MSA101/MVE187 2017 Lecture 9

Petter Mostad

Chalmers University

September 26, 2017

Some superficial notes from RC chapter 4 sections 1-3

In the context of Monte Carlo integration using IID samples:

- We have looked at how to obtain a "confidence band" using cumulative averages and cumulative computations of the sample variance. (Example 3.3. Figure 3.3)
- A more stable "confidence band" can be produced by sampling k parallell chains. (Example 4.1. Figure 4.1)
- As we often only know the posterior density up to a constant, computing a posterior expectation may involve computing the quotient of two approximations of integrals. (Example 4.2). There are ways to obtain adjusted estimates for the accuracy of the estimates of such quotients.

Multivariate normal approximations

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

$$\pi(\theta) \propto_\theta \exp(h(\theta))$$

for some function h. 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

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

If we integrate both sides with respect to θ *(and interpret the local approximation above as a global approximation), we get that the integration constant for* $\pi(\theta)$ *is approximately equal to*

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

Examples

- Example 6.4: Target density Normal$(0, 1)$, proposal function is the uniform distribution on $[x - \delta, x + \delta]$.
  - The only parameter in the method is δ.
  - We see that too small or too large values for δ gives slow convergence of the Markov chain.
- Example 6.5: The likelihood is a mixture:

$$\frac{1}{4}\,\text{Normal}(\mu_1, 1) + \frac{3}{4}\,\text{Normal}(\mu_2, 1)$$

- We simulate 400 data values using $\mu_1 = 0$, and $\mu_2 = 2.5$.
- With a prior for (μ_1, μ_2) that is uniform on $[-2, 5] \times [-2, 5]$ we get a posterior density as in Figure 6.8.
- R-code for log-likelihood function on page 128.
- R-code for simulation from posterior on page 184.
- Result very dependent on "scale" parameter. Can you think of alternative approaches?

The Langevin algorithm

- The idea: Use not only the density value at $X^{(t)}$ but also the gradient of the density at that point to make a smart proposal Y^t.
- Concrete proposal function

$$Y^t = X^{(t)} + \frac{\sigma^2}{2} \nabla \log f(X^{(t)}) + \sigma \epsilon_t$$

- Nice to implement when formulas for the gradient can be computed analytically.
- BUT: In many cases, the convergence of the Markov chain is not improved: (One can get too easily stuck at a mode). Example 6.7.

Acceptance rates

- In a number of cases, a high acceptance rate gives a better sample.
- Example 6.9: Using a double-exponential independent proposal to simulate from Normal$(0, 1)$.
- However, maximizing the acceptance rate does not necessarily improve the sample when you don't have independent proposals, as it might also increase the autocorrelation in the sample.
- Example 6.10

Missing data

- Idea: Simulate the missing data given the parameter, and then simulate the parameters given the missing data: Gibbs sampling idea!
- Example: Censored data, for example in survival analysis: 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 density.
- Simulating alternatively the missing data and distribution for the parameters given *all* data may be easier than dealing with the likelihood above.
- Example 7.6: A Normal$(\theta, 1)$ model with data truncated at a.

Augmented data

(or latent variables)

- Idea: Sometimes the model had been much simpler to handle if we had observed certain parameters. So: Pretend that these are missing data!
- Example 7.7: The model is the multinomial distribution

$$\mathcal{M}_4(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4})$$

- The likelihood for θ is not easy to deal with.
- We extend the data (x_1, x_2, x_3, x_4) with a latent variable z, so that

$$(x_1 - z, z, x_2, x_3, x_4) \sim \mathcal{M}_5(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4})$$

- What is the posterior probability of θ given the extended data and a Beta prior?
- What is the conditional probability of z given θ and the actual data?

Mixture models

- Assume likelihood has form

$$\pi(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i \mid \xi_j)$$

  where $\theta = (\xi_1, \ldots, \xi_k)$ are the parameters.
- Improved model: Add latent variables $Z = (Z_1, \ldots, Z_n)$, where $Z_i = j$ indicates the distribution x_i comes from:

$$x_i \mid z_i \sim f(x_i \mid \xi_{z_i}) \text{ and } z_i \mid \text{Multinomial}(p_1, \ldots, p_k)$$

- The full conditional $\pi(Z_i \mid x_i, \theta)$ can be computed as the probabilities that x_i belongs to the various distributions $f(x_i \mid \xi_j)$, when the parameters θ are given: $P(Z_i = j \mid x, \theta) \propto p_j f(x_i \mid \xi_j)$.
- The full conditional $\pi(\theta \mid x_1, \ldots, x_n, Z_1, \ldots, Z_n)$ can be much easier to handle than the original likelihood: No sums occur.