Markov chain Monte Carlo

Oswaldo Gressani

Practical information

Oswaldo Gressani (Data Science Institute)

Email: oswaldo.gressani@uhasselt.be

Slides and additional course material (R scripts, references,…) available on my GitHub repository (https://github.com/oswaldogressani/BayesianCourse) and on Blackboard.

References used for this lecture (not exhaustive):

[1] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

[2] Lesaffre, E., & Lawson, A. B. (2012). *Bayesian Biostatistics*. John Wiley & Sons.

[3] Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

[4] Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press.

[5] Box, G. E., & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley Classics Library.

Bayes' theorem and the power of sampling

- Bayesian paradigm
- Monte Carlo integration
- Probability integral transformation

Gibbs sampling

- The two-stage Gibbs sampler
- Properties of the Gibbs sampler
- The multi-stage Gibbs sampler
- Gibbs sampling through completion
- Remarks

Metropolis algorithm

- Idea behind the Metropolis algorithm
- Graphical illustration
- Metropolis (random walk) pseudo-code
- Acceptance rates

Metropolis-Hastings algorithm

- Idea behind the Metropolis-Hastings algorithm
- Metropolis-Hastings (random walk) pseudo-code
- The independent Metropolis-Hastings algorithm
- Avoiding numeric overflow
- Remarks

Why does MCMC work?

Exercise

Take home messages

Bayes' theorem and the power of sampling

Bayesian paradigm

Let $\mathcal{D} = \{y_1, \ldots, y_n\}$ denote a sample of n i.i.d. observations with $y_i \sim f(y|\boldsymbol{\theta})$.

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top \in \mathbb{R}^K$ parameter vector with $K \in \mathbb{N}$.

Bayes' theorem $p(\boldsymbol{\theta}|\mathcal{D}) = \dfrac{\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})p(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})p(\boldsymbol{\theta})$.

"Posterior is proportional to the prior times the likelihood"

The goal of Bayesian inference is to explore the (conditional) probability density $p(\boldsymbol{\theta}|\mathcal{D})$ and compute certain features that characterize the posterior distribution.

Bayesian paradigm

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\, p(\boldsymbol{\theta})$$

The likelihood $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ is seen as a function of $\boldsymbol{\theta}$ given the data $\mathcal{D}$ (Fisherian view).

The prior $p(\boldsymbol{\theta})$ summarizes the belief about plausible values of $\boldsymbol{\theta}$ before sampling.

The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ updates knowledge about $\boldsymbol{\theta}$ by making use of the data $\mathcal{D}$.

Knowing the posterior only up to a multiplicative constant is fine as the normalizing constant only acts as a scaling effect and does not distort the shape of $p(\boldsymbol{\theta}|\mathcal{D})$.

Beta-Binomial model and conjugacy

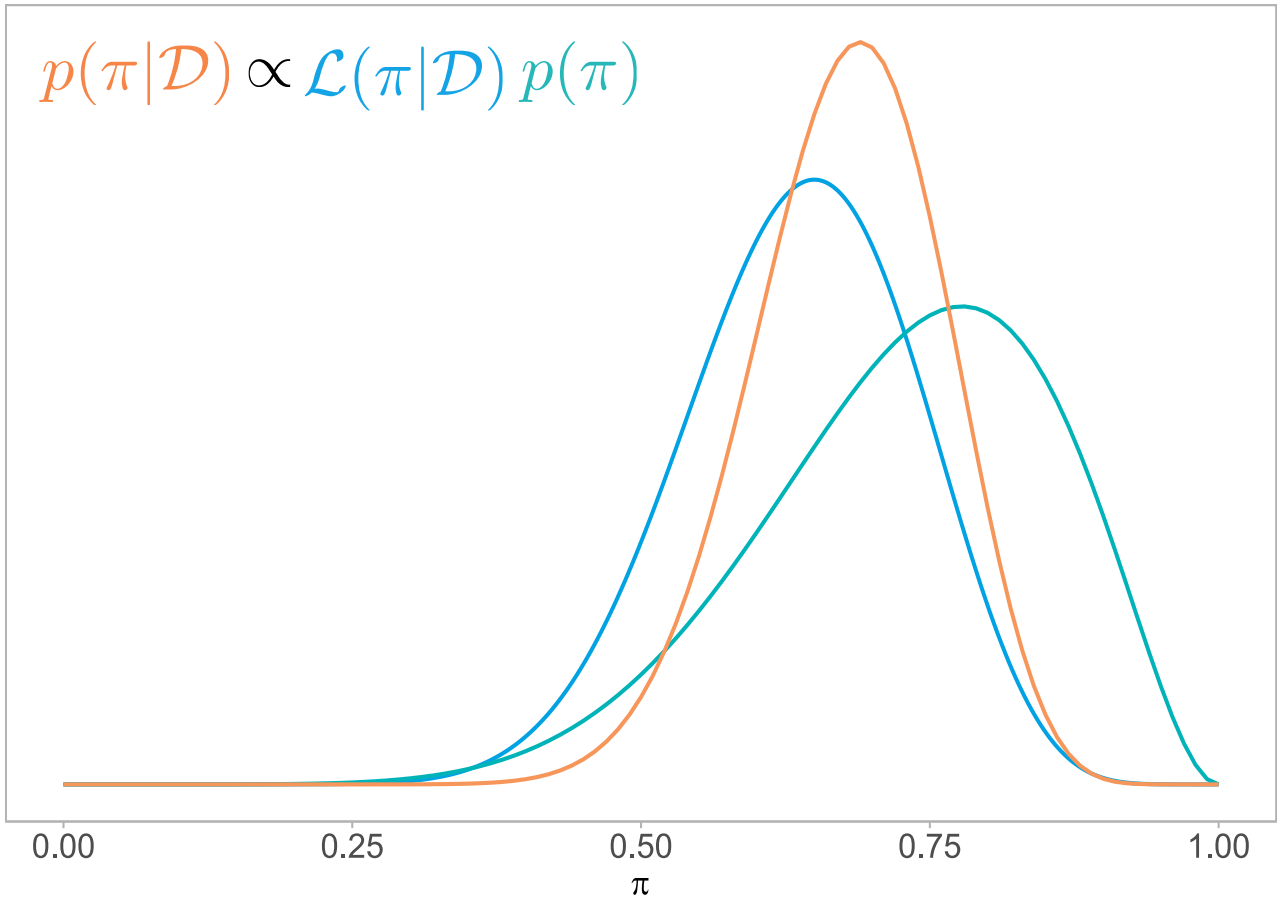


Fig 1. Posterior, likelihood and prior for the Beta-Binomial model.

Let y be a binomial variable representing the number of successes in n trials.

Each trial has success probability $\pi \in [0, 1]$.

Assume a Beta prior $\pi \sim \text{Beta}(a, b)$ with $a > 0, b > 0$.

Posterior $(\pi|\mathcal{D}) \sim \text{Beta}(a + y, b + n - y)$.

Posterior mean $\widehat{\pi} = \mathbb{E}(\pi|\mathcal{D}) = \dfrac{a + y}{a + b + n}$.

Cauchy model

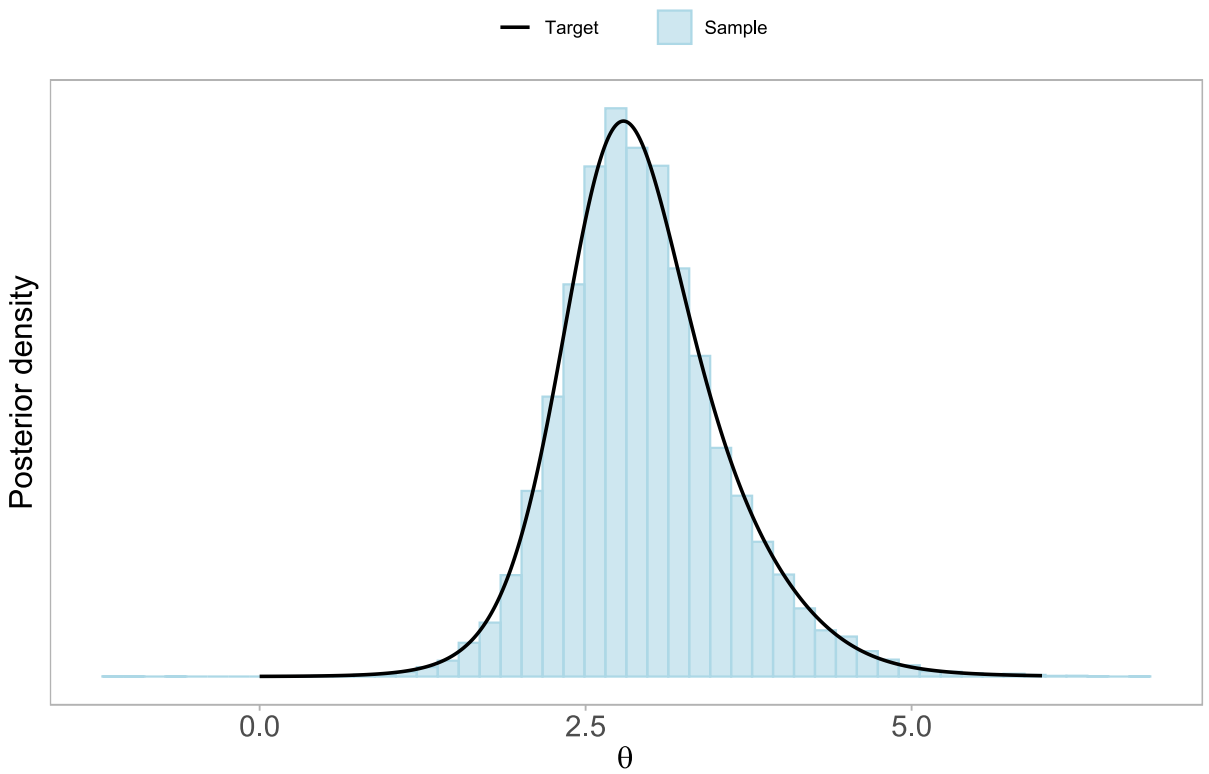


Fig 2. Normalized posterior density (solid) and histogram of the samples drawn from $p(\theta|\mathcal{D})$ in the Cauchy model.

Consider an i.i.d. sample from a Cauchy distribution $\mathcal{D} = \{y_1, \ldots, y_n\} \sim \mathcal{C}(\theta, 1)$.

Location parameter $\theta \in \mathbb{R}$ and scale equal to 1.

Gaussian prior $\theta \sim \mathcal{N}(0, \tau^{-1})$.

By Bayes' theorem the posterior is shown to be

$$p(\theta|\mathcal{D}) \propto \frac{\exp(-0.5\tau\theta^2)}{\prod_{i=1}^{n}(1 + (y_i - \theta)^2)}$$

Posterior mean is not analytically available!

Sampling and the Monte Carlo principle is the solution!

Monte Carlo integration

In Bayesian statistics estimators of interest often appear as posterior expectations, (e.g. posterior mean, variance, standard deviation).

These expectations are usually expressed as integrals without analytic solution in most cases.

Let $\theta \sim f$ and assume we are interested in solving generic integrals of the form

$$\mathbb{E}(g(\theta)) = \int g(\theta) f(\theta) d\theta$$

Given a sample $\theta^{(1)}, \ldots, \theta^{(M)} \underset{\text{i.i.d.}}{\sim} f$, the above expectation can be approximated by the Monte Carlo estimator:

$$\widehat{\mathbb{E}}(g(\theta)) = \frac{1}{M} \sum_{m=1}^{M} g\left(\theta^{(m)}\right) \overset{a.s.}{\to} \mathbb{E}(g(\theta))$$

Probability integral transformation

Sampling $X \sim f$ plays an important role in Bayesian analysis.

Theorem (Probability integral transform)

Let X be a continuous random variable with (strictly) increasing cdf F and inverse F^{-1}. Define $U = F(X) \in (0,1)$. Then U has a uniform distribution on $(0,1)$ and $F^{-1}(U)$ is distributed like X, that is, $F^{-1}(U) \stackrel{d}{=} X$.

Inversion method to sample $X \sim f$

---

1. Draw $U \sim \mathcal{U}(0,1)$
2. Compute $X = F^{-1}(U)$
3. Return X

Proof (Probability integral transformation)

To show that $U = F(X)$ has a uniform distribution in $(0, 1)$, note that for $0 < u < 1$:

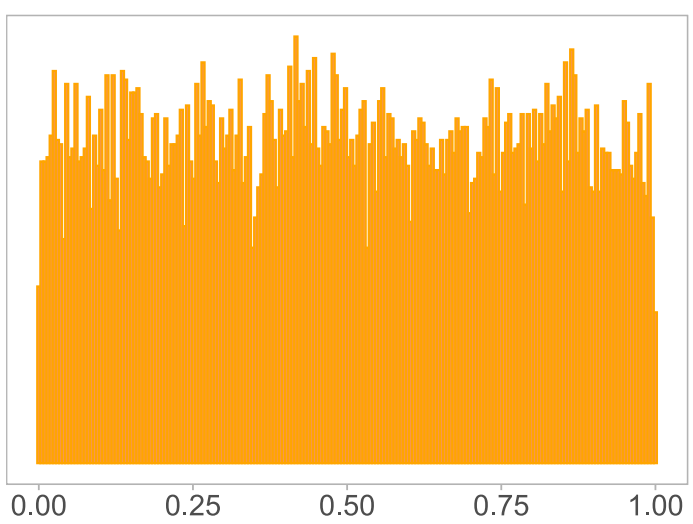
$$\begin{aligned} P(U \leq u) = P(F(X) \leq u) &= P(F^{-1}(F(X)) \leq F^{-1}(u)) \\ &= P(X \leq F^{-1}(u)) \\ &= F(F^{-1}(u)) = u \end{aligned}$$

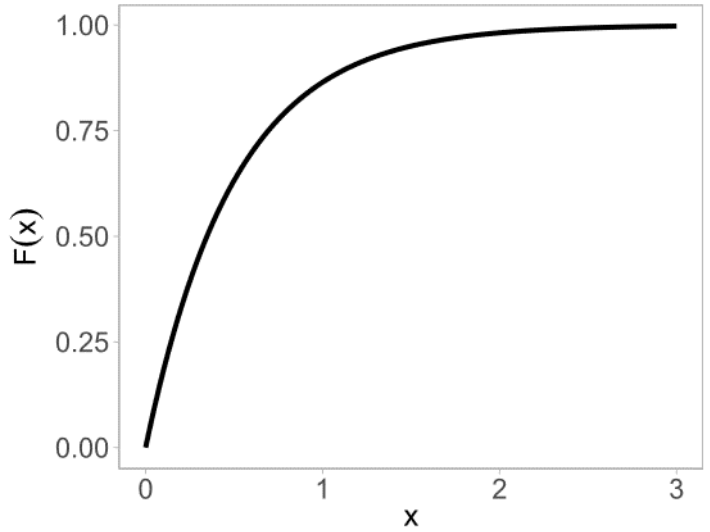
It is also easily shown that $P(U \leq u) = 0$ for $u \leq 0$ and $P(U \leq u) = 1$ for $u \geq 1$, hence U has a uniform distribution in $(0, 1)$.

To show that $F^{-1}(U)$ is distributed like X write:

$$\begin{aligned} & P(U \leq u) = P(U \leq F(x)) = F(x) \\ \Rightarrow\ & P(F^{-1}(U) \leq x) = F(x) = P(X \leq x) \\ \Rightarrow\ & X \stackrel{d}{=} F^{-1}(U) \quad \square \end{aligned}$$

Illustration of the inversion method





$$X \sim \mathcal{E}(2)$$

$$F^{-1}(U) = -0.5 \log(1 - U)$$

Cauchy model

We are interested in computing the posterior feature $\mathbb{E}(\theta|\mathcal{D}) = \int_{\mathbb{R}} \theta \; p(\theta|\mathcal{D}) d\theta$.

Using Bayes' theorem $\mathbb{E}(\theta|\mathcal{D}) = \int_{\mathbb{R}} \theta \; \frac{\mathcal{L}(\theta|\mathcal{D})p(\theta)}{\left(\int_{\mathbb{R}} \mathcal{L}(\theta|\mathcal{D})p(\theta) \; d\theta \right)} \; d\theta$.

Can use numerical integration methods, e.g. Riemann numerical integration (RNI).

$$\widehat{\mathbb{E}}(\theta|\mathcal{D}) \overset{RNI}{=} \sum_{m=0}^{M-1} \theta^{(m)} \; \frac{\mathcal{L}(\theta^{(m)}|\mathcal{D})p(\theta^{(m)})}{\left(\sum_{m=0}^{M-1} \mathcal{L}(\theta^{(m)}|\mathcal{D})p(\theta^{(m)}) \; \Delta \right)} \; \Delta \, ,$$

with an equidistant grid $\Delta = \theta^{(m)} - \theta^{(m-1)}, \; m = 1, \ldots, M$ on a compact support capturing most of the posterior probability mass.

Cauchy model

Using an equidistant grid of size $M = 50000$ on $[-10, 10]$, we get $\widehat{E}(\theta|\mathcal{D}) \overset{RNI}{=} 2.9401$.

From a sample (Metropolis) of size $M = 50000$, we get $\widehat{E}(\theta|\mathcal{D}) = \frac{1}{M}\sum_{m=1}^{M}\theta^{(m)} = 2.9442$.

The true location parameter in the data generating mechanism is $\theta = 3$.

Viewing $\widehat{E}(\theta|\mathcal{D})$ as a point estimate of θ, the sampling approach and the numerical integration approach seem to deliver similar performance (in terms of "closeness" to θ).

However, the computational burden of numerical integration rises drastically with the dimension of $\boldsymbol{\theta}$, while it is typically much smaller with sampling techniques.

The Gibbs sampler

Hammersley-Clifford theorem

Conditional distributions contain sufficient information to summarize the joint distribution.

This allows to set up an iterative algorithm where a sample from the joint distribution is obtained by sampling the conditional distributions.

If obtaining random draws from conditionals is fairly simple, Gibbs sampling is convenient.

Theorem (Hammersley-Clifford)

If a joint distribution has conditional densities $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, then the joint density $p(\theta_1, \theta_2)$ can be reconstructed from the conditional densities:

$$p(\theta_1, \theta_2) = \frac{p(\theta_2|\theta_1)}{\int \left(p(\theta_2|\theta_1)/p(\theta_1|\theta_2)\right)\, d\theta_2}$$

Proof (Hammersley-Clifford)

$$p(\theta_1, \theta_2) = p(\theta_1|\theta_2)p(\theta_2) = p(\theta_2|\theta_1)p(\theta_1)$$

$$\frac{p(\theta_2|\theta_1)}{p(\theta_1|\theta_2)} = \frac{p(\theta_2)}{p(\theta_1)} \Rightarrow \int \frac{p(\theta_2|\theta_1)}{p(\theta_1|\theta_2)}\, d\theta_2 = \int \frac{p(\theta_2)}{p(\theta_1)}\, d\theta_2 = \frac{1}{p(\theta_1)} \int p(\theta_2)\, d\theta_2 = \frac{1}{p(\theta_1)}$$

$$\Rightarrow p(\theta_1) = \left(\int \left(p(\theta_2|\theta_1)/p(\theta_1|\theta_2) \right)\, d\theta_2 \right)^{-1}$$

$$p(\theta_1, \theta_2) = p(\theta_2|\theta_1)p(\theta_1)$$

$$= \frac{p(\theta_2|\theta_1)}{\int \left(p(\theta_2|\theta_1)/p(\theta_1|\theta_2) \right)\, d\theta_2} \quad \square$$

The Gibbs sampler

Introduced by Geman and Geman (1984) in models for image-processing

The Gibbs sampler was revived by Gelfand and Smith (1990) who outlined its potential use in Bayesian statistics.

Gibbs sampling is an iterative technique to obtain a random sample from (marginal) posterior distributions without having to analytically/numerically compute the corresponding densities.

The Gibbs sampler is a Markov chain Monte Carlo method.

The two-stage Gibbs sampler

Using the Hammersley-Clifford theorem in a Bayesian context, we know that the joint posterior distribution $p(\theta_1, \theta_2|\mathcal{D})$ can entirely be summarized by the conditional posterior distributions $p(\theta_1|\theta_2, \mathcal{D})$ and $p(\theta_2|\theta_1, \mathcal{D})$.

This motivates the following iterative scheme to generate a Markov chain $\left(\theta_1^{(m)}, \theta_2^{(m)}\right)$:

Two-stage Gibbs sampler

---

1. Choose a starting value $\theta_1^{(0)}$
2. for m in 1 to M do
3. Sample $\theta_2^{(m)} \sim p(\theta_2|\theta_1^{(m-1)}, \mathcal{D})$
4. Sample $\theta_1^{(m)} \sim p(\theta_1|\theta_2^{(m)}, \mathcal{D})$
5. end for

Properties of the Gibbs sampler

The generated sequence of vectors $\boldsymbol{\theta}^{(m)} = \left(\theta_1^{(m)}, \theta_2^{(m)}\right)^\top$ for $m = 1, \ldots, M$ are **dependent** and the chain has the Markov property:

$$p(\boldsymbol{\theta}^{(m)} | \boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta}^{(m-2)}, \ldots, \boldsymbol{\theta}^{(1)}, \mathcal{D}) = p(\boldsymbol{\theta}^{(m)} | \boldsymbol{\theta}^{(m-1)}, \mathcal{D})$$

The chain is dependent on the initial seed (starting value).

In the early phase of the chain, the generated sequence of vectors may be located in an "uninformative" support of the target posterior distribution.

This is a region in which the probability mass under the curve of the target posterior is low.

We typically leave out this "uninformative" part of the chain when computing posterior summary statistics ⟶ burn-in or warm-up.

Under mild regularity conditions Gibbs sampler generates samples from the target distribution.

Two-stage Gibbs sampler Example I

Consider the following bivariate family of distributions (Geman and Meng 1991):

$$p(\theta_1, \theta_2|\mathcal{D}) \propto \exp\left\{-\frac{1}{2}\left(A\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 2B\theta_1\theta_2 - 2C_1\theta_1 - 2C_2\theta_2\right)\right\}$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$ and scalars $A > 0, B, C_1, C_2 \in \mathbb{R}$.

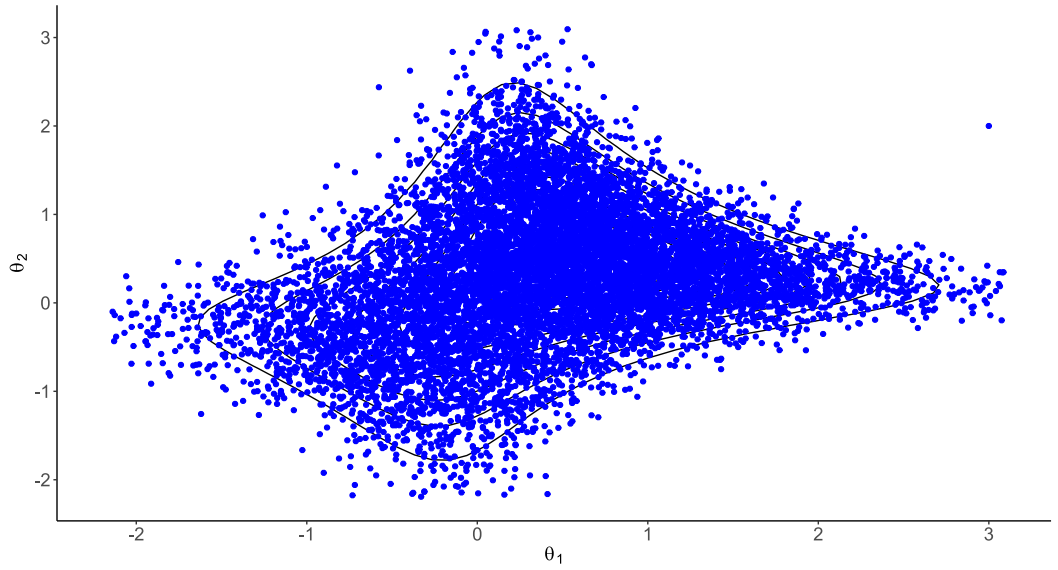
$$p(\theta_1|\theta_2, \mathcal{D}) \propto \exp\left\{-\frac{1}{2}\left(A\theta_1^2\theta_2^2 + \theta_1^2 - 2B\theta_1\theta_2 - 2C_1\theta_1\right)\right\}$$

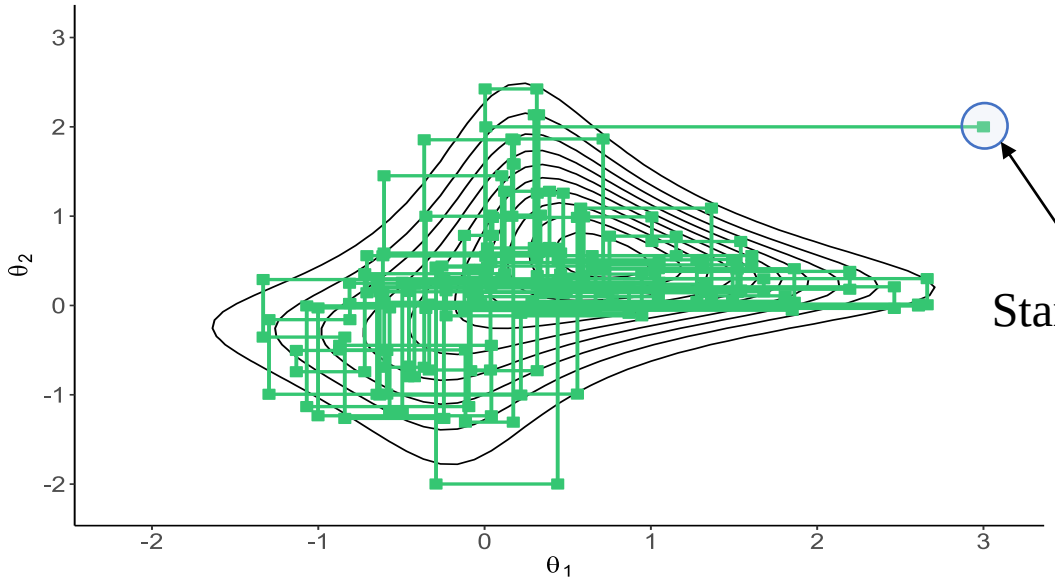
$$\Rightarrow (\theta_1|\theta_2, \mathcal{D}) \sim \mathcal{N}\left(\frac{B\theta_2 + C_1}{1 + A\theta_2^2}, \frac{1}{1 + A\theta_2^2}\right)$$

$$p(\theta_2|\theta_1, \mathcal{D}) \propto \exp\left\{-\frac{1}{2}\left(A\theta_1^2\theta_2^2 + \theta_2^2 - 2B\theta_1\theta_2 - 2C_2\theta_2\right)\right\}$$

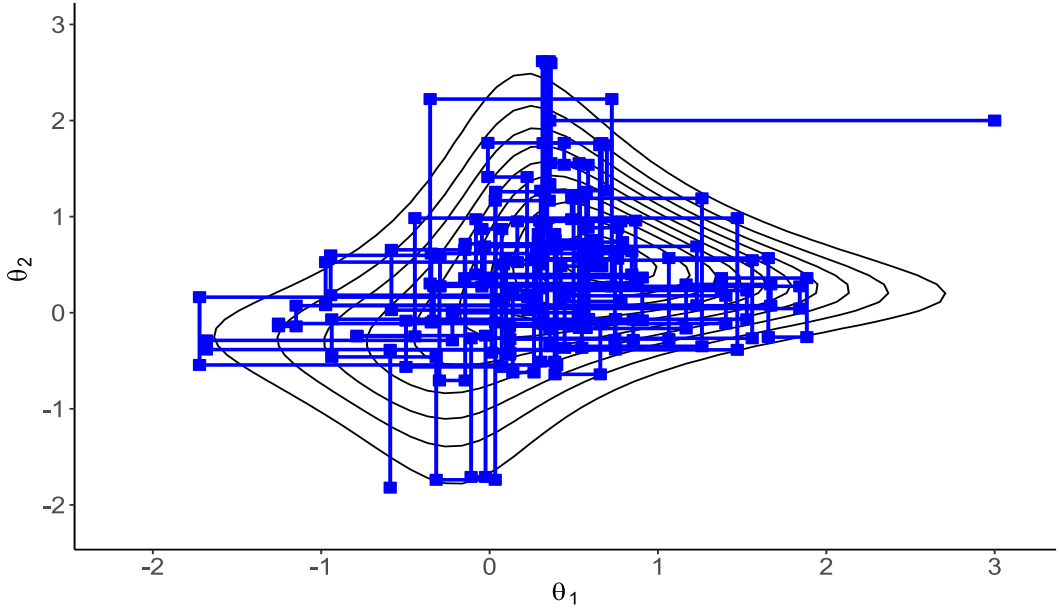
$$\Rightarrow (\theta_2|\theta_1, \mathcal{D}) \sim \mathcal{N}\left(\frac{B\theta_1 + C_2}{1 + A\theta_1^2}, \frac{1}{1 + A\theta_1^2}\right)$$

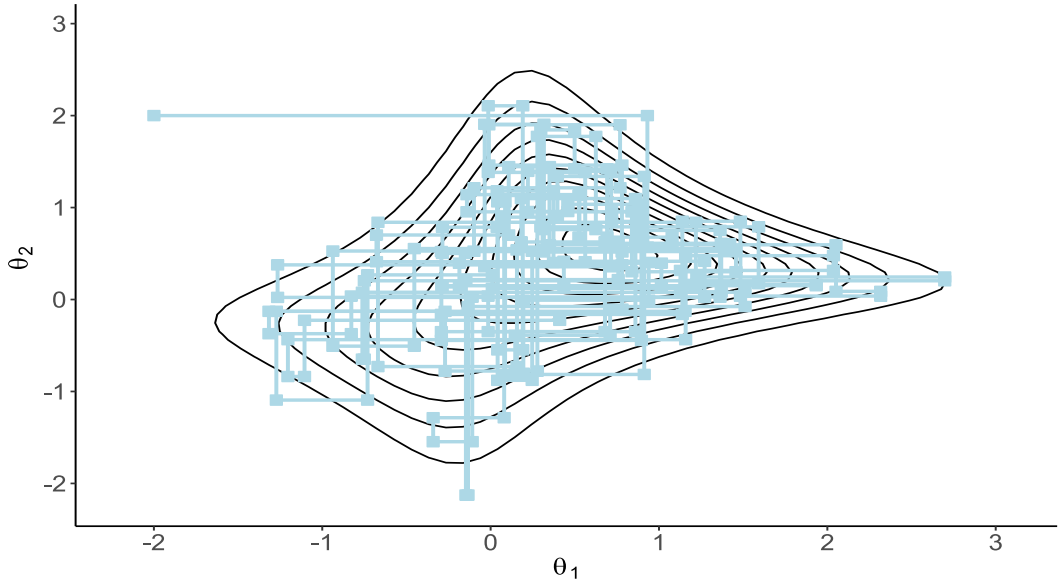
What happens in the (θ_1, θ_2) space?





Starting point





Example from Casella and George (1992)

Given the following joint distribution:

$$p(\theta_1, \theta_2|\mathcal{D}) \propto \frac{n!}{\theta_1!(n-\theta_1)!}\theta_2^{\theta_1+\alpha-1}(1-\theta_2)^{n-\theta_1+\beta-1} \text{ with } \theta_1 \in \{0, 1, \ldots, n\}, \theta_2 \in [0, 1].$$

Assume we are interested in summarizing the marginal posterior $p(\theta_1|\mathcal{D})$.

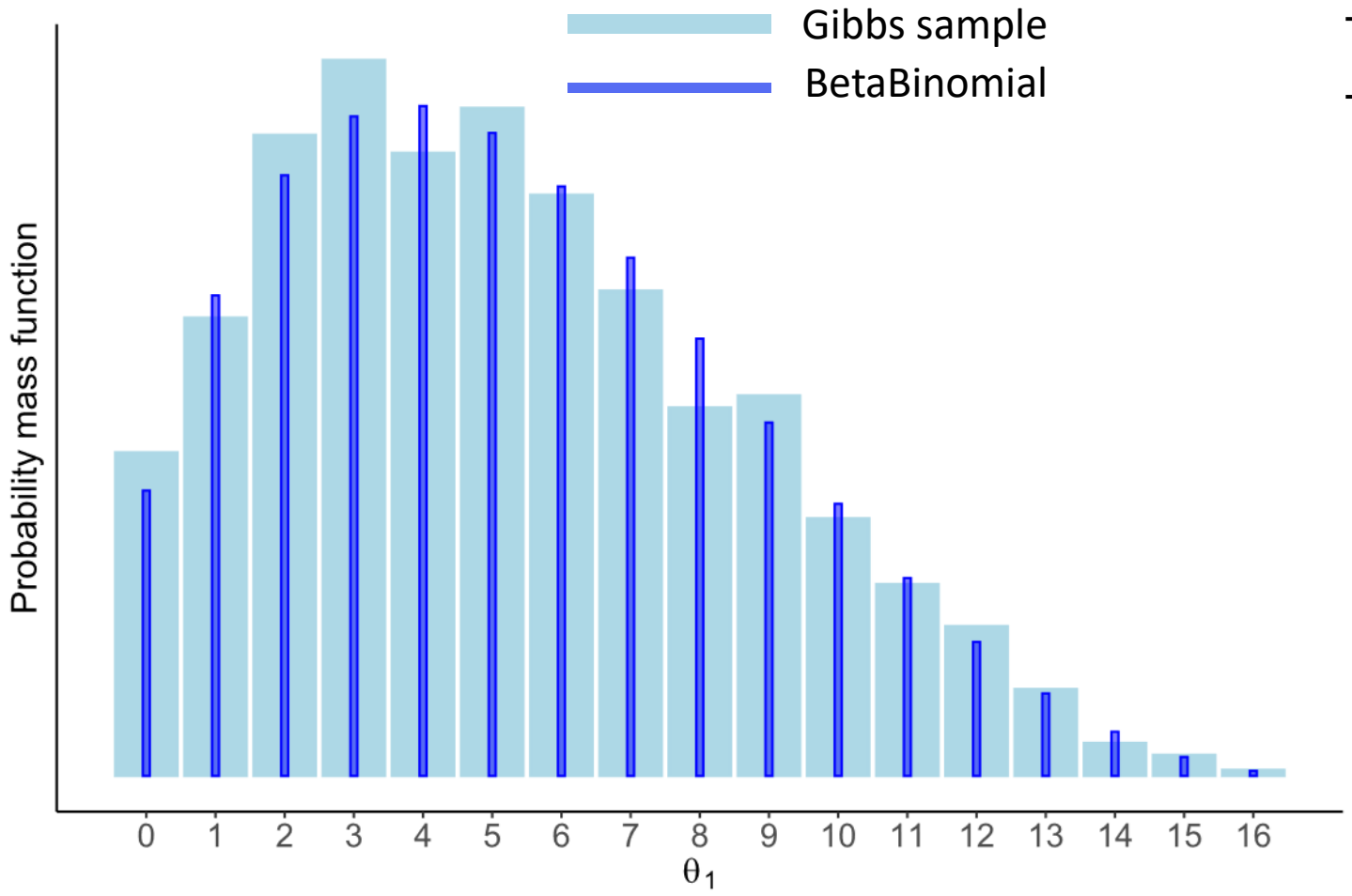
The following conditional distributions can be derived:

$$(\theta_2|\theta_1, \mathcal{D}) \sim \text{Beta}(\theta_1 + \alpha, n - \theta_1 + \beta)$$
$$(\theta_1|\theta_2, \mathcal{D}) \sim \text{Bin}(n, \theta_2)$$

Note also that the marginal is analytically available $(\theta_1|\mathcal{D}) \sim \text{BetaBinomial}(n, \alpha, \beta)$.

Example from Casella and George (1992)

With $n = 16$, $\alpha = 2$ and $\beta = 4$.



Gibbs sampler

1. Fix $\theta_2^{(0)} = 0.5$.
2. for m in 1 to M do
3. $\left(\theta_1^{(m)} | \theta_2^{(m-1)}, \mathcal{D}\right) \sim \text{Bin}\left(n, \theta_2^{(m-1)}\right)$
4. $\left(\theta_2^{(m)} | \theta_1^{(m)}, \mathcal{D}\right) \sim \text{Beta}\left(\theta_1^{(m)} + \alpha, n - \theta_1^{(m)} + \beta\right)$
5. end for

The multi-stage Gibbs sampler

Given data $\mathcal{D}$, we are interested in posterior features of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top \in \mathbb{R}^K$.

Assume that the following (univariate) conditional distributions are available

$$\underbrace{p(\theta_k | \boldsymbol{\theta}_{-k}, \mathcal{D})}_{\text{“full conditionals”}} = p(\theta_k | \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K, \mathcal{D})$$

and that a random draw can be obtained, i.e. $\theta_k \sim p(\theta_k | \boldsymbol{\theta}_{-k}, \mathcal{D})$ for $k = 1, \ldots, K$.

The multi-stage Gibbs sampler

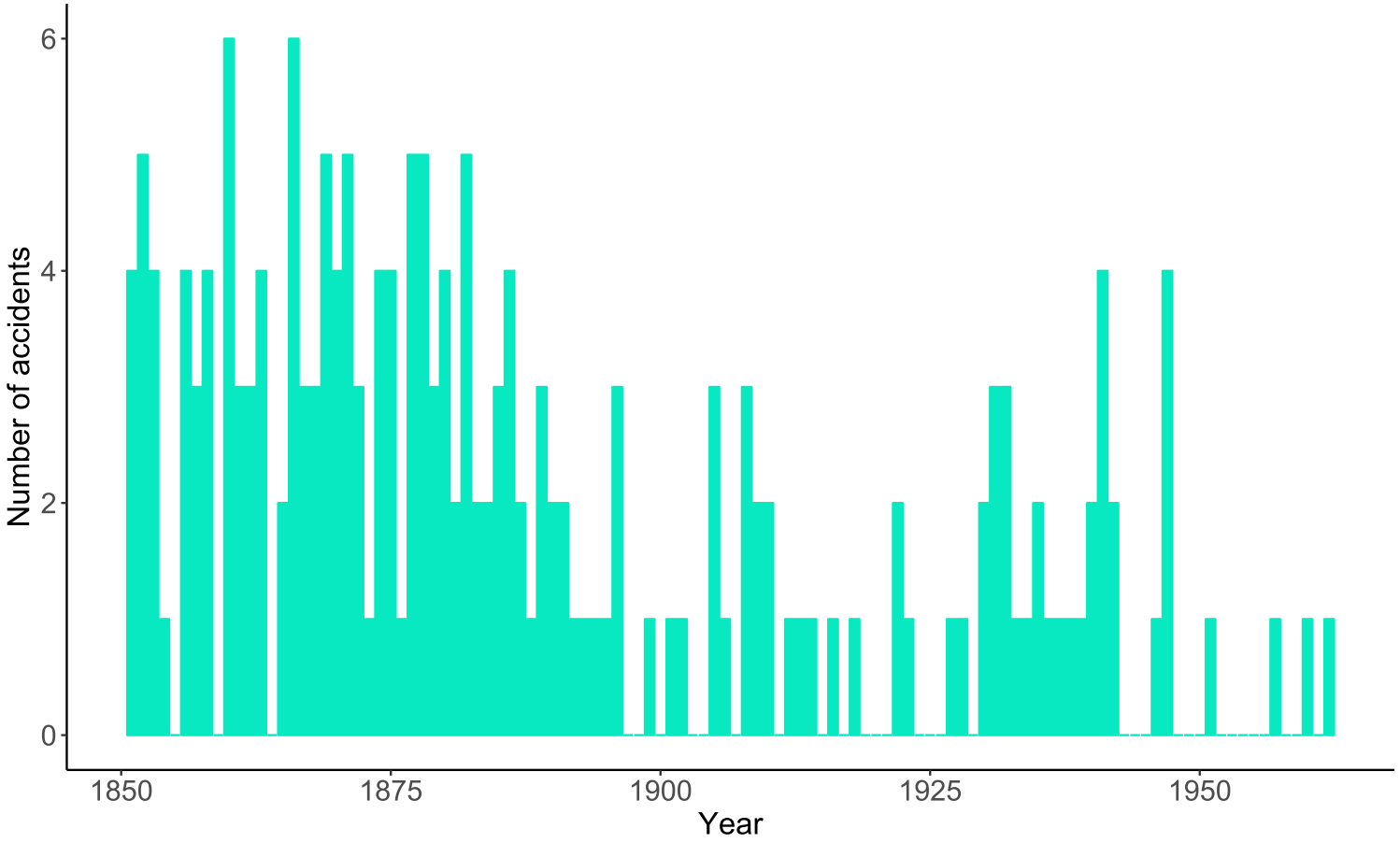
Gibbs sampler

1. Choose a starting value $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^K$
2. for m in 1 to M do
   - 2.1. $\theta_1^{(m)} \sim p\left(\theta_1 | \theta_2^{(m-1)}, \ldots, \theta_K^{(m-1)}, \mathcal{D}\right)$
   - 2.2. $\theta_2^{(m)} \sim p\left(\theta_2 | \theta_1^{(m)}, \theta_3^{(m-1)}, \ldots, \theta_K^{(m-1)}, \mathcal{D}\right)$
   - $\vdots$
   - 2.K. $\theta_K^{(m)} \sim p\left(\theta_K | \theta_1^{(m)}, \ldots, \theta_{K-1}^{(m)}, \mathcal{D}\right)$
3. end for

Under mild regularity conditions $\boldsymbol{\theta}^{(m)}$, $\boldsymbol{\theta}^{(m+1)}$,… can be viewed as draws from $p(\boldsymbol{\theta}|\mathcal{D})$.

Coal mining accidents in Britain

Data on counts of coal mining accidents in Great Britain from 1851 to 1962.



Year	Number of accidents
1851	4
1852	5
1853	4
⋮	⋮
1961	0
1962	1

Poisson model with change point

Observed counts $\mathcal{D} = \{y_1, \ldots, y_n\}$ with $y_i \in \mathbb{N}$ and sample size $n = 112$.

Observation process assumed to be Poisson with change point at $k \in \{1, \ldots, n\}$.

$$y_i \sim \text{Poisson}(\theta_1), \; i = 1, \ldots, k \text{ with } \theta_1 > 0\,.$$

$$y_i \sim \text{Poisson}(\theta_2), \; i = k+1, \ldots, n \text{ with } \theta_2 > 0.$$

Priors

$$\theta_1 \sim \mathcal{G}(a_1, b_1)$$
$$\theta_2 \sim \mathcal{G}(a_2, b_2)$$
$$a_1 = a_2 = 0.5$$

$$b_1 \sim \mathcal{G}(c_1, d_1)$$
$$b_2 \sim \mathcal{G}(c_2, d_2)$$
$$c_1 = c_2 = 0$$
$$d_1 = d_2 = 1$$

Discrete uniform prior

$$p(k) = \frac{1}{n}$$

Poisson model with change point

Bayes' theorem

$$p(\theta_1, \theta_2, b_1, b_2, k|\mathcal{D}) \propto \mathcal{L}(\theta_1, \theta_2, k|\mathcal{D}) p(\theta_1|b_1) p(\theta_2|b_2) p(b_1) p(b_2) p(k)$$

Full conditional posterior distributions

$$(\theta_1|\theta_2, b_1, b_2, k, \mathcal{D}) \sim \mathcal{G}\left(\sum_{i=1}^{k} y_i + a_1, b_1 + k\right)$$

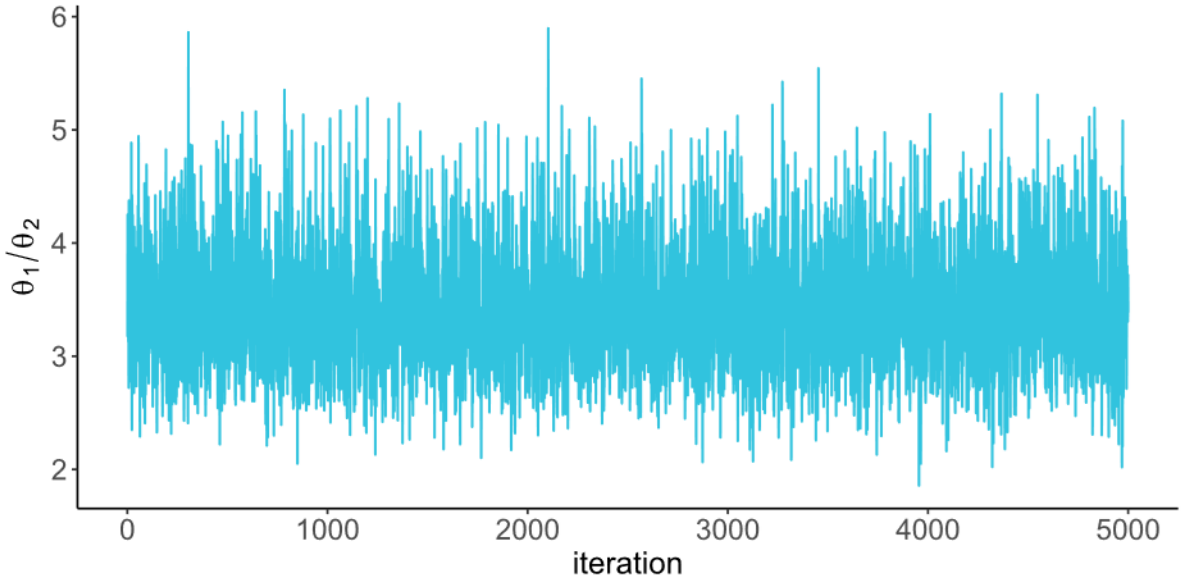
$$(\theta_2|\theta_1, b_1, b_2, k, \mathcal{D}) \sim \mathcal{G}\left(\sum_{i=k+1}^{n} y_i + a_2, n - k + b_2\right)$$

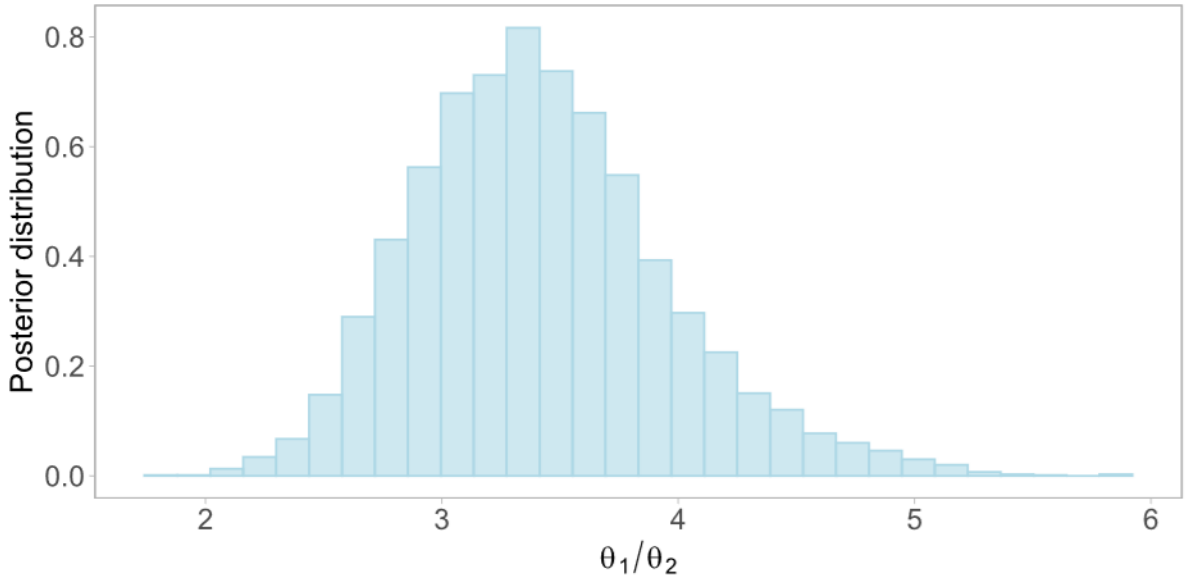
$$(b_1|\theta_1, \theta_2, b_2, k, \mathcal{D}) \sim \mathcal{G}\left(a_1 + c_1, d_1 + \theta_1\right)$$

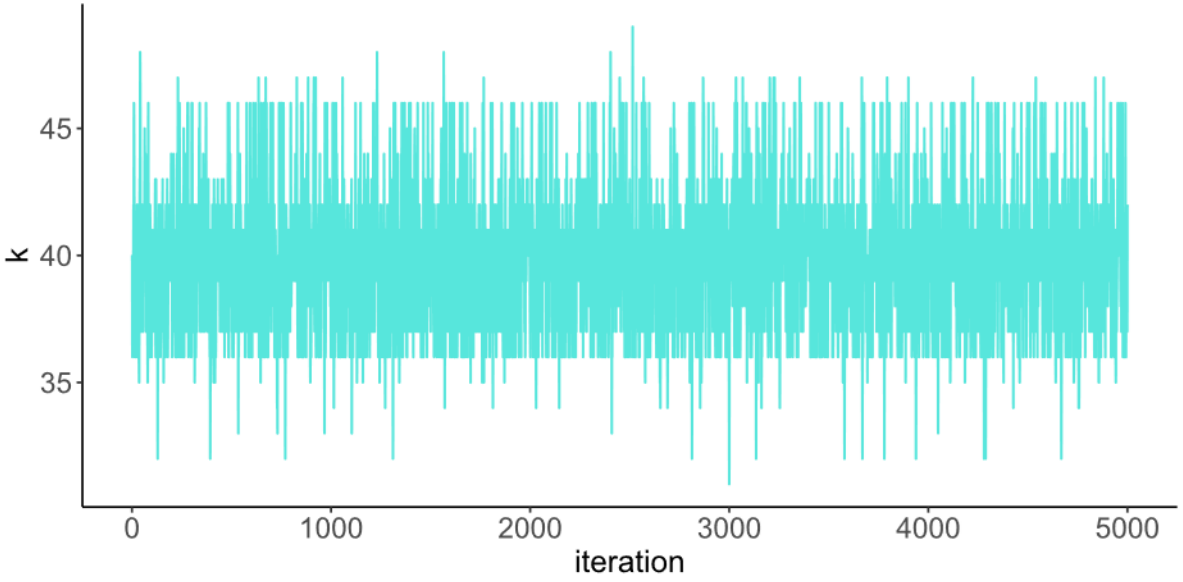
$$(b_2|\theta_1, \theta_2, b_1, k, \mathcal{D}) \sim \mathcal{G}\left(a_2 + c_2, d_2 + \theta_2\right)$$

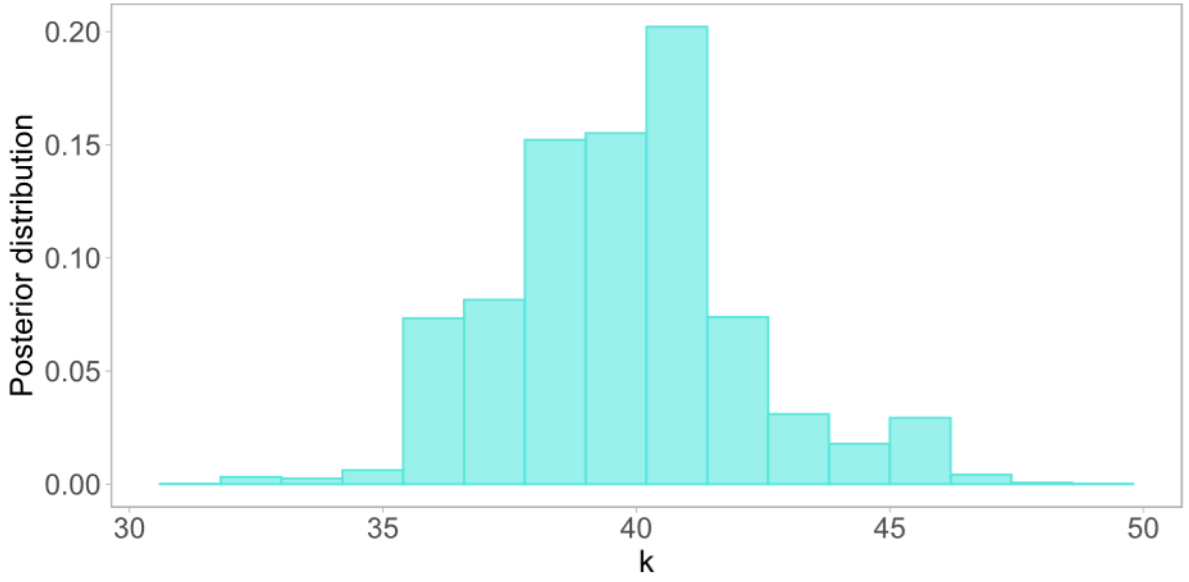
$$p(k|\theta_1, \theta_2, b_1, b_2, \mathcal{D}) \propto \left(\frac{\theta_1}{\theta_2}\right)^{\sum_{i=1}^{k} y_i} \exp(k(\theta_2 - \theta_1))$$

Trace plot and posterior distribution









Posterior summary features

Parameter	Mean	SD	95% CI
θ_1	3.1212	0.2908	[2.5731, 3.7412]
θ_2	0.9271	0.1193	[0.7056, 1.1779]
θ_1/θ_2	3.4210	0.5370	[2.5123 , 4.6472]
k	1890	2.4532	[1886, 1896]

- The year 1890 can be seen as a turning point for the number of accidents.
- After 1890 the number of coal mining accidents was reduced. The average number of accidents is approximately 3.5 times higher before 1890.

Smooth estimation of a regression curve

Consider the following regression model

$$\begin{aligned} y_i &= f(x_i) + \varepsilon_i,\ i = 1, \ldots, n, \\ \varepsilon &\sim \mathcal{N}(0, \tau^{-1}) \end{aligned}.$$

Function f is approximated with B-spline basis functions $f(x) = \sum_k \theta_k b_k(x)$.

λ is a parameter responsible for tuning the amount of smoothness in the fitted curve.

δ and τ are further hyperparameters on which a prior has to be imposed.

$$\begin{aligned} (y_i|\boldsymbol{\theta}, \tau) &\sim \mathcal{N}(\boldsymbol{\theta}^\top \mathbf{b}(x_i), \tau^{-1}), \\ (\boldsymbol{\theta}|\lambda, \tau) &\sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left(0, (\lambda\tau P)^{-1}\right), \\ (\lambda|\delta) &\sim \mathcal{G}(\nu/2, (\nu\delta)/2), \\ \delta &\sim \mathcal{G}(a_\delta, b_\delta), \\ p(\tau) &\propto \tau^{-1}. \end{aligned}$$

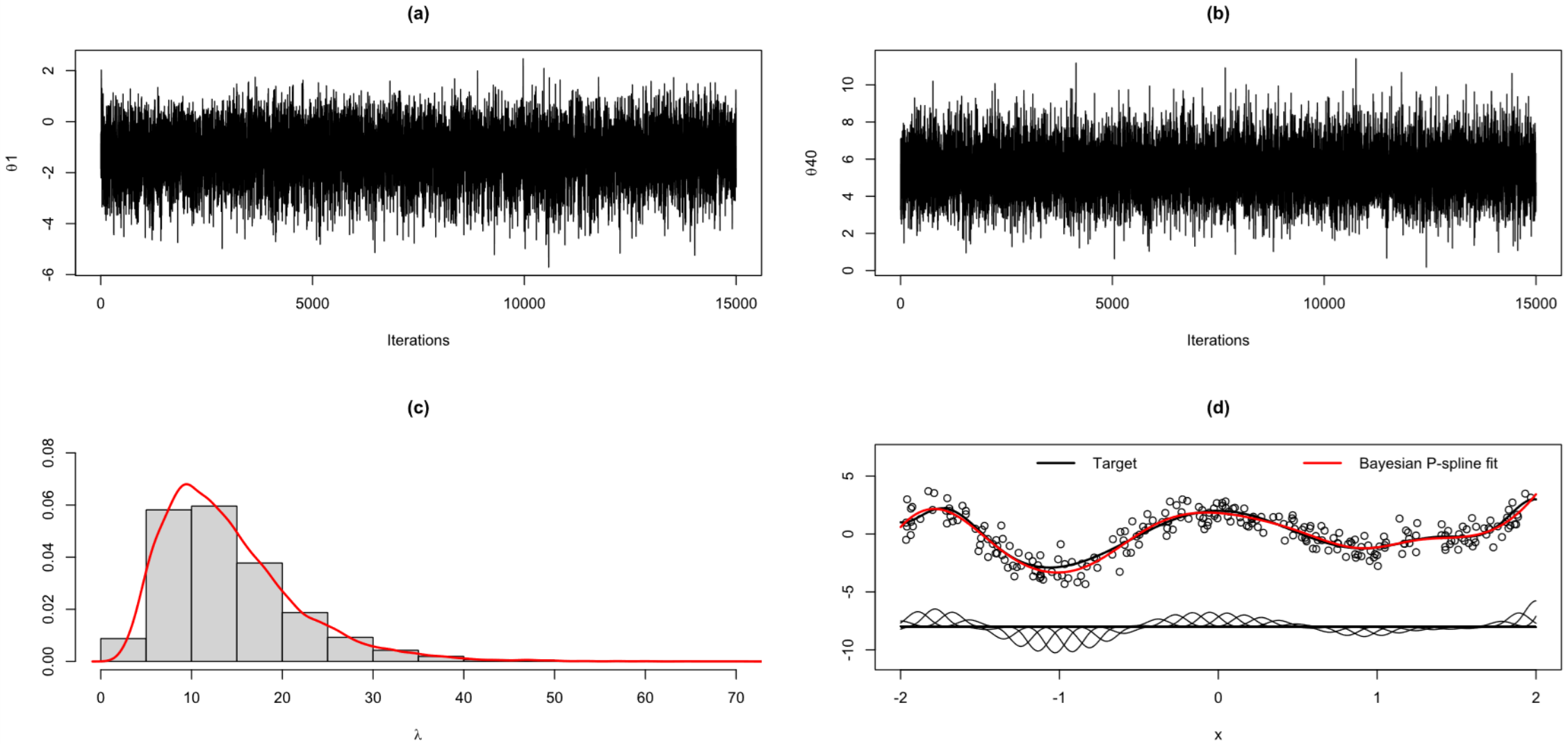
Smooth estimation of a regression curve

Full conditionals are available !

Pseudo-code: Gibbs sampler to draw from $p(\boldsymbol{\theta}, \tau, \delta, \lambda | \mathcal{D})$

1: Fix initial value $\lambda^{(0)}$.
2: **for** $m = 1, \ldots, M$ **do**
3: $\quad \boldsymbol{\theta}^{(m)} \sim \mathcal{N}_{\dim(\boldsymbol{\theta})} \left((B^\top B + \lambda^{(m-1)} P)^{-1} B^\top \mathbf{y}, \Sigma_{\boldsymbol{\theta}}^{(m-1)} \right)$.
4: $\quad \tau^{(m)} \sim \mathcal{G} \left(0.5(n+K), 0.5 \left(\|\mathbf{y} - B\boldsymbol{\theta}^{(m)}\|^2 + \lambda^{(m-1)} \boldsymbol{\theta}^{(m)\top} P \boldsymbol{\theta}^{(m)} \right) \right)$.
5: $\quad \delta^{(m)} \sim \mathcal{G} \left(0.5\nu + a_\delta, 0.5\nu\lambda^{(m-1)} + b_\delta \right)$.
6: $\quad \lambda^{(m)} \sim \mathcal{G} \left(0.5(K+\nu), 0.5 \left(\tau^{(m)} \boldsymbol{\theta}^{(m)\top} P \boldsymbol{\theta}^{(m)} + \nu\delta^{(m)} \right) \right)$.
7: **end for**

Smooth estimation of a regression curve



Gibbs sampling through completion

Consider the density (Robert, 1996):

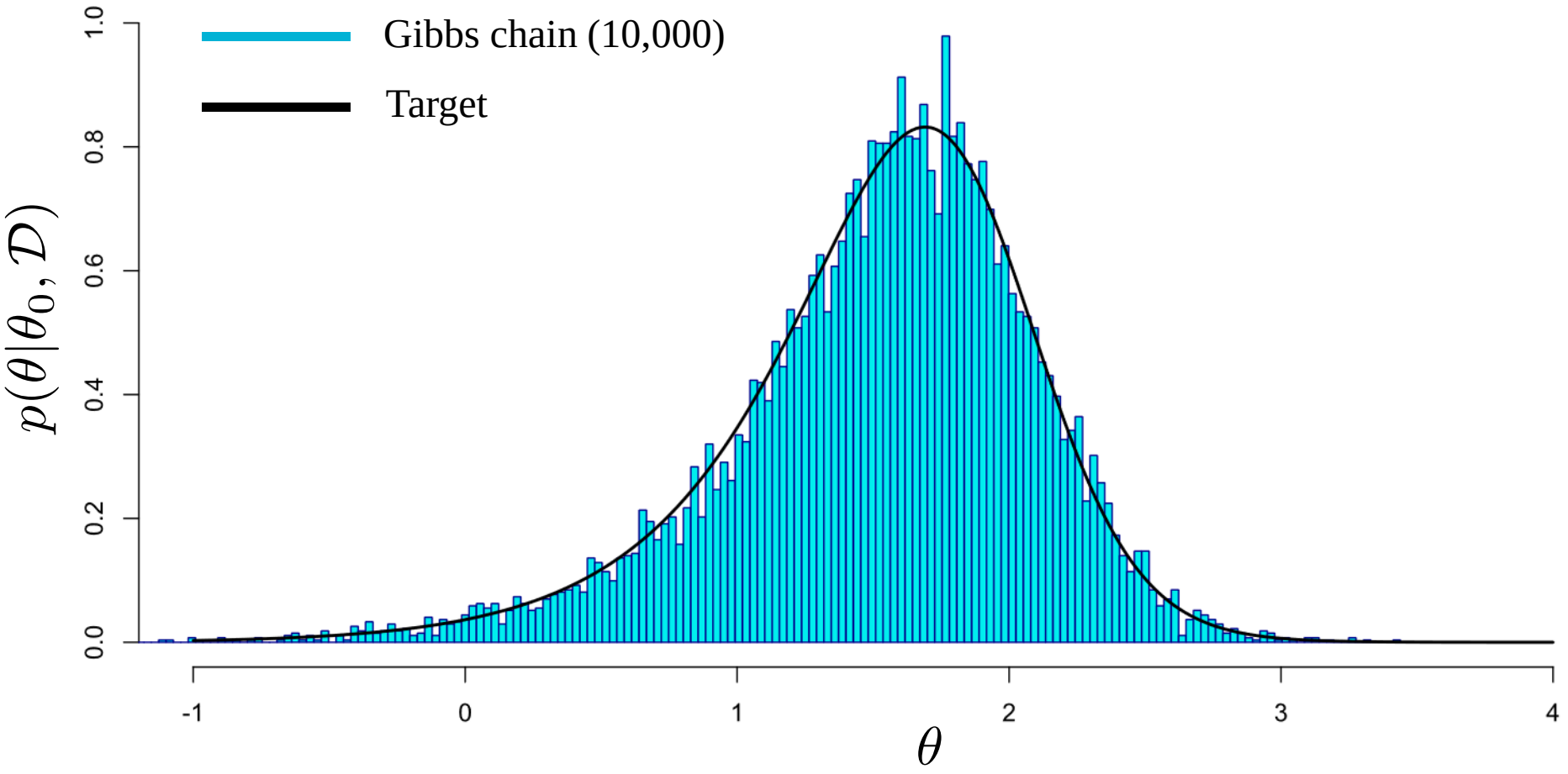
$$p(\theta|\theta_0, \mathcal{D}) \propto \frac{\exp(-\theta^2/2)}{(1+(\theta-\theta_0)^2)^\nu}, \ \theta \in \mathbb{R}, \nu > 0.$$

Samples from $p(\theta|\theta_0, \mathcal{D})$ can be obtained by using a completion technique, i.e. find a joint density $g(\theta, \eta|\theta_0, \mathcal{D})$, such that the target p appears as the marginal density of g.

The full conditionals are:

$$\begin{aligned}(\eta|\theta, \theta_0, \mathcal{D}) &\sim \mathcal{G}\left(\nu, 0.5(1+(\theta-\theta_0)^2)\right), \\ (\theta|\eta, \theta_0, \mathcal{D}) &\sim \mathcal{N}\left(\frac{\eta}{1+\eta}\theta_0, \frac{1}{1+\eta}\right).\end{aligned}$$

Gibbs sampling through completion



Remarks

If the following conditional densities $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ are known, they can be used to determine the marginal density $p(\theta_1)$.

$$
\begin{aligned}
p(\theta_1) &= \int p(\theta_1, \theta_2) d\theta_2 = \int p(\theta_1|\theta_2) p(\theta_2) d\theta_2 \\
&= \int p(\theta_1|\theta_2) \left[\int p(\theta_2|\phi) p(\phi) d\phi \right] d\theta_2 \\
&= \int \left[\int p(\theta_1|\theta_2) p(\theta_2|\phi) d\theta_2 \right] p(\phi) d\phi \\
&= \int K(\theta_1, \phi) p(\phi) d\phi
\end{aligned}
$$

$K(\theta_1, \phi)$ is called transition kernel.

It expresses the probability of a move(ment) from θ_1 to ϕ.

Equation on the left is a fixed point integral equation with solution $p(\theta_1)$.

Gibbs sampling stochastically solves it.

Gibbs sampler in practice

Full conditionals can be sampled in various ways.

- There are many software available: SAS, R, WinBUGS, JAGS, C++,…
- Deterministic Gibbs sampler: dimensions are explored following a fixed sequence.
- Random Gibbs sampler: dimensions are explored in random order.
- Reversible Gibbs sampler: dimensions are explored in order and reversed order.
- Block Gibbs sampler: conditionals are sampled in blocks.

Metropolis algorithm

Idea behind the Metropolis algorithm

Objective: sample from a target distribution $f(\theta)$ with $\theta \in \mathbb{R}$.

Problem: direct simulation from $f(\theta)$ is difficult/impossible.

Idea (Metropolis et al. 1953): Use a symmetric "surrogate" distribution $q(\cdot)$ and indirectly sample from $f(\theta)$ via the surrogate distribution.

The surrogate distribution is called proposal distribution or jumping distribution.

The proposal distribution must be symmetric and easy to sample.

Idea behind the Metropolis algorithm

Let $\theta^{(m-1)} \in \mathbb{R}$ be an arbitrary starting point satisfying $f(\theta^{(m-1)}) > 0$.

Denote by $q(\cdot|\theta^{(m-1)})$ the symmetric proposal density centered at $\theta^{(m-1)}$.

Sample $\widetilde{\theta}^{(m)} \sim q(\cdot|\theta^{(m-1)})$ a candidate from the proposal.

Note that by symmetry, we have $q\big(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\big) = q\big(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\big)$.

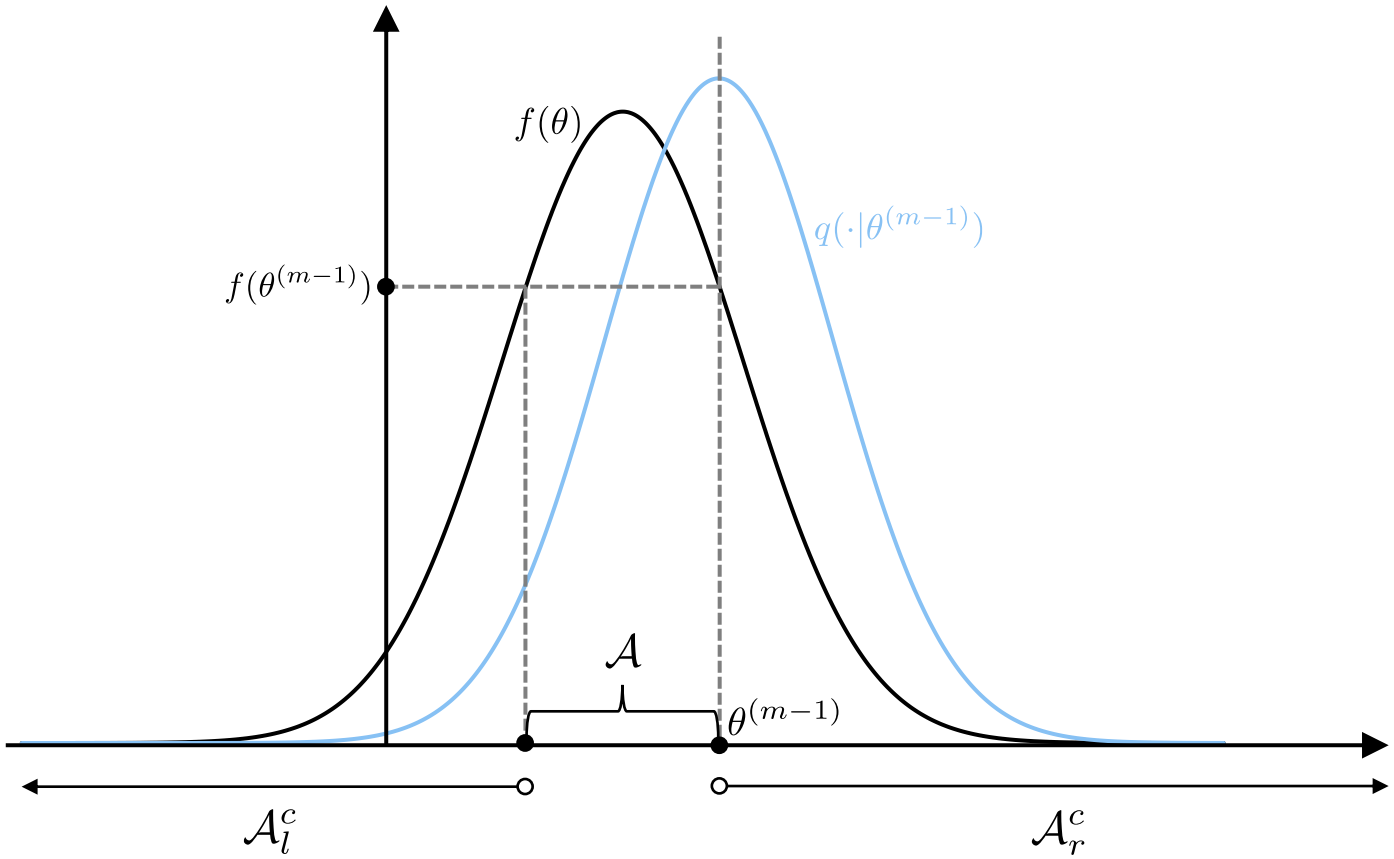
When the proposal depends on the current state we call it a random walk algorithm.

Compute the ratio $\varrho(\theta^{(m-1)}, \widetilde{\theta}^{(m)}) = \dfrac{f(\widetilde{\theta}^{(m)})}{f(\theta^{(m-1)})}$.

If $\varrho \geq 1$, make the move to the simulated candidate $\theta^{(m)} = \widetilde{\theta}^{(m)}$.

If $\varrho < 1$, make the move to the simulated candidate with probability $0 \leq \varrho < 1$.

Graphical illustration



$$\mathcal{A} = \left\{\theta \in \mathbb{R} : \frac{f(\theta)}{f(\theta^{(m-1)})} \geq 1\right\}$$

$$= \left\{\theta \in \mathbb{R} : \varrho(\theta^{(m-1)}, \theta) \geq 1\right\}$$

$$\mathcal{A}_l^c \cup \mathcal{A}_r^c = \left\{\theta \in \mathbb{R} : \frac{f(\theta)}{f(\theta^{(m-1)})} < 1\right\}$$

$$= \left\{\theta \in \mathbb{R} : \varrho(\theta^{(m-1)}, \theta) < 1\right\}$$

Graphical illustration

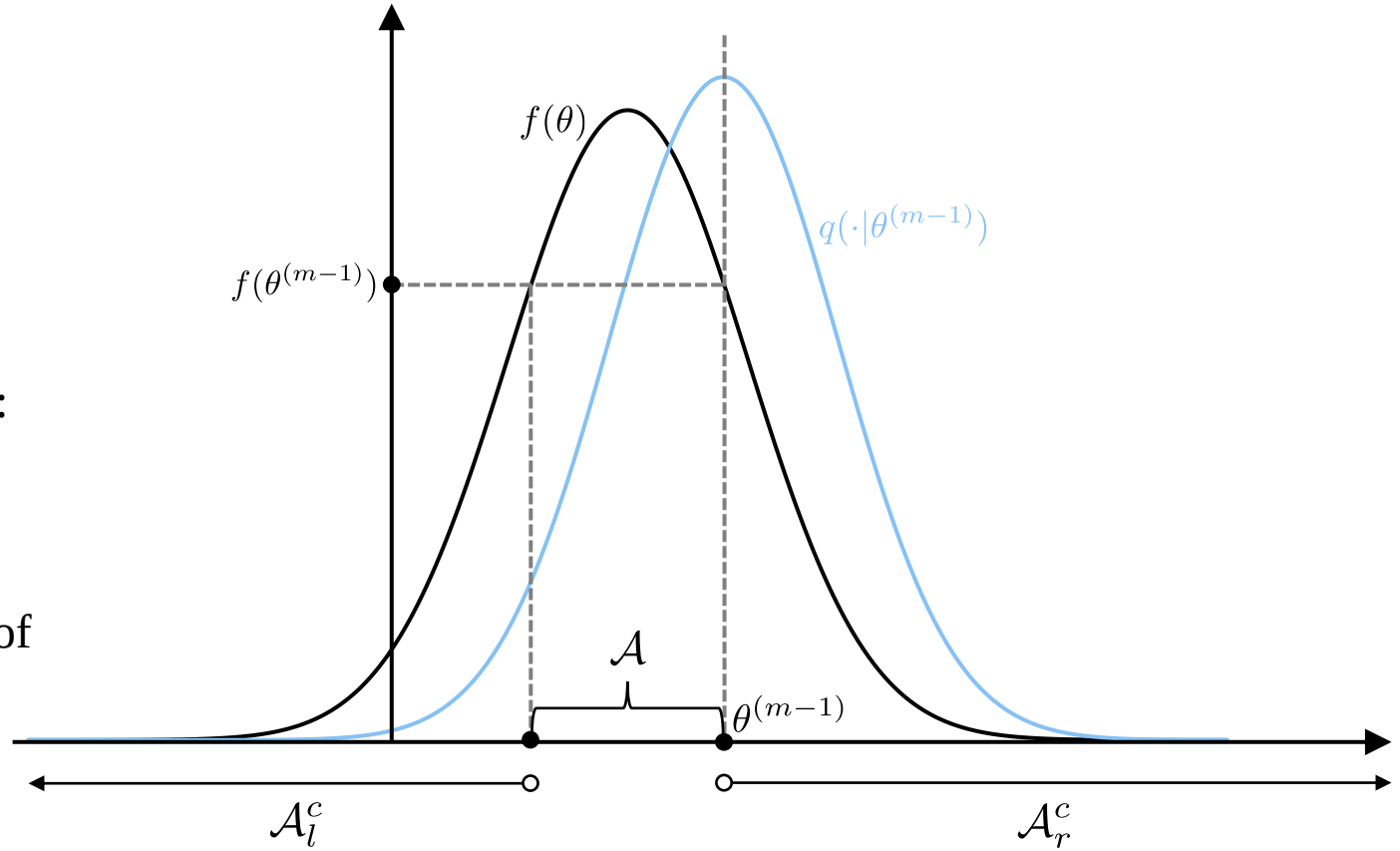
Draw $\widetilde{\theta}^{(m)} \sim q(\cdot|\theta^{(m-1)})$.

If $\widetilde{\theta}^{(m)} \in \mathcal{A}$, accept the candidate with probability 1.

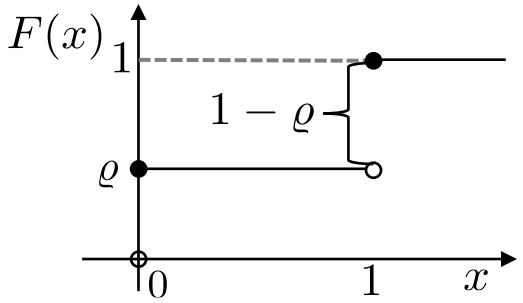
If $\widetilde{\theta}^{(m)} \in \mathcal{A}_l^c \cup \mathcal{A}_r^c$ define the (binary) random variable:

$$X = \mathbb{I}(\theta^{(m)} = \theta^{(m-1)})$$

i.e. $X = 1$ if $\theta^{(m)} = \theta^{(m-1)}$ indicates a rejection of the candidate (absence of movement) with probability $1 - \varrho$ and $X = 0$ (movement) with probability ϱ.



The cumulative distribution function of X is:



$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$$

$$u \in (0,1) \longrightarrow U \sim U(0,1)$$

If $0 < U \leq \varrho$ then $X = F^{-1}(U) = 0$ (move)
If $\varrho < U < 1$ then $X = F^{-1}(U) = 1$ (no move)

Remarks

The algorithm generates a Markov sample (successive draws are statistically dependent).

The ratio $\varrho(\theta^{(m-1)}, \widetilde{\theta}^{(m)}) = f(\widetilde{\theta}^{(m)})/f(\theta^{(m-1)})$ can be computed even if f is known only up to a constant multiple since the normalizing constant appears both in the numerator and denominator .

At each iteration of the algorithm it must be possible to:

- compute the ratio $f(\widetilde{\theta}^{(m)})/f(\theta^{(m-1)})$ for all $\theta, \widetilde{\theta}$.
- draw $\widetilde{\theta}$ from the proposal distribution for all θ.
- draw a uniform random number in $(0, 1)$.

If a candidate is rejected at iteration $m-1$, the current value at m coincides with the current value at $m-1$.

Metropolis algorithm in a Bayesian framework

Metropolis algorithm (random walk)

1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do
   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot|\boldsymbol{\theta}^{(m-1)})$.
   2.2. Compute the ratio $\varrho(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D})/p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D})$.
   2.3. if $\varrho \geq 1$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate).
   2.4. if $\varrho < 1$ do
      2.4.1. Sample $u \sim U(0, 1)$.
      2.4.2. if $u \leq \varrho$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate), else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).
3. end for .

Metropolis algorithm in a Bayesian framework

Metropolis algorithm (random walk) compact version

1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do
   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot|\boldsymbol{\theta}^{(m-1)})$.
   2.2. Compute the *probability of move* $\alpha(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \min\left(p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D})/p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}), 1\right)$.
   2.3. Sample $u \sim U(0,1)$.
   2.4. if $u \leq \alpha$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate), else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).
3. end for .

Metropolis in action (univariate)

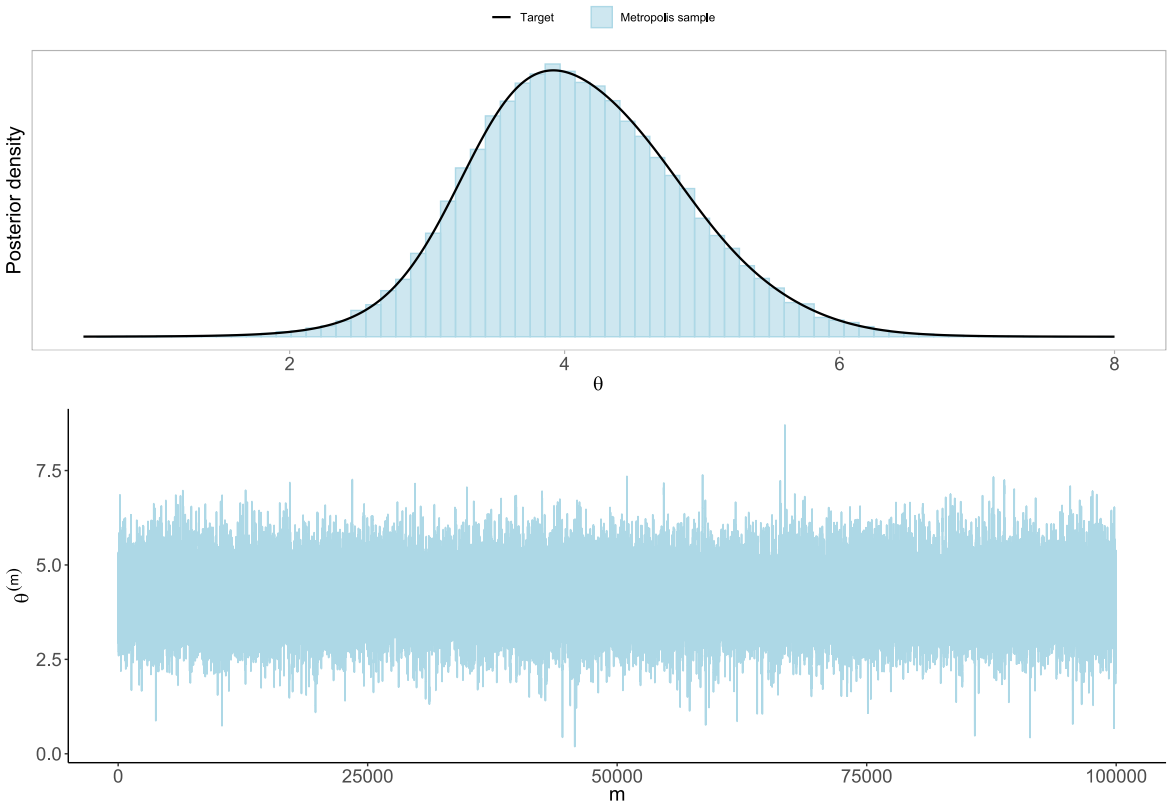


Fig 2. (Top) Normalized posterior density (solid) and histogram of the samples drawn from $p(\theta|\mathcal{D})$ in the Cauchy model. (Bottom) Trace plot of the Markov sample of size 100,000.

i.i.d. sample from a Cauchy distribution
$\mathcal{D} = \{y_1, \ldots, y_n\} \sim \mathcal{C}(\theta, 1)$.

Location parameter $\theta \in \mathbb{R}$ and scale equal to 1.

Gaussian prior $\theta \sim \mathcal{N}(0, \tau^{-1})$.

By Bayes' theorem the posterior is shown to be

$$p(\theta|\mathcal{D}) \propto \frac{\exp(-0.5\tau\theta^2)}{\prod_{i=1}^{n}(1+(y_i-\theta)^2)}\,.$$

Use the (random walk) Metropolis algorithm to obtain a sample from the posterior.

Metropolis in action (bivariate)

Consider the following bivariate family of distributions (Geman and Meng 1991):

$$p(\theta_1, \theta_2|\mathcal{D}) \propto \exp\left\{-\frac{1}{2}\left(A\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 2B\theta_1\theta_2 - 2C_1\theta_1 - 2C_2\theta_2\right)\right\}$$

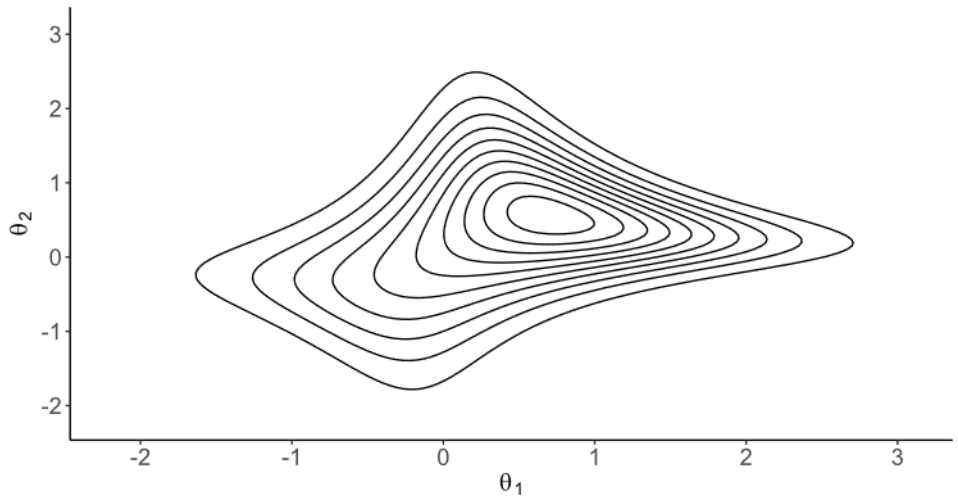
with $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$ and scalars $A > 0, B, C_1, C_2 \in \mathbb{R}$.

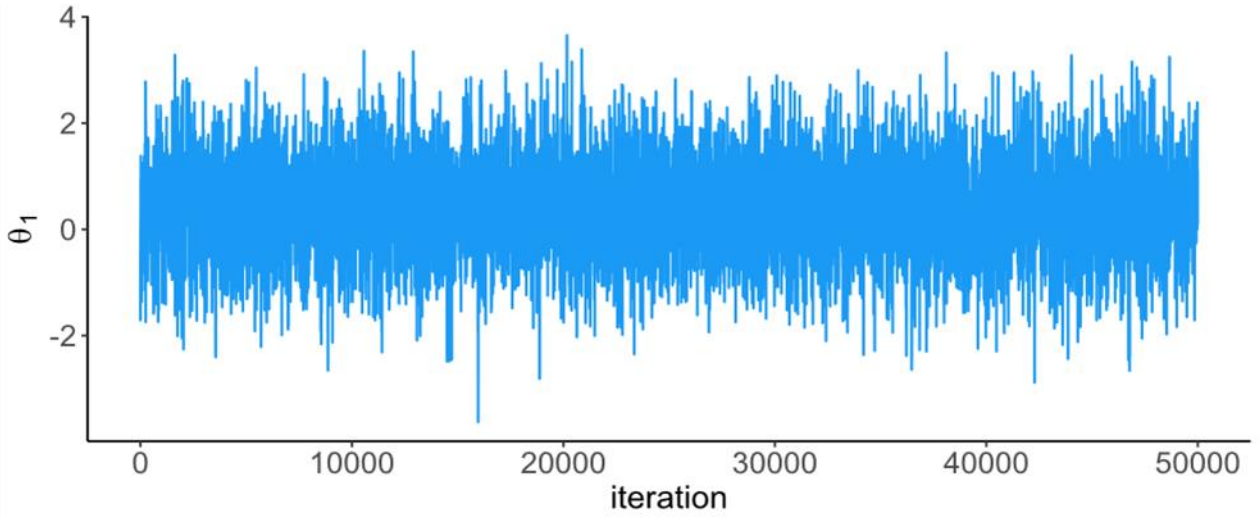
Metropolis algorithm with a Gaussian candidate-generating density:

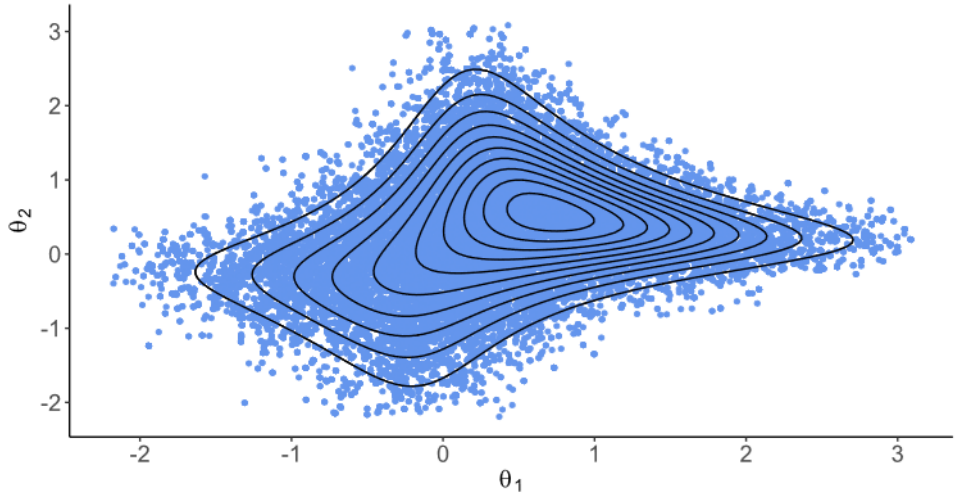
$$q(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = (2\pi)^{\frac{-\dim(\tilde{\boldsymbol{\theta}})}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \Sigma^{-1} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\right)$$

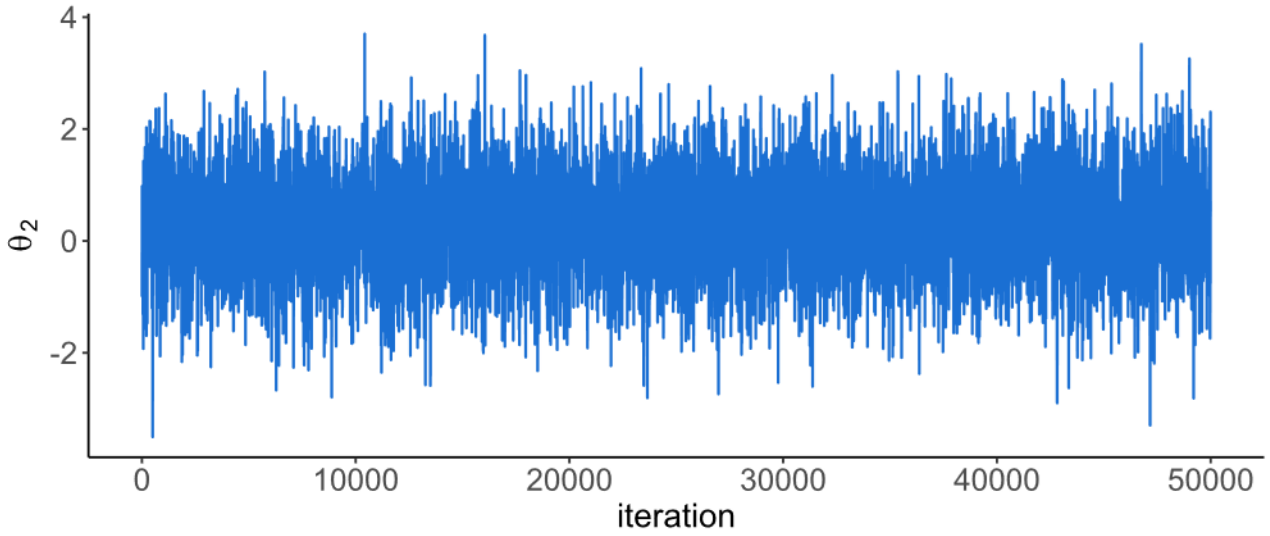
and covariance matrix $\Sigma = \begin{pmatrix} 2.82 & 0 \\ 0 & 2.82 \end{pmatrix}$.

Metropolis in action (bivariate)





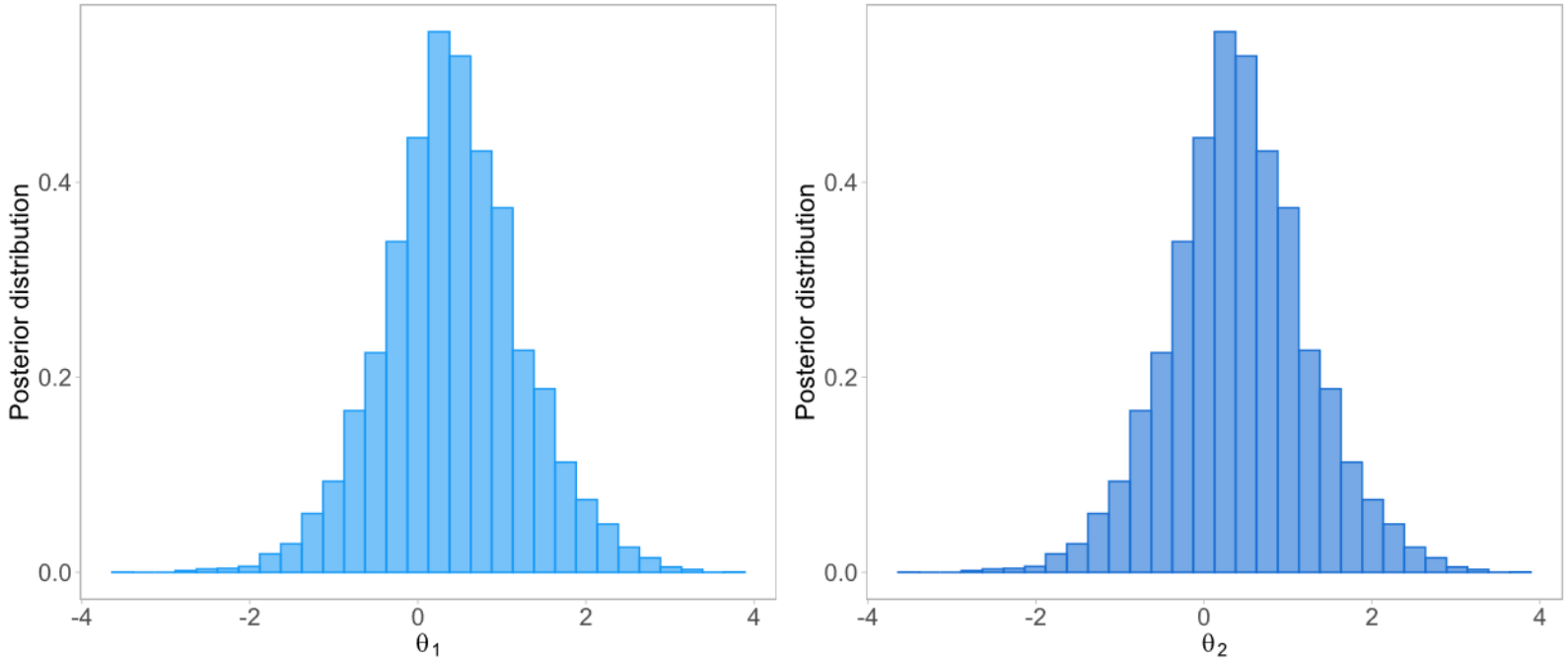




Acceptance rate: **23.48%**

Metropolis in action (bivariate)

Marginal posterior distributions



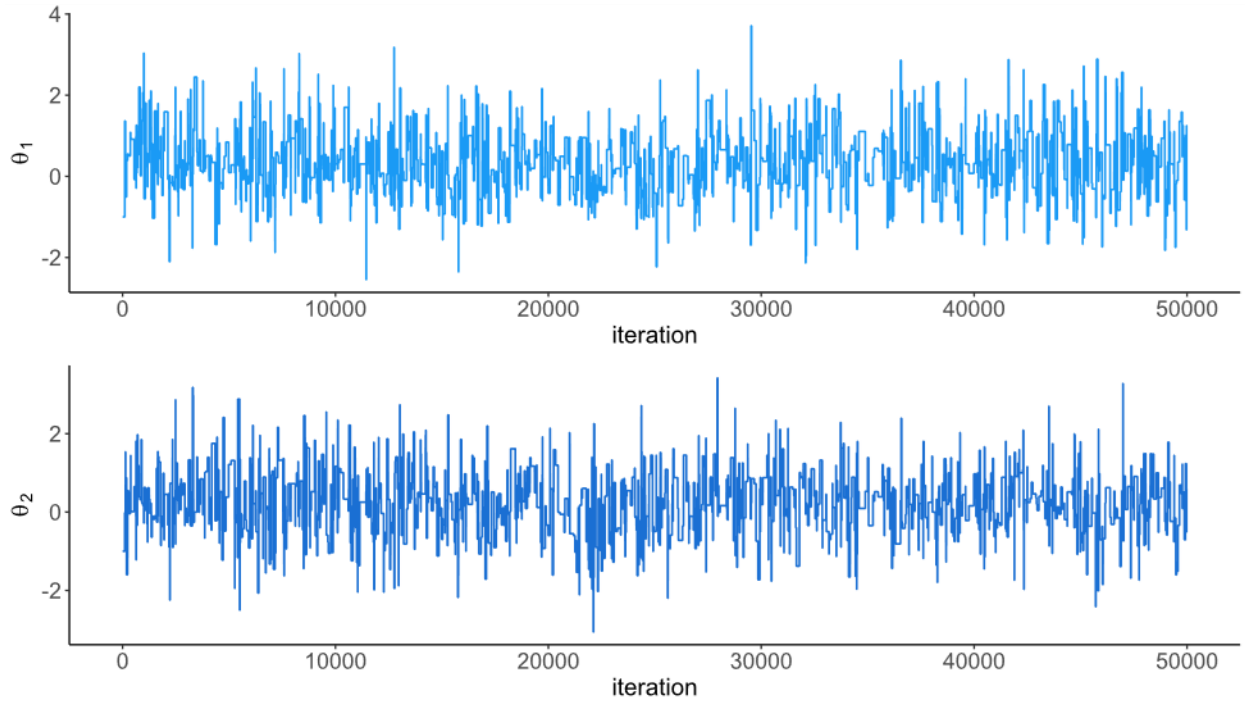
Posterior summary statistics

Parameter	Mean	Median	SD	95% CI
θ_1	0.4164	0.3970	0.8231	[-1.2159 2.1237]
θ_2	0.3095	0.3095	0.7786	[-1.2728 1.8920]

Acceptance rates

Consider the proposal with covariance matrix: $\Sigma = \begin{pmatrix} 40 & 0 \\ 0 & 40 \end{pmatrix}$.

Acceptance rate ~ **2%**.



What is a “good” acceptance rate?

Acceptance rates

Efficiency of the Metropolis algorithm depends on the dispersion of the proposal density.

If the variance of the proposal is too small → slow moving chain yielding a poor and sluggish exploration of the posterior parameter space and high acceptance rates.

If the variance of the proposal is too large → the sampler will explore regions where the posterior has little support and this will cause abundant rejection of the proposed candidates.

A general rule for greatest efficiency in random walk sampling schemes is to reach an acceptance rate around **44%** in one-dimensional problems and roughly **23%** in higher dimensions.

If Gaussian proposal, the covariance matrix Σ is tuned to reach those "optimal" rates.

Metropolis-Hastings algorithm

Idea behind the Metropolis-Hastings algorithm

Generalization of the Metropolis algorithm proposed by Hastings (1970).

This general algorithm is called the Metropolis-Hastings (M-H) algorithm.

In M-H algorithm, generalization is at the level of the proposal distribution.

The proposal distribution $q(\cdot)$ is no longer restricted to the set of symmetric distributions.

Proposals satisfying $q\left(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\right) \neq q\left(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\right)$ are allowed.

This generalization requires to update the ratio:

$$\varrho(\theta^{(m-1)}, \widetilde{\theta}^{(m)}) = \frac{f(\widetilde{\theta}^{(m)})}{f(\theta^{(m-1)})} \quad \Rightarrow \quad \varrho(\theta^{(m-1)}, \widetilde{\theta}^{(m)}) = \frac{f(\widetilde{\theta}^{(m)})}{f(\theta^{(m-1)})} \frac{q\left(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\right)}{q\left(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\right)}$$

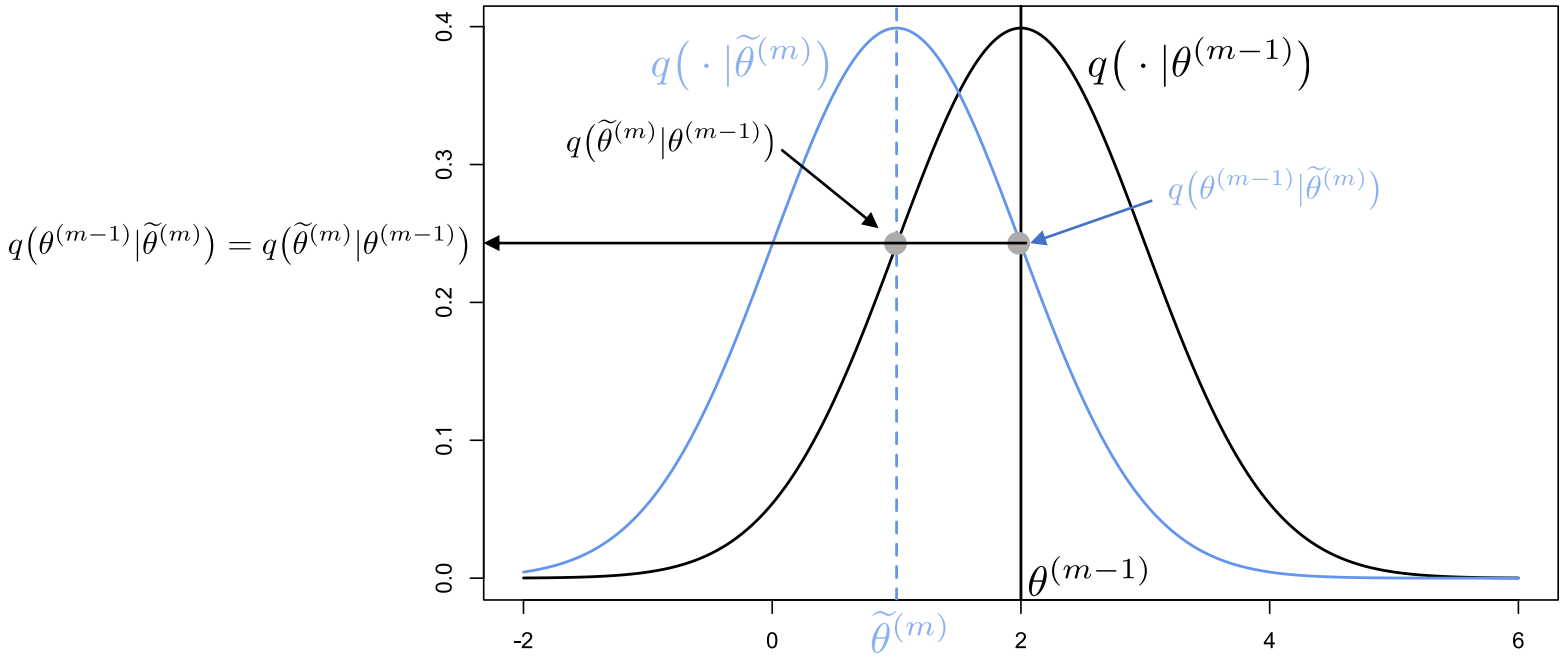
Metropolis — Metropolis-Hastings

Idea behind the Metropolis-Hastings algorithm

If the proposal is symmetric, we have $q\left(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\right) = q\left(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\right)$.

The ratio simplifies to $\varrho(\theta^{(m-1)}, \widetilde{\theta}^{(m)}) = \dfrac{f\left(\widetilde{\theta}^{(m)}\right)}{f\left(\theta^{(m-1)}\right)} \dfrac{q\left(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\right)}{q\left(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\right)} = \dfrac{f\left(\widetilde{\theta}^{(m)}\right)}{f\left(\theta^{(m-1)}\right)}$ and

we recover the Metropolis algorithm.



$q(\cdot|\theta) = \mathcal{N}(\cdot, \mu = \theta, \sigma^2 = 1)$
$\theta^{(m-1)} = 2$
$\widetilde{\theta}^{(m)} = 1$

$$q\left(\theta^{(m-1)}|\widetilde{\theta}^{(m)}\right) = \mathcal{N}(\theta^{(m-1)}; \mu = \widetilde{\theta}^{(m)}, \sigma^2 = 1) = \mathcal{N}(2; \mu = 1, \sigma^2 = 1) = 0.242$$

$$q\left(\widetilde{\theta}^{(m)}|\theta^{(m-1)}\right) = \mathcal{N}(\widetilde{\theta}^{(m)}; \mu = \theta^{(m-1)}, \sigma^2 = 1) = \mathcal{N}(1; \mu = 2, \sigma^2 = 1) = 0.242$$

Metropolis-Hastings algorithm in a Bayesian framework

Metropolis-Hastings algorithm (random walk)

1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do

   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot|\boldsymbol{\theta}^{(m-1)})$.

   2.2. Compute the ratio $\varrho(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \dfrac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D})q(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D})q(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)})}$.

   2.3. if $\varrho \geq 1$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate).

   2.4. if $\varrho < 1$ do

      2.4.1. Sample $u \sim U(0, 1)$.

      2.4.2. if $u \leq \varrho$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate), else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).

3. end for .

Metropolis-Hastings algorithm in a Bayesian framework

Metropolis-Hastings algorithm (random walk) compact version

1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do
   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot|\boldsymbol{\theta}^{(m-1)})$.
   2.2. Compute the *probability of move* $\alpha(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \min\left(\frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D})q(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D})q(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)})}, 1\right)$.
   2.3. Sample $u \sim U(0,1)$.
   2.4. if $u \leq \alpha$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate), else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).
3. end for .

The independent Metropolis-Hastings algorithm

Arises when the candidate-generating density $q(\cdot)$ is independent of the current state:

$$q(\cdot|\theta^{(m-1)}) = q(\cdot)$$

Although the new candidate is generated independently of the current state, the resulting sample preserves its Markovian structure as the *probability of move* still depends on the current state.

Contrary to the random-walk algorithm we do not seek an "optimal" acceptance rate.

Here, the proposal $q(\cdot)$ should be chosen so as to have an acceptance rate as high as possible. This will be the case if $q(\cdot)$ is a good approximation of the target distribution.

The candidate-generating density can for instance be a Gaussian density (obtained via Laplace approximation) or a *t*-distribution with the mode centered at the mode of the target distribution (provided the target is unimodal).

Independent Metropolis-Hastings algorithm

Metropolis-Hastings algorithm (independent) compact version

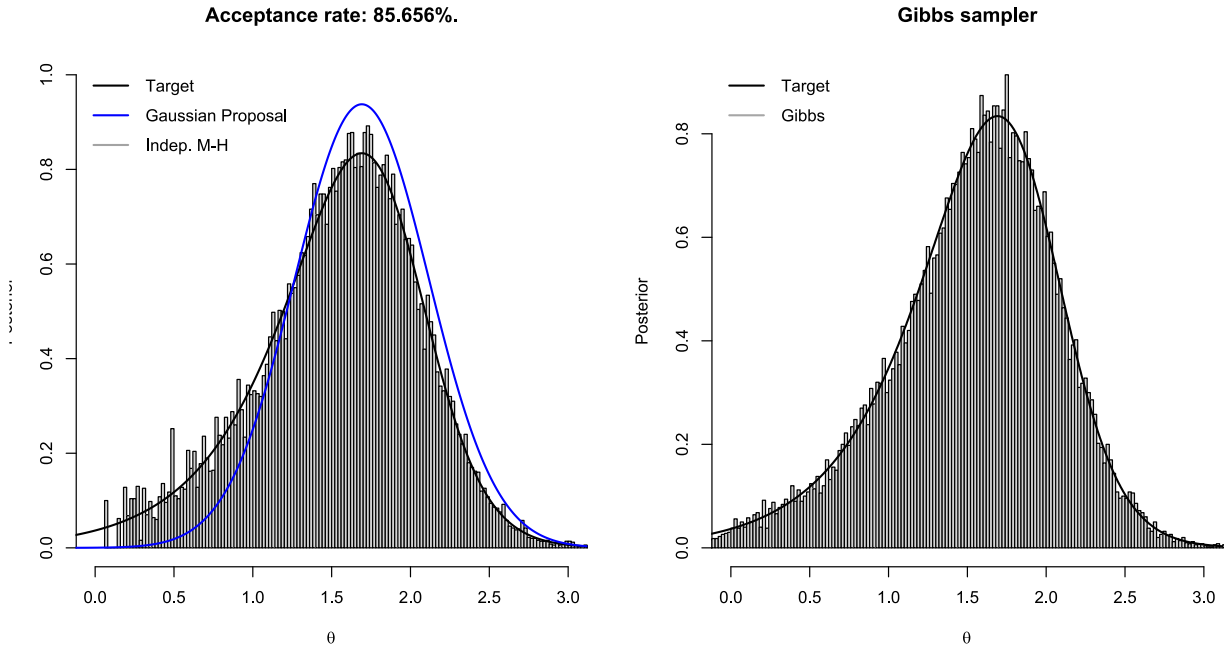
1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do
   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot)$.
   2.2. Compute the *probability of move* $\alpha(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \min\left(\frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D})q(\boldsymbol{\theta}^{(m-1)})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D})q(\widetilde{\boldsymbol{\theta}}^{(m)})}, 1\right)$.
   2.3. Sample $u \sim U(0,1)$.
   2.4. if $u \leq \alpha$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate), else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).
3. end for .

How do independent samplers perform?

Consider the density (Robert, 1996):

$$p(\theta|\theta_0, \mathcal{D}) \propto \frac{\exp(-\theta^2/2)}{(1+(\theta-\theta_0)^2)^\nu}, \ \theta \in \mathbb{R}, \nu > 0.$$

Performance of the independent Metropolis-Hastings algorithm with a Gaussian proposal obtained from a Laplace approximation around mode of $p(\theta|\theta_0, \mathcal{D})$.



A note on numeric overflow

Careful about numeric overflow when coding the Metropolis/Metropolis-Hastings algorithms.

A good idea is to compute the Metropolis/Metropolis-Hastings ratio in log scale:

$$\log \varrho(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \log \left(\frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) q\big(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)}\big)}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) q\big(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)}\big)} \right)$$

$$= \log \left(p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) q\big(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)}\big) \right) - \log \left(p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) q\big(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)}\big) \right)$$

$$= \log p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) + \log q\big(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)}\big) - \log p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) - \log q\big(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)}\big) \ .$$

A note on numeric overflow

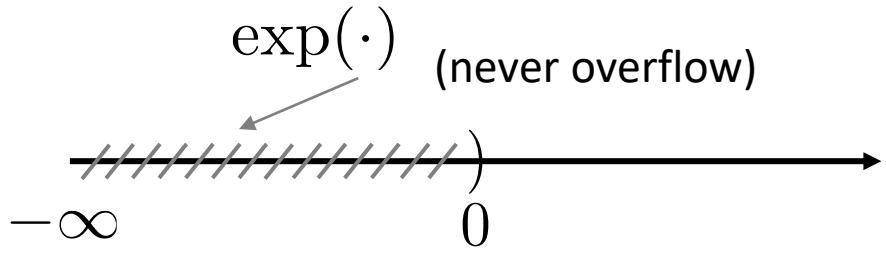
Recall that a new candidate $\widetilde{\theta}^{(m)}$ is **accepted** under two conditions:

(C1) $\varrho \geq 1 \Leftrightarrow \log(\varrho) \geq \log(1) \Leftrightarrow \log(\varrho) \geq 0$.

(C2) $u \leq \varrho \Leftrightarrow u \leq \exp(\log(\varrho))$ where $u \sim U(0,1)$.

Using the OR "||" operator allows to avoid overflow through the IF statement:

if $\underbrace{\log(\varrho) \geq 0}_{\text{C1}}$ || $\underbrace{u \leq \exp(\log(\varrho))}_{\text{C2}}$ then accept $\widetilde{\theta}^{(m)}$



With the || operator, C2 will only be evaluated if C1 is FALSE (i.e. if $\log(\varrho) < 0$).

Applying $\exp(\cdot)$ on any negative number (even infinite) will never overflow.

Metropolis-Hastings algorithm in a Bayesian framework

Metropolis-Hastings algorithm (random walk) avoiding overflow

1. Choose initial value $\boldsymbol{\theta}^{(0)}$ satisfying $p(\boldsymbol{\theta}^{(0)}|\mathcal{D}) > 0$.
2. for m in 1 to M do

   2.1. Sample $\widetilde{\boldsymbol{\theta}}^{(m)} \sim q(\cdot|\boldsymbol{\theta}^{(m-1)})$.

   2.2. Compute the log ratio $\log \varrho(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \log \left(\dfrac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) q(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) q(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)})} \right)$.

   2.3. Sample $u \sim U(0,1)$.

   2.4. if $\log(\varrho) \geq 0$ || $u \leq \exp(\log(\varrho))$, set $\boldsymbol{\theta}^{(m)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(m)}$ (accept candidate).

   2.5 else $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ (reject).
3. end for .

Gibbs sampler as a special case of M-H

Let $\boldsymbol{\theta} \in \mathbb{R}^K$ and define the subvector $\boldsymbol{\theta}_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K)^\top$.

Gibbs sampler can be seen as a sequence of K M-H algorithms where the generated candidate is always accepted.

Consider an algorithm, where iteration m has a total of K sub steps.

At step $k = 1, \ldots, K$ only a jump in the k^{th} dimension of $\boldsymbol{\theta}$ will be made and all other parameters remain frozen at their current state.

Gibbs sampler as a special case of M-H

At step k of iteration m, the proposal distribution is given by $q_k^G(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)}) = p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})$ if $\widetilde{\boldsymbol{\theta}}_{-k}^{(m)} = \boldsymbol{\theta}_{-k}^{(m-1)}$ and $q_k^G(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)}) = 0$ otherwise.

The Hastings ratio is thus given by:

$$\varrho_k^G(\boldsymbol{\theta}^{(m-1)}, \widetilde{\boldsymbol{\theta}}^{(m)}) = \frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) q_k^G(\boldsymbol{\theta}^{(m-1)}|\widetilde{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) q_k^G(\widetilde{\boldsymbol{\theta}}^{(m)}|\boldsymbol{\theta}^{(m-1)})} = \frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) p(\boldsymbol{\theta}_k^{(m-1)}|\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}, \mathcal{D})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})} = \frac{p(\widetilde{\boldsymbol{\theta}}^{(m)}|\mathcal{D}) / p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) / p(\boldsymbol{\theta}_k^{(m-1)}|\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}, \mathcal{D})}$$

$$= \frac{\left(p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}, \mathcal{D}) p(\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}|\mathcal{D})\right) \Big/ p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})}{\left(p(\boldsymbol{\theta}_k^{(m-1)}|\boldsymbol{\theta}_{-k}^{(m-1)}\mathcal{D}) p(\boldsymbol{\theta}_{-k}^{(m-1)}|\mathcal{D})\right) \Big/ p(\boldsymbol{\theta}_k^{(m-1)}|\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}, \mathcal{D})} = \frac{\left(p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D}) p(\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}|\mathcal{D})\right) \Big/ p(\widetilde{\boldsymbol{\theta}}_k^{(m)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})}{\left(p(\boldsymbol{\theta}_k^{(m-1)}|\boldsymbol{\theta}_{-k}^{(m-1)}\mathcal{D}) p(\boldsymbol{\theta}_{-k}^{(m-1)}|\mathcal{D})\right) \Big/ p(\boldsymbol{\theta}_k^{(m-1)}|\boldsymbol{\theta}_{-k}^{(m-1)}, \mathcal{D})}$$

$$= \frac{p(\widetilde{\boldsymbol{\theta}}_{-k}^{(m)}|\mathcal{D})}{p(\boldsymbol{\theta}_{-k}^{(m-1)}|\mathcal{D})} = \frac{p(\boldsymbol{\theta}_{-k}^{(m-1)}|\mathcal{D})}{p(\boldsymbol{\theta}_{-k}^{(m-1)}|\mathcal{D})} = 1$$

$\Longrightarrow$ candidate is always accepted (like in Gibbs sampling).

Transition kernel of M-H

Let $\boldsymbol{\theta} := \boldsymbol{\theta}^{(m-1)}$ denote the current state of the chain and $\boldsymbol{\phi} := \widetilde{\boldsymbol{\theta}}^{(m)}$ denote a proposed candidate with both terms belonging to the parameter space.

The transition kernel of the M-H algorithm gives the probability to move from $\boldsymbol{\theta}$ to $\boldsymbol{\phi}$ and can be decomposed in two elements.

The first element describes the move towards $\boldsymbol{\phi}$, the candidate suggested by the proposal:

$$K(\boldsymbol{\theta}, \boldsymbol{\phi}) = \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) q(\boldsymbol{\phi}|\boldsymbol{\theta})$$

The second element expresses the absence of movement arising with probability:

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}) = 1 - \int \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) q(\boldsymbol{\phi}|\boldsymbol{\theta}) d\boldsymbol{\phi}$$

Transition kernel of M-H

Combining the two elements in a single expression yields:

$$p(\boldsymbol{\theta}, \mathcal{S}) = \int_{\mathcal{S}} \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) q(\boldsymbol{\phi}|\boldsymbol{\theta}) d\boldsymbol{\phi} + \mathbb{I}(\boldsymbol{\theta} \in \mathcal{S}) \left[1 - \int \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) q(\boldsymbol{\phi}|\boldsymbol{\theta}) d\boldsymbol{\phi} \right]$$

for any subset $\mathcal{S}$ of the parameter space, where $\mathbb{I}(\cdot)$ is the indicator function.

Having a Markov chain satisfying reversibility is a sufficient condition for the chain to converge to the target posterior distribution. Reversibility requires for any two subsets $\mathcal{S}, \mathcal{S}'$:

$$\int_{\mathcal{S}'} p(\boldsymbol{\theta}, \mathcal{S}) d\boldsymbol{\theta} = \int_{\mathcal{S}} p(\boldsymbol{\phi}, \mathcal{S}') d\boldsymbol{\phi} \, ,$$

which is satisfied if the *detailed balance condition* holds for any two subsets $\mathcal{S}, \mathcal{S}'$:

$$\int_{\mathcal{S}'} \int_{\mathcal{S}} K(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\phi} d\boldsymbol{\theta} = \int_{\mathcal{S}} \int_{\mathcal{S}'} K(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\theta} d\boldsymbol{\phi} \, .$$

Versions of the Metropolis-Hastings algorithm

- Random walk Metropolis-Hasting algorithm.
- Random walk Metropolis algorithm.
- Independent Metropolis-Hastings algorithm.
- Metropolis-within-Gibbs algorithm.
- Metropolis-adjusted Langevin algorithm (MALA).
- Reversible jump Metropolis-Hastings algorithm.

Why does MCMC work?

Markov chains generated by the algorithms we have seen here (if iterated long enough) will provide samples form the posterior distribution.

Summary features computed from the generated Markov chains provide consistent estimates of the true (and unknown) features.

The above results can be shown by making use of Markov chain theory.

The property of ergodicity of a Markov chain is what mainly drives the validity of MCMC methods.

A Markov chain is ergodic if it satisfies:

1. *Irreducibility.*
2. *Aperiodicity.*
3. *Positive recurrence.*

Why does MCMC work?

Law of Large Numbers (LLN) and Central Limit Theorem (CLT) are fundamental but based on i.i.d. samples.

Markov chains produced by MCMC do not satisfy independence.

However, LLN and CLT can be generalized to the case of dependent random variables yielding the *Markov chain Law of Large Numbers* and the *Markov chain Central Limit Theorem*.

These theorems from Markov chain theory ensure that the Markov chains obtained via MCMC can be used to infer the (unknown) features of the target distribution.

Exercise

The table below reports the number of dead mice y_i out of the n_i exposed to a dangerous virus through the air during 3 hours at different concentration levels $w_i, i = 1, \ldots, 8$.

w_i	1.583	1.712	1.774	1.843	1.875	1.892	1.902	1.930
y_i	7	12	18	50	59	60	61	64
n_i	58	61	63	55	61	68	63	64

Table: Number of dead mice exposed to concentration levels of a virus through the air.

- Number of dead mice conditional on concentration level follow $y_i|\pi_i \sim \text{Bin}(n_i, \pi_i)$.
- Death probability modeled via a logit transformation $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta w_i$.
- Use non-informative priors $p(\alpha) \propto 1$ and $p(\beta) \propto 1$ to obtain the joint posterior of (α, β).
- Implement the (random-walk) Metropolis algorithm to obtain a point estimate + 95%CI for α and β .

Take home messages

Sampling is a powerful and invaluable tool for inference in Bayesian models.

Direct sampling methods providing i.i.d. samples are typically not available ...

Markov chain Monte Carlo is the way to go!

If full conditional distributions are available, use the Gibbs sampler.

Otherwise, use the (random walk) Metropolis-Hastings.

Even if Markov chains are dependent, theory ensures that LLN and CLT still hold.

This allows to use MCMC algorithm seen here for inference on parameters of interest.

Thank you!