MARKOV CHAIN MONTE CARLO METHODS

Hedibert Freitas Lopes

The University of Chicago Booth School of Business
5807 South Woodlawn Avenue, Chicago, IL 60637
http://faculty.chicagobooth.edu/hedibert.lopes

hlopes@ChicagoBooth.edu

Outline

Historical facts

Dongarra and Sullivan (2000) list the top algorithms with the greatest influence on the development and practice of science and engineering in the 20th century (in chronological order):

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform

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.

MH algorithms

A sequence $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$ is drawn from a Markov chain whose *limiting equilibrium distribution* is the posterior distribution, $\pi(\theta)$.

Algorithm

1. Initial value: $\theta^{(0)}$
2. Proposed move: $\theta^* \sim q(\theta^*|\theta^{(i-1)})$
3. Acceptance scheme:

$$\theta^{(i)} = \begin{cases} \theta^* & \text{com prob. } \alpha \\ \theta^{(i-1)} & \text{com prob. } 1-\alpha \end{cases}$$

where

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta^{(i-1)})} \frac{q(\theta^{(i-1)}|\theta^*)}{q(\theta^*|\theta^{(i-1)})}\right\}$$

Special cases

1. Symmetric chains: $q(\theta|\theta^*) = q(\theta^*|\theta)$

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

2. Independence chains: $q(\theta|\theta^*) = q(\theta)$

$$\alpha = \min\left\{1, \frac{\omega(\theta^*)}{\omega(\theta)}\right\}$$

where $\omega(\theta^*) = \pi(\theta^*)/q(\theta^*)$.

Random walk Metropolis

The most famous symmetric chain is the random walk Metropolis:

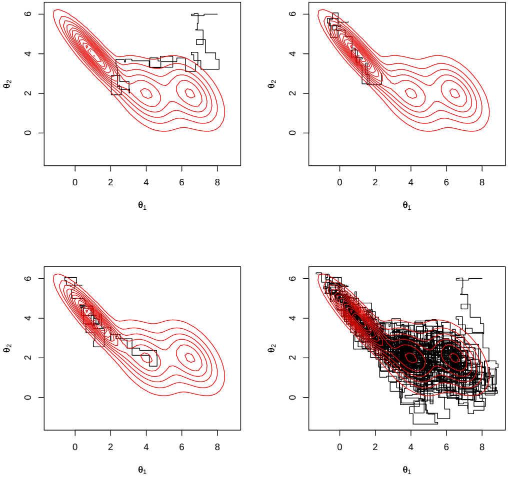
$$q(\theta|\theta^*) = q(|\theta - \theta^*|)$$

Hill climbing: when

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

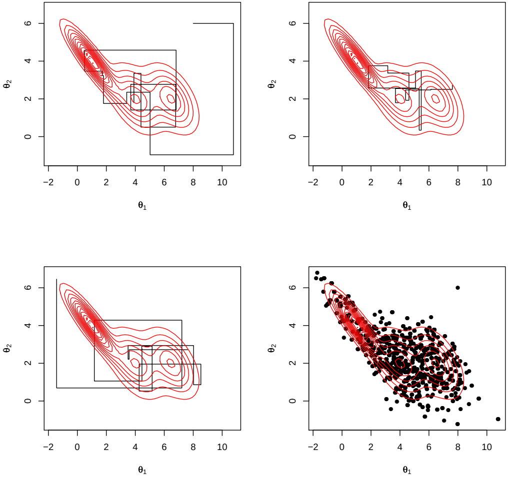
a value θ^* with higher density $\pi(\theta^*)$ greater than $\pi(\theta)$ is automatically accepted.

Example iv. RW Metropolis



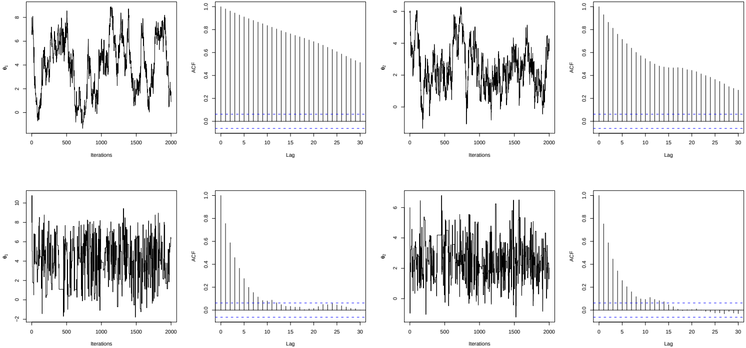
$q(\theta|\theta_i) \sim N(\theta_i, 0.25\Sigma_2)$.

Example iv. Ind. Metropolis

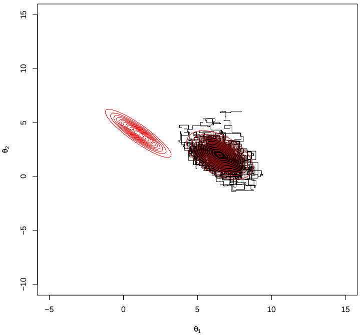


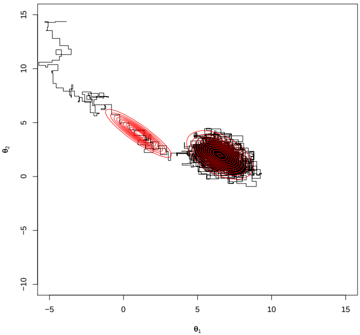
$$q(\theta) \equiv q_{SIR}(\theta) \sim N(\mu, \Sigma).$$

Example iv. Autocorrelations

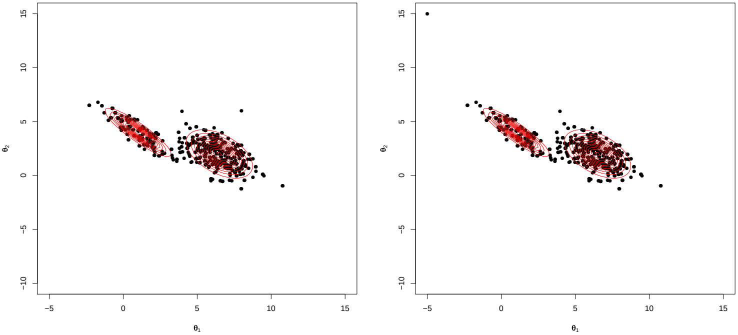


Example v. RW Metropolis

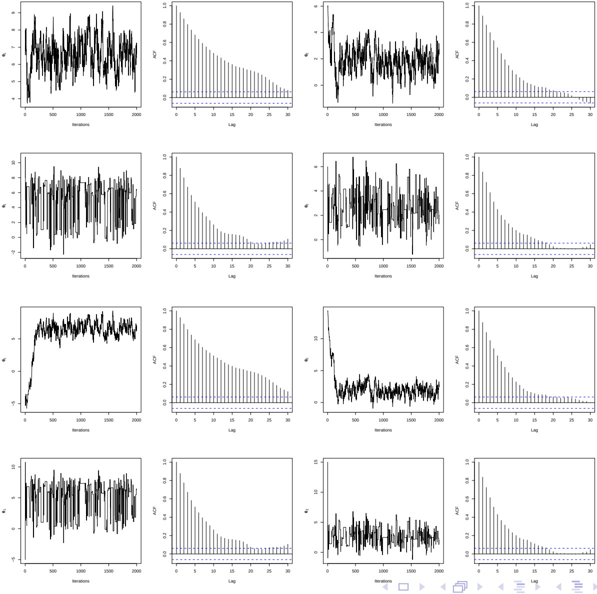




Example v. Ind. Metropolis



Example v. Autocorrelations



Example vi. tuning selection

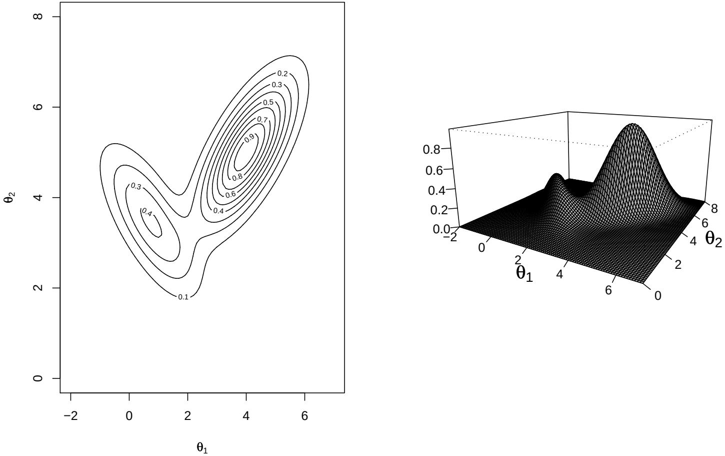
Tthe target distribution is a two-component mixture of bivariate normal densities, ie:

$$\pi(\theta) = 0.7 f_N(\theta; \mu_1, \Sigma_1) + 0.3 f_N(\theta; \mu_2, \Sigma_2).$$

where

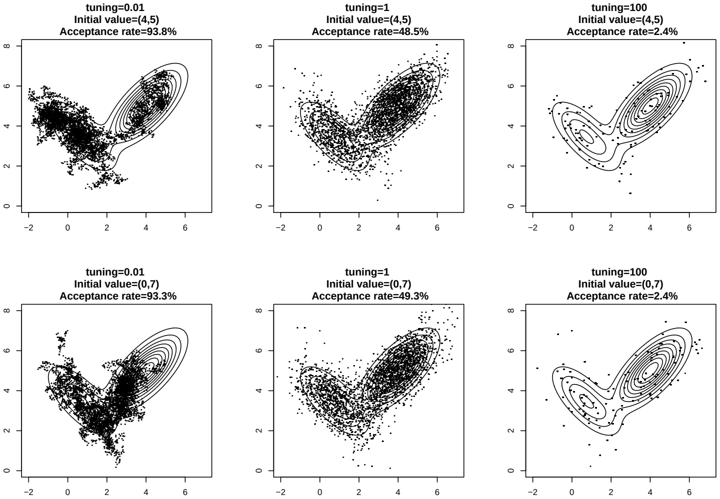
$$\begin{aligned} \mu_1' &= (4.0, 5.0) \\ \mu_2' &= (0.7, 3.5) \\ \Sigma_1 &= \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix} \\ \Sigma_2 &= \begin{pmatrix} 1.0 & -0.7 \\ -0.7 & 1.0 \end{pmatrix}. \end{aligned}$$

Target distribution

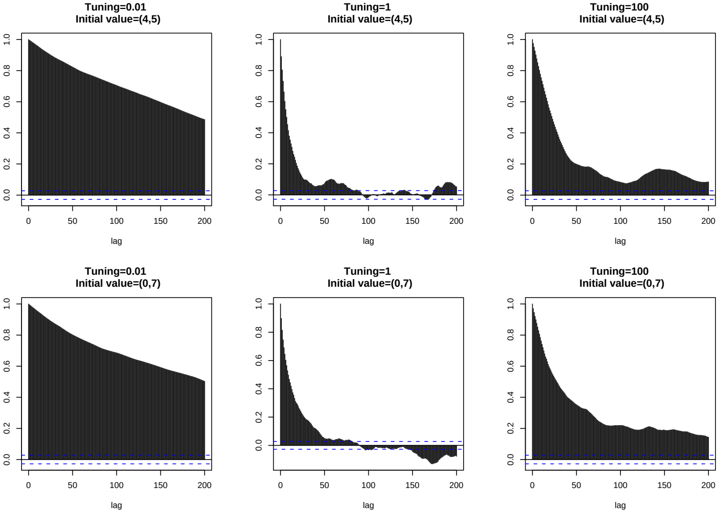


RW Metropolis

$q(\theta, \phi) = f_N(\phi; \theta, \nu I_2)$ and $\nu =$tuning.

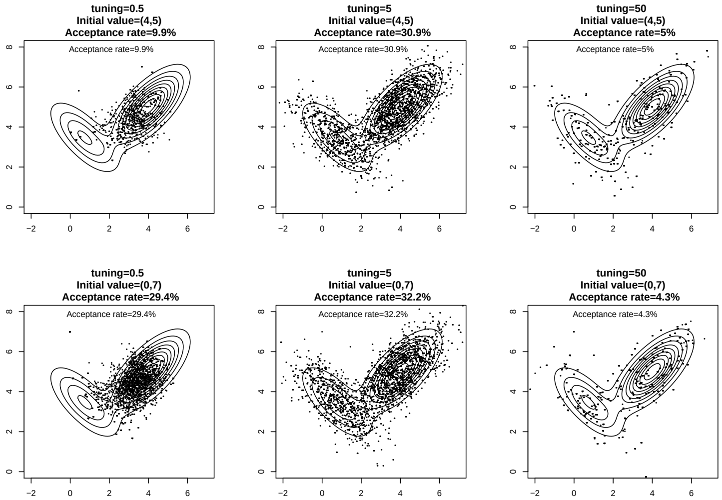


Autocorrelations

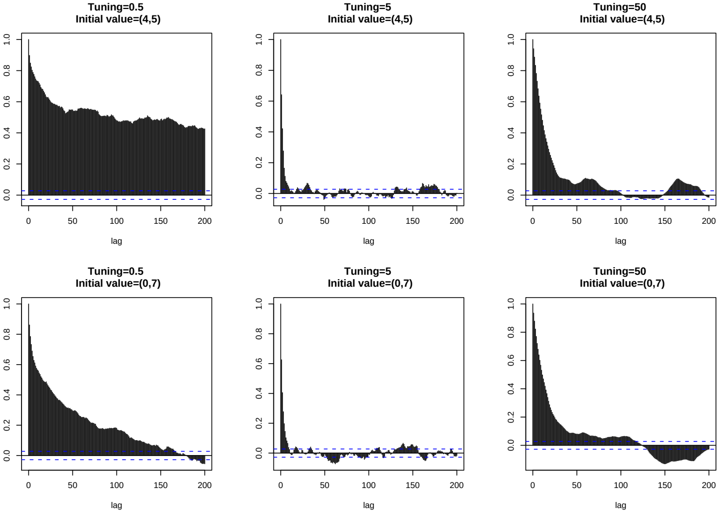


Independent Metropolis

$q(\theta, \phi) = f_N(\phi; \mu_3, \nu I_2)$ and $\mu_3 = (3.01, 4.55)'$.



Autocorrelations



Simulated annealing

Simulated annealing[1] is an optimization technique designed to find maxima of functions.

It can be seen as a M-H algorithm that *tempers* with the target distribution:

$$q(\theta) \propto \pi(\theta)^{1/T}$$

where the constant $T > 1$ receives the physical interpretation of system temperature, hence the nomenclature used (Jennison, 1993).

The *heated* distribution q is flattened with respect to π and its density gets closer to the uniform distribution, which is particularly relevant for the case of a distribution with distant modes.

By flattening the modes, the moves required to cover adequately the parameter space become more likely.

[1] Kirkpatrick, Gelatt and Vecchi (1983)

Example vii: Nonlinear surface

Assume that the goal is to find the mode/maximum of

$$\pi(\beta_1, \beta_2) \propto \prod_{i=1}^{4} \frac{e^{(\beta_1+\beta_2 x_i)y_i}}{(1+e^{\beta_1+\beta_2 x_i})^5},$$

with $x = (-0.863, -0.296, -0.053, 0.727)$ and $y = (0, 1, 3, 5)$.

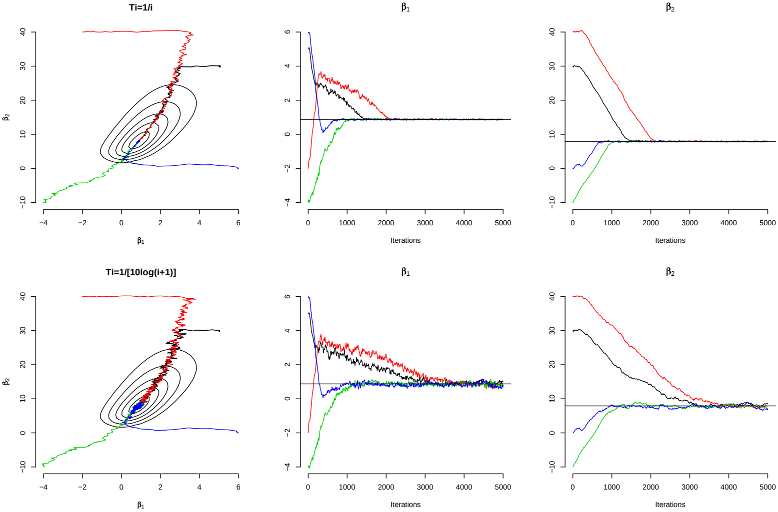
The simulated annealing algorithm is implemented for four initial values:

$$(5, 30) \quad (-2, 40) \quad (-4, -10) \quad (6, 0)$$

and two cooling schedules:

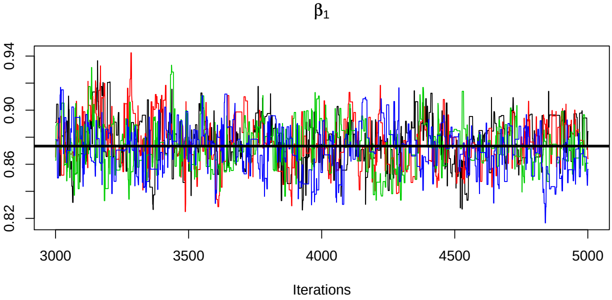
$$T_i = 1/i \quad \text{and} \quad T_i = 1/[10\log(1+i)].$$

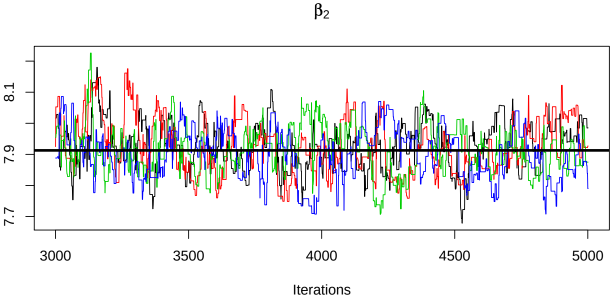
The proposal distribution is $q(\beta|\beta^{(i)}) = f_N(\beta; \beta^{(i)}, 0.05^2 I_2)$.



Newton-Raphson mode: $(0.87, 7.91)$.

$T_i = 1/i$: mode is $(0.88, 7.99)$ when $(\beta_1^{(0)}, \beta_2^{(0)}) = (5, 30)$.


β_1
0.82
0.86
0.90
0.94
3000
3500
4000
4500
5000
Iterations


β_2
7.7
7.9
8.1
3000
3500
4000
4500
5000
Iterations

Gibbs sampler

Technically, the Gibbs sampler is an MCMC scheme whose transition kernel is the product of the full conditional distributions.

Algorithm

1. Start at $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots)$
2. Sample the components of $\theta^{(j)}$ iteratively:

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots) \\
\theta_2^{(j)} &\sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots) \\
\theta_3^{(j)} &\sim \pi(\theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \ldots) \\
&\vdots
\end{aligned}$$

The Gibbs sampler opened up a new way of approaching statistical modeling by combining simpler structures (the full conditional models) to address the more general structure (the full model).

Example viii: Bivariate normal

Assume that the target distribution is the bivariate normal with mean vector and covariance matrix given by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

respectively.

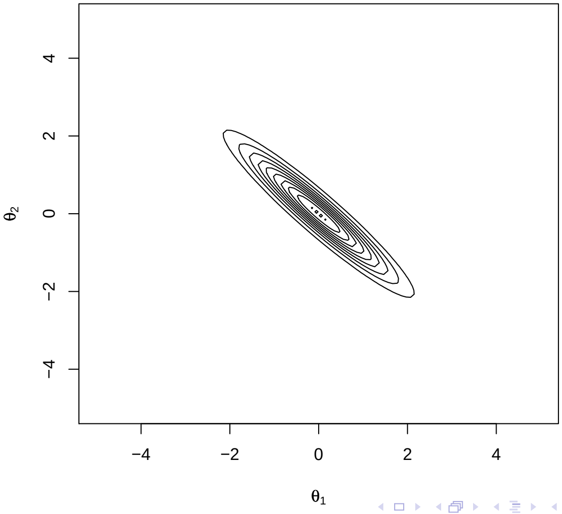
In this case, the two full conditionals are given by

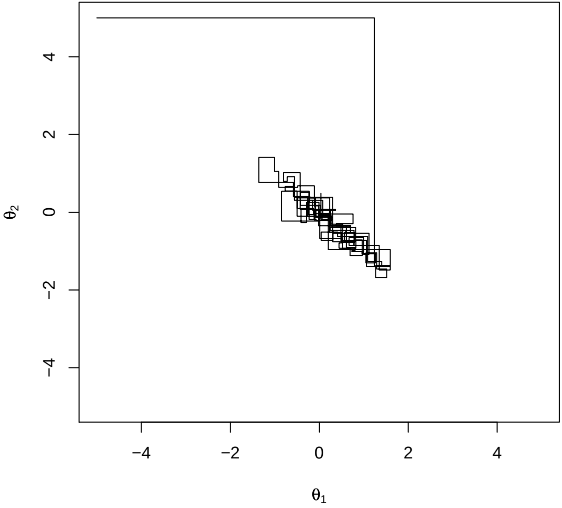
$$\theta_1|\theta_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(\theta_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right)$$

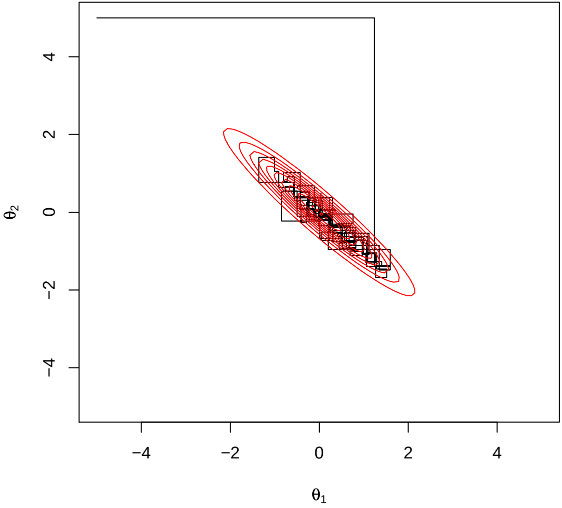
and

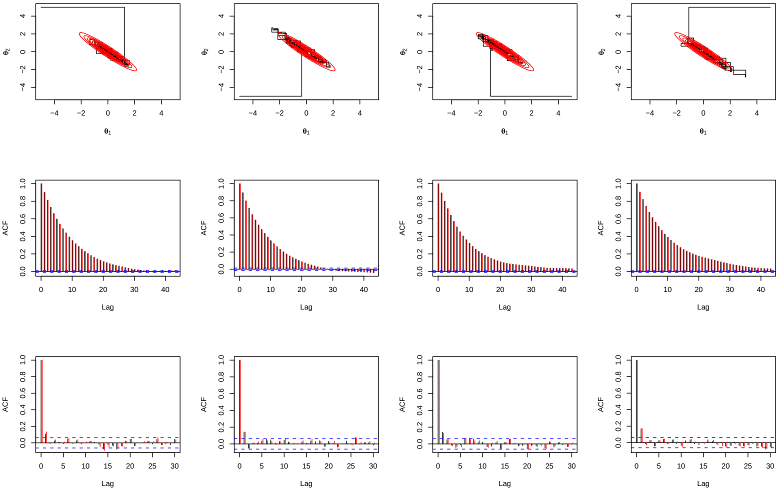
$$\theta_2|\theta_1 \sim N\left(\mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(\theta_1 - \mu_1), \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}\right)$$

$\mu_1 = \mu_2 = 0$
$\sigma_1^2 = \sigma_2^2 = 1$
$\sigma_{12} = -0.95$

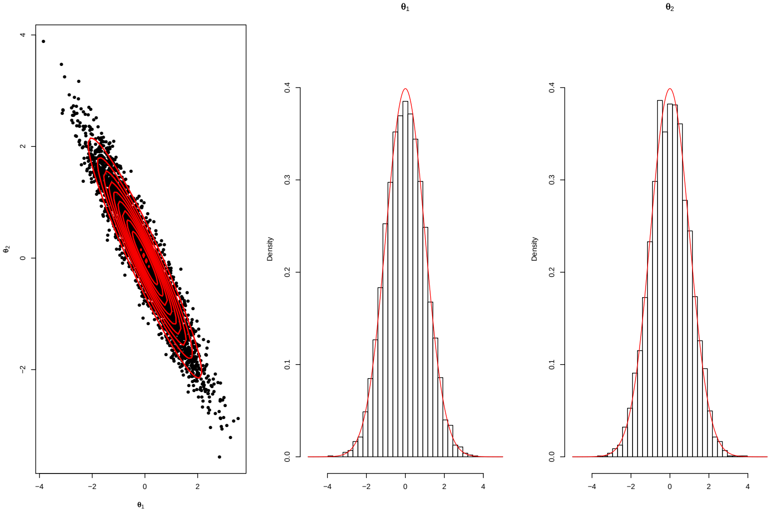



4
2
0
−2
−4
θ2
−4
−2
0
2
4
θ1


θ_2
θ_1
4
2
0
−2
−4



Middle frame: Based on $M = 21,000$ consecutive draws.
Bottom frame: Based on $M = 1000$ draws, after initial $M_0 = 1000$ draws and saving every 20th draws.


θ_1
θ_2
θ_2
θ_1
Density
Density
4
2
0
−2
−4
0.0
0.1
0.2
0.3
0.4

Simple random effects model

Let response (dependent variable) y_i be linearly related to regressor (explanatory variable) x_i in the following way:

$$y_i = \beta x_i + \theta_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where θ_i is the random effect modeled by

$$\theta_i \sim N(\mu, \tau^2)$$

with ϵ_i and θ_i uncorrelated and x_i known.

The parameters of the model are

$$(\beta, \sigma^2, \mu, \tau^2) \quad \text{and} \quad \theta = (\theta_1, \ldots, \theta_n).$$

Prior and simulated data

We assume the following independent prior distributions:

$$\begin{array}{rcll} \beta & \sim & N(b_0, C_0) & \sigma^2 \sim IG(a, b) \\ \mu & \sim & N(\mu_0, V_0) & \tau^2 \sim IG(c, d) \end{array}$$

We simulated $n = 300$ pairs (y_i, x_i) based on $\beta = 5$, $\sigma^2 = 0.2$, $\mu = 0$, $\tau^2 = 0.1$ and $x_i \sim U(0, 1)$.

In this case, the hyperparameters are $a = 6, b = 1$, $c = 3$, $d = 0.2$, $b_0 = 0$, $C_0 = 100$, $\mu_0 = 0$ and $V_0 = 100$, such that $E(\sigma^2) = 0.2$, $E(\tau^2) = 0.1$ and $V(\sigma^2) = V(\tau^2) = 0.01$.

Full conditional distributions

$(\sigma^2|\beta, \theta, y, x) \sim IG\left(a + n/2, b + \sum_{i=1}^n (y_i - \beta x_i - \theta_i)^2/2\right)$

$(\tau^2|\theta, \mu) \sim IG\left(c + n/2, d + \sum_{i=1}^n (\theta_i - \mu)^2/2\right)$

$(\beta|\theta, \sigma^2, x, y) \sim N(m, C)$, where $C = 1/(1/C_0 + \sum_{i=1}^n x_i^2/\sigma^2)$ and $m = C(b_0/C_0 + \sum_{i=1}^n (y_i - \theta_i)x_i/\sigma^2)$

$(\mu|\theta, \tau^2) \sim N(m, C)$, where $C = 1/(1/V_0 + n/\tau^2)$ and $m = C(\mu_0/V_0 + \sum_{i=1}^n \theta_i/\tau^2)$

For $i = 1, \ldots, n$, $(\theta_i|y_i, x_i, \beta, \sigma^2, \mu, \tau^2) \sim N(m_i, C)$, where $C = 1/(1/\tau^2 + 1/\sigma^2)$ and $m_i = C(\mu/\tau^2 + (y_i - \beta x_i)/\sigma^2)$

Posterior predictive

Let $\gamma = (\beta, \sigma^2, \mu, \tau^2)$, $D = (y, x)$, $\tilde{y} = y_{n+1}$, $\tilde{x} = x_{n+1}$ and $\tilde{\theta} = \theta_{n+1}$. Then

$$\begin{aligned} p(\tilde{y}|\tilde{x}, D) &= \int p(\tilde{y}|\tilde{x}, \tilde{\theta}, \gamma, D) p(\tilde{\theta}|\gamma, D) p(\gamma|D) d\tilde{\theta} d\gamma \\ &= \int p(\tilde{y}|\tilde{x}, \tilde{\theta}, \gamma) p(\tilde{\theta}|\mu, \tau^2) p(\gamma|D) d\tilde{\theta} d\gamma \\ &= \int f_N(\tilde{y}|\beta\tilde{x} + \tilde{\theta}, \sigma^2) f_N(\tilde{\theta}|\mu, \tau^2) p(\gamma|D) d\tilde{\theta} d\gamma \end{aligned}$$

with a Monte Carlo approximation given by

$$\hat{p}_1(\tilde{y}|\tilde{x}, D) = \frac{1}{M} \sum_{m=1}^{M} f_N(\tilde{y}|\beta^{(m)}\tilde{x} + \tilde{\theta}^{(m)}, \sigma^{2(m)})$$

where $\tilde{\theta}^{(m)} \sim p(\tilde{\theta}|\mu^{(m)}, \tau^{2(m)})$ (random effects distribution) and $\gamma^{(m)} \sim p(\gamma|D)$ (MCMC draws), for $m = 1, \ldots, M$.

Posterior predictive mean

Similarly, a MC approximation to the mean of the posterior predictive is given by

$$\begin{aligned}
\hat{E}_1(\tilde{y}|\tilde{x}, D) &= \frac{1}{M}\sum_{m=1}^{M} E(\tilde{y}|\tilde{x}, \tilde{\theta}^{(m)}, \gamma^{(m)}) \\
&= \left(\frac{1}{M}\sum_{m=1}^{M}\beta^{(m)}\right)\tilde{x} + \left(\frac{1}{M}\sum_{m=1}^{M}\tilde{\theta}^{(m)}\right) \\
&= \bar{\beta}\tilde{x} + \bar{\tilde{\theta}}
\end{aligned}$$

Posterior predictive draws

Draws $\tilde{y}^{(m)}$, for $m = 1, \ldots, M$, from the posterior predictive can be obtained from draws $\gamma^{(m)}$ and $\tilde{\theta}^{(m)}$ by simply sampling $\tilde{y}^{(m)}$ from $N(\beta^{(m)}\tilde{x} + \tilde{\theta}^{(m)}, \sigma^{2(m)})$.

An alternative MC approximation to $E(\tilde{y}|\tilde{x}, D)$ is given by

$$\hat{E}_2(\tilde{y}|\tilde{x}, D) = \frac{1}{M}\sum_{m=1}^{M} \tilde{y}^{(m)}.$$

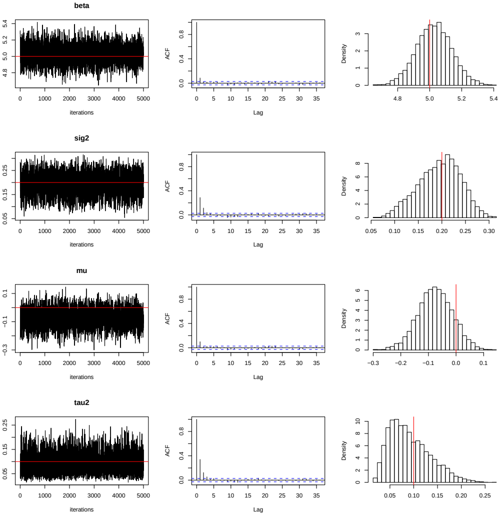
It is easy to show that the MC variance of $\hat{E}_1$ is at most as large as the variance of $\hat{E}_2$ (Rao-Blackwell result) - *Raoblackwellization*.

Similarly, an alternative MC approximation to $p(\tilde{y}|\tilde{x}, D)$ is given by the histogram (kernel approximation) of the draws $\tilde{y}^{(m)}$, namely $\hat{p}_2(\tilde{y}|\tilde{x}, D)$.

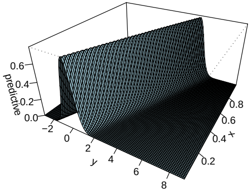
Posterior summary

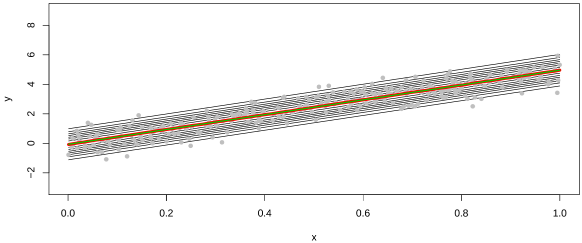
Initial values: $\beta = 5.0$, $\mu = 0$, $\theta = \theta_{true}$

MCMC set up: $(M_0, L, M) = (50000, 50, 5000)$

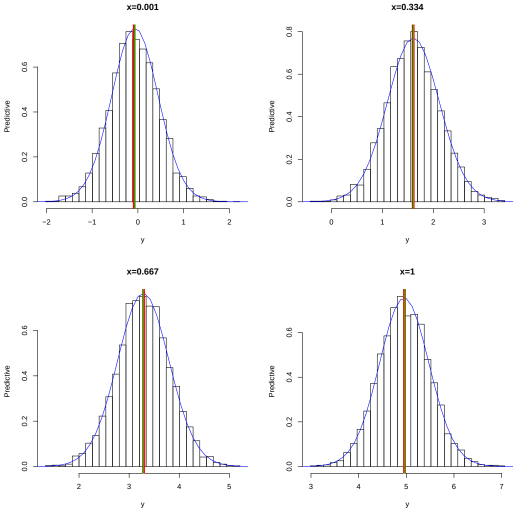


Predictive: $\hat{p}_1(\tilde{y}|\tilde{x}, D)$





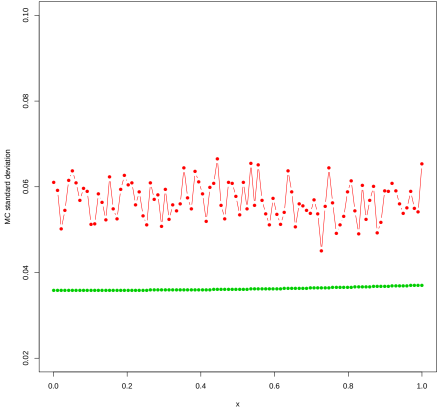
Predictive: $\hat{p}_i(\tilde{y}|\tilde{x}, D)$ for $i = 1, 2$



$\hat{E}_1$, $\hat{E}_2$, $\hat{p}_1$ and $\hat{p}_2$ (histogram).

Monte Carlo error

$\hat{E}_1$ and $\hat{E}_2$ are computed based on 50 sets of 100 MCMC draws for each value of $\tilde{x}$. The picture plots the standard deviations over the 50 sets.



Book references

1. Ripley (1987) Stochastic Simulation. Wiley. (Chapters 5 and 7).
2. Thisted (1988) Elements of Statistical Computing. Chapman & Hall/CRC. (Chapter 5).
3. Gentle (1998) Random Number Generation and Monte Carlo Methods. Springer. (Chapter 5).
4. Liu (2001) Monte Carlo Strategies in Scientific Computing. Springer. (Chapters 2, 5 and 6).
5. Robert and Casella (2004) Monte Carlo Statistical Methods. Springer. (Chapters 3, 4 and 7).
6. Givens and Hoeting (2005) Computational Statistics. Wiley. (Chapters 6, 7 and 8).
7. **Gamerman and Lopes (2006) MCMC: Stochastic Simulation for Bayesian Inference. Chapman & Hall/CRC. (Chapters 3, 5 and 6).**
8. Rizzo (2008) Statistical Computing with R. Chapman & Hall/CRC. (Chapters 5 and 9).
9. Rubinstein and Kroese (2008) Simulation and the Monte Carlo Method. Wiley. (Chapters 5 and 6).
10. Robert and Casella (2010) Introducing Monte Carlo Methods with R. Springer. (Chapters 3, 6 and 7).

Classical papers

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. Articial 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.
17. Dongarra and Sullivan (2000) Guest editors' introduction: The top 10 algorithms. Computing in Science and Engineering, 2, 22-23.