Lecture 06: Bayesian Inference Lecture # 3

- Overview: Outside of very simple Bayesian models, it is hard to rely on doing things analytically. However, as we will see today, in most cases it is posisble to sample from posterior distributions using MCMC. These samples can then be used to approximate posterior quantities of interest such as posterior means, medians, intervals etc.

- Recap: General Gibbs Sampling Algorithm:

  1. Start at $(x_1^{(0)}, x_2^{(0)}, \ldots, x_p^{(0)}))$ and set $t = 0$.
  2. Sample $x_1^{(t+1)}$ from $p(x_1|x_2^{(t)}, \ldots, x_p^{(t)})$
  3. Sample $x_2^{(t+1)}$ from $p(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$
  4. $(\ldots$ Sample $x_k^{(t+1)}$ from $p(x_k|x_{1:(k-1)}^{(t+1)}, x_{(k+1):p}^{(t)}) \ldots)$
  5. Sample $x_p^{(t+1)}$ from $p(x_p|x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$
  6. Increment $t \mapsto t+1$ and return to 2.

  Let $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_p^{(t)})$ i.e., $x^{(t)}$ is the vector of all components at time t. The Gibbs sampler generates a sequence:

  $$x^{(1)}, x^{(2)}, x^{(3)}, \ldots.$$

  It is straightforward to see that the sequence forms a Markov Chain (next sample only depends on the current state). To see what the chain converges to, we need some general Markov Chain theory.

- Recap: Markov Chains

  – For ergodic Markov chains, the time spent in each state over the long-run converges to a constant value. In other words, the long-run time average of the chain converges to a stationary distribution π that satisfies the following:

  $$\pi = \pi P \quad \text{(discrete)}, \qquad \pi(y) = \int \pi(x)p(x,y)dx, \quad \forall\ y \quad \text{(continuous)}$$

  Ergodicity gives:

  $$\mathbb{P}(X^{(t)} = j) \longrightarrow \pi_j, \qquad \text{as } t \to \infty, \quad \forall\ j.$$

  Time-averaged state of chain converges to the stationary distribution (regardless of the starting point!).

- Applications of Markov Chain Theory

  – Homework: prove that Gibbs sampler has stationary distribution $p(x_1, \ldots, x_p)$. What is the transition kernel?

  – In a Bayesian context, suppose we can construct a Markov Chain (e.g., a Gibbs sampler) to obtain samples from $p(\theta|y)$. How can we estimate, say, $\mathbb{E}[\theta|y]$ (the posterior mean)?

Theorem: Let $\theta^{(1)}, \theta^{(2)}, \ldots$ be an ergodic Markov Chain with stationary distribution π, g be a bounded function and $\mathbb{E}_\pi\left[g(\theta)\right] < \infty$. Then with probability 1:

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta^{(i)}) \to \int g(\theta)\pi(\theta)d\theta = \mathbb{E}_\pi\left[g(\theta)\right].$$

as $M \to \infty$. This generalizes the earlier Monte Carlo integration result to allow for *dependent* samples. This is crucial, as independent samples are often impossible to obtain.

- The Gibbs sampler provides a simple way to construct an ergodic Markov chain with stationary distribution $p(\theta|y)$.

Example: Let:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

then standard normal theory tells us that:

$$\begin{aligned} X_1|X_2 = x_2 &\sim N\left(\mu_1 + \rho\sigma_1\frac{(x_2-\mu_2)}{\sigma_2}, (1-\rho^2)\sigma_1^2\right) \\ X_2|X_1 = x_1 &\sim N\left(\mu_2 + \rho\sigma_2\frac{(x_1-\mu_1)}{\sigma_1}, (1-\rho^2)\sigma_2^2\right). \end{aligned}$$

These results lead to a simple Gibbs sampler (see code on course website).

- A Markov Chain with transition density $p(x, y)$ is said to be *reversible* if:

$$\pi(x)p(x,y) = \pi(y)p(y,x), \qquad \forall\ x, y.$$

This is also known as the *detailed balance* condition. For general transition kernels this condition ensures that the MC has stationary distribution π.

- There are many methods to construct ergodic Markov chains with a specified stationary distrbution, the most famous is the *Metropolis-Hastings algorithm.*

1. Select the starting value $\theta^{(0)}$ and set $t = 0$.

2. Given the current state $\theta^{(t)}$, propose a new value θ^* from a proposal density $p(\theta^{(t)}, \cdot)$.

3. Accept the proposed value with probability α, where:

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)p(\theta^*, \theta^{(t)})}{\pi(\theta^{(t)})p(\theta^{(t)}, \theta^*)}\right\}.$$

If accepted, set $\theta^{(t+1)} = \theta^*$, else set $\theta^{(t+1)} = \theta^{(t)}$.

4. Increment $t \mapsto t + 1$ and return to step 2.

Example: Let $X_i \sim N(e^\theta, 1)$ with prior $\theta \sim N(0, 1)$. The posterior distribution is seen to be:

$$p(\theta|x) \propto \exp\left\{-\frac{1}{2}\left[\theta^2 + \sum_{i=1}^{n}\left(x_i - e^\theta\right)^2\right]\right\}.$$

We want to obtain samples from $p(\theta|x)$, and will use the Metropolis-Hastings to do so.

1. Select the starting value $\theta^{(0)} = \log\left(\min\{0.1, \bar{x}\}\right)$ and set $t = 0$.

2. Given the current state $\theta^{(t)}$, propose a new value θ^* from $N(\theta^{(t)}, v^2)$ where v^2 (the proposal variance) is a constant.

3. Sample $U \sim \text{Unif}(0, 1)$. If:

$$\begin{aligned}\log(U) &< \log \pi(\theta^*) + \log p(\theta^*, \theta^{(t)}) - \log \pi(\theta^{(t)}) - \log p(\theta^{(t)}, \theta^*) \\ &< \frac{1}{2}\left((\theta^{(t)})^2 + \sum_{i=1}^{n}(x_i - e^{\theta^{(t)}})^2\right) - \frac{1}{2}\left((\theta^*)^2 + \sum_{i=1}^{n}(x_i - e^{\theta^*})^2\right),\end{aligned}$$

then set $\theta^{(t+1)} = \theta^*$, else set $\theta^{(t+1)} = \theta^{(t)}$.

4. Increment $t \mapsto t + 1$ and return to step 2.

$$\theta^* \sim N(\theta^{(t)}, v^2)$$

This is called a random walk proposal.
What is $p(\theta^{(t)}, \theta^*)$? $[dnorm(\theta^*, mean = \theta^{(t)}, sd = v)]$.
What is $p(\theta^{(*)}, \theta^t)$? $[dnorm(\theta^t, mean = \theta^{(*)}, sd = v)]$.
Here we have a symmetric proposal, this two equals. $\theta^{(t)} = 10.0$, $\pi(\theta^t)$=0.60, $\theta^{(*)}$=12.5, $\pi(\theta^*)$=0.12. So $\alpha = 0.2$.
Generate $U \sim \text{Unif}(0, 1)$, if $U < 0.2$, set $\theta^{(t+1)} = \theta^*$, else set $\theta^{(t+1)} = \theta^{(t)}$.
Notes:

- We only know $p(\theta|x)$ up to proportionality: does it matter?
- Here the proposal distribution is symmetric: $p(\theta^{(t)}, \theta^*) = p(\theta^{(t)}, \theta)$ so the proposal terms cancel! This special case is called the *Metropolis algorithm.*
- *Always* compute posterior densities on the log-scale!
- By our theory, the samples $\left\{\theta^{(1)}, \theta^{(2)}, \ldots\right\}$ will converge to the stationary distribution $p(\theta|y)$. In practice, we usually want to throw away an initial *burnin* period of the samples. For example, if we collect $11,000$ samples, then we might throw away the first 1000 and keep the next $10,000$.
- How to select the proposal variance, v^2? Trial and error! In the normal setting, theory tells us that the optimal acceptance rate for the MH algorithm is between roughly 30% and 60%.
- It is possible to tweak the variance, v^2 within the MCMC run, but you must not change v^2 once the burnin period has been completed. If v continues to change the transition kernel is not constant and the convergence results are not guaranteed. This strategy of constantly adapting the transition kernel is called *Adaptive MCMC* and is an area of active research (Moral: be careful, even simple adaptive schemes can be shown to fail!).

Now we have the posterior samples, we can compute things such as posterior means, posterior intervals etc. Remaining questions:

- The samples from the posterior distribution are dependent: do they form a reliable sample from the posterior distribution?

- **MCMC Diagnostics** All of the following diagnostics are available in the `coda` package within `R`. In **Python**, see the module `pymc` for convergence diagnostics (not all listed below are available).

- **Effective Sample Size (ESS):** Quantifies the approximate number of independent samples that your dependent samples correspond to. For example, if the ESS for your 10,000 dependent samples is 100, this means that, you are using the equivalent of approximately 100 independent samples to compute your posterior quantities of interest. If possible (and depending on the accuracy required), it is recommend to always strive for an ESS in the thousands.
- **Gelman-Rubin $\hat{R}$:** The idea behind the Gelman-Rubin $\hat{R}$ statistic is to run multiple independent MCMC chains starting from different values. Ideally all of the chains will converge to the same stationary distribution and thus look very similar in practice. The $\hat{R}$ statistic quantifies the 'similarity' of the multiple chains (it is recommended to run at least 3 chains). $\hat{R}$ values much larger than 1.0 are indicative of a lack of convergence, although values close to 1.0 do not necessarily guarantee convergence! Implemented in `R` in `gelman.diag`.
- **Geweke Diagnostic:** Test that compares properties of different segments of the chain. If the chain has reached stationarity and mixed well, then the segments should be 'similar', if not then it can be indicative of a lack of convergence (`geweke.diag`).
- **Hiedelberger-Welch Diagnostic:** Half-width test based on the accuracy of the estimate of the mean of the target distribution (`heidel.diag`)

- **Validating MCMC Code for Bayesian Inference:**

While convergence diagnostics are helpful to some extent, they never provide a complete picture of whether the MCMC algorithm has correctly explored the full posterior distribution. For example, it is possible to be stuck in a local mode, and for the convergence diagnostics to look just fine, as illustrated below.

Example: Mixture normal. Sample with MH.

Fortunately, we can use some theory to construct a validation simulation for any Bayesian model. The idea:

- For $i = 1, 2, \ldots, B$:
- Simulate $\theta^{(i)}$ from the prior $p(\theta)$
- Simulate a dataset $y^{(i)}$ from the model $p(y|\theta^{(i)})$
- Obtain posterior percentiles for $\theta^{(i)}$
- For each posterior percentile, record whether the percentile was greater than the true value $\theta^{(i)}$

The theory: For the p^{th} percentile, approximately p% of the posterior intervals should cover the truth. For example, if we simulate 200 datasets, then approximately 50% of the posterior medians should be greater than the true value of the parameter for the corresponding dataset. Similarly, approximately 95% of central 95% posterior credible intervals should contain the true value (which will be different for each dataset). The methodology presented in Cook, Gelman & Rubin (2006) is a generalization of this approach.

You will be doing this for homework 1. (Ex: Prove the coverage property)