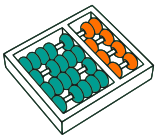

Markov Chain Monte Carlo

Lesson No. 12

Alana de Santana Correia [*]

1 Introduction

Probabilistic inference involves estimating an expected value or density using a probabilistic model. However, directly inferring values is not treatable with probabilistic models and, instead, approximation methods are useful.

Markov chain Monte Carlo sampling (MCMC) provides a class of algorithms for systematic random sampling of high-dimensional probability distributions. Unlike Monte Carlo sampling, which can be intractable for high-dimensional probabilistic models, the MCMC provides an alternative approach where the next sample depends on the current sample. The two most common approaches to MCMC are **Gibbs sampling** and the more general **Metropolis-Hastings algorithm**. These two approaches are in the following sections of this summary.

2 Markov Chain Monte Carlo (MCMC)

MCMC methods are an alternative to non-iterative sampling methods for complex problems. The basic idea is to build a Markov chain in the state space θ, whose stationary distribution is the target density of interest (this may be a prior or a posterior). This section presents the most used MCMC methods, the Gibbs sampling, and the Metropolis-Hastings algorithm. The basic idea is to simulate a random walk in the θ space that converges to a stationary distribution, which is the distribution of interest in the problem.

A Markov chain is a stochastic process $\{X_0, X_1, \ldots\}$ such that the distribution of X_t given all previous values $X_0, \ldots, X_{t-1}$ depends only on X_{t-1}. Mathematically,

$$P(X_t \in A | X_0, \ldots, X_{t-1}) = P(X_t \in A | X_{t-1}) \tag{1}$$

for any A subset.

MCMC methods also require the chain to be:

- homogeneous, i.e. the probabilities of transition from one state to another are invariant.
- irreducible, i.e. each state can be reached from any other in a finite number of iterations.
- aperiodic, i.e. there are no absorbent states.

It is important to note that the algorithms in section § 2.1 and § 2.2 satisfy all of these conditions.

2.1 Metropolis-Hastings

Metropolis-Hastings algorithms use the same idea as rejection methods: a value is generated from an auxiliary distribution and accepted with a given probability. This correction mechanism guarantees the convergence of the chain for the stationary distribution, which in this case, is the posterior distribution.

Suppose the chain is in the state θ and a value θ' is generated from a proposed distribution $q(\cdot|\theta)$. The proposed distribution may depend on the current state of the chain, for example $q(\cdot|\theta)$ can be a normal distribution centered on θ. The new value θ' is accepted with probability:

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\, q(\theta|\theta')}{\pi(\theta)\, q(\theta'|\theta)}\right) \tag{2}$$

[*] RA: 229999 - a229999@dac.unicamp.br

where π is the distribution of interest. Besides, it is only necessary to know π partially.
The following steps define the Metropolis-Hastings algorithm:

- Initialize the iteration counter $t = 0$ and specify an initial value $\theta^{(0)}$.
- Generate a new value $\theta^{'}$ from the distribution $q(\cdot|\theta)$.
- Calculate the probability of acceptance $\alpha(\theta,\theta^{'})$ and generate $u \sim U(0,1)$.
- If $u \leq \alpha$ accept the new value and $\theta^{(t+1)} = \theta^{'}$, otherwise reject and $\theta^{(t+1)} = \theta$.
- Increment the counter t for $t+1$ and go back to step 2.

The independence sampler is a Metropolis-Hastings algorithm whose proposal distribution does not depend on the current state of the chain, i.e., $q(\theta,\theta^{'}) = q(\theta^{'})$. In general, $q(\cdot)$ should be a good approximation of $\pi(\cdot)$, but it is safest if $q(\cdot)$ is heavier-tailed than $\pi(\cdot)$.

The Metropolis algorithm considers only symmetric proposals, i.e., $q(\theta,\theta^{'}) = q(\theta^{'},\theta)$ for all values of θ and $\theta^{'}$, and the acceptance probability reduces to

$$\alpha(\theta,\theta^{'}) = \min\left(1, \frac{\pi(\theta^{'})}{\pi(\theta)}\right) \tag{3}$$

A special important case is the random-walk Metropolis for which $q(\theta,\theta^{'}) = q(|\theta - \theta^{'}|)$, so that the probability of generating a move from θ to $\theta^{'}$ depends only on the distance between them. Using a proposal distribution with variance σ^2, very small values of σ^2 will lead to small jumps which are almost all accepted but it will be difficult to traverse the whole parameter space and it will take many iterations to converge. On the other hand, large values of σ^2 will lead to an excessively high rejection rate since the proposed values are likely to fall in the tails of the posterior distribution.

2.2 Gibbs Sampling

Gibbs sampling is an MCMC method where the transition kernel is formed by the full conditional distributions, $\pi(\theta_i|\theta_{-i})$, where $\theta_{-i} = (\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_d)'$. In general, each one of the components θ_i can be either uni- or multi-dimensional. So, the full conditional distribution is the distribution of the $i-th$ component of θ conditioning on all the remaining components, and it is derived from the joint distribution as follows:

$$\pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta)}{\int \pi(\theta) d\theta_i} \tag{4}$$

If generation schemes to draw a sample directly from $\pi(\theta)$ are costly, complicated or simply unavailable but the full conditional distributions are completely known and can be sampled from, then Gibbs sampling proceeds as follows:

- Initialize the iteration counter of the chain $t = 1$ and set initial values $\theta^{(0)} = (\theta_1^{(0)},\ldots,\theta_d^{(0)})'$.
- Obtain a new value of $\theta^{(t)}$ from $\theta^{(t-1)}$ through successive generation of values:

$$\begin{gathered}
\theta_1^{(t)} \sim \pi(\theta_1|\theta_2^{(t-1)},\theta_3^{(t-1)},\ldots,\theta_d^{(t-1)}) \\
\theta_2^{(t)} \sim \pi(\theta_2|\theta_1^{(t)},\theta_3^{(t-1)},\ldots,\theta_d^{(t-1)}) \\
\vdots \\
\theta_d^{(t)} \sim \pi(\theta_d|\theta_1^{(t)},\theta_2^{(t)},\ldots,\theta_{d-1}^{(t)})
\end{gathered}$$

- Increment the counter t to $t+1$ and return to step 2 until convergence is reached.

So, each iteration is completed after d moves along the coordinates axes of the components of θ. When convergence is reached, the resulting value θ is a draw from $\pi(\theta)$. It is worth noting that, even in a high-dimensional problem, all of the simulations may be univariate, which is usually a computational advantage. However, the Gibbs sampler does not apply to problems where the number of parameters varies because of the lack of irreducibility of the resulting chain. When the length of θ is not fixed and its elements need not have a fixed interpretation across all models, to resample some components conditional on the remainder would rarely be meaninful. Note also that the Gibbs sampler is a special case of the Metropolis-Hastings algorithm, in which individual elements of θ are updated one at a time (or in blocks).

3 Conclusion

This summary presents the necessary steps of the Gibbs sampling method and the Metropolis algorithm. Besides, these approaches are a good sampling alternative in situations where the simple Monte Carlo method does not fit (especially in problems with high dimensions). These methods differ from previous methods of sampling because they are iterative, and it is possible to adapt them to diverse and challenging issues.