_{data}) = \pi(Y_{pred} \mid \theta)$).
- If we cannot approximate the integral, we may instead simulate from

  $$\pi(Y_{pred}, \theta \mid Y_{data})$$

  and simply use the part of the simulated vectors that represent Y_{pred}.

Hierarchical models: examples

- Heart transplant example:
  - **Data**: Exposures e_i and deaths y_i at $i = 1, \dots, 94$ hospitals. **Parameters**: Mortality rates $\lambda_1, \dots, \lambda_{94}$ at the 94 hospitals. Underlying parameters μ, α. **Prediction**: E.g., prob that $\lambda_1 < \lambda_2$.
  - Model: $y_i \sim \text{Poisson}(e_i\lambda_i)$, $\lambda_i \sim \text{Gamma}(\alpha, \alpha/\mu)$, $\mu \propto_\mu 1/\mu$, $\alpha \propto_\alpha 1/\alpha$.
  - Draw the model as a graph!
- More examples:
  - The examples of conjugacy we started the course with
  - Second assignment!

Review: A hierarchical example

Data $x_1, \ldots, x_8$ and $y_1, \ldots, y_6$ are organized into groups, and we want to predict a value z_1 in a third group. We assume a model

$$\begin{aligned}
x_1, \ldots, x_8 &\sim \text{Normal}(\mu_1, \tau_1^{-1}) \\
y_1, \ldots, y_6 &\sim \text{Normal}(\mu_2, \tau_1^{-1}) \\
z_1 &\sim \text{Normal}(\mu_3, \tau_1^{-1}) \\
\mu_1, \mu_2, \mu_3 &\sim \text{Normal}(10, \tau_0^{-1}) \\
\tau_0 &\sim \text{Gamma}(1, 4) \\
\tau_1 &\sim \text{Gamma}(7, 3)
\end{aligned}$$

- Draw the model!
- We simulate from the conditional distribution of all unknowns $\tau_0, \tau_1, \mu_1, \mu_2, \mu_3, z_1$ after fixing the data values $x_1, \ldots, x_8, y_1, \ldots, y_6$.
- We then use the simulated values for z_1 to make predictions.
- For many hierarchical models, Gibbs sampling is a good way to simulate from the conditional distribution: See next overhead.

Inference for hierarchical models

- The joint density for all variables in a hierarchical model is a product with one factor for each node in its graph.
- To find a function proportional to the posterior density for Y_{pred}, θ given Y_{data}, simply fix in this function all values corresponding to the data Y_{data}.
- To simulate from the resulting posterior density using Gibbs sampling, the conditional distribution for each unknown variable given all the others must be found.
- A function proportional to each such conditional density can be found by fixing all the other variables, and removing constant factors.
- The resulting functions will only have terms from original factors corresponding to edges in or out of the variable in the graph!

Inference for hierarchical models

1. Write down the joint probability of all variables (Y_{data}, Y_{pred}, and θ) as a product over all the nodes in the hierarchical model.
2. If possible, try to analytically marginalize over some of the components of θ.
3. In the expression from (1), fix all values in Y_{data}, to obtain a function proportional to the joint posterior $Y_{pred}, \theta \mid Y_{data}$. (Remove constant factors).
4. Generate an (approximate) sample from the remaining variables; the dimensions corresponding to Y_{pred} represent an approximate sample for your prediction.

- It may be convenient to use a Metropolis Hastings with symmetric proposals, or:
- Many or all of the conditional densities needed for a Gibbs sampler can often be found easily.
- If some conditional densities cannot be found easily, one may for these use a standard Metropolis Hastings proposal function (and acceptance probability).

Review: Conditional distributions for the example

The conditional distributions become (prove yourself!)

$$\begin{aligned}
\mu_1 \mid x_1, \ldots, x_8, \tau_1, \tau_0 &\sim \text{Normal}\left(\frac{10\tau_0 + 8\overline{x}\tau_1}{\tau_0 + 8\tau_1}, \frac{1}{\tau_0 + 8\tau_1}\right) \\
\mu_2 \mid y_1, \ldots, y_6, \tau_1, \tau_0 &\sim \text{Normal}\left(\frac{10\tau_0 + 6\overline{y}\tau_1}{\tau_0 + 6\tau_1}, \frac{1}{\tau_0 + 6\tau_1}\right) \\
\mu_3 \mid z_1, \tau_1, \tau_0 &\sim \text{Normal}\left(\frac{10\tau_0 + z_1\tau_1}{\tau_0 + \tau_1}, \frac{1}{\tau_0 + \tau_1}\right) \\
\tau_0 \mid \mu_1, \mu_2, \mu_3 &\sim \text{Gamma}\left(1 + \frac{3}{2}, 4 + \frac{1}{2}\sum_{i=1}^{3}(\mu_i - 10)^2\right) \\
\tau_1 \mid \mu_1, \mu_2, \mu_3, x_1 \ldots x_8, y_1 \ldots y_6, z_1 &\sim \text{Gamma}\left(7 + \frac{15}{2}, 3 + \frac{1}{2}\sum_{i=1}^{8}(x_i - \mu_1)^2\right. \\
\left. + \frac{1}{2}\sum_{i=1}^{6}(y_i - \mu_2)^2 + \frac{1}{2}(z_1 - \mu_3)^2\right) \\
z_1 \mid \mu_3, \tau_1 &\sim \text{Normal}(\mu_3, \tau_1^{-1})
\end{aligned}$$

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 $x_1, \dots, x_k$ are observed values and $c_1, \dots, c_n$ are observations that the corresponding x_i is greater than some a_i. The likelihood becomes

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

  where $F(\cdot \mid \theta)$ is the cumulative distribution function.
- 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.

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.

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:

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

- 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

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

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.

Part 2. 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: Idea

We are given a posterior density $\pi(q) \propto_q \exp(-U(q))$ for vectors $q = (q_1, \dots, q_d)$. We want to find a smart proposal function that utilizes U:

- Look at $U(q)$ as some kind of "potential energy" for a particle that can move between different q's.
- If the particle moves so that it looses potential energy, it gains kinetic energy, i.e., it moves faster.
- If the particle moves in this way, it will move faster in the direction of higher density for $\pi(q)$.
- Idea: As a proposal function, randomly generate a direction and a speed for the particle to move from the current q. Then let the particle move according to dynamics above for time period s.
- Below, we use pairs (p, q) of particle momentum $p = (p_1, \dots, p_d)$ and particle position q, moving the particle so that the total energy

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

  is kept constant (where we may set $\sigma_i^2 = m$ where m is the "mass").

A Metropolis Hastings step using transformations

Assume $\pi(x)$ is a density and T_1, T_2 are invertible transformations on the set of possible x values satisfying for all x

$$\pi(T_1(x)) = \pi(x) \qquad \pi(T_2(x)) = \pi(x) \qquad T_1(T_2(x)) = T_2^{-1}(T_1^{-1}(x)).$$

Then the deterministic M-H proposal function $T_2(T_1(x))$ has acceptance probability 1.

- Proof: We have symmetry

$$T_2(T_1(T_2(T_1(x)))) = T_2(T_2^{-1}(T_1^{-1}(T_1(x)))) = x$$

and invariance of the density

$$\pi(T_2(T_1(x))) = \pi(T_1(x)) = \pi(x).$$

Hamiltonian dynamics

Given a "Hamiltonian" function $H(p, q)$, with $p, q \in \mathbb{R}^d$, $H(p, q) \in \mathbb{R}$.

- A particle $p : \mathbb{R} \to \mathbb{R}^{2d}$, with $p(t) = (p, q)$, is said to have "position" q and "momentum" p at time t.
- The particle follows Hamiltonian dynamics if, for $i = 1, \ldots, d$,

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \text{and} \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \tag{1}$$

- After time s, the particle starting at position q with momentum p will have position q^* and momentum p^*. So the solution to Equations 1 defines a mapping T_1 sending (p, q) to (p^*, q^*).
- We have $H(T_1(p, q)) = H(p, q)$ because

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

- If $H(-p, q) = H(p, q)$ for all (p, q) then Equations 1 are unchanged if we simultaneously switch the signs of t and p. Defining $T_2(p, q) = (-p, q)$ we get

$$T_1(T_2(p, q)) = T_2^{-1} T_1^{-1}(p, q).$$

Hamiltonian Monte Carlo

Assume $\pi(p,q) \propto \exp(-H(p,q))$ with $H(-p,q) = H(p,q)$ as above.

- A Metropolis Hastings algorithm alternating between the proposals
  - simulate p using $\pi(p \mid q) \propto \exp(-H(p,q))$
  - compute $T_2(T_1(p,q))$

  and always accepting will provide an approximate sample from $\pi(p,q)$ provided the Markov chain is ergodic.
- If $H(p,q) = V(p) + U(q)$ with $V(-p) = V(p)$ then p and q are independent, and $\pi(p \mid q) \propto \exp(-V(p))$ is simulated independently in each iteration.
- By throwing away the part of the sample concerning p we get an approximate sample from $\pi(q) \propto \exp(-U(q))$.

Hamiltonian Monte Carlo algorithm

- Start with a density $\pi(q) \propto \exp(-U(q))$ you want to simulate from.
- Define $V(p) = \frac{1}{2}\sum_{i=1}^{d} p_i^2/\sigma_i^2$ for some $\sigma_1, \ldots, \sigma_d$.
- Find an initial value $q^{(0)}$ for q.
- For each iteration, given $q^{(j)}$:
    - Simulate $p_i \sim \text{Normal}(0, \sigma_i^2)$
    - Compute $(p^*, q^*) = T_1(p, q^{(j)})$
    - Set $q^{(j+1)} = q^*$ and throw away p^*.
- It "just" remains to find an efficient way to compute $T_1(p, q^{(j)})$.
- Note that

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \text{becomes} \qquad \frac{dq_i}{dt} = \frac{p_i}{\sigma_i^2}$$

and

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \qquad \text{becomes} \qquad \frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}.$$

The Leapfrog algorithm: A numerical approximation of T_1

- For simplicity set all $\sigma_i = 1$ and use vector notation: We need that

$$\frac{dq}{dt} = p \qquad \text{and} \qquad \frac{dp}{dt} = -\nabla U(q)$$

- Let $q_0, q_1, q_2 \ldots, q_n$ be the values of q along the particle path at times $0, \frac{s}{n}1, \frac{s}{n}2, \ldots, \frac{s}{n}n = s$, respectively.
- Let $p_0, p_1, p_2, \ldots, p_{n+1}$ be the values of p along the particle path at times $0, \frac{s}{n}(1-\frac{1}{2}), \frac{s}{n}(2-\frac{1}{2}), \ldots, \frac{s}{n}(n-\frac{1}{2}), s$, respectively.
- Approximate $\frac{dq}{dt} = p$ with

$$\frac{q_{j+1} - q_j}{s/n} = p_{j+1} \qquad j = 0, \ldots, n-1.$$

- Approximate $\frac{dp}{dt} = -\nabla U(q)$ with

$$\frac{p_{j+1} - p_j}{s/n} = -\nabla U(q_i) \qquad j = 1, \ldots, n-1.$$

while using half stepsize for $j = 0$ and $j = n$.

- We get $p_1 = p_0 - (s/2n)\nabla U(q_0)$, $p_{n+1} = p_n - (s/2n)\nabla U(q_n)$, and

$$\begin{aligned} q_{j+1} &= q_j + (s/n)p_{j+1} \\ p_{j+1} &= p_j - (s/n)\nabla U(q_j) \end{aligned}$$

Hamiltonian MCMC: Summary

- Note: n computations of the gradient ∇U must be done: Possible? Time consuming?
- Note: As this is an approximation, we only have that $H(p^*, q^*) \approx H(p, q)$. But this is no problem, as we can compute and use the standard acceptance probability for Metropolis Hastings proposals.
- Note: You must still check that the Markov chain is Ergodic: In practice, that the algorithm can reach any q from any q.
- Can give great fast convergence in the cases where the gradient of the logged density is easily available and computable.
- For more information see for example Neal (2011) “MCMC Using Hamiltonian Dynamics”.