Bayesian Computation:
A brief introduction

Hedibert Freitas Lopes[1]
hedibert.org

[1]Professor of Statistics and Econometrics at Insper, São Paulo.

Outline

Monte Carlo: a toy example

Monte Carlo method: a bit of history

Monte Carlo integration

Monte Carlo simulation

Gibbs sampler

Metropolis-Hastings algorithm

References

Monte Carlo: a toy example

In what follows, we will see how to approximate integrals and sample from unknown distributions via the well known *Monte Carlo* method.

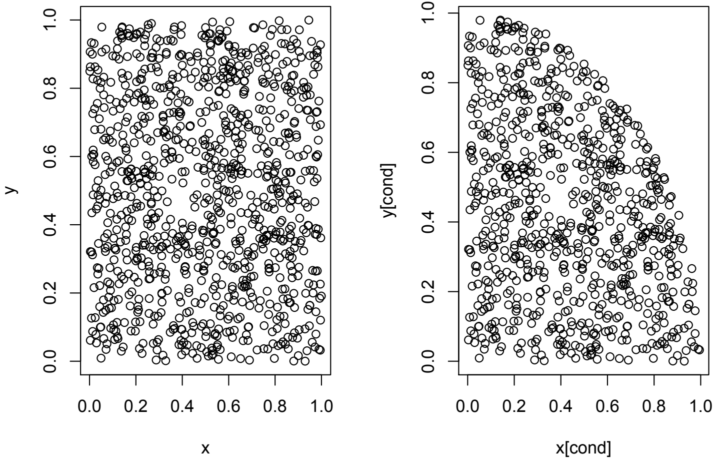
Let us think about calculating $\pi = 3.141593\ldots$

We could sample a bunch ($i = 1, \ldots, M$) of pairs (x_i, y_i) in the unit square $(0,1) \times (0,1)$ and compute the fraction α of those pairs satisfying the condition $x_i^2 + y_i^2 < 1$. In this case, $pi = 4\alpha$.

```
M = 1000
x = runif(M)
y = runif(M)
cond = (x^2+y^2)<1
par(mfrow=c(1,2))
plot(x,y)
plot(x[cond],y[cond])
pi.mc = 4*sum(cond)/M
```

$\pi_{mc} = 3.1292$

$$\frac{\pi}{4} = \int_0^1 \int_0^{\sqrt{1-x^2}} dydx$$

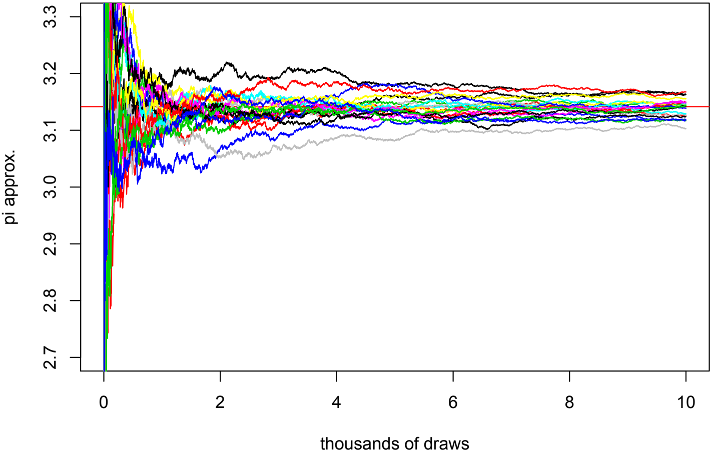


Monte Carlo: Let us play with M

```
set.seed(12345)
M = 20000
x = runif(M)
y = runif(M)
cond = (x^2+y^2)<1
pi.mc = 4*cumsum(cond)/(1:M)
plot(1:M/1000,pi.mc,ylim=c(2.7,3.3),type="l",
     xlab="thousands of draws",ylab="pi approx.")
abline(h=pi,col=2)

for (i in 1:20){
  x = runif(M)
  y = runif(M)
  cond = (x^2+y^2)<1
  pi.mc = 4*cumsum(cond)/(1:M)
  lines(1:M/1000,pi.mc,col=i)
}
```

MC error



MC in the 40s and 50s

Stan Ulam soon realized that computers could be used in this fashion to answer questions of neutron diffusion and mathematical physics;

He contacted John Von Neumann and they developed many Monte Carlo algorithms (importance sampling, rejection sampling, etc);

In the 1940s Nick Metropolis and Klari Von Neumann designed new controls for the state-of-the-art computer (ENIAC);

Metropolis and Ulam (1949) The Monte Carlo method. *Journal of the American Statistical Association*.
Metropolis *et al.* (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*.

70s and 80s

Metropolis-Hastings:
Hastings (1970) and his student Peskun (1973) showed that Metropolis and the more general Metropolis-Hastings algorithm are particular instances of a larger family of algorithms.

Gibbs sampler:

Besag (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems.
Geman and Geman (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.
Pearl (1987) Evidential reasoning using stochastic simulation.
Tanner and Wong (1987). The calculation of posterior distributions by data augmentation.
Gelfand and Smith (1990) Sampling-based approaches to calculating marginal densities.

A few references

- MC integration (Geweke, 1989)
- Rejection methods (Gilks and Wild, 1992)
- SIR (Smith and Gelfand, 1992)
- Metropolis-Hastings algorithm (Hastings, 1970)
- Simulated annealing (Metropolis *et al.*, 1953)
- Gibbs sampler (Gelfand and Smith, 1990)

Two main tasks

1. Compute high dimensional integrals:

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

2. Obtain

$$\textit{a sample } \{\theta_1, \ldots, \theta_n\} \textit{ from } \pi(\theta)$$

   when only

$$\textit{a sample } \{\tilde{\theta}_1, \ldots, \tilde{\theta}_m\} \textit{ from } q(\theta)$$

   is available.

$q(\theta)$ is known as the *proposal/auxiliary* density.

Bayes via MC

MC methods appear frequently, but not exclusively, in modern Bayesian statistics.

Posterior and predictive densities are hard to sample from:

$$\text{Posterior} \;:\; \pi(\theta) = \frac{f(x|\theta)p(\theta)}{f(x)}$$

$$\text{Predictive} \;:\; f(x) = \int f(x|\theta)p(\theta)d\theta$$

Other important integrals and/or functionals of the posterior and predictive densities are:

- Posterior modes: $\max_\theta \pi(\theta)$;
- Posterior moments: $E_\pi[g(\theta)]$;
- Density estimation: $\hat{\pi}(g(\theta))$;
- Bayes factors: $f(x|M_0)/f(x|M_1)$;
- Decision: $\max_d \int U(d,\theta)\pi(\theta)d\theta$.

Monte Carlo integration

The integrals

$$
\begin{aligned}
E_{p(\theta|x)}\{g(\theta)\} &= \int g(\theta)p(\theta|x)d\theta \\
E_{p(\theta)}\{p(x|\theta)\} &= \int p(x|\theta)p(\theta)d\theta = p(x)
\end{aligned}
$$

can be approximated, respectively, by

$$
\frac{1}{M}\sum_{i=1}^{M} g(\tilde{\theta}^{(i)}) \quad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M} p(x|\theta^{(i)}),
$$

where

$$
\{\theta^{(1)},\ldots,\theta^{(M)}\} \sim p(\theta|x) \quad \text{and} \quad \{\tilde{\theta}^{(1)},\ldots,\tilde{\theta}^{(M)}\} \sim p(\theta)
$$

Monte Carlo simulation via SIR

Sampling importance resampling (SIR) is a well-known MC tool that resamples draws from a candidate density $q(\cdot)$ to obtain draws from a target density $\pi(\cdot)$.

SIR Algorithm:

1. Draws $\{\theta^{(i)}\}_{i=1}^{M}$ from candidate density $q(\cdot)$
2. Compute resampling weights: $w^{(i)} \propto \pi(\theta^{(i)})/q(\theta^{(i)})$
3. Sample $\{\tilde{\theta}^{(j)}\}_{j=1}^{N}$ from $\{\theta^{(i)}\}_{i=1}^{M}$ with weights $\{w^{(i)}\}_{i=1}^{M}$.

Result: $\{\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(N)}\} \sim \pi(\theta)$

Bayesian bootstrap

When . . .

- the target density is the posterior $p(\theta|x)$, and
- the candidate density is the prior $p(\theta)$, then
- the weight is the likelihood $p(x|\theta)$:

$$w^{(i)} \propto \frac{p(\theta^{(i)})p(x|\theta^{(i)})}{p(\theta^{(i)})} = p(x|\theta^{(i)})$$

Note: We used $M = 10^6$ and $N = 0.1M$ in the previous two plots.

MC is expensive!

Exact solution

$$I = \int_{-\infty}^{\infty} \exp\{-0.5\theta^2\} d\theta = \sqrt{2\pi} = 2.506628275$$

MC is expensive!

Exact solution

$$I = \int_{-\infty}^{\infty} \exp\{-0.5\theta^2\} d\theta = \sqrt{2\pi} = 2.506628275$$

Let us assume that

$$I = \int_{-\infty}^{\infty} \exp\{-0.5\theta^2\} d\theta = \int_{-5}^{5} \exp\{-0.5\theta^2\} d\theta$$

MC is expensive!

Exact solution

$$I = \int_{-\infty}^{\infty} \exp\{-0.5\theta^2\} d\theta = \sqrt{2\pi} = 2.506628275$$

Let us assume that

$$I = \int_{-\infty}^{\infty} \exp\{-0.5\theta^2\} d\theta = \int_{-5}^{5} \exp\{-0.5\theta^2\} d\theta$$

Grid approximation (less than 0.01 seconds to run)
For $\theta_1 = -5$ $\theta_2 = -5 + \Delta$, $\ldots, \theta_{1001} = 5$ and $\Delta = 0.01$,

$$\hat{I}_{hist} = \sum_{i=1}^{1001} \exp\{-0.5\theta_i^2\}\Delta = 2.506626875$$

MC integration

It is easy to see that

$$\begin{aligned}\int_{-5}^{5} \exp\{-0.5\theta^2\} d\theta &= \int_{-5}^{5} 10 \exp\{-0.5\theta^2\} \frac{1}{10} d\theta \\ &= E_{U(-5,5)} \left[10 \exp\{-0.5\theta^2\}\right]\end{aligned}$$

MC integration

It is easy to see that

$$\int_{-5}^{5} \exp\{-0.5\theta^2\} d\theta = \int_{-5}^{5} 10 \exp\{-0.5\theta^2\} \frac{1}{10} d\theta$$
$$= E_{U(-5,5)} \left[10 \exp\{-0.5\theta^2\}\right]$$

Therefore, for $\{\theta^{(i)}\}_{i=1}^{M} \sim U(-5,5)$,

$$\hat{I}_{MC} = \frac{1}{M} \sum_{i=1}^{M} 10 \exp\{-0.5\theta^{(i)2}\}$$

M	$\hat{I}_{MC}$	MC error
1,000	2.505392026	0.10640840352
10,000	2.507470696	0.03380205878
100,000	2.506948869	0.01067906810

To improve on digital point, one needs M^2 draws!

It takes about 0.02 seconds to run.

Monte Carlo methods

- They are expensive.
- They are scalable.
- Readily available MC error bounds.

Why not simply use deterministic approximations?

Let us consider the bidimensional integral, for $\theta = (\theta_1, \theta_2, \theta_3)$,

$$I = \int \exp\{-0.5\theta'\theta\} d\theta = (2\pi)^{3/2} = 15.74960995$$

Grid approximation (20 seconds)

$$\hat{I}_{hist} = \sum_{i=1}^{1001} \sum_{j=1}^{1001} \sum_{k=1}^{1001} \exp\{-0.5(\theta_{1i}^2 + \theta_{2j}^2 + \theta_{3k}^2)\}\Delta^3 = 15.74958355$$

Monte Carlo approximation (0.02 seconds)

M	$\hat{I}_{MC}$	MC error
1,000	15.75223328	2.2768286659
10,000	15.72907660	0.7515860214
100,000	15.75368350	0.2236006764

Gibbs sampler

The Gibbs sampler is the most famous of the Markov chain Monte Carlo methods.

Roughly speaking, one can sample from the joint posterior of $(\theta_1, \theta_2, \theta_3)$

$$p(\theta_1, \theta_2, \theta_3 | y)$$

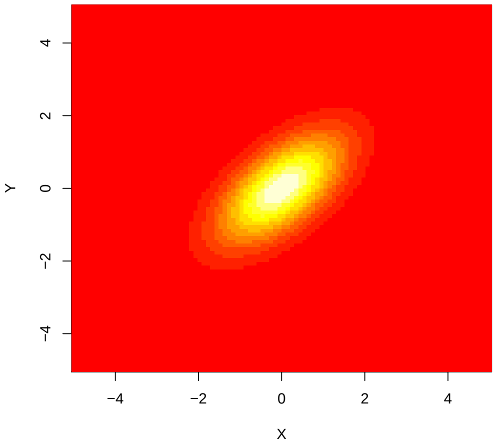
by iteratively sampling from the full conditional distributions

$$p(\theta_1 | \theta_2, \theta_3, y)$$
$$p(\theta_2 | \theta_1, \theta_3, y)$$
$$p(\theta_3 | \theta_1, \theta_1, y)$$

After a *warm up* phase, the draws will behave as coming from posterior distribution.

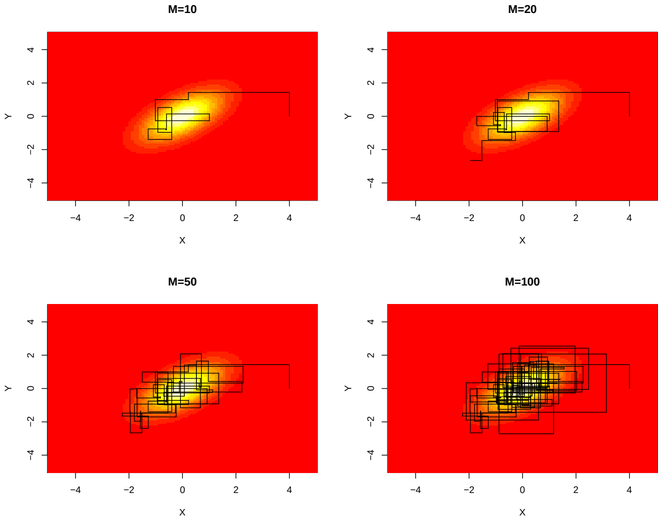
Taget distribution: bivariate normal with $\rho = 0.6$

$$p(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{x^2 - 2\rho xy - y^2}{2(1-\rho^2)} \right\}$$



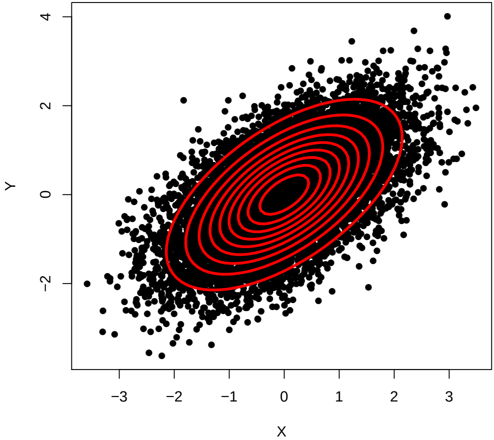
Full conditional distributions

Easy to see that $x|y \sim N(\rho y, 1-\rho^2)$ and $y|x \sim N(\rho x, 1-\rho^2)$.
Initial value: $x^{(0)} = 4$

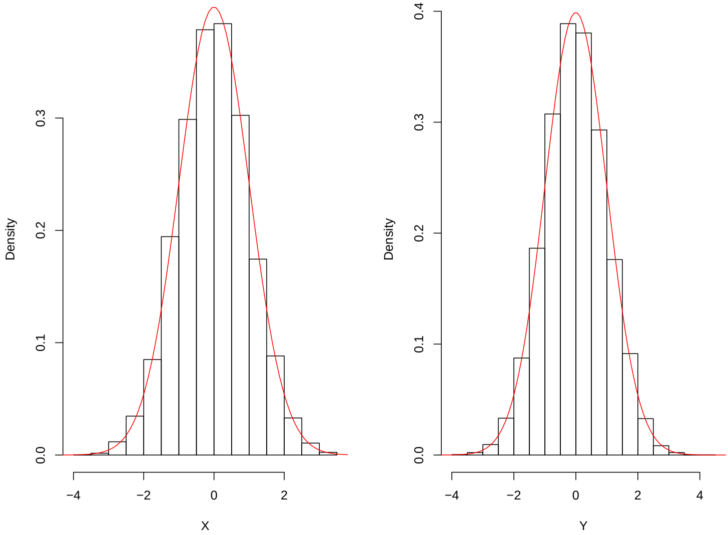


Posterior draws

Running the Gibbs sampler for 11,000 iterations and discarding the first 1,000 draws.



Marginal posterior distributions



Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is, in fact, more general than the Gibbs sampler and older (1950's).

One can sample from the joint posterior $p(\theta_1, \theta_2, \theta_3|y)$ by iteratively sampling θ_1^* from a proposal density $q_1(\cdot)$ and accepting the draw with probability

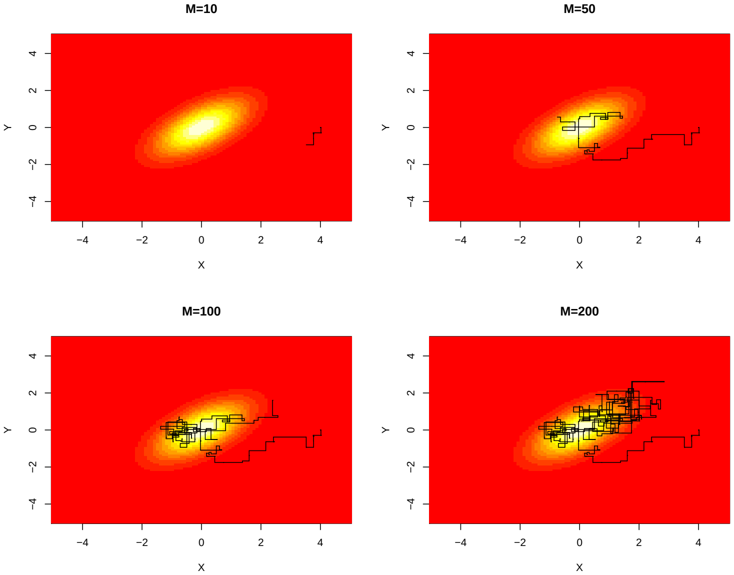
$$\min\left\{1, \frac{p(\theta_1^*, \theta_2, \theta_3|y)}{p(\theta_1, \theta_2, \theta_3|y)} \frac{q_1(\theta_1)}{q_1(\theta_1^*)}\right\},$$

with θ_2 and θ_3 fixed at the final draws from the previous iteration. The steps are repeated for θ_2^* and θ_3^*.

After a *warm up* phase, the draws will behave as coming from posterior distribution.

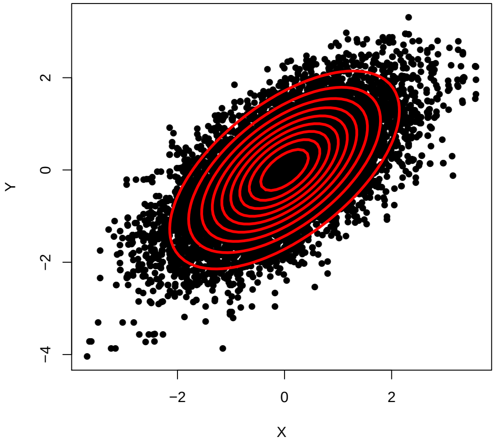
Random-walk Metropolis algorithm

The proposals are $x^* \sim N(x^{old}, 0.25)$ and $y^* \sim N(y^{old}, 0.25)$

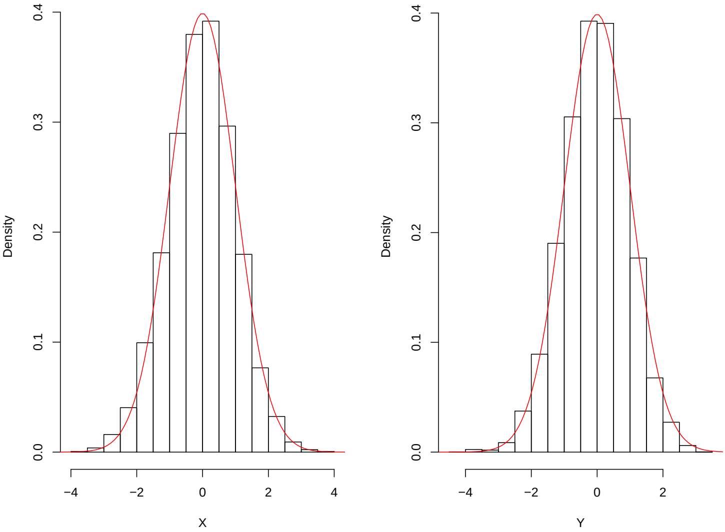


Posterior draws

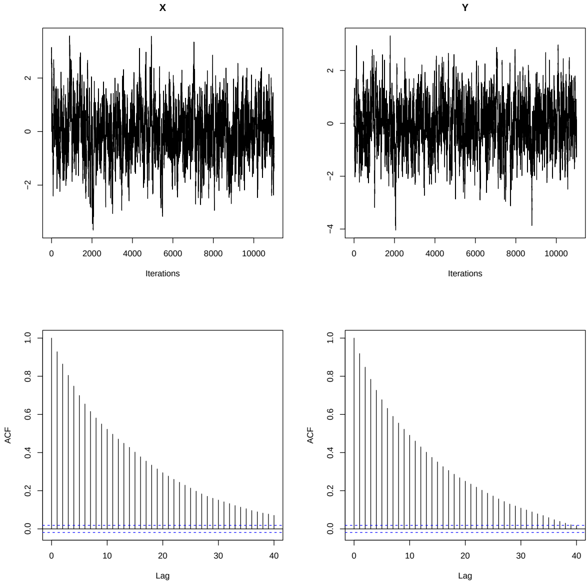
Running the Metropolis-Hastings algorithm for 11,000 iterations and discarding the first 1,000 draws.



Marginal posterior distributions

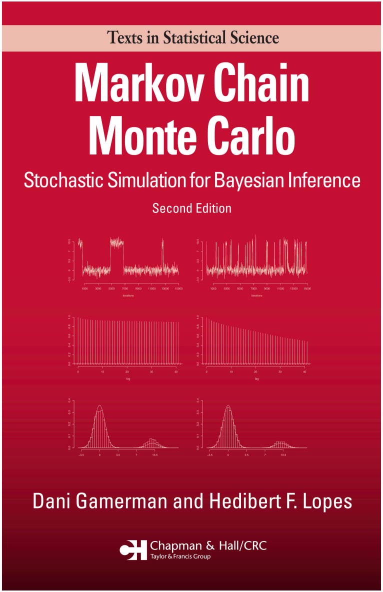


Markov chains and autocorrelation



Want to learn more?

hedibert.org has a link to book webpage.



References

1. Metropolis and Ulam (1949) The Monte Carlo method. JASA, 44, 335-341.
2. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) Equation of state calculations by fast computing machines. Journal of Chemical Physics, Number 21, 1087-1092.
3. Hastings (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97-109.
4. Peskun (1973) Optimum Monte-Carlo sampling using Markov chains. Biometrika, 60, 607-612.
5. Besag (1974) Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B, 36, 192-236.
6. Geman and Geman (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.
7. Jennison (1993) Discussion of the meeting on Gibbs sampling and other Markov chain Monte Carlo methods. JRSS-B, 55, 54-6.
8. Kirkpatrick, Gelatt and Vecchi (1983) Optimization by simulated annealing. Science, 220, 671-80.
9. Pearl (1987) Evidential reasoning using stochastic simulation. Arti[U+FFFD]ial Intelligence, 32, 245-257.
10. Tanner and Wong (1987) The calculation of posterior distributions by data augmentation. JASA, 82, 528-550.
11. Geweke (1989) Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57, 1317-1339.
12. Gelfand and Smith (1990) Sampling-based approaches to calculating marginal densities, JASA, 85, 398-409.
13. Casella and George (1992) Explaining the Gibbs sampler. The American Statistician,46,167-174.
14. Smith and Gelfand (1992) Bayesian statistics without tears: a sampling-resampling perspective. American Statistician, 46, 84-88.
15. Gilks and Wild (1992) Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337-48.
16. Chib and Greenberg (1995) Understanding the Metropolis-Hastings algorithm. The American Statistician, 49, 327-335.