MSA101/MVE187 2020 Lecture 7.1
Laplace approximation
MCMC convergence

Petter Mostad

Chalmers University

September 16, 2020

The Laplace approximation

- For many models, the posterior $\pi(\theta \mid \text{data})$ for the parameter will have a shape that is close to a normal distribution. (A reason for this is the Central Limit Theorem).
- So, sometimes using some (multivariate) normal approximation for the true posterior distribution is a good enough approximation.
- If we use the normal distribution that has the same mode as the actual posterior and the same second derivatives (the first derivatives are zero at the mode), we call it the *Laplace approximation*.
- The Laplace approximation can be found for example by numerical differentiation of the logged posterior density, which needs to be known only up to a constant. See the R function `laplace` in the R package `LearnBayes`.

The Laplace multivariate normal approximation

It is sometimes useful to consider the following approximation, when we have a density written

$$\pi(\theta) = C \cdot \exp(h(\theta))$$

for some known function h and unknown constant C. If $\hat{\theta}$ is the mode of the density, the second-degree Taylor approximation gives

$$h(\theta) \approx h(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^t H(\hat{\theta})(\theta - \hat{\theta})$$

where $H(\theta)$ is the Hessian matrix of second derivatives. We get

$$\pi(\theta) \approx C \cdot \exp(h(\hat{\theta})) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^t((-H(\hat{\theta}))^{-1})^{-1}(\theta - \hat{\theta})\right).$$

This means that $\pi(\theta)$ might be approximated by a multivariate normal distribution with expectation $\hat{\theta}$ and covariance matrix $-H(\hat{\theta})^{-1}$. If we integrate both sides with respect to θ we get

$$C \approx \frac{1}{\exp(h(\hat{\theta}))|2\pi(-H(\hat{\theta}))^{-1}|^{1/2}}.$$

A return to the "cars" example

- Remember examample from Lecture 6: Speeds and braking distances for 50 cars are given. Predict the braking distance for a car with speed 21 mph.
- We decided on a model and on using Metropolis Hastings random walk for simulation. Some more decisions to make:
  - Generating the starting point for the Markov chain.
  - What step lengths should be used in the random walk?
  - How many steps should be simulated?
  - Should we remove a burn-in? How long?
- Some tools for making such decisions:
  - How to find good starting points and reasonable step lengths (more discussion now).
  - Trace plots (discussed last time).
  - Acceptance rates (discussed last time).
  - Autocorrelation (more discussion now).
  - Using multiple starting points and parallell chains (more discussion now).

MSA101/MVE187 2020 Lecture 7.2
Gibbs sampling

Petter Mostad

Chalmers University

September 16, 2020

Simulating from conditional densities

- Assume you have defined a density as proportional to some complicated function of your parameters, for example

$$\pi(\theta_1, \theta_2, \theta_3) \propto_{\theta_1,\theta_2,\theta_3} \theta_1^{\theta_2} \exp\left(-\theta_1(\theta_3 - \theta_2)^4\right) sin^2(\theta_2 + \theta_3)$$

  What can you say about conditional densities, such as $\pi(\theta_1 \mid \theta_2, \theta_3)$?
- As $\pi(\theta_1 \mid \theta_2, \theta_3) = \pi(\theta_1, \theta_2, \theta_3)/\pi(\theta_2, \theta_3) \propto_{\theta_1} \pi(\theta_1, \theta_2, \theta_3)$, we get that we can use exactly the same formula, just fixing the values of the parameters we condition on:

$$\pi(\theta_1 \mid \theta_2, \theta_3) \propto_{\theta_1} \theta_1^{\theta_2} \exp\left(-\theta_1(\theta_3 - \theta_2)^4\right) sin^2(\theta_2 + \theta_3)$$

- The remaining function, regarded as a function of (in our case) just θ_1 may be a simpler density, maybe even recognizable as a standard density.
- In our case we see that

$$\theta_1 \mid \theta_2, \theta_3 \sim \text{Gamma}\left(\theta_2 + 1, (\theta_3 - \theta_2)^4\right)$$

Gibbs sampling

- In many multivariate models, it may be difficult to simulate directly from the posterior $\pi(x_1, x_2, \ldots, x_n)$, but easy to simulate from each of the conditional distributions $\pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ for $j = 1, \ldots, n$.
- Idea for simulation method: Iterate between the different j's, simulating each time from the conditional distribution given previously simulated values for the other coordinates.
- We formalize this as a Metropolis Hastings algorithm iterating between n different proposal functions: For each $j = 1, \ldots, n$, fix all x_i with $i \neq j$ and for the j'th variable simulate x_j^* using $\pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$.

Gibbs sampling, continued

- The acceptance probability in the MH algorithm is computed with

$$
\begin{aligned}
& \frac{\pi(x^*)q(x \mid x^*)}{\pi(x)q(x^* \mid x)} \\
= \quad & \frac{\pi(x_1, \ldots, x_j^*, \ldots, x_n)\pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)}{\pi(x_1, \ldots, x_j, \ldots, , x_n)\pi(x_j^* \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)} \\
= \quad & \frac{\pi(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)}{\pi(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)} = 1
\end{aligned}
$$

  So accept always!
- This algorithm is called *Gibbs sampling*.
- For many models it is easy to implement and program.
- However, the convergence may be too slow unless the density is nice and unimodal.

Gibbs sampling: Examples

- Example: Simulate from a bivariate normal distribution. The conditional distributions are normal, formulas are given in a previous lecture. See R code.
- Example: Data $y_1, y_2, \ldots, y_n$ are from a Normal(μ, τ^{-1}) distribution, with independent priors $\mu \sim$ Normal$(0, 1)$ and $\tau \sim$ Gamma$(3, 4)$.
  - When τ is fixed we get

$$\mu \mid \tau, \text{data} \sim \text{Normal}\left(\frac{n\bar{y}\tau}{n\tau + 1}, \frac{1}{n\tau + 1}\right).$$

  - When μ is fixed we get

$$\tau \mid \mu, \text{data} \sim \text{Gamma}\left(3 + \frac{n}{2}, 4 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

- When τ is fixed, the formula above is a result of the formula for the posterior in the Normal-Normal conjugacy with fixed precision.
- When μ is fixed, the formula above is a result of the formula for the posterior in the Normal-Gamma conjugacy with fixed expectation.

MSA101/MVE187 2020 Lecture 7.3
Hierarchical models

Petter Mostad

Chalmers University

September 17, 2020

Hierarchiclal models

- Sometimes, observed data have dependencies that can best be described using a hierarchy.
- Example: Test results for students may depend on the class they are in, the school they attend, and the country they live in.
- A statistical model for the data should then contain a random variable for each "source of infuence"; they would depend on each other in a hierarchy, which can be drawn as an upside-down tree, or more generally as a network.
- When making computations, the tree structure can be very useful, for example when using Gibbs sampling.

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

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

- We can make predictions for z_1 given data $x_1, \ldots, x_8$ and $y_1, \ldots, y_6$ by simulating with Gibbs sampling from the model where the data is fixed and the remaining variables $\mu_1, \mu_2, \mu_3, \tau_0, \tau_1, z_1$ are simulated.
- Note: The exact form for the conditional distributions of each of these variables can be found using conjugacy.

Conditional distributions for the example

The conditional distributions become (prove yourself!)

$$\mu_1 \mid x_1, \ldots, x_8, \tau_1, \tau_0 \sim \text{Normal}\left(\frac{10\tau_0 + 8\bar{x}\tau_1}{\tau_0 + 8\tau_1}, \frac{1}{\tau_0 + 8\tau_1}\right)$$

$$\mu_2 \mid y_1, \ldots, y_6, \tau_1, \tau_0 \sim \text{Normal}\left(\frac{10\tau_0 + 6\bar{y}\tau_1}{\tau_0 + 6\tau_1}, \frac{1}{\tau_0 + 6\tau_1}\right)$$

$$\mu_3 \mid z_1, \tau_1, \tau_0 \sim \text{Normal}\left(\frac{10\tau_0 + z_1\tau_1}{\tau_0 + \tau_1}, \frac{1}{\tau_0 + \tau_1}\right)$$

$$\tau_0 \mid \mu_1, \mu_2, \mu_3 \sim \text{Gamma}\left(1 + \frac{3}{2}, 4 + \frac{1}{2}\sum_{i=1}^{3}(\mu_i - 10)^2\right)$$

$$\tau_1 \mid \mu_1, \mu_2, \mu_3, x_1 \ldots x_8, y_1 \ldots y_6, z_1 \sim \text{Gamma}\left(7 + \frac{15}{2}, 3 + \frac{1}{2}\sum_{i=1}^{8}(x_i - \mu_1)^2 + \frac{1}{2}\sum_{i=1}^{6}(y_i - \mu_2)^2 + \frac{1}{2}(z_1 - \mu_3)^2\right)$$

$$z_1 \mid \mu_3, \tau_1 \sim \text{Normal}(\mu_3, \tau_1^{-1})$$

Hierarchical models

- In most hierarchical models, there are conditional distributions that do not have nice analytic forms.
- Using the posterior density over all the variables and removing factors that do not involve the variable we want to simulate, we still get a function proportional to its conditional density.
- We may update this variable using another type of Metropolis Hastings proposal (like random walk).
- Note: It may often be better to work with the logged posterior density: Then one may remove additive terms not involving the variable one wants to simulate over.

MSA101/MVE187 2020 Lecture 7.4
Slice sampling

Petter Mostad

Chalmers University

September 17, 2020

The slice sampler

- Idea: Do Gibbs sampling from "the area under the density curve".
- More formally, given density $f_x(x)$, simulate from the joint density

$$f(x,u) = I(0 < u < f_x(x))$$

- Works even if the density f_x is known only up to a constant.
- The challenge is to simulate x uniformly on $\{x : u < f_x(x)\}$. This is most easily done if for example f_x is a decreasing function, so that it is invertible.
- Example: Simulate from the density $\pi(x) = \frac{1}{2}\exp\left(-\sqrt{x}\right)$. We iterate between the following steps:
  - Given an x value, simulate $u \sim \text{Uniform}\left(0, \frac{1}{2}\exp\left(-\sqrt{x}\right)\right)$.
  - Given a u value simulate $x \sim \text{Uniform}\left(0, (\log(2u))^2\right)$: Note that $u = \frac{1}{2}\exp\left(-\sqrt{x}\right)$ if and only if $x = (\log(2u))^2$ and that $\pi(x)$ is decreasing as a function of x.

Generalization to more dimensions

- The theory can easily be extended to more dimensions: When we want to simulate from the density

$$f(x) = \prod_{i=1}^{n} g_i(x)$$

we can define the joint density

$$h(x, u_1, \ldots, u_n) = \prod_{i=1}^{n} I\left(0 < u_i < g_i(x)\right)$$

- We see that the marginal density for x is $f(x)$.
- We simulate from the joint density using Gibbs sampling. This is very easy for the variables $u_1, \ldots, u_n$.
- The conditional distribution of x given $u_1, \ldots, u_n$ is the uniform distribution on the set

$$\cap_{i=1}^{n} \{x : u_i < g_i(x)\}.$$

If it is easy to compute this set, slice sampling works well. One example: If all the $g_i(x)$ functions are decreasing and invertible.

Example: The Challenger disaster

- The goal is to compute the probability that a space shuttle "o-ring" fails at a specific temperature. (An o-ring failing because of cold weather was the cause of the Challenger space shuttle disaster).
- Data $(x_1, y_1), \ldots, (x_n, y_n)$ where x_i denotes the temperature (in Farenheit) and y_i is 1 if there is a failure, 0 otherwise.
- We use a logistic regression model:

$$y_i \sim \text{Bernoulli}(p(x_i)) \qquad p(x_i) = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)}.$$

- The posterior becomes (using flat priors on a and b)

$$\begin{aligned} \pi(a, b \mid \text{data}) &\propto \prod_{i=1}^{n} \left(\frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(a + bx_i)} \right)^{1 - y_i} \\ &= \prod_{i=1}^{n} \frac{\exp(a + bx_i)^{y_i}}{1 + \exp(a + bx_i)} \end{aligned}$$

Example continued

- Simulate from posterior for parameters (a, b) using slice sampling:
  - For $i = 1, \ldots, n$, simulate $u_i \sim \text{Uniform}\left[0, \frac{\exp(a+bx_i)^{y_i}}{1+\exp(a+bx_i)}\right]$.
  - Simulate (a, b) uniformly on set satisfying, for all i, $u_i < \frac{\exp(a+bx_i)^{y_i}}{1+\exp(a+bx_i)}$.
- Corresponds to $a + bx_i > \log(u_i/(1-u_i))$ for i with $y_i = 1$, and $a + bx_i < \log((1-u_i)/u_i)$ for i with $y_i = 0$.
- To simulate (a, b) uniformly on this set, we first simulate a with

$$a \sim \text{Uniform}\left[\max_{y_i=1}\left(\log\frac{u_i}{1-u_i} - bx_i\right), min_{y_i=0}\left(\log\frac{1-u_i}{u_i} - bx_i\right)\right]$$

- Then for b, we need to be more careful, simulating b uniformly in the interval of numbers
  - Greater than $\left(\log\frac{u_i}{1-u_i} - a\right)/x_i$ for i with $y_i = 1$ and $x_i > 0$.
  - Smaller than $\left(\log\frac{u_i}{1-u_i} - a\right)/x_i$ for i with $y_i = 1$ and $x_i < 0$.
  - Smaller than $\left(\log\frac{1-u_i}{u_i} - a\right)/x_i$ for i with $y_i = 0$ and $x_i > 0$.
  - Greater than $\left(\log\frac{1-u_i}{u_i} - a\right)/x_i$ for i with $y_i = 0$ and $x_i < 0$.

Example continued

- This is actually Example 7.11 in RC, but the book contains some errors:
  - Confusion beween (a, b) and (α, β)
  - Second and fourth formulas on page 220 are wrong.
  - No need to use a prior for a and b to get this to work; use centering instead.
- Note that a and b are highly correlated in the posterior if we implement the code directly. Much improved convergence and accuracy is obtained by *centering* the data: Subtracting the average value from the temperature values, performing the analysis, and then adding back the average value.