Monte Carlo Methods

Hedibert Freitas Lopes
Professor of Statistics
hedibert.org

Insper

Outline

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*.

Monte Carlo methods

We introduce several Monte Carlo (MC) methods for integrating and/or sampling from nontrivial densities.

- MC integration
  - Simple MC integration
  - MC integration via importance sampling (IS)
- MC sampling
  - Rejection method
  - Sampling importance resampling (SIR)
- Iterative MC sampling
  - Metropolis-Hastings algorithms
  - Simulated annealing
  - Gibbs sampler

Based on the book by Gamerman and Lopes (1996).

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

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

when only

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

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:

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

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

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: $\max_d \int U(d,\theta)\pi(\theta)d\theta$.

MC integration

The objective here is to compute moments

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

If $\theta_1, \ldots, \theta_n$ is a random sample from $\pi(\cdot)$ then

$$\bar{h}_{mc} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i) \rightarrow E_\pi[h(\theta)] \qquad \text{as } n \rightarrow \infty.$$

If, additionally, $E_\pi[h^2(\theta)] < \infty$, then

$$V_\pi[\bar{h}_{mc}] = \frac{1}{n}\int \{h(\theta) - E_\pi[h(\theta)]\}^2 \pi(\theta)d\theta$$

and

$$v_{mc} = \frac{1}{n^2}\sum_{i=1}^{n}(h(\theta_i) - \bar{h}_{mc})^2 \rightarrow V_\pi[\bar{h}_{mc}] \qquad \text{as } n \rightarrow \infty.$$

Example i.

The objective here is to compute[1]

$$p = \int_0^1 [cos(50\theta) + sin(20\theta)]^2 d\theta$$

by noticing that the above integral can be rewritten as

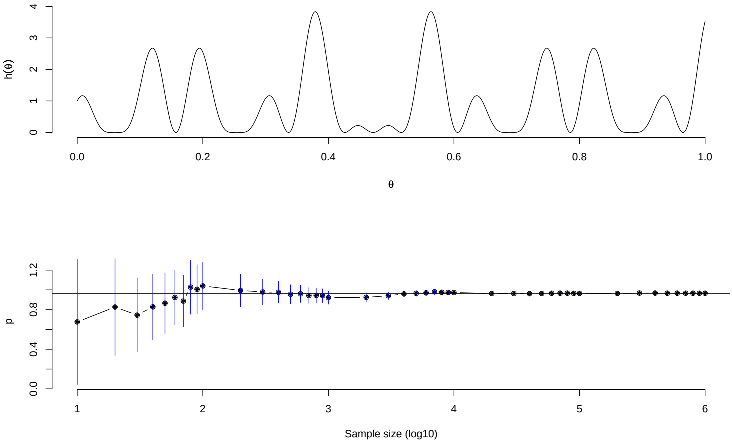
$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

where $h(\theta) = [cos(50\theta) + sin(20\theta)]^2$ and $\pi(\theta) = 1$ is the density of a $U(0,1)$. Therefore

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i)$$

where $\theta_1, \ldots, \theta_n$ are i.i.d. from $U(0,1)$.

[1] True value is 0.965.


h(θ)
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
θ
p
0.0
0.4
0.8
1.2
1
2
3
4
5
6
Sample size (log10)

MC via IS

The objective is still the same, ie to compute

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

by noticing that

$$E_\pi[h(\theta)] = \int \frac{h(\theta)\pi(\theta)}{q(\theta)} q(\theta)d\theta$$

where $q(\cdot)$ is an *importance function*.

If $\theta_1, \ldots, \theta_n$ is a random sample from $q(\cdot)$ then

$$\Rightarrow \bar{h}_{is} = \frac{1}{n} \sum_{i=1}^{n} \frac{h(\theta_i)\pi(\theta_i)}{q(\theta_i)} \to E_\pi[h(\theta)]$$

as $n \to \infty$.

Ideally, $q(\cdot)$ should be

- As *close* as possible to $h(\cdot)\pi(\cdot)$, and
- Easy to sample from.

Example ii.

The objective here is to estimate

$$p = Pr(\theta > 2) = \int_2^{\infty} \frac{1}{\pi(1+\theta^2)} d\theta = 0.1475836$$

where θ is a standard Cauchy random variable.

A natural MC estimator of p is

$$\hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} I\{\theta_i \in (2, \infty)\}$$

where $\theta_1, \ldots, \theta_n \sim$ Cauchy(0,1).

A more elaborated estimator based on a change of variables from θ to $u = 1/\theta$ is

$$\hat{p}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{u_i^{-2}}{2\pi[1 + u_i^{-2}]}$$

where $u_1, \ldots, u_n \sim U(0, 1/2)$.

The true value is $p = 0.147584$.

n	$\hat{p}_1$	$\hat{p}_2$	$v_1^{1/2}$	$v_2^{1/2}$
100	0.100000	0.1467304	0.030000	0.001004
1000	0.137000	0.1475540	0.010873	0.000305
10000	0.148500	0.1477151	0.003556	0.000098
100000	0.149100	0.1475591	0.001126	0.000031
1000000	0.147711	0.1475870	0.000355	0.000010

With only $n = 1000$ draws, $\hat{p}_2$ has roughly the same precision that $\hat{p}_1$, which is based on $1000n$ draws, ie. three orders of magnitude.

Rejection method

The objective is to draw from a target density

$$\pi(\theta) = c_\pi \tilde{\pi}(\theta)$$

when only draws from an auxiliary density

$$q(\theta) = c_q \tilde{q}(\theta)$$

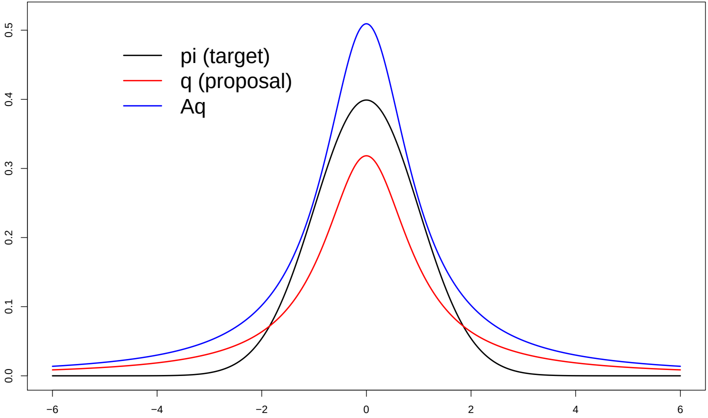
is available, for normalizing constants c_π and c_q.

If there exist a constant $A < \infty$ such that

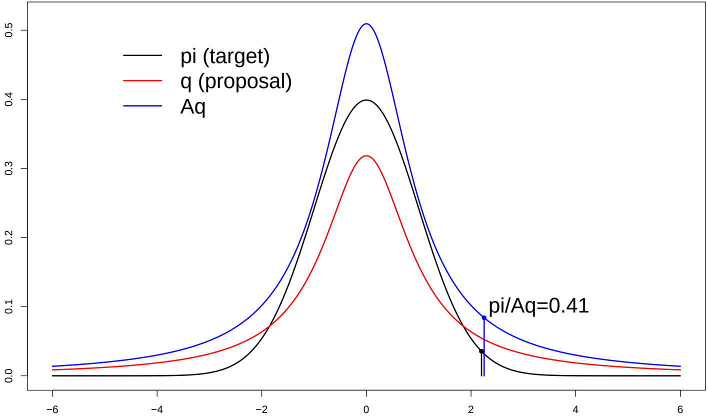
$$0 \leq \frac{\tilde{\pi}(\theta)}{A\tilde{q}(\theta)} \leq 1 \quad \text{for all } \theta$$

then $q(\theta)$ becomes a *blanketing density* or an *envelope* and A the *envelope constant*.

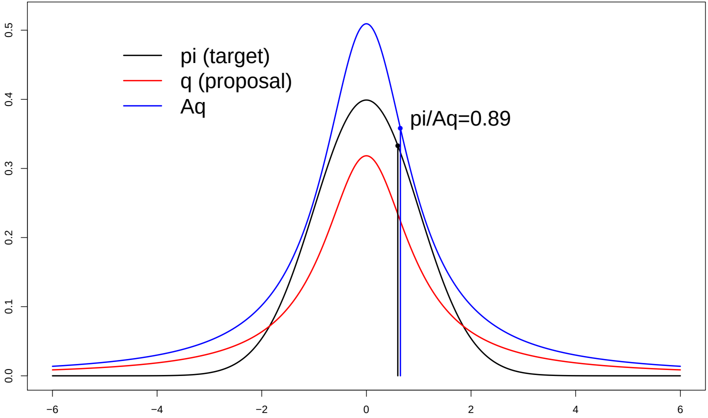
Blanket distribution



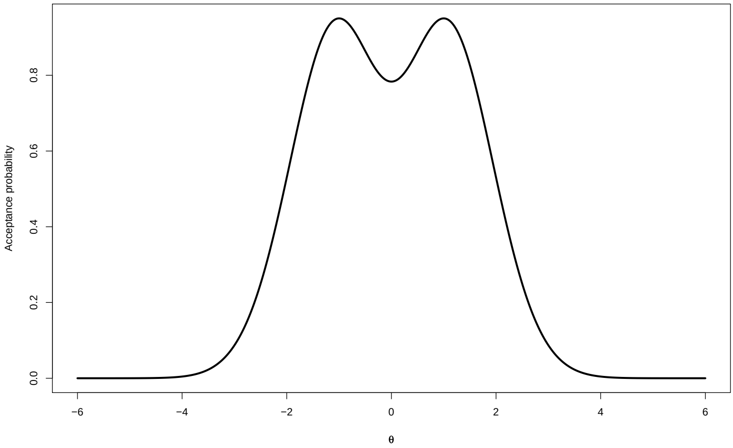
Bad draw



Good draw



Acceptance probability



Algorithm

Drawing from $\pi(\theta)$.

1. Draw θ^* from $q(\cdot)$;
2. Draw u from $U(0,1)$;
3. Accept θ^* if $u \leq \frac{\tilde{\pi}(\theta^*)}{A\tilde{q}(\theta^*)}$;
4. Repeat 1, 2 and 3 until n draws are accepted.

Normalizing constants c_π and c_q are not needed.

The theoretical acceptance rate is $\frac{c_q}{Ac_\pi}$.

The smaller the A, the larger the acceptance rate.

Example iii.

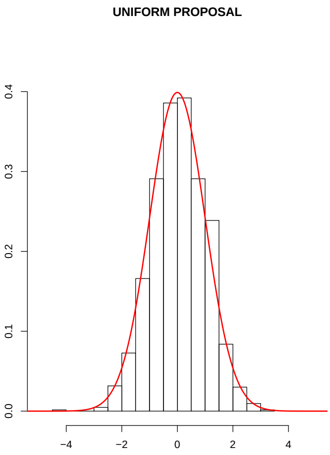
Enveloping the standard normal density

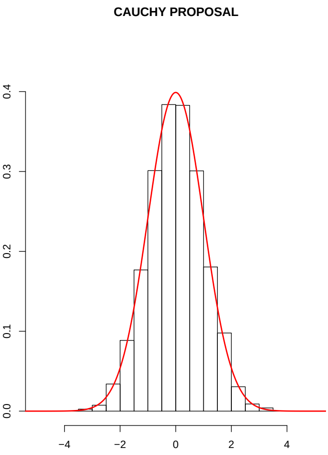
$$\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\{-0.5\theta^2\}$$

by a Cauchy density $q_C(\theta) = 1/(\pi(1+\theta^2))$, or a uniform density $q_U(\theta) = 0.05$ for $\theta \in (-10, 10)$.

Bad proposal:The maximum of $\pi(\theta)/q_U(\theta)$ is roughly $A_U = 7.98$ for $\theta \in (-10, 10)$. The theoretical acceptance rate is 12.53%.

Good proposal: The max of $\pi(\theta)/q_C(\theta)$ is equal to $A_C = \sqrt{2\pi/e} \approx 1.53$. The theoretical acceptance rate is 65.35%.





Empirical rates: 0.1265 (Uniform) and 0.6483 (Cauchy)
Theoretical rates: 0.1253 (Uniform) and 0.6535 (Cauchy)

SIR method

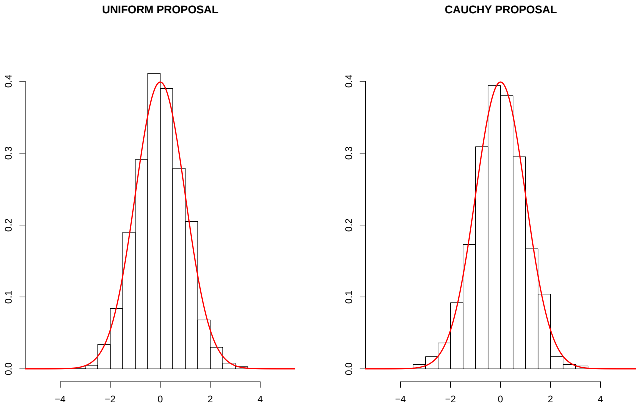
No need to rely on the existance of A!

Algorithm

1. Draw $\theta_1^*, \ldots, \theta_n^*$ from $q(\cdot)$
2. Compute (unnormalized) weights
$$\omega_i = \pi(\theta_i^*)/q(\theta_i^*) \qquad i = 1, \ldots, n$$
3. Sample θ from $\{\theta_1^*, \ldots, \theta_n^*\}$ such that
$$Pr(\theta = \theta_i^*) \propto \omega_i \qquad i = 1, \ldots, n.$$
4. Repeat m times step 3.

Rule of thumb: $n/m = 20$.
Ideally, $\omega_i = 1/n$ and $Var(\omega) = 0$.

Example iii. revisited



Fraction of redraws: 0.391 (Uniform) and 0.1335 (Cauchy)
Variance of weights: 4.675 (Uniform) and 0.332 (Cauchy)

Example iv. 3-component mixture

Assume that we are interested in sampling from

$$\pi(\theta) = \alpha_1 p_N(\theta; \mu_1, \Sigma_1) + \alpha_2 p_N(\theta; \mu_2, \Sigma_2) + \alpha_3 p_N(\theta; \mu_3, \Sigma_3)$$

where $p_N(\cdot; \mu, \Sigma)$ is the density of a bivariate normal distribution with mean vector μ and covariance matrix Σ. The mean vectors are

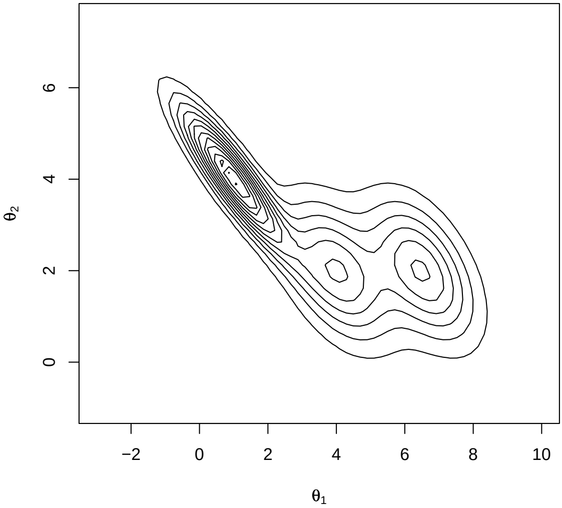
$$\mu_1 = (1, 4)' \quad \mu_2 = (4, 2)' \quad \mu_3 = (6.5, 2),$$

the covariance matrices are

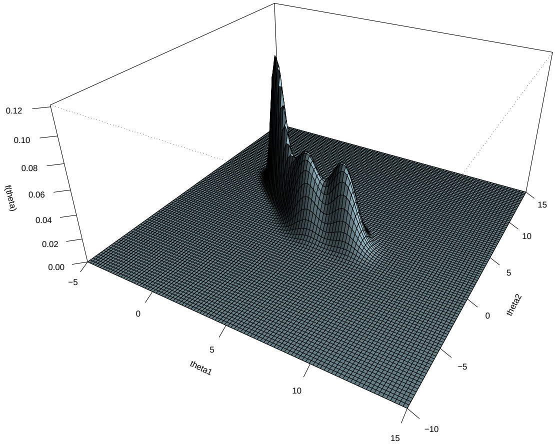
$$\Sigma_1 = \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix} \text{ and } \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{pmatrix},$$

and weights $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$.

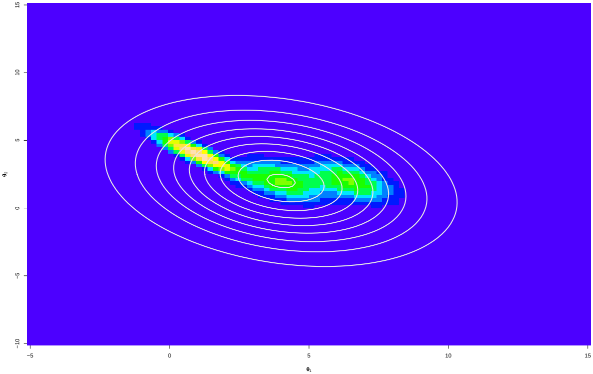
Target $\pi(\theta)$



Target $\pi(\theta)$



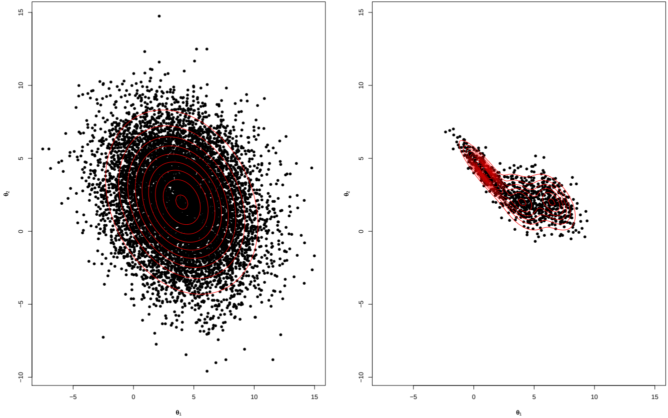
Proposal $q(\theta)$



$q(\theta) \sim N(\mu, \Sigma)$ where

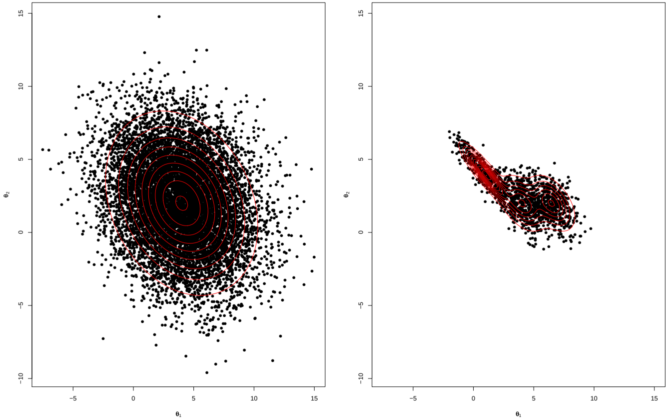
$$\mu_2 = (4,2)' \quad \text{and} \quad \Sigma = 9 \begin{pmatrix} 1.0 & -0.25 \\ -0.25 & 1.0 \end{pmatrix}$$

Rejection method



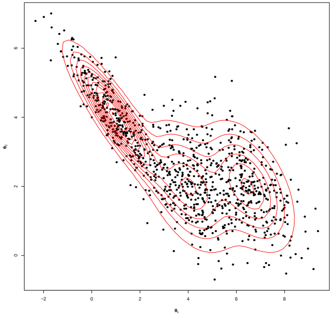
Acceptance rate: 9.91% of $n = 10,000$ draws.

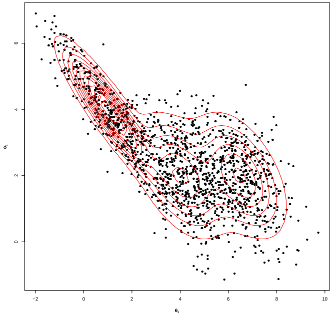
SIR method



Fraction of redraws: 29.45% of ($n = 10,000$, $m = 2,000$).

Rejection & SIR





Example v. 2-component mixture

Let us now assume that

$$\pi(\theta) = \alpha_1 p_N(\theta; \mu_1, \Sigma_1) + \alpha_3 p_N(\theta; \mu_3, \Sigma_3)$$

where mean vectors are

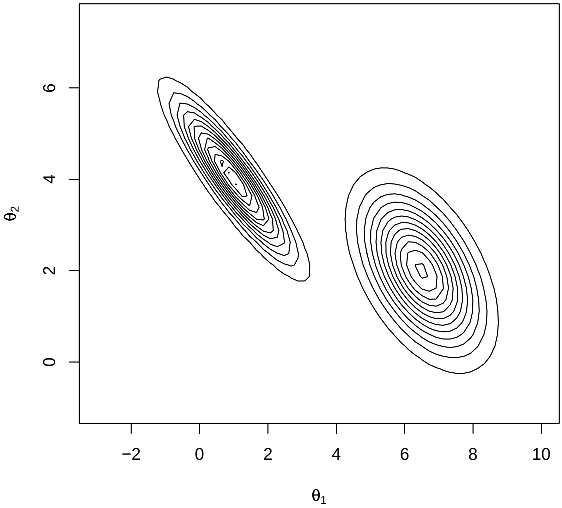
$$\mu_1 = (1, 4)' \quad \mu_3 = (6.5, 2),$$

the covariance matrices are

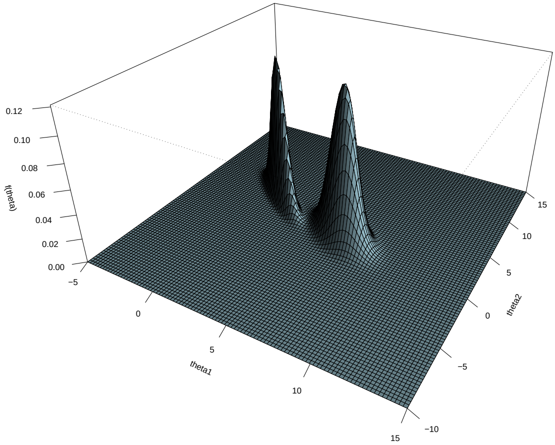
$$\Sigma_1 = \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix} \text{ and } \Sigma_3 = \begin{pmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{pmatrix},$$

and weights $\alpha_1 = 1/3$ and $\alpha_3 = 2/3$.

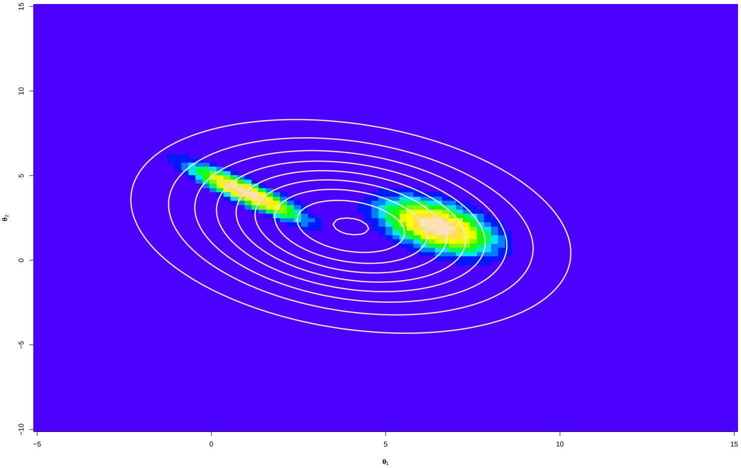
Target $\pi(\theta)$



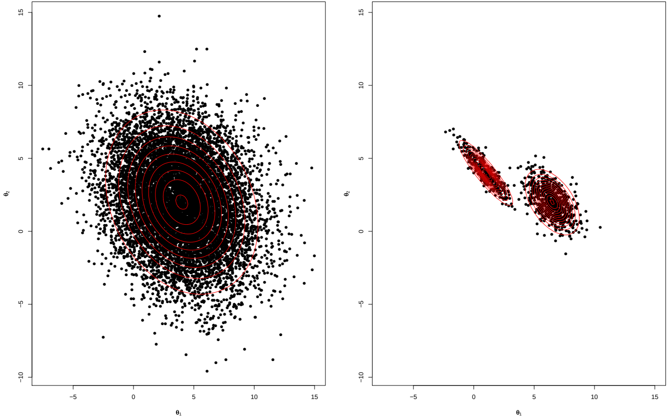
Target $\pi(\theta)$



Proposal $q(\theta)$

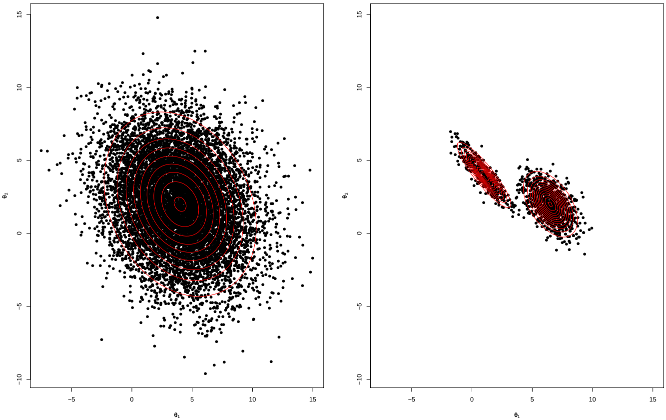


Rejection method



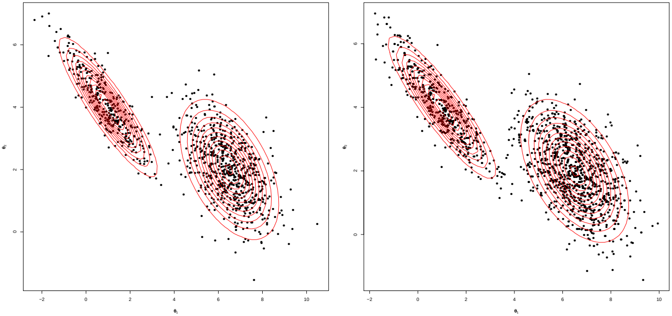
Acceptance rate: 10.1% of $n = 10,000$ draws.

SIR method



Fraction of redraws: 37.15% of $(n = 10,000, m = 2,000)$.

Rejection & SIR



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. 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.