1.5 Simulation

Ordinary Monte Carlo

Theorem 1.2 *Law of Large Numbers: Suppose $\{X_i\}$ is a sequence of iid random variables with $E[X_i] = \mu$. Then, with $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$*

$$\bar{X}_M \xrightarrow{a.s.} \mu$$

- One common application is to justify the approximation of $E[X]$ by $\bar{x}_M$ when $x_1, \ldots, x_M$ are observed data
- Another application: represent approximately one probability distribution by a computer-generated sample $x_1, \ldots, x_M$ simulated from this distribution
- (Almost) all the aspects of this probability distribution can be arbitrarily approximated using exclusively $x_1, \ldots, x_M$ for large enough M

Facts

- Expectations: $E[\psi(X)] \approx \frac{1}{M}\sum_{i=1}^{M}\psi(x_i)$
- Probabilities:
$$P(X \in A) \approx \frac{1}{M}\#\{i : x_i \in A\}$$
- Densities: for small enough $\delta > 0$
$$f(a) \approx \frac{1}{\delta}\ \frac{1}{M}\#\{i : \delta < x_i \leq a + \delta\}$$
that is, the histogram is a good approximation to the density

Facts (ctd.)

- $\psi(x_1), \ldots, \psi(x_M)$ is a sample from the distribution of $\psi(X)$
- Suppose we can obtain a sample $(x_1, y_1), \ldots, (x_M, y_M)$ from the joint distribution $f(x, y)$ of (X, Y). Then, $x_1, \ldots, x_M$ is a sample from the marginal distribution of X
- If y is a draw from the distribution of Y and x is a draw from the distribution of $X \mid y$, then (x, y) is a draw from the joint (X, Y)
- Very important for prediction: if $Y \amalg \boldsymbol{X} \mid \theta$

$$f(y \mid \boldsymbol{x}) = \int_{\Theta} f(y \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta$$

  To obtain a sample from $f(y \mid \boldsymbol{x})$ we need a sample from $\pi(\theta \mid \boldsymbol{x})$, $\theta_1, \ldots, \theta_M$, and to be able to simulate y_i from $f(y \mid \theta_i)$

Example 1.9 *Generating from a t distribution with ν degrees of freedom: $X \sim t_\nu$ can be written as mixture:*

$$X \mid Y = y \sim N(0, \nu/y) \textit{ and } Y \sim \chi^2_\nu$$

Algorithm: for $i = 1, \ldots, M$

- Generate y_i from χ^2_ν
- Generate x_i from $\mathrm{N}(0, \nu/y_i)$

$(x_1, \ldots, x_M)$ is a sample from t_ν

Statistical models are sometimes written in the form

$$f(x \mid \theta) = \int f(x, y \mid \theta)\, dy$$

either artificially (data augmentation) or as a natural consequence of the modeling strategy (eg, latent variable models) and that can be explored in order to facilitate sampling.

Example 1.10 *Probit regression:* $Y_i \mid \theta_i \sim B(1, \theta_i)$ *independently, with* $\theta_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$*, where* $\boldsymbol{x}_i$ *corresponds to known covariate information.*

If we let $Z_i \mid \beta \sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ and $Y_i = I_{(0,+\infty)}(Z_i)$ it's easy to see that

$$P(Y_i = 1) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$$

so that if we obtain a sample from $\boldsymbol{\beta}, \boldsymbol{Z} \mid \boldsymbol{y}$ we obtain also a sample from $\boldsymbol{\beta} \mid \boldsymbol{y}$.

1.6 Markov chain Monte Carlo

- Problem: in most cases, it will be very difficult to obtain a sample of simulated iid observations from $\pi(\theta \mid \boldsymbol{x})$, especially if $m(x)$ is unknown
- MCMC methods allow us to construct (even in situations where $m(\boldsymbol{x})$ is unknown) a Markov chain $\{\theta_n\}$ whose stationary (limiting) distribution is $\pi(\theta \mid \boldsymbol{x})$
- Additionally it is still the case that

$$\frac{1}{M}\sum_{n=1}^{M}\psi(\theta_n) \stackrel{as}{\to} E[\psi(\theta) \mid \boldsymbol{x}]$$

- Robert and Casella (2004). *Monte Carlo Statistical Methods*. Springer.

Gibbs Sampler

- Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$
- Let $\boldsymbol{\theta}_{(-i)} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_p)$
- Let $\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)}, \boldsymbol{x} \sim f_i(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)})$
- the density f_i is called the full-conditional of $\boldsymbol{\theta}_i$
- the Gibbs sampler proceeds by iteratively sampling from each of these full-conditionals to transition from the current state $\boldsymbol{\theta}^{(t)}$ to state $\boldsymbol{\theta}^{(t+1)}$

The Gibbs sampler algorithm:

Start at $\boldsymbol{\theta}^{(0)}$. For $t = 1, 2, \ldots$, generate

1- $\boldsymbol{\theta}_1^{(t+1)} \sim f_1(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_p^{(t)})$

2- $\boldsymbol{\theta}_2^{(t+1)} \sim f_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_p^{(t)})$

3- $\boldsymbol{\theta}_3^{(t+1)} \sim f_3(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_4^{(t)} \ldots, \boldsymbol{\theta}_p^{(t)})$

$\ldots$

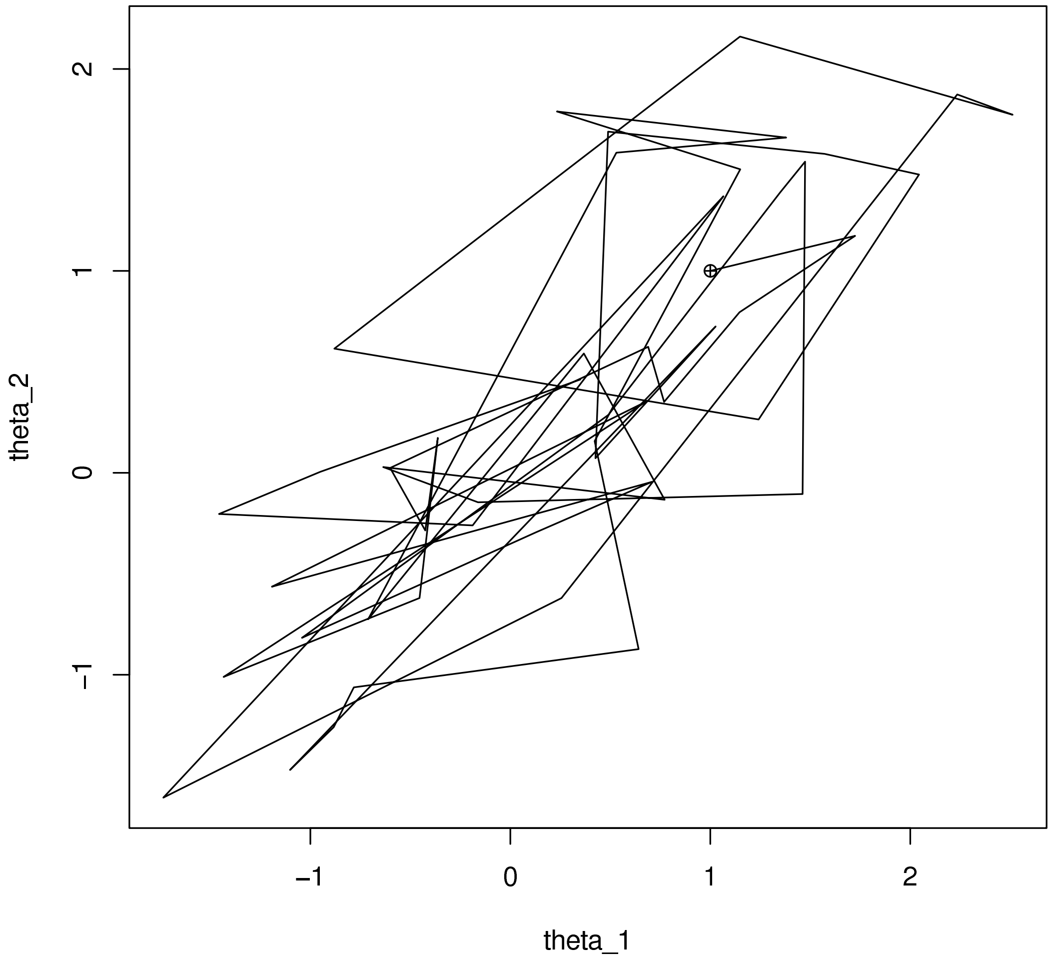
p- $\boldsymbol{\theta}_p^{(t+1)} \sim f_p(\boldsymbol{\theta}_p \mid \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{p-1}^{(t+1)})$

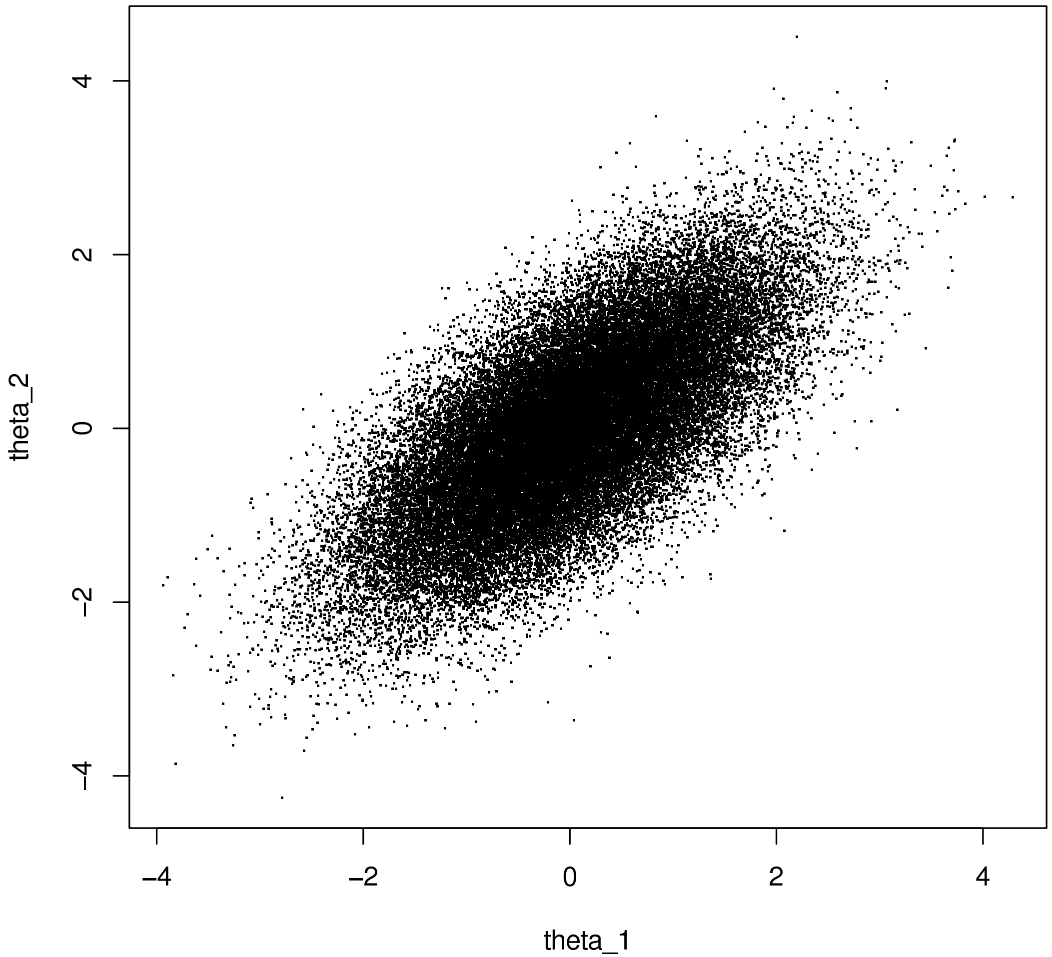
Example 1.11 *Let* $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ *where*

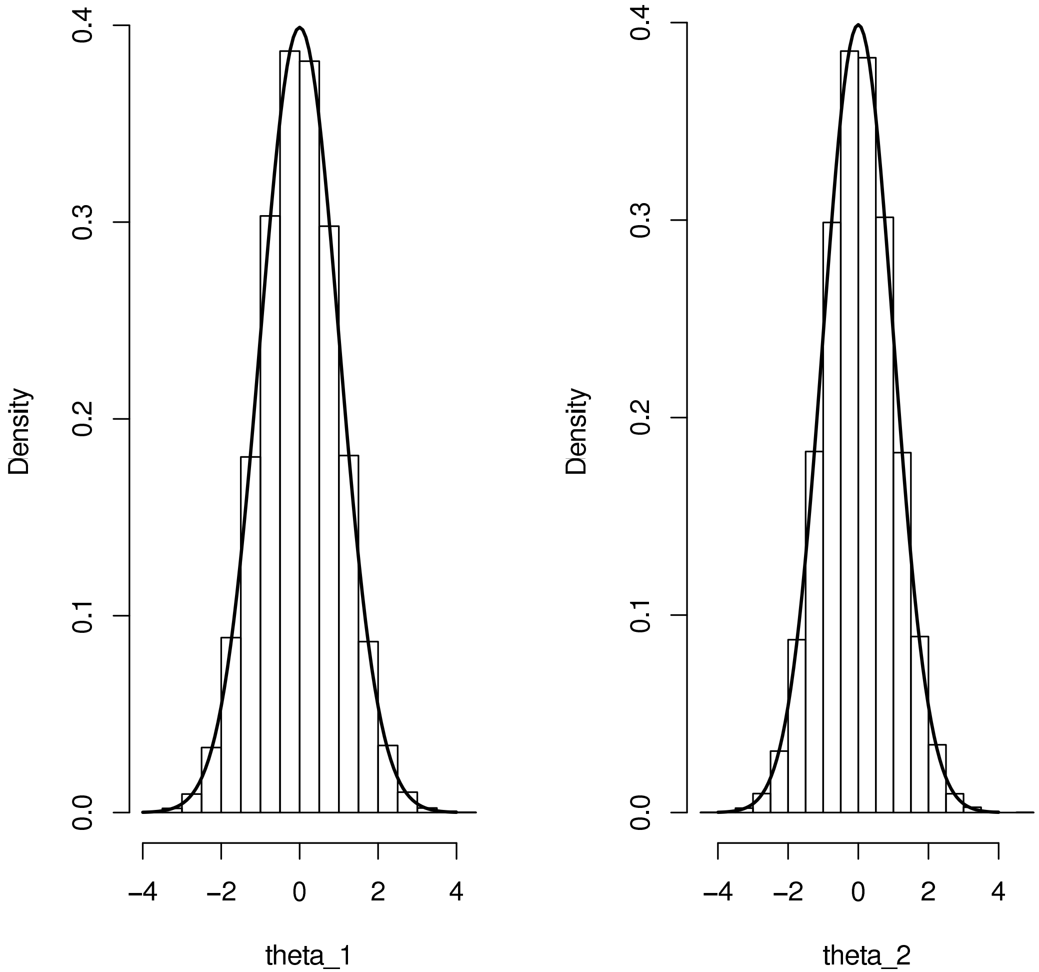
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Gibbs sampler to obtain a sample from this probability distribution: if the current state is $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$ to obtain the next state generate

$$\theta_1^{(t+1)} \sim \mathrm{N}(\rho\theta_2^{(t)}, 1-\rho^2)$$
$$\theta_2^{(t+1)} \sim \mathrm{N}(\rho\theta_1^{(t+1)}, 1-\rho^2)$$


2
1
0
−1
theta_2
−1
0
1
2
theta_1


theta_2
theta_1
4
2
0
−2
−4
−4
−2
0
2
4


Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
theta_1
theta_2

The Metropolis-Hastings Algorithm

We need a conditional density $q(\theta \mid \theta')$ called the instrumental or proposal density. The target is the posterior $\pi(\theta \mid \boldsymbol{x})$.

Start at $\theta^{(0)}$. For $t = 1, 2, \ldots$,

1. Generate $\theta^* \sim q(\theta \mid \theta^{(t)})$
2. Take

$$\theta^{(t+1)} = \begin{cases} \theta^* \text{ with probability } \rho(\theta^{(t)}, \theta^*) \\ \theta^{(t)} \text{ with probability } 1 - \rho(\theta^{(t)}, \theta^*) \end{cases}$$

   where

$$\rho(\theta^{(t)}, \theta^*) = \min\left\{ \frac{\pi(\theta^* \mid \boldsymbol{x})}{\pi(\theta^{(t)} \mid \boldsymbol{x})} \; \frac{q(\theta^{(t)} \mid \theta^*)}{q(\theta^* \mid \theta^{(t)})}, 1 \right\}$$

Observations:

- To compute the acceptance ratio ρ we do not need to know $m(\boldsymbol{x})$
- The algorithm is implementable in practice if $q(\cdot \mid \theta')$ is easy to simulate from and is either available explicitly (up to a constant independent of θ') or symmetric, ie $q(\theta \mid \theta') = q(\theta' \mid \theta)$
- with very minor restrictions on the support of the proposal, the algorithm works in *theory*

Independent Metropolis-Hastings:

- $q(\theta \mid \theta') = q(\theta)$
- close connections to the accept-reject method
- $q(\theta)$ is typically designed to closely approximate the target (eg, analytic approximations to the posterior)

Random walk Metropolis-Hastings:

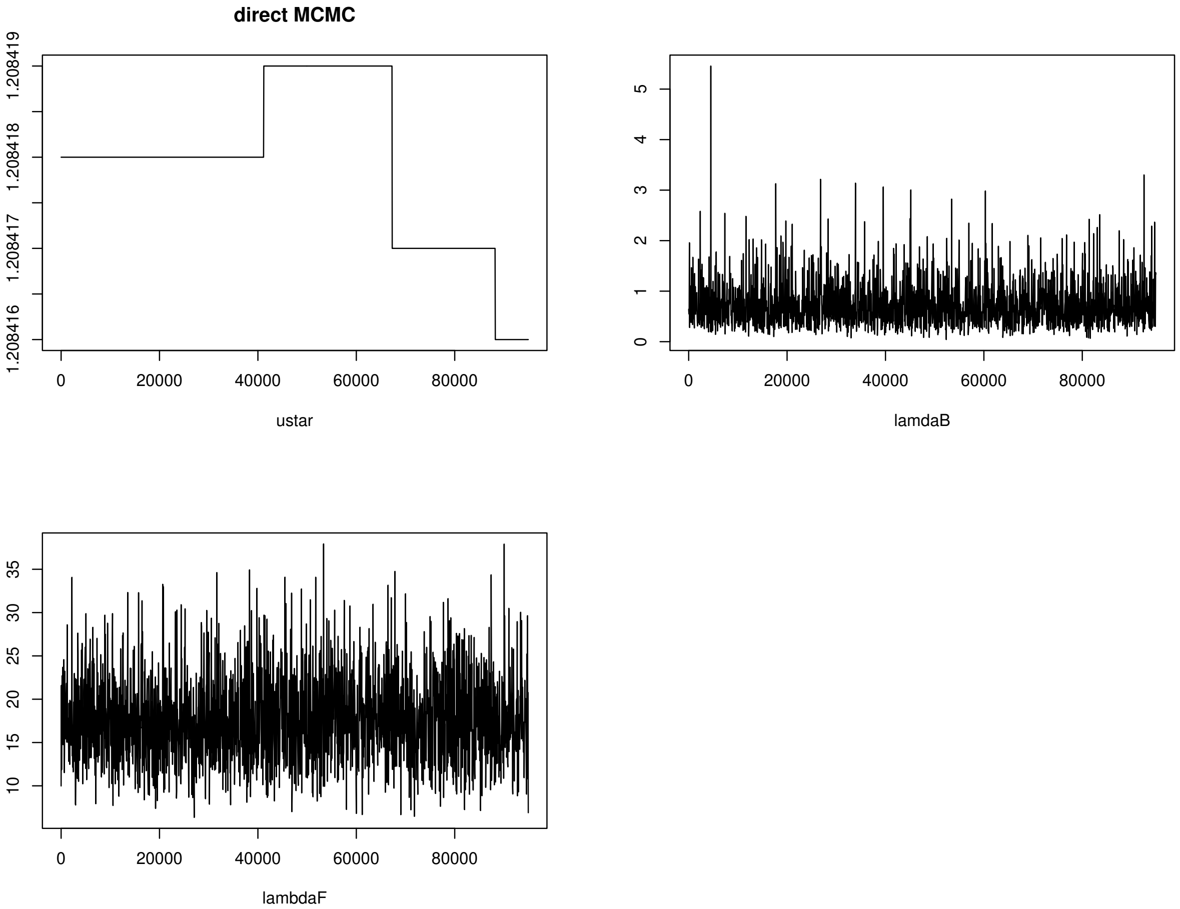
- $q(\theta \mid \theta') = q(\theta - \theta')$, ie $\theta^* = \theta^{(t)} + \varepsilon_t$ with ε_t a random perturbation with density q independent of $\theta^{(t)}$
- Typical choices for q are uniform, normal or t centered at the origin and appropriately scaled

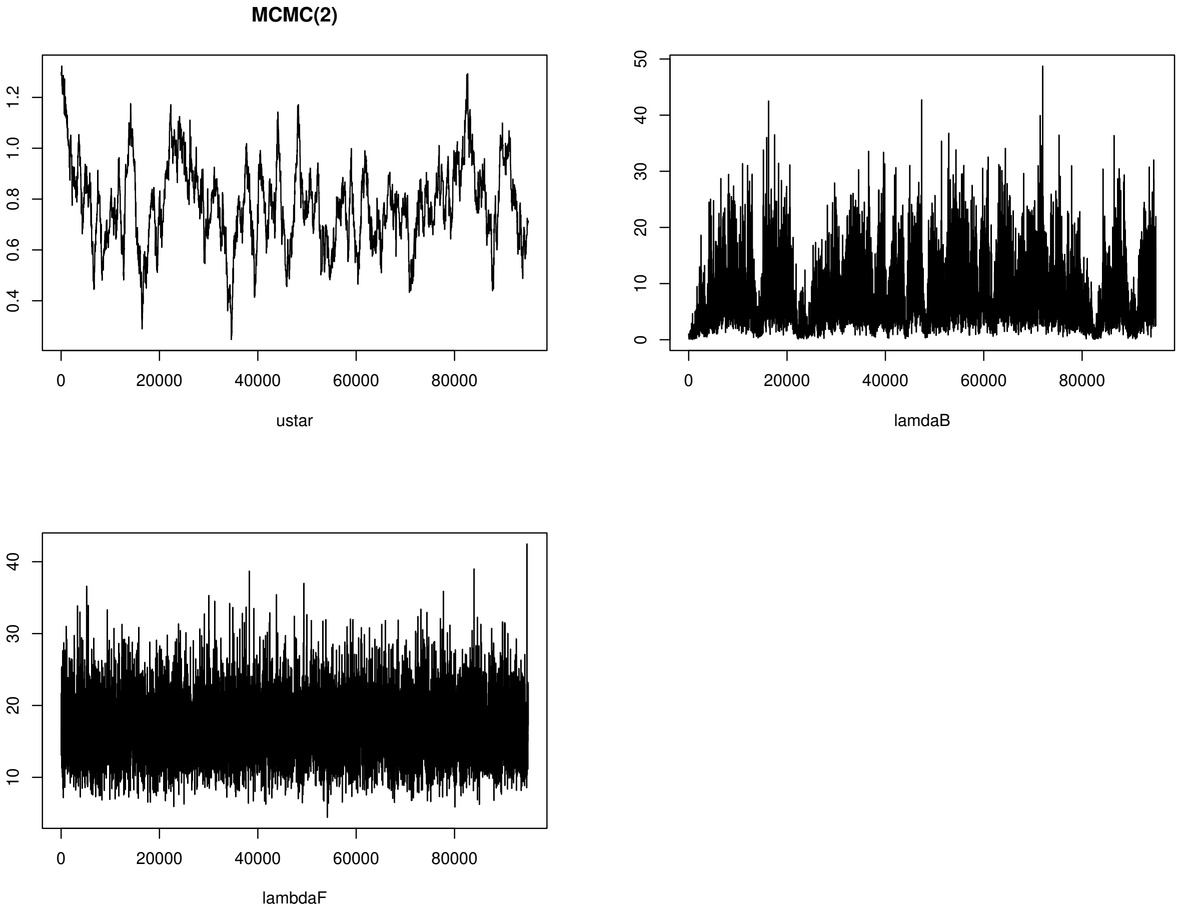
Metropolis-within-Gibbs or Hybrid MCMC:

- The Gibbs sampler as described can only be implemented if we can directly generate from all the full-conditionals $f_i(\theta_i \mid \theta_{(-i)})$
- However, the algorithm is still valid if simulation from the ith full conditional is replaced by a Metropolis-Hastings step, that is, a simulation from a proposal which is accepted according to a M-H ratio
- Typically, a number of M-H steps are done and only the last is retained (to reduce auto-correlation)

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values
- Look at traceplots to empirically ascertain convergence and decide about the length of burn-in
- Thinning: retaining only the mth iteration
- Plots of autocorrelation functions to identify highly correlated chains
- WinBUGS/Stan are software packages which automatically implement Bayesian analysis via MCMC


direct MCMC
1.208416
1.208417
1.208418
1.208419
0
20000
40000
60000
80000
ustar
0
1
2
3
4
5
0
20000
40000
60000
80000
lamdaB
10
15
20
25
30
35
0
20000
40000
60000
80000
lambdaF


MCMC(2)
0.4
0.6
0.8
1.0
1.2
0
20000
40000
60000
80000
ustar
0
10
20
30
40
50
0
20000
40000
60000
80000
lamdaB
10
20
30
40
0
20000
40000
60000
80000
lambdaF

1.7 Bibliography

- Robert, C. (2001). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Spinger
- Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer
- Gelman et al. (2004). *Bayesian Data Analysis*. Chapman & Hall.
- Marin, JM and Robert, C. (2007). *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. Springer.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.