Lecture 2
Markov chain Monte Carlo algorithms

olivier.francois@imag.fr

Dakar Fevrier 2010

Outline

- Markov chain Monte Carlo algorithms
- Metropolis-Hastings
- Multi-parametric models: Gibbs sampling

Model definition.

- Parameter $\theta = (\theta_1, \ldots, \theta_J)$, $J \geq 1$.
- Data $y = (y_1, \ldots, y_n)$, $n \geq 1$.
- A model is a joint distribution

$$p(y, \theta) = p(y|\theta)p(\theta)$$

- $p(\theta)$ is the prior distribution.
- $p(y|\theta)$ is the likelihood or sampling distribution.

Inference.

- Use the Bayes formula to compute the posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

- Monte Carlo simulation methods can sample from (probability) distribution that are defined up to a constant

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Markov chain Monte Carlo methods

- Principle: A target distribution π is the invariant distribution of some ergodic Markov chain with transition kernel $K(\theta, \varphi)d\varphi$

$$\pi(\varphi) = \int K(\theta, \varphi)\pi(\theta)d\theta$$

- Applying this to $\pi(\theta) = p(\theta|y)$ should avoid computing the marginal distribution $p(y)$

Metropolis-Hastings algorithm

- Define a proposal transition kernel $Q(\theta, \theta^*)$
  1. Initialize θ^0, $t = 0$
  2. Sample θ^* according to $Q(\theta^t, \theta^*)$
  3. Compute
  $$r = \frac{\pi(\theta^*)Q(\theta^*, \theta^t)}{\pi(\theta^t)Q(\theta^t, \theta^*)}$$
  4. With probability $\min(1, r)$, do $\theta^{t+1} \leftarrow \theta^*$, otherwise $\theta^{t+1} \leftarrow \theta^t$
  5. Increment t and go to 2

Why does it work?

- Let $\pi(\theta) = p(\theta|y)$
- The Markov transition kernel is

$$K(\theta_0, \theta_1) = Q(\theta_0, \theta_1) \min\left(\frac{\pi(\theta_1)Q(\theta_1, \theta_0)}{\pi(\theta_0)Q(\theta_0, \theta_1)}, 1\right)$$

- Time-reversibility

$$\pi(\theta_0)K(\theta_0, \theta_1) = \pi(\theta_1)K(\theta_1, \theta_0)$$

$\Longrightarrow$ π is an invariant distribution (Exercise).

Comments

- The MCMC algorithm simulates from an approximate posterior distribution
- Stationarity is reached after a burn-in period which determination has led to several methods in the literature.
- Its advantage is that it only requires computing the Metropolis-Hasting ratio r and avoids $p(y)$.

Beta-binomial model

- Proposal kernel = Prior distribution
- Metropolis-Hastings ratio

$$r = \left(\frac{\theta^*}{\theta^t}\right)^y \times \left(\frac{1-\theta^*}{1-\theta^t}\right)^{(n-y)}$$

- Exercise: Implement the Markov chain in the R language.

R scripts

- MCMC algorithm (core):

```
theta.1 = runif(1)
ratio = (theta.1/theta.0)^ y *((1 - theta.1)/(1 -
theta.0))^ (n-y)
if (runif(1) < ratio) theta.0 = theta.1
```

- Exercise: Compute a 95% credible interval for θ and a histogram of the posterior predictive distribution given y.

Random walk proposal

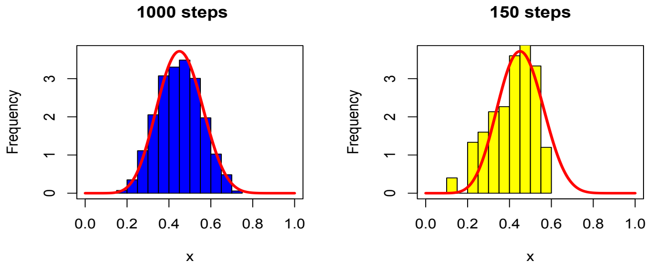
- Proposal kernel = Local search (depends on θ_0)
- Example: $Q(\theta^0, .)$ is the beta$(100\theta^0/(1-\theta^0))$, 100) distribution
- Expected move = θ_0 (ie, slowly move to a neighborhood of θ_0).

Random walk proposal

- Proposal kernel $Q(\theta^0, .)$ is the beta$(100\theta^0/(1-\theta^0)$, $b)$ distribution $(b = 100)$

```
theta.1 <- rbeta(1, b*theta.0/(1-theta.0), b)
ratio1 <- (theta.1/theta.0)^ y * ( (1 -
theta.1)/(1 - theta.0) )^ (n-y)
ratio2 <- dbeta(theta.0, b*theta.1/(1-theta.1),
b)/dbeta(theta.1, b*theta.0/(1-theta.0), b)
ratio <- ratio1*ratio2
```

Random walk proposal



- Effect of a burnin period (right figure): The chain did not converge after 150 steps.

Multi-dimensional parameters: A basic algorithm

- $\theta = (\theta_1, \theta_2)$

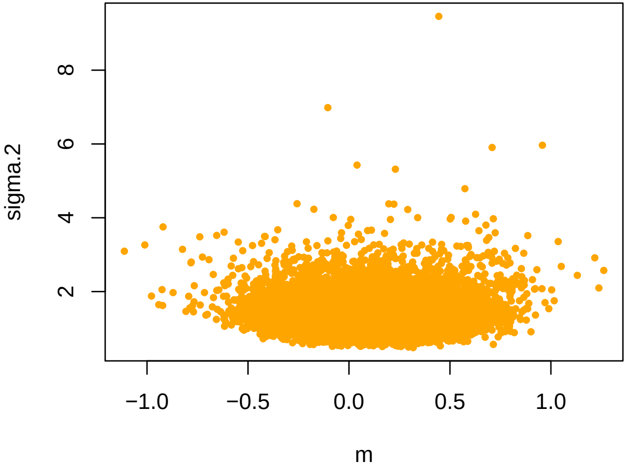
1. Simulate θ_1 from the marginal distribution $p(\theta_1)$
2. Given θ_1, simulate θ_2 from the conditional distribution $p(\theta_2|\theta_1)$.

Example: Posterior distribution of $\theta = (m, \sigma^2)$

1. $\sigma^2|y \sim (n-1)\mathrm{var}(y)/\chi^2_{n-1}$
2. $m|\sigma^2, y \sim N(\mathrm{mean}(y), \sigma^2/n)$

```
# Simulated data
n = 20; y = rnorm(n)
sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```

Posterior distribution (m, σ^2)



The Gibbs sampler

- $\theta^t = (\theta_1^t, \theta_2^t)$
- Repeat the following cycle (or sweep)

GS1. Given θ_2^t, simulate θ_1^{t+1} from the conditional distribution $p(\theta_1|\theta_2^t)$.

GS2. Given θ_1^{t+1}, simulate θ_2^{t+1} from the conditional distribution $p(\theta_2|\theta_1^{t+1})$.

Example: Posterior distribution of $\theta = (m, \sigma^2)$

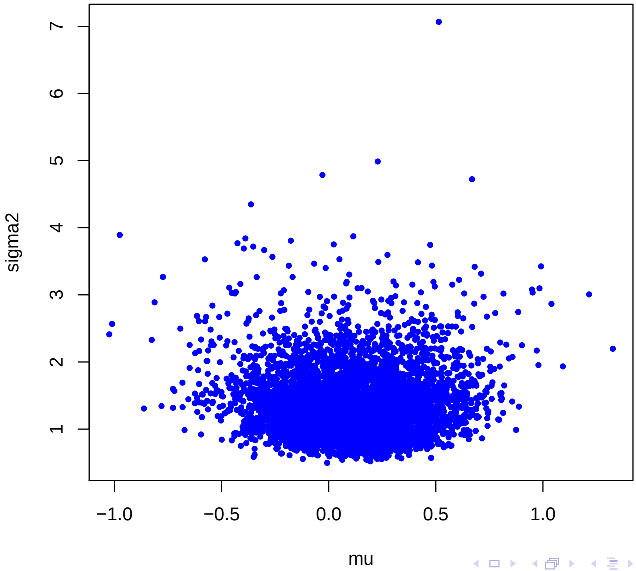
GS1. $\sigma^2 | m, y \sim n s_n^2 / \chi_n^2$

GS2. $m | \sigma^2, y \sim N(\mathrm{mean}(y), \sigma^2/n)$

sigma.2 = sum((y -m)^2)/$rchisq(10000, n)$

m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))

Posterior distribution (m, σ^2)



Convergence of the Gibbs sampler

- Markov kernel K

$$K(\theta^t, \theta^{t+1}) = p(\theta_2^{t+1} | \theta_1^t, y) p(\theta_1^{t+1} | \theta_2^{t+1}, y)$$

- The posterior distribution is a stationary distribution (Exercise)

$$p(\theta^{t+1} | y) = \int p(\theta^t | y) K(\theta^t, \theta^{t+1}) d\theta^t$$

- Warning: Not always ergodic!

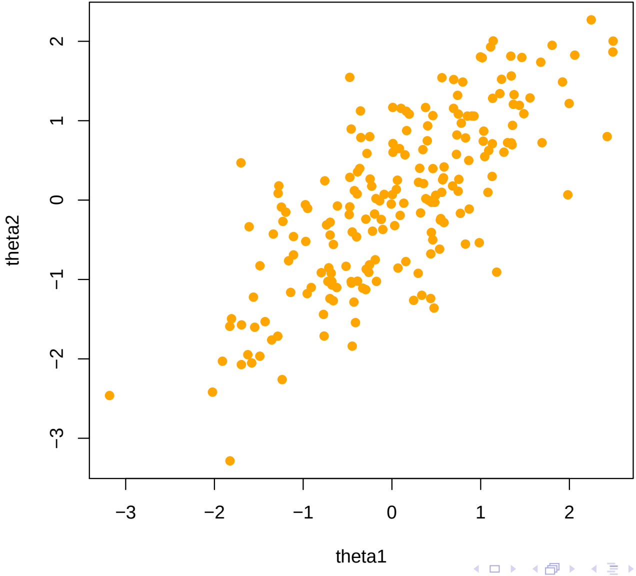
Yet another example

- Simulate from a two dimensional Gaussian distribution of mean = $(0, 0)$ and covariance matrix

$$\Lambda = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

```
rho = .75
theta1 = rnorm(200)
theta2 = rnorm(200, rho*theta1, sd = sqrt(1 -
rho^ 2))
```

Two-dimensional Gaussian distribution

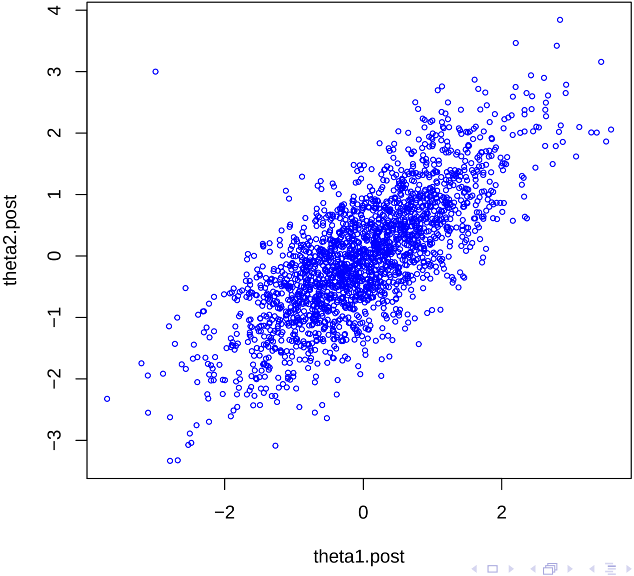


Gibbs sampler

- Gibbs sampling sweeps

```
theta1 <- rnorm(1, rho*theta2, sd = sqrt(1 -
rho^ 2))
theta2 <- rnorm(1, rho*theta1, sd = sqrt(1 -
rho^ 2))
```

Two-dimensional Gaussian distribution



Take-home messages

- The Metropolis-Hastings and Gibbs sampler algorithms are useful because they avoid the computation of the marginal distribution p(y)
- But the convergence of the algorithm can be hard to ascertain in some cases.

Exercises

Ex1. Compute the 95% credible interval for θ and the posterior predictive distribution given y from the MCMC algorithm for the beta-binomial model

Ex2. Compute the Metropolis-Hastings Markov chain transition kernel and prove that $\pi = p(\theta|y)$ is invariant for the corresponding Markov chain

Ex3. Implement the Metropolis-Hastings algorithm with a non-uniform proposal. Evaluate the convergence rate of the above algorithm experimentally.

Ex4. Implement the Gibbs sampler for (m, σ^2) and for the two dimensional Gaussian distribution. Evaluate the convergence rate of the above algorithm for distinct values of ρ experimentally.

Bibliography and resources

- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis 2nd ed. Chapman & Hall, New-York.
- E. Paradis (2005) R pour les débutants. Univ. Montpellier II.

- R website: `http://cran.r-project.org/`