ADVANCED BIOSTATISTICS ABSTAT18

Gibbs Sampling

Lisete Sousa

Department of Statistics and Operations Research | CEAUL
Faculty of Sciences of Lisbon University

IGC, April 3rd - 6th, 2018

Contents

- Introduction
- The Gibbs Sampler

Introduction

In the continuous case, the integration operation plays a fundamental role in Bayesian Statistics.

Several numerical approximation strategies have been suggested, such as,

- Laplace Approximation,
- Sampling-importance-resampling (SIR),
- Markov Chain Monte Carlo Methods (MCMC).

The origin of MCMC

The original Monte Carlo approach was a method developed by physicists to use random number generation to compute integrals:

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_{p(x)}\left[f(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

by decomposing $h(x)$ into the production of a function $f(x)$ and a probability density function $p(x)$, defined over the interval (a, b).

One problem with applying Monte Carlo integration is in obtaining samples from some complex probability distribution. The attempts to solve this problem are the roots of MCMC methods.

Another motivation for the development of MCMC was the calculation of the posterior distribution necessary for Bayesian approaches. This often requires the integration of high-dimensional functions, which can be computationally very difficult. In this case, one would have the approximation:

$$p(\theta) = \int f(x|\theta)p(\theta)d\theta \approx \frac{1}{n}\sum_{i=1}^{n} f(x|\theta_i)$$

The Gibbs Sampler

Geman and Geman (1984) introduced the GIBBS sampler as a way of simulating from high-dimensional complex distributions arising in image restoration.

Gelfand and Smith (1990) showed how the algorithm can be used to simulate from posterior distributions, and hence how to be used to solve problems in Bayesian Statistics.

In this situation, the algorithm is based on the fact that, if the joint distribution $p(\boldsymbol{\theta}|\mathbf{x})$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is positive over its entire domain, then it is uniquely determined by the k full conditional distributions (Besag, 1974)

$$p(\theta_i|\mathbf{x}, \boldsymbol{\theta}_{(-i)}), i = 1, \ldots, k,$$

where

$$\boldsymbol{\theta}_{(-i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots \theta_k).$$

The algorithm is then a Markovian updating scheme which requires sampling from these full conditional distributions as follows.

Suppose we are given an arbitrary set of initial values

$$\boldsymbol{\theta}^0 = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$$

then we proceed with the following iterative procedure:

- draw $\theta_1^{(1)}$ from $p(\theta_1 | \mathbf{x}, \theta_2^{(0)}, \dots, \theta_k^{(0)})$,
- draw $\theta_2^{(1)}$ from $p(\theta_2 | \mathbf{x}, \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$,
- $\dots$
- draw $\theta_k^{(1)}$ from $p(\theta_k | \mathbf{x}, \theta_1^{(1)}, \dots, \theta_{k-1}^{(1)})$.

This completes one iteration of the scheme and a transition from $\boldsymbol{\theta}^0$ to $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \dots, \theta_k^{(1)})$. Iteration of this cycle produces a sequence $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(t)}, \dots$, which is a realization of a Markov chain.

Checking Convergence

Since convergence usually occurs regardless of our starting point, we can usually pick any feasible (for example, picking starting draws that are in the parameter space) starting point.

However, the time it takes for the chain to converge varies depending on the starting point.

As a matter of practice, most people **throw out a certain number of the first draws**, known as the burn-in. This is to make our draws closer to the stationary distribution and less dependent on the starting point.

Lam, P., MCMC Methods: Gibbs Sampling and the Metropolis-Hastings Algorithm

However, it is unclear how much we should burn-in since our draws are all slightly dependent and we do not know exactly when convergence occurs.

In order to break the dependence between draws in the Markov chain, some have suggested only **keeping every dth draw of the chain**.

This is known as thinning.

Lam, P., MCMC Methods: Gibbs Sampling and the Metropolis-Hastings Algorithm

Example - Univariate Normal

Consider a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$, whose components are independent and normally distributed:

$$Y_i \frown N(\mu, \sigma^2),\ \mu \in \mathbb{R},\ i = 1, \ldots, n,\ \sigma > 0.$$

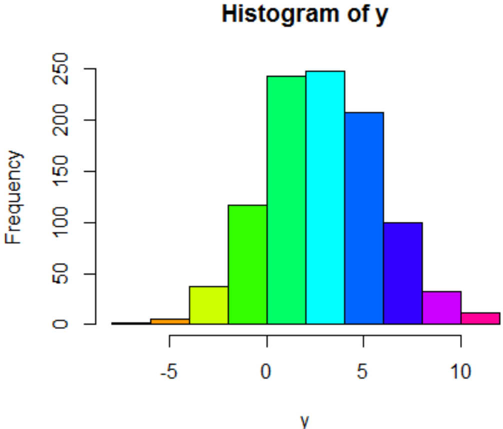
```
> n<-1000
> mu<-3
> sigma<-3
> sigma2<-sigma^2
> y<-rnorm(n,mu,sigma); summary(y)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -7.127   1.015   3.086   3.018   5.081  13.260
```

```
> head(y)

[1] 4.008252 7.581140 2.693658 3.924530 3.763089 1.599321

> hist(y,col=rainbow(10))
```



Considering a non-informative prior distribution, according to Jeffrey, we have:

$$p(\mu, \sigma^2) \propto \sigma^{-3}$$

and the posterior distributions for the parameters μ and $\tau = \frac{1}{\sigma^2}$ are:

$$\mu|\tau, \mathbf{y} \frown N\left(\overline{y}, \frac{1}{n\tau}\right)$$

$$\tau|\mathbf{y} \frown Gamma\left(\frac{n}{2}, \frac{\sum(y_i - \overline{y})^2}{2}\right)$$

$$\tau|\mu, \mathbf{y} \frown Gamma\left(\frac{n}{2}, \frac{\sum(y_i - \mu)^2}{2}\right)$$

Now it is possible to sampling from the full conditional distributions and

- draw $\mu^{(1)}$ from $p(\mu|\tau^{(0)}, \mathbf{y})$,
- draw $\tau^{(1)}$ from $p(\tau|\mu^{(1)}, \mathbf{y})$,
- and so on...

In order to simulate $\mu^{(1)}$, we have to initialize τ (represented by $\tau^{(0)}$) and calculate $\overline{y}$:

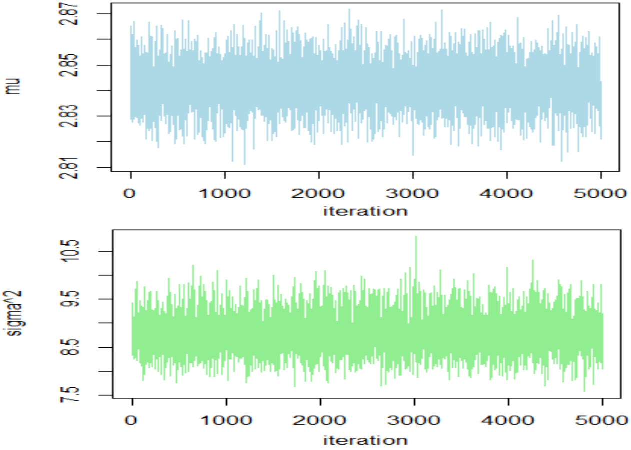
```
> m<-mean(y)
> tau<-rgamma(1,n/2,sum((y-m)^2)/2)
```

The Gibbs Sampler code, for 5000 simulations, is as follows:

```
> ns<-5000
> Mu<-Sigma2<-rep(0,ns) # store parameters' updates

> for (i in 1:ns) {
>   mu<- Mu[i]<-rnorm(1,m,1/(n*tau))
>   tau<-rgamma(1,n/2,sum((y-mu)^2)/2)
>   Sigma2[i]<-1/tau
> }
```

The convergence of both parameters is checked graphically:



By the trace, we can see that there is no need of burn-in, and the estimates of μ and σ^2 correspond to the mean of the simulated parametrs over the 5000 iterations:

```
> mean(Mu[1:ns])
[1] 2.739533
> sd(Mu[1:ns])
[1] 0.009119242
```

$\hat{\mu} = 2.74 \approx 3$

```
> mean(Sigma2[1:ns])
[1] 8.999175
> sd(Sigma2[1:ns])
[1] 0.4005737
```

$\hat{\sigma}^2 = 8.999 \approx 9$

Acknowledgements:

Antónia Turkman (DEIO – FCUL) and Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by us in previous courses.

Bibliography:

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pat. Anal. Mach. Intel.* 6, 721–741.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. New York: Springer.

Walsh, B. (2004). *Lecture Notes for EEB 581*, v. 26th April.