Gibbs Sampling

Adam Smith, Spring 2017

1 Introduction

Gibbs sampling is a widely used method for generating random samples from a joint distribution when this distribution does not belong to a known parametric family. Gibbs sampling is especially useful for Bayesian inference, where the goal is to sample from unnamed and/or high-dimensional posterior distributions $p(\theta|y)$.

Rather than attempt to sample directly from the the joint posterior of $\theta|y$, the Gibbs sampler only considers sampling from a sequence of conditional distributions $p(\theta_j|\theta_{-j}, y)$ for each $j = 1, \ldots, k$. Thus, the only requirement for using a Gibbs sampler is that the full conditional distribution of each model parameter (or block of parameters) is known and easy to sample from.

To understand why the full conditional distributions fully summarize the joint distribution, we first review the Hammersley-Clifford Theorem.

2 Hammersley-Clifford Theorem

The Hammersley-Clifford Theorem proves that we can write out a joint distribution $p(\theta_1, ..., \theta_k)$ in terms of only the full conditional distributions $p(\theta_j|\theta_{-j})$ for $j = 1, ..., k$. For example, consider the bivariate joint distribution $p(\theta_1, \theta_2)$. It follows that

$$
\begin{aligned}
p(\theta_1, \theta_2) &= p(\theta_2|\theta_1) \times p(\theta_1) \\
&= p(\theta_2|\theta_1) \times \frac{1}{\frac{1}{p(\theta_1)}} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_2)}{p(\theta_1)} d\theta_2} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_1,\theta_2)/p(\theta_1)}{p(\theta_1,\theta_2)/p(\theta_2)} d\theta_2} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_2|\theta_1)}{p(\theta_1|\theta_2)} d\theta_2}.
\end{aligned}
$$

Hence, the set of full conditional distributions, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, summarize all information in the joint distribution.

3 Full Conditional Distributions

A full conditional distribution is the distribution of a model parameter conditional on all the other parameters: $p(\theta_j|\theta_{-j})$. To find the full conditional distributions of an k-dimensional joint distribution $p(\theta_1, ..., \theta_k)$, start by picking a single parameter, or block of parameters, θ_j. Next, write out the full joint distribution, and drop anything that does not depend on θ_j. Finally, match what is left to the kernel of a known distribution in order to determine the form of $p(\theta_j|\theta_{-j})$.

4 The Gibbs Sampling Algorithm

We outline the basic steps of the algorithm which samples from the joint posterior distribution $p(\theta|y)$ where $\theta = (\theta_1, ..., \theta_n)$.

For $r = 1, \ldots, R$

1. sample $\theta_1^{[r]} \sim p(\theta_1|\theta_1^{[r-1]}, ..., \theta_k^{[r-1]}, y)$
2. sample $\theta_2^{[r]} \sim p(\theta_2|\theta_1^{[r]}, \theta_3^{[r-1]}, ..., \theta_k^{[r-1]}, y)$
3. sample $\theta_3^{[r]} \sim p(\theta_3|\theta_1^{[r]}, \theta_2^{[r]}, \theta_4^{[r-1]}, ..., \theta_k^{[r-1]}, y)$

$\vdots$

k. sample $\theta_n^{[r]} \sim p(\theta_n|\theta_1^{[r]}, \theta_2^{[r]}, ..., \theta_{k-1}^{[r]}, y)$

These steps will result in a Markov chain with R samples of $(\theta_1, ..., \theta_n)$, which will converge to the stationary distribution that is the posterior distribution $p(\theta|y)$. Note that the order in which the elements of θ are drawn in step (2) does not matter.

References

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005), *Bayesian Statistics and Marketing.* New York: John Wiley and Sons.

Casella, G. and E. George (1992), 'Explaining the Gibbs Sampler', *The American Statistician*, *46*, 167-174.