

---

# Advanced Bayesian Inference with Case Studies

Mathieu Ribatet—Full Professor of Statistics



# References

---

▷ 0. Introduction

1. Bayesian Refresher

1.5 Bayesian asymptotics

2. Intractable posterior

3. Hierarchical models

4. Finite mixture models

5. Approximate Bayesian Computation

# 0. Introduction

- 
- In your first Bayesian course, we were mainly concerned with simple models where the posterior distribution were know explicitly
  - It won't be the case anymore and thus we will need computational tools to bypass this hurdle.

- In your first Bayesian course, we were mainly concerned with simple models where the posterior distribution were know explicitly
- It won't be the case anymore and thus we will need computational tools to bypass this hurdle.



Always bring your laptop during the lectures!!!

# What will I learn in this course?

---

- Be able to work with more realistic Bayesian models
- Extend your Monte Carlo knowledge with Monte Carlo Markov Chain techniques
- Learn a bit of graphical models, a.k.a., Bayesian networks
- Feel confident with hierarchical models
- Write code from scratch, i.e., know exactly of all the machinery actually works!

# More realistic Bayesian models

X-rays of the children's skulls were shot by orthodontists to measure the distance from the hypophysis to the pterygomaxillary fissure. Shots were taken every two years from 8 years of age until 14 years of age.

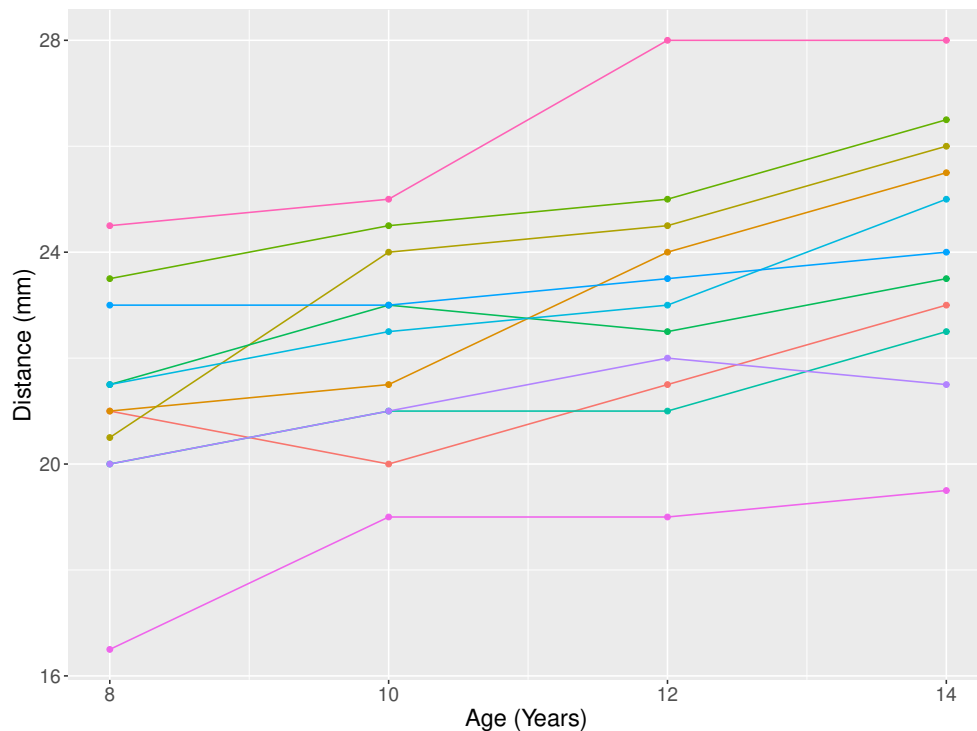


Figure 1: The data collected by the orthodontists.

$$Y_{ij} = \beta_1 + b_j + \beta_2 x_{ij} + \varepsilon_{ij},$$
$$b_j \sim N(0, \sigma_b^2),$$
$$\varepsilon_{ij} \sim N(0, \sigma^2),$$

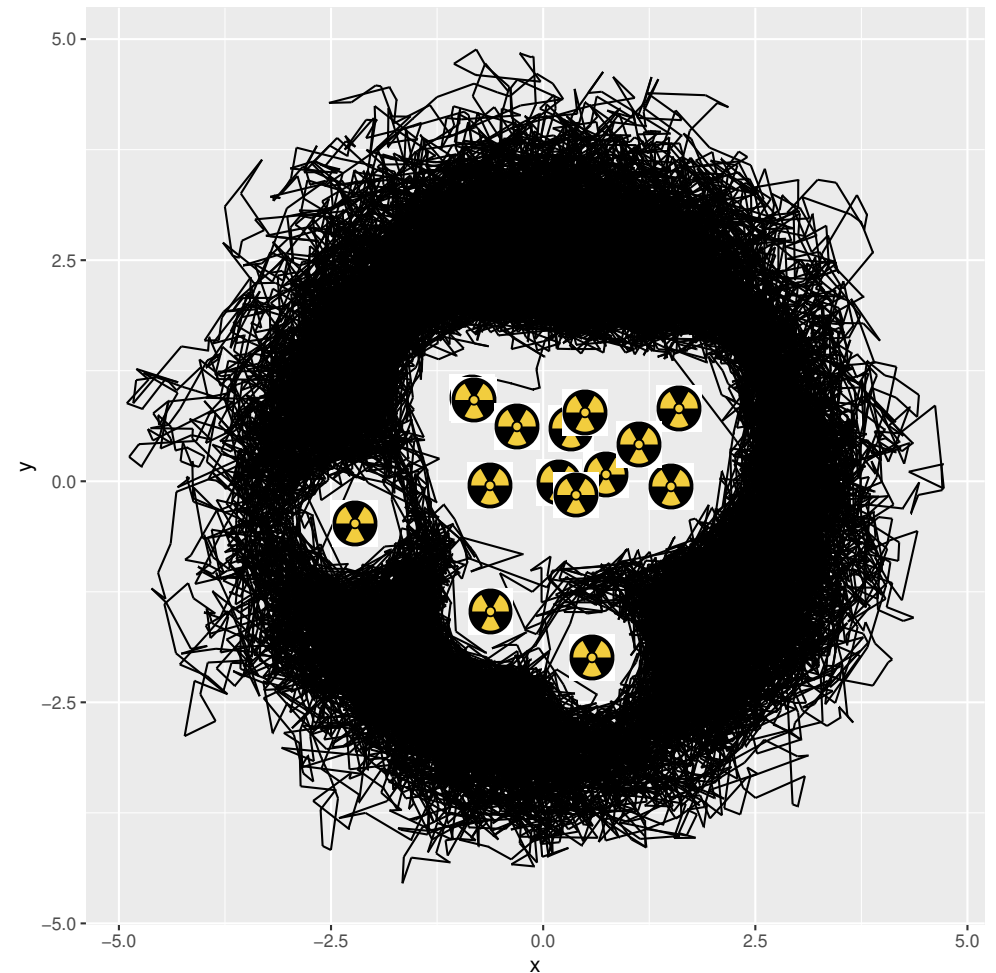
- $Y_{ij}$  is the distance for obs.  $i$  on subject  $j$ ;
- $x_{ij}$  is the age of the subject when the  $i$ -th obs. is made on subject  $j$ ;
- Bayesian: prior on  $\beta_1, \sigma_b^2, \sigma^2$ .

# Extend your Monte Carlo abilities

**Exercise 1** (C. P., Robert (2007)). Let  $\mu_1, \dots, \mu_p \in \mathbb{R}^2$  be  $p$  fixed repulsive points. We aim at sampling from

$$g(\theta) \propto \exp\left(-\frac{\|\theta\|_2^2}{2}\right) \prod_{j=1}^p \exp\left(-\frac{1}{\|\theta - \mu_j\|_2^2}\right).$$

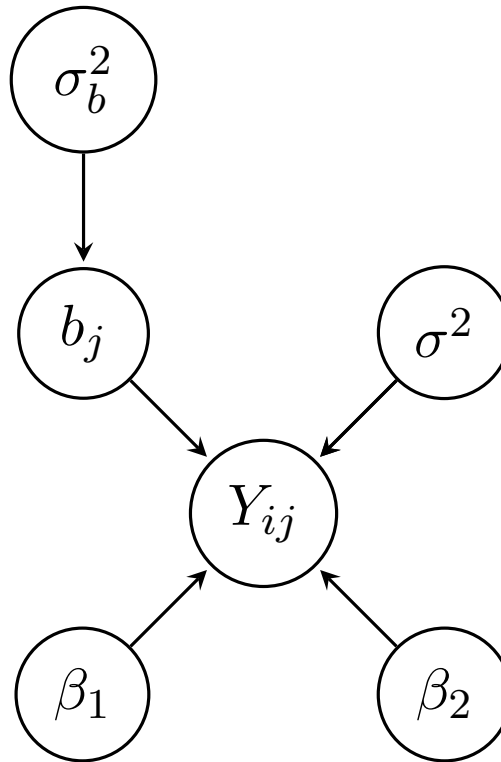
Write an R / Python code to sample from this distribution using a gaussian random walk M.-H. algorithm with innovations  $N(0, \sigma \text{Id}_2)$ .



**Figure 2:** *Sample path of the Markov chain. The repulsive points are represented as ☢. Settings:  $p = 15$ ,  $\theta_0 = (-1, 1)^\top$ ,  $\sigma = 0.1$*



# Learn a bit of graphical models



0. Introduction

---

1. Bayesian  
▷ Refresher

---

1.5 Bayesian  
asymptotics

---

2. Intractable  
posterior

---

3. Hierarchical  
models

---

4. Finite mixture  
models

---

5. Approximate  
Bayesian  
Computation

---

# 1. Bayesian Refresher

# Bayesian statistical models

---

**Definition 1.** A parametric family of functions  $\{f(x; \theta) : x \in E, \theta \in \Theta\}$  is a **statistical model** if, for any  $\theta \in \Theta$ ,  $x \mapsto f(x; \theta)$  is a probability density function on  $E$ .

The sets  $\Theta$  and  $E$  are respectively called **parameter space** and **observational space**.

The above model is said to be **parametric** if  $\dim(\Theta) < \infty$ .

If we further place a **prior distribution**  $\pi$  on the parameter  $\theta$  we are dealing with a **Bayesian statistical model**  $(f, \pi)$ .

The parameters of the prior distribution  $\pi$  are called the **hyper-parameters**.

**Example 1.** The Gaussian model with known variance  $\sigma^2$  and a Normal prior on  $\mu$ , i.e.,

$$\begin{aligned} Y \mid \mu &\sim N(\mu, \sigma^2) \\ \mu \mid \mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2). \end{aligned}$$

# Posterior distributions

---

**Definition 2.** Given a sample  $\mathbf{x}_{1:n} = (x_1, \dots, x_n)$  and a Bayesian model  $(f, \pi)$ . The main focus in Bayesian inference is on the **posterior distribution**

$$\pi(\theta \mid \mathbf{x}_{1:n}) = \frac{f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)}{\int f(\mathbf{x}_{1:n}; \theta)\pi(\theta)d\theta},$$

provided that the **marginal distribution (normalizing constant)**

$$m(\mathbf{x}_{1:n}) = \int f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)d\theta < \infty.$$

# Posterior distributions

**Definition 2.** Given a sample  $\mathbf{x}_{1:n} = (x_1, \dots, x_n)$  and a Bayesian model  $(f, \pi)$ . The main focus in Bayesian inference is on the **posterior distribution**

$$\pi(\theta \mid \mathbf{x}_{1:n}) = \frac{f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)}{\int f(\mathbf{x}_{1:n}; \theta)\pi(\theta)d\theta},$$

provided that the **marginal distribution (normalizing constant)**

$$m(\mathbf{x}_{1:n}) = \int f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)d\theta < \infty.$$

 It is often very convenient to work up to a multiplicative factor independent of  $\theta$  since it will cancel out in the above expression. In such situations we will write

$$\pi(\theta \mid \mathbf{x}_{1:n}) \propto f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta).$$

# Prior distributions

---

**Definition 3.** A family  $\mathcal{F}$  of probability distribution on  $\Theta$  is **conjugate** for the statistical model  $\{f(x | \theta): x \in E, \theta \in \Theta\}$  if, for any  $\pi \in \mathcal{F}$ , the posterior distribution  $\pi(\theta | \mathbf{x}_{1:n}) \in \mathcal{F}$ .

**Definition 4.** A measure  $\pi$  on  $\Theta$  is an **improper prior** if it is actually **not** a probability measure but only a  **$\sigma$ -finite** distribution on  $\Theta$ .

# Prior distributions

---

**Definition 3.** A family  $\mathcal{F}$  of probability distribution on  $\Theta$  is **conjugate** for the statistical model  $\{f(x | \theta): x \in E, \theta \in \Theta\}$  if, for any  $\pi \in \mathcal{F}$ , the posterior distribution  $\pi(\theta | \mathbf{x}_{1:n}) \in \mathcal{F}$ .

**Definition 4.** A measure  $\pi$  on  $\Theta$  is an **improper prior** if it is actually **not** a probability measure but only a  $\sigma$ -finite distribution on  $\Theta$ .

 Watch out when using improper priors, there is no guarantee that the posterior distribution will exist!

# Non informative priors

---

- Non informative priors, although quite controversial, try to mitigate the impact of the prior distribution on the posterior distribution
- Two main types of non informative priors:
  - Laplace prior for which

$$\pi(\theta) \propto 1_{\{\theta \in \Theta\}}, \quad (\text{might be improper})$$

- Jeffreys' prior for which

$$\pi(\theta) \propto \sqrt{\det I(\theta)},$$

where for any (non random!)  $\theta \in \Theta$ ,  $I(\theta) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X; \theta) \right]$  with  $X \sim f(\cdot; \theta)$ .



# Point estimates

---

- Given a parametric Bayesian model, it is common practice to summarize the posterior distribution.
- Possible choices for these point estimate are
  - posterior mean ( $\ell^2$  loss);
  - posterior median ( $\ell^1$  loss);
  - posterior mode (no loss-based estimate);
  - posterior quantiles (weighted  $\ell^1$  loss).

# Credible intervals

**Definition 5.** Given a Bayesian model  $(f, \pi)$ , a interval  $I_{\mathbf{x}_{1:n}}$  is said to be a **credible interval** of level  $\alpha$  if

$$\Pr_{\pi}(\theta \in I_{\mathbf{x}_{1:n}} \mid \mathbf{x}_{1:n}) = \int_{I_{\mathbf{x}_{1:n}}} \pi(\theta \mid \mathbf{x}_{1:n}) d\theta = \alpha.$$

Credible intervals are not unique but typical version of them are:

- symmetric credible interval for which

$$I_{\mathbf{x}_{1:n}} = \left[ q_{\pi} \left( \frac{1 - \alpha}{2}, \mathbf{x}_{1:n} \right), q_{\pi} \left( 1 - \frac{1 - \alpha}{2}, \mathbf{x}_{1:n} \right) \right];$$

- high posterior density interval for which

$$I_{\mathbf{x}_{1:n}} = \{ \theta \in \Theta : \pi(\theta \mid \mathbf{x}_{1:n}) \geq u_{\alpha} \}.$$

# Predictive distribution

---

- Often one wish to estimate a future observation based on the past data  $\mathbf{x}_{1:n} = (x_1, \dots, x_n)^\top$ .
- Since we are Bayesian,  $\theta$  is a random variable and the predictor has to integrate w.r.t. the posterior distribution.

**Definition 6.** The [posterior predictive distribution](#) is defined by

$$\pi(x_{n+1} \mid \mathbf{x}_{1:n}) = \int f(x_{n+1} \mid \theta, \mathbf{x}_{1:n})\pi(\theta \mid \mathbf{x}_{1:n})d\theta.$$

# Predictive distribution

- Often one wish to estimate a future observation based on the past data  $\mathbf{x}_{1:n} = (x_1, \dots, x_n)^\top$ .
- Since we are Bayesian,  $\theta$  is a random variable and the predictor has to integrate w.r.t. the posterior distribution.

**Definition 6.** The **posterior predictive distribution** is defined by

$$\pi(x_{n+1} \mid \mathbf{x}_{1:n}) = \int f(x_{n+1} \mid \theta, \mathbf{x}_{1:n})\pi(\theta \mid \mathbf{x}_{1:n})d\theta.$$

 In particular one could estimate the future observation  $x_{n+1}$  with

$$\hat{x}_{n+1} = \int x_{n+1}\pi(x_{n+1} \mid \mathbf{x}_{1:n})dx_{n+1}.$$

0. Introduction

---

1. Bayesian Refresher

---

1.5 Bayesian  
▷ asymptotics

---

2. Intractable  
posterior

---

3. Hierarchical  
models

---

4. Finite mixture  
models

---

5. Approximate  
Bayesian  
Computation

---

# 1.5 Bayesian asymptotics

## Why a section 0.5?

- Talking about **asymptotics** in Bayesian statistics is a bit **awkward**.
- Indeed the core concept in Bayesian statistics is to base inference on the **actual observed sample**.
- For instance, think about **credible intervals**

$$\Pr_{\pi}(\theta \in I \mid \mathbf{x}_{1:n}) = 1 - \alpha,$$

which states that,<sup>1</sup> **given the observation  $\mathbf{x}_{1:n}$** , the “true parameter  $\theta_0$ ” belongs to  $I$  with probability  $1 - \alpha$ .

- This has to be contrasted with (usually asymptotics) **confidence intervals** for which we have

$$\Pr(\theta_0 \in I(\hat{\theta})) \longrightarrow 1 - \alpha, \quad n \rightarrow \infty,$$

which states that, provided  $n$  is large enough,  $100(1 - \alpha)\%$  of the time, the “true parameter  $\theta_0$ ” is **expected** to lie into intervals of the form  $I(\hat{\theta})$ .

---

<sup>1</sup>if our model is correct

# Bayesian asymptotics

---

**Definition 7.** The sequence of posterior distributions  $\{\pi(\cdot \mid \mathbf{x}_{1:n}) : n \geq 1\}$  is said to be **consistent** at some  $\theta_0 \in \Theta$ , if

$$\pi(\cdot \mid \mathbf{x}_{1:n}) \xrightarrow{\text{proba}} \delta_{\theta_0}(\cdot), \quad n \rightarrow \infty,$$

where convergence in probability is under the p.d.f.  $f(\cdot; \theta_0)$ .

**Proposition 1.** *We assume that:*

- *the prior distribution is  $O(1)$ , i.e., for any  $\theta \in \Theta$ ,  $n^{-1}\pi(\theta) \rightarrow 0$ ;*
- *there exists a neighbourhood  $\mathcal{N}$  of  $\theta_0$  such that  $\pi(\theta) > 0$  for all  $\theta \in \mathcal{N}$ ;*
- *the observations  $\mathbf{x}_{1:n}$  are iid realizations from the “true” p.d.f.  $f(\cdot; \theta_0)$ .*

*Then the posterior distribution  $\pi(\theta \mid \mathbf{x}_{1:n})$  is consistent at  $\theta_0$ .*

*Proof.* Investigate the behaviour of  $\ln \pi(\theta \mid \mathbf{x}_{1:n}) / \pi(\theta_0 \mid \mathbf{x}_{1:n})$ . □

# Asymptotic Normality

---

**Proposition 2** (No proof (a bit too long and not essential I think)). *With the same assumptions as before and the usual regularity conditions to have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N \{0, -H(\theta_0)^{-1}\}, \quad n \rightarrow \infty,$$

where  $\hat{\theta}$  denotes the MLE and  $H(\theta_0) = \mathbb{E}\{\nabla_{\theta}^2 \ln f(X; \theta_0)\}$ , then

$$\pi(\sqrt{n}(\theta - \hat{\theta}) \mid \mathbf{x}_{1:n}) \xrightarrow{d.} N \{0, -H(\theta_0)^{-1}\}, \quad n \rightarrow \infty.$$




# Asymptotic Normality

**Proposition 2** (No proof (a bit too long and not essential I think)). *With the same assumptions as before and the usual regularity conditions to have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N \{0, -H(\theta_0)^{-1}\}, \quad n \rightarrow \infty,$$

where  $\hat{\theta}$  denotes the MLE and  $H(\theta_0) = \mathbb{E}\{\nabla_{\theta}^2 \ln f(X; \theta_0)\}$ , then

$$\pi(\sqrt{n}(\theta - \hat{\theta}) \mid \mathbf{x}_{1:n}) \xrightarrow{d} N \{0, -H(\theta_0)^{-1}\}, \quad n \rightarrow \infty.$$

 This result indicates that, for any (sensible) prior distribution, and provided  $n$  is large enough, the posterior distribution will approximately be equal to that of the maximum likelihood estimator.

0. Introduction

1. Bayesian Refresher

1.5 Bayesian asymptotics

2. Intractable  
▷ posterior

The magic of M.-H.  
A more interesting application

3. Hierarchical models

4. Finite mixture models

5. Approximate Bayesian Computation

## 2. Intractable posterior distribution

## When things goes wrong?

---

$$\pi(\theta \mid \mathbf{x}_{1:n}) = \frac{f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)}{\int f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)d\theta}$$

- Bayesian analysis require to characterize this posterior distribution.
- But if we don't have closed form expressions for  $\pi(\theta \mid \mathbf{x}_{1:n})$ ?

## When things goes wrong?

$$\pi(\theta \mid \mathbf{x}_{1:n}) = \frac{f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)}{\int f(\mathbf{x}_{1:n} \mid \theta)\pi(\theta)d\theta}$$

- Bayesian analysis require to characterize this posterior distribution.
- But if we don't have closed form expressions for  $\pi(\theta \mid \mathbf{x}_{1:n})$ ?
- Why not trying to generate a  $N$ -sample, say  $(\theta_1, \dots, \theta_N)$ , from this **posterior distribution** and base (Bayesian) inference on this sample?
- Such approach is part of **Monte Carlo techniques** which heavily rely on the Law of Large Numbers

$$\frac{1}{N} \sum_{i=1}^N h(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}\{h(X)\}, \quad N \rightarrow \infty, \quad X_1, X_2, \dots \stackrel{\text{iid}}{\sim} X,$$

provided that  $\mathbb{E}\{|h(X)|\} < \infty$ .



[BAYESIAN MODE OFF]

(We aim at sampling from a given target density  $g$ )

# Monte Carlo Markov Chains

---

- In this course, we will restrict our attention to a subclass of Monte Carlo techniques: [Monte Carlo Markov Chain](#) algorithms, or MCMC for short.
- Please note that, although taught within a Bayesian course, MCMC techniques is [not](#) specific to Bayesian inference.
- MCMC techniques are just a collection of sampling schemes that produce a Markov chain whose [stationary distribution](#) is a [pre-specified](#) distribution.

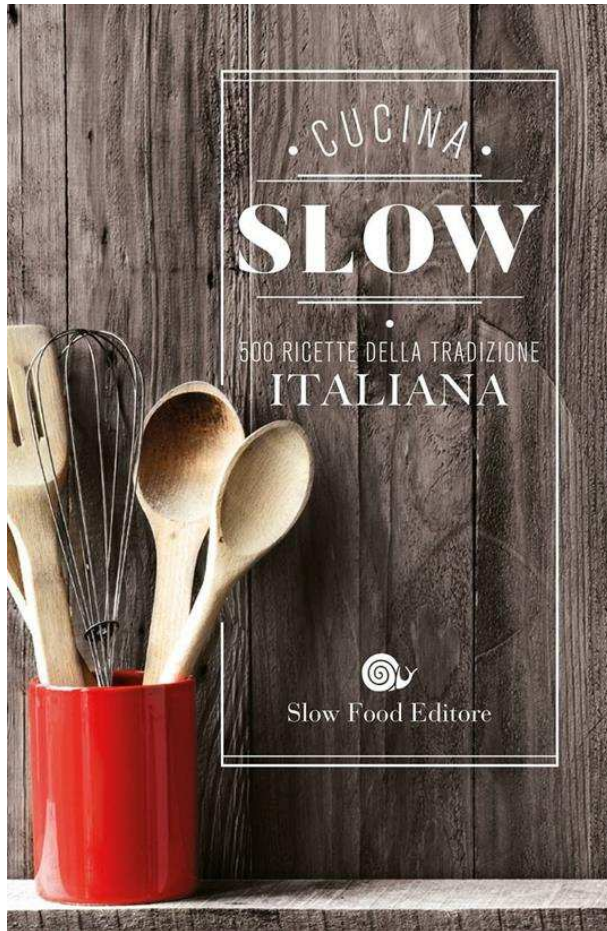
# Monte Carlo Markov Chains

---

- In this course, we will restrict our attention to a subclass of Monte Carlo techniques: [Monte Carlo Markov Chain](#) algorithms, or MCMC for short.
- Please note that, although taught within a Bayesian course, MCMC techniques is [not](#) specific to Bayesian inference.
- MCMC techniques are just a collection of sampling schemes that produce a Markov chain whose [stationary distribution](#) is a [pre-specified](#) distribution.

 Hence in Bayesian inference, this pre-specified distribution will most often be our posterior distribution.

# Metropolis–Hastings recipe



## Ingredients

- a proposal kernel  $K(\cdot, \cdot): \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p$  such that for any  $x \in \mathbb{R}^p$ ,  $K(x, \cdot)$  is a p.d.f.
- A target p.d.f.  $g$ .

**Idea** Start with some fixed  $x \in \mathbb{R}^p$  and add perturbation using  $K(x, \cdot)$ .

**Results** A Markov chain whose stationary distribution is  $g$ .



# Metropolis–Hastings algorithm

---

**Algorithm 1:** The Metropolis–Hastings algorithm.

---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ , initial state  $X_0 \in \mathbb{R}^p$ , proposal kernel  $K(\cdot, \cdot)$ ,  
 $N \in \mathbb{N}_*$ .

**output:** A Markov chain whose stationary distribution is  $g$ .

1 **for**  $t \leftarrow 1$  **to**  $N$  **do**

2     Draw a **proposal**  $X_*$  from the proposal kernel  $K(X_{t-1}, \cdot)$ ;

3     Compute the **acceptance probability**

$$\alpha(X_{t-1}, X_*) = \min \left\{ 1, \frac{g(X_*)K(X_*, X_{t-1})}{g(X_{t-1})K(X_{t-1}, X_*)} \right\}$$

4     Draw  $U \sim U(0, 1)$  and let

$$X_t = \begin{cases} X_*, & \text{if } U \leq \alpha(X_{t-1}, X_*) \\ X_{t-1}, & \text{otherwise} \end{cases}$$

5 **Return** the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;

---

## Reversibility and detailed balance condition

---

**Definition 8.** A Markov chain  $\{X_t: t \geq 0\}$  with transition kernel  $P$  satisfies the **detailed balance condition** if there exists a function  $f$  satisfying

$$f(x)P(x, y) = f(y)P(y, x), \quad x, y \in \mathbb{R}^p.$$

**Theorem 1.** *Suppose that a Markov chain with transition kernel  $P$  satisfies the detailed balance condition for some p.d.f.  $g$ . Then  $g$  is the invariant density of the chain.*

*Proof.* We have to show that for any  $y \in \mathbb{R}^p$ ,  $\int g(x)P(x, y)dx = g(y)$ . □

# Reversibility and detailed balance condition

**Definition 8.** A Markov chain  $\{X_t: t \geq 0\}$  with transition kernel  $P$  satisfies the **detailed balance condition** if there exists a function  $f$  satisfying

$$f(x)P(x, y) = f(y)P(y, x), \quad x, y \in \mathbb{R}^p.$$

**Theorem 1.** *Suppose that a Markov chain with transition kernel  $P$  satisfies the detailed balance condition for some p.d.f.  $g$ . Then  $g$  is the invariant density of the chain.*

*Proof.* We have to show that for any  $y \in \mathbb{R}^p$ ,  $\int g(x)P(x, y)dx = g(y)$ . □

 In the Markov chain literature, chains satisfying the detailed balance condition are said **reversible**.

## Justification of the M.-H. algorithm

---

**Theorem 2.** *Let  $\{X_t: t \geq 0\}$  be the Markov chain produced by the M.-H. algorithm. For every proposal kernel  $K$  whose support includes that of  $g$ ,*

- 1. the transition kernel of the chain satisfies the detailed balance condition for  $g$ ;*
- 2.  $g$  is a stationary distribution of the chain.*

## Justification of the M.-H. algorithm

---

**Theorem 2.** *Let  $\{X_t: t \geq 0\}$  be the Markov chain produced by the M.-H. algorithm. For every proposal kernel  $K$  whose support includes that of  $g$ ,*

- 1. the transition kernel of the chain satisfies the detailed balance condition for  $g$ ;*
- 2.  $g$  is a stationary distribution of the chain.*

*Proof.* Start by writing the transition kernel of the M.-H. algorithm and then show the detailed balance condition for  $g$  so that  $g$  is the invariant distribution.  $\square$

# The magic of M.-H.



The M.-H. algorithm is appealing as :

- very versatile, i.e., widely applicable
- easy to implement
- normalizing constant free, i.e., only ratios

$$\frac{g(X_*)}{g(X_t)}, \quad \frac{K(X_*, X_t)}{K(X_t, X_*)}.$$

# The magic of M.-H.



The M.-H. algorithm is appealing as :

- very versatile, i.e., widely applicable
- easy to implement
- normalizing constant free, i.e., only ratios

$$\frac{g(X_*)}{g(X_t)}, \quad \frac{K(X_*, X_t)}{K(X_t, X_*)}.$$

👉 Now you know why M.-H. is widely used in Bayesian inference, e.g., when

$$m(\mathbf{x}_{1:n}) = \int f(\mathbf{x}_{1:n} | \theta) \pi(\theta) d\theta$$

has no closed form!

## Application : Naive hard-shell ball model for gas

---

**Exercise 2.** We aim at sampling  $K$  non overlapping hard-shell balls, with equal diameters  $d$ , uniformly on  $[0, 1] \times [0, 1]$ .

Write a pseudo-code to sample from this model using the M.-H. algorithm.



**Figure 3:** *Click me!* (Note that the chain was thinned as you might have guessed because of weird jumps)



## Ergodicity // Law of large numbers

---

**Theorem 3.** *Suppose that the M.-H. chain  $\{X_t: t \geq 0\}$  is  $(g-)$  irreducible.*

1. *If  $h \in L^1(g)$ , then*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N h(X_t) = \int h(x)g(x)dx, \quad g\text{-a.e.}$$

2. *If, in addition, the chain is aperiodic, then*

$$\lim_{n \rightarrow \infty} \left\| \int P^n(x, \cdot) \mu(dx) - g \right\|_{TV} = 0,$$

*for every initial distribution  $\mu$  and where  $\|\nu\|_{TV} = \sup_B |\nu(B)|$ .*

*Proof.* Admitted. See your Markov chains' lecture notes. Essentially show Harris recurrence. □

## Ergodicity // Law of large numbers

**Theorem 3.** Suppose that the M.-H. chain  $\{X_t: t \geq 0\}$  is  $(g-)$  irreducible.


1. If  $h \in L^1(g)$ , then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N h(X_t) = \int h(x)g(x)dx, \quad g\text{-a.e.}$$

2. If, in addition, the chain is aperiodic, then

$$\lim_{n \rightarrow \infty} \left\| \int P^n(x, \cdot) \mu(dx) - g \right\|_{TV} = 0,$$

for every initial distribution  $\mu$  and where  $\|\nu\|_{TV} = \sup_B |\nu(B)|$ .

 This result allows us to estimate  $I = \int h(x)g(x)dx$  from the empirical mean  $\hat{I}_N = N^{-1} \sum_{t=1}^N h(X_t)$ . Convergence was not clear as the  $X_t$ 's are **dependent!**

# Famous type of M.-H. algorithms

---

- Independent M.-H., i.e.,

$$X_* \sim q \quad X_* \text{ independent from } X_t.$$

- Random walk M.-H., i.e., the proposal state is given by

$$X_* = X_t + \varepsilon_t, \quad \varepsilon_t \stackrel{\text{iid}}{\sim} q,$$

e.g.,  $q$  is the p.d.f. of a centered Gaussian distribution with (proposal) covariance  $\Sigma \text{Id}$ .

- Log-scale random walk M.-H., i.e., the proposal state satisfies

$$\ln X_* = \ln X_t + \varepsilon_t, \quad \varepsilon_t \stackrel{\text{iid}}{\sim} q,$$

- The Gibbs sampler that we will focus later on...

## Focus on the random walk M.-H.

---

**Proposition 3.** Consider the random walk M.-H. updating scheme  $X_* = X_t + \varepsilon_t$  with  $\varepsilon_t \sim q$ . If  $q$  is symmetric around 0, then the acceptance probability simplifies to

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)}{g(X_t)} \right\}.$$


*Proof.* Just write the proposal kernel and simplify the acceptance probability.  $\square$

## Focus on the random walk M.-H.

**Proposition 3.** Consider the random walk M.-H. updating scheme  $X_* = X_t + \varepsilon_t$  with  $\varepsilon_t \sim q$ . If  $q$  is symmetric around 0, then the acceptance probability simplifies to

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)}{g(X_t)} \right\}.$$

*Proof.* Just write the proposal kernel and simplify the acceptance probability.  $\square$

 This case corresponds actually to the original definition of the Metropolis algorithm (1953) later generalized by Hastings (1970).

## Focus on the log-scale random walk M.-H.

---

**Proposition 4.** *Consider the random walk M.-H. updating scheme  $\ln X_* = \ln X_t + \varepsilon_t$  with  $\varepsilon_t \sim q$ . If  $q$  is symmetric around 0, then the acceptance probability is given by*

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)X_*}{g(X_t)X_t} \right\}.$$

*Proof.* Give the p.d.f. of  $X_*$  conditionally on  $X_t$  and simplify. □

## Focus on the log-scale random walk M.-H.

**Proposition 4.** Consider the random walk M.-H. updating scheme  $\ln X_* = \ln X_t + \varepsilon_t$  with  $\varepsilon_t \sim q$ . If  $q$  is symmetric around 0, then the acceptance probability is given by

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)X_*}{g(X_t)X_t} \right\}.$$

*Proof.* Give the p.d.f. of  $X_*$  conditionally on  $X_t$  and simplify. □

 The log-scale random walk is often used when  $X_t$  has to be **positive**.

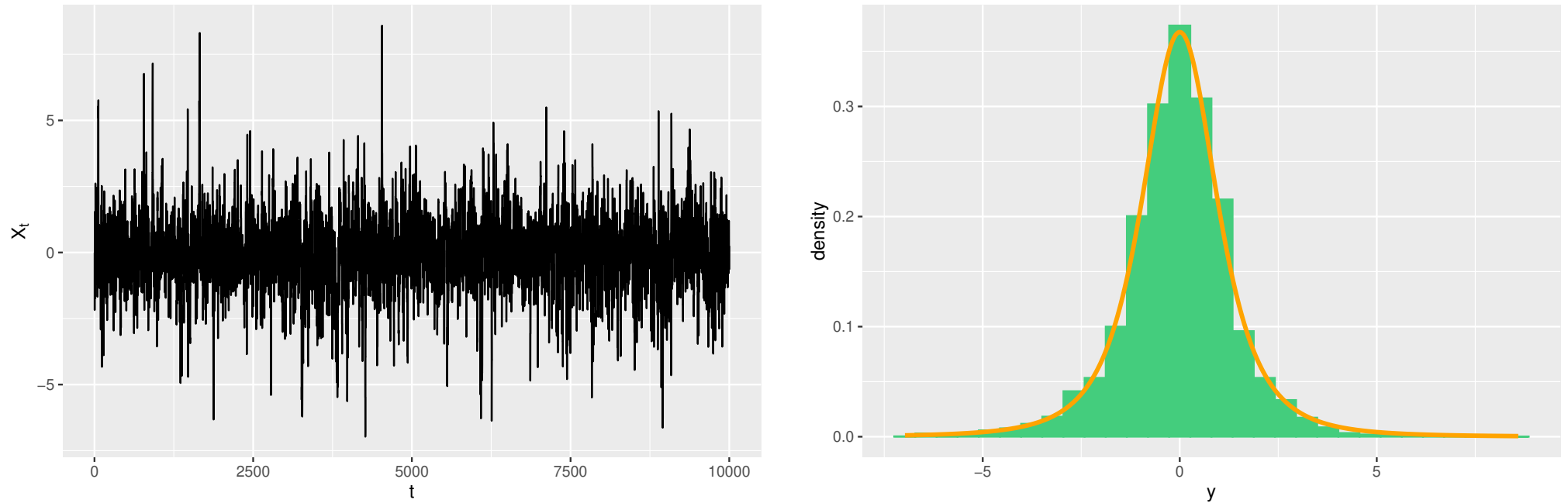
## Toy (and very stupid) example

---

**Exercise 3.** We aim at sampling from a  $t_\nu$  using a random walk M.-H. with Gaussian innovations. Write a pseudo-code for this. Do an implementation in R or Python.



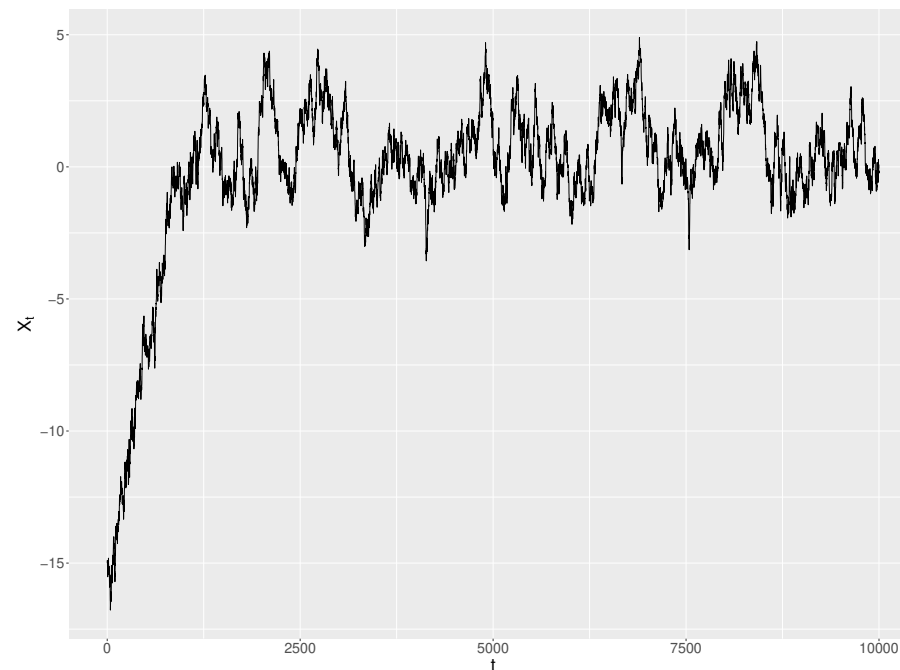
# A chain and the associated histogram



**Figure 4:** *Left : Sample path of a simulated M.-H. chain on our toy example with a  $t_3$  target distribution. Right : Associated histogram of the chain and true target density (solid line).*

# Burnin period

- By construction of the M.-H. algorithm, if there exists  $t_0 \geq 0$  such that  $X_{t_0} \sim g$ , then for all  $t \geq t_0$ ,  $X_t \sim g$ .
- But it may long to reach the **stationary regime** and we typically discard the first  $K$  states, i.e., **removing the burnin period**.



**Figure 5:** *Illustration of the burnin period. Here we set  $X_0 = -15$ . It took around 1250 iterations to reach the stationary regime.*

## A closer look at $\alpha(X_t, X_*)$

---

□ Since

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)K(X_*, X_t)}{g(X_t)K(X_t, X_*)} \right\},$$

to accept  $X_*$  with high probability we have two options:

1. For some  $\varepsilon > 0$ ,  $\Pr(\|X_* - X_t\| > \varepsilon \mid X_t = x_t) \ll 1$
2.  $K(x, y) \approx g(y)$ .

## A closer look at $\alpha(X_t, X_*)$

---

□ Since

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)K(X_*, X_t)}{g(X_t)K(X_t, X_*)} \right\},$$

to accept  $X_*$  with high probability we have two options:

1. For some  $\varepsilon > 0$ ,  $\Pr(\|X_* - X_t\| > \varepsilon \mid X_t = x_t) \ll 1$
2.  $K(x, y) \approx g(y)$ .

□ Unfortunately, both options have undesirable side effects:

1. The chain will explore the state space, i.e., support of  $g$ , very slowly;
2. Since  $g$  is not known explicitly, finding  $K(x, \cdot) \approx g(\cdot)$  is hopeless.

## A closer look at $\alpha(X_t, X_*)$

□ Since

$$\alpha(X_t, X_*) = \min \left\{ 1, \frac{g(X_*)K(X_*, X_t)}{g(X_t)K(X_t, X_*)} \right\},$$

to accept  $X_*$  with high probability we have two options:

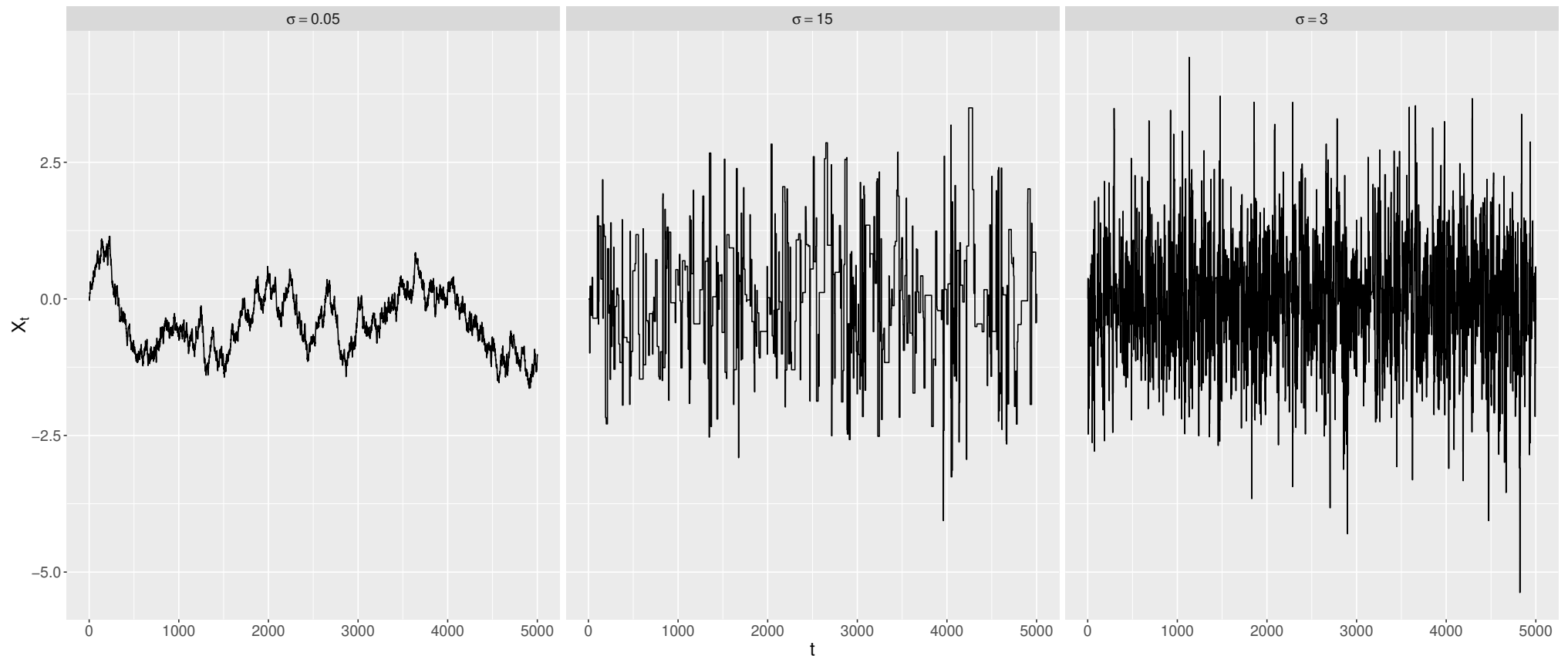
1. For some  $\varepsilon > 0$ ,  $\Pr(\|X_* - X_t\| > \varepsilon \mid X_t = x_t) \ll 1$
2.  $K(x, y) \approx g(y)$ .

□ Unfortunately, both options have undesirable side effects:

1. The chain will explore the state space, i.e., support of  $g$ , very slowly;
2. Since  $g$  is not known explicitly, finding  $K(x, \cdot) \approx g(\cdot)$  is hopeless.

 It is highly recommended to assess the **mixing properties** of the simulated chain.

# Pathological examples of mixing properties



**Figure 6:** Illustration of the mixing properties of a simulated chain. Left: the chain is poorly mixing due to “small moves”.  $\sigma$  is too small. Middle : The chain is poorly mixing due to “large proposal moves” that are thus often rejected so that the chain get piecewise constant.  $\sigma$  is too large. Right: A quite good mixing chain.  $\sigma$  is just right.

# Acceptance rate recommendations

---

**Definition 9.** Consider a Markov chain  $\{X_t: t = 0, \dots, N\}$  (with continuous state space) obtained from a M.-H. algorithm with proposal kernel  $K$ . The **acceptance rate** is given by

$$\rho = \frac{1}{N} \sum_{t=1}^N 1_{\{X_{t-1} \neq X_t\}} = \frac{\# \text{ accepted proposals}}{N}.$$

- Numerical simulations shows that defining  $K$  to reach a
  - 50% acceptance rate for low dimensional problem, i.e.,  $X \in \mathbb{R}^d$ ,  $d = 1, 2$ ;
  - 25% acceptance rate for high dimensional problems, i.e.,  $d > 2$ .
- These a just guidance and should not be considered as a gold standard!

# Acceptance rate recommendations

**Definition 9.** Consider a Markov chain  $\{X_t: t = 0, \dots, N\}$  (with continuous state space) obtained from a M.-H. algorithm with proposal kernel  $K$ . The **acceptance rate** is given by

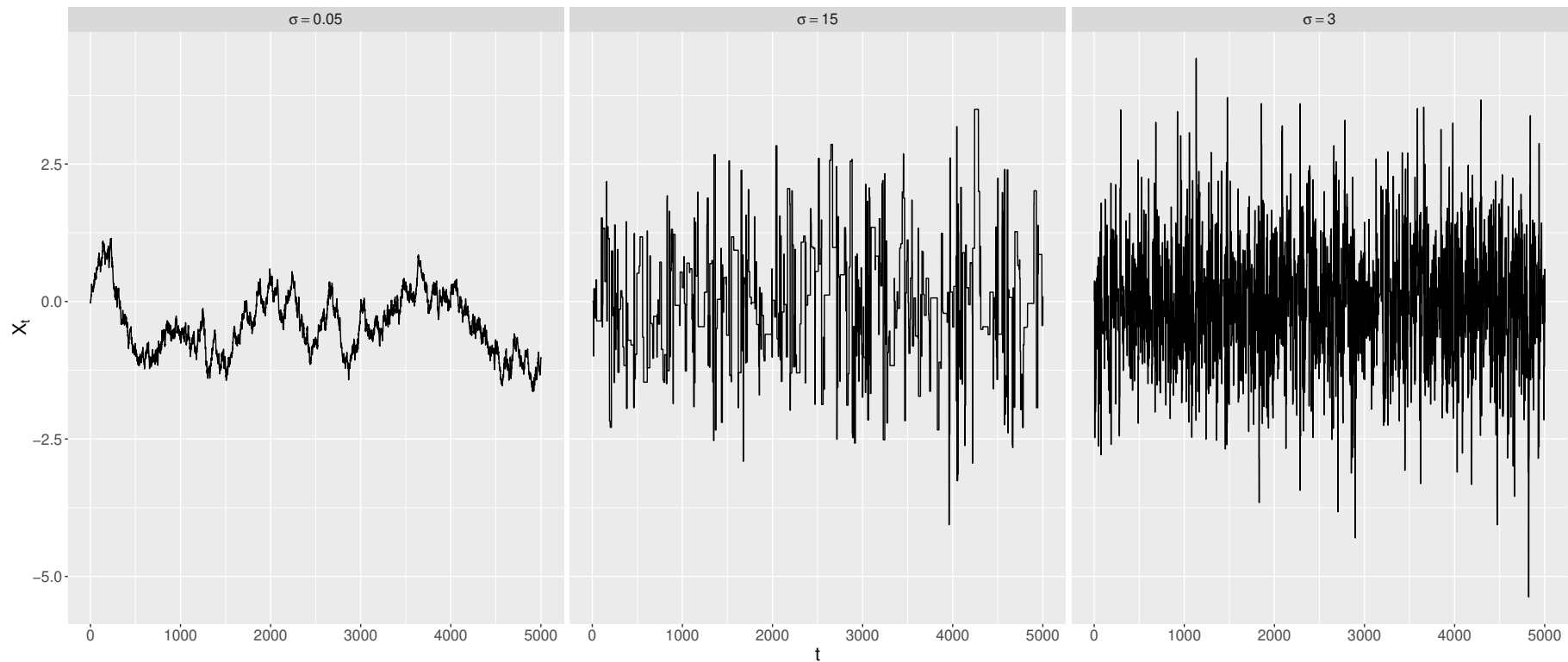
$$\rho = \frac{1}{N} \sum_{t=1}^N 1_{\{X_{t-1} \neq X_t\}} = \frac{\# \text{ accepted proposals}}{N}.$$

- Numerical simulations shows that defining  $K$  to reach a
  - 50% acceptance rate for low dimensional problem, i.e.,  $X \in \mathbb{R}^d$ ,  $d = 1, 2$ ;
  - 25% acceptance rate for high dimensional problems, i.e.,  $d > 2$ .
- These a just guidance and should not be considered as a gold standard!

 Always have a look at the sample path of your simulated chain!



# Pathological examples of poor mixing chains



**Figure 7:** Illustration of the mixing properties of a simulated chain. Left: the chain is poorly mixing due to “small moves”.  $\sigma$  is too small. Middle : The chain is poorly mixing due to “large proposal moves” that are thus often rejected so that the chain get piecewise constant.  $\sigma$  is too large. Right: A quite good mixing chain.  $\sigma$  is just right.



Here the acceptance ratio were respectively: 0.99, 0.09 and 0.39.

# Thinning a chain

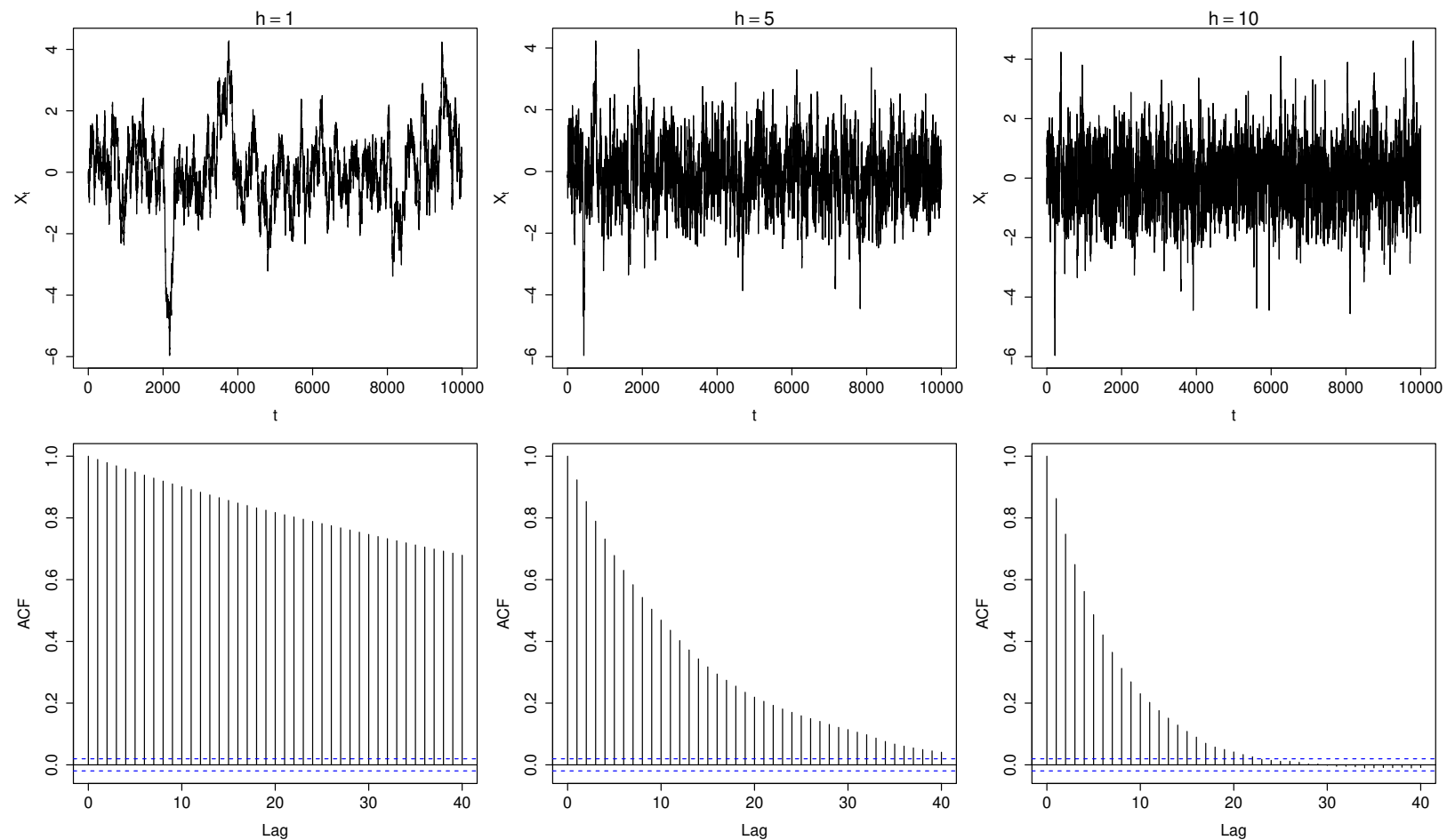
---

**Definition 10.** **Thinning** a chain  $\{X_t : t = 0, \dots, N\}$  by a lag  $h$  consists in taking only the  $h$ -lagged states, i.e.,

$$\{X_{th} : t = 0, \dots, [N/h]\}.$$

- The motivation for thinning a chain is to mitigate the serial dependence within the original chain, i.e., get closer to our beloved “iid” case.
- However from a probabilistic point of view, thinning is useless as far as our chain is ergodic.

# Illustration of thinning a chain



**Figure 8:** *Thinning a chain. Top sample path of the chain and its thinned version—all of length 10000. Bottom: Associated ACF.*

## A more interesting application

---

**Exercise 4** (C. P., Robert (2007)). Let  $\mu_1, \dots, \mu_p \in \mathbb{R}^2$  be  $p$  fixed repulsive points. We aim at sampling from

$$g(\theta) \propto \exp\left(-\frac{\|\theta\|_2^2}{2}\right) \prod_{j=1}^p \exp\left(-\frac{1}{\|\theta - \mu_j\|_2^2}\right).$$

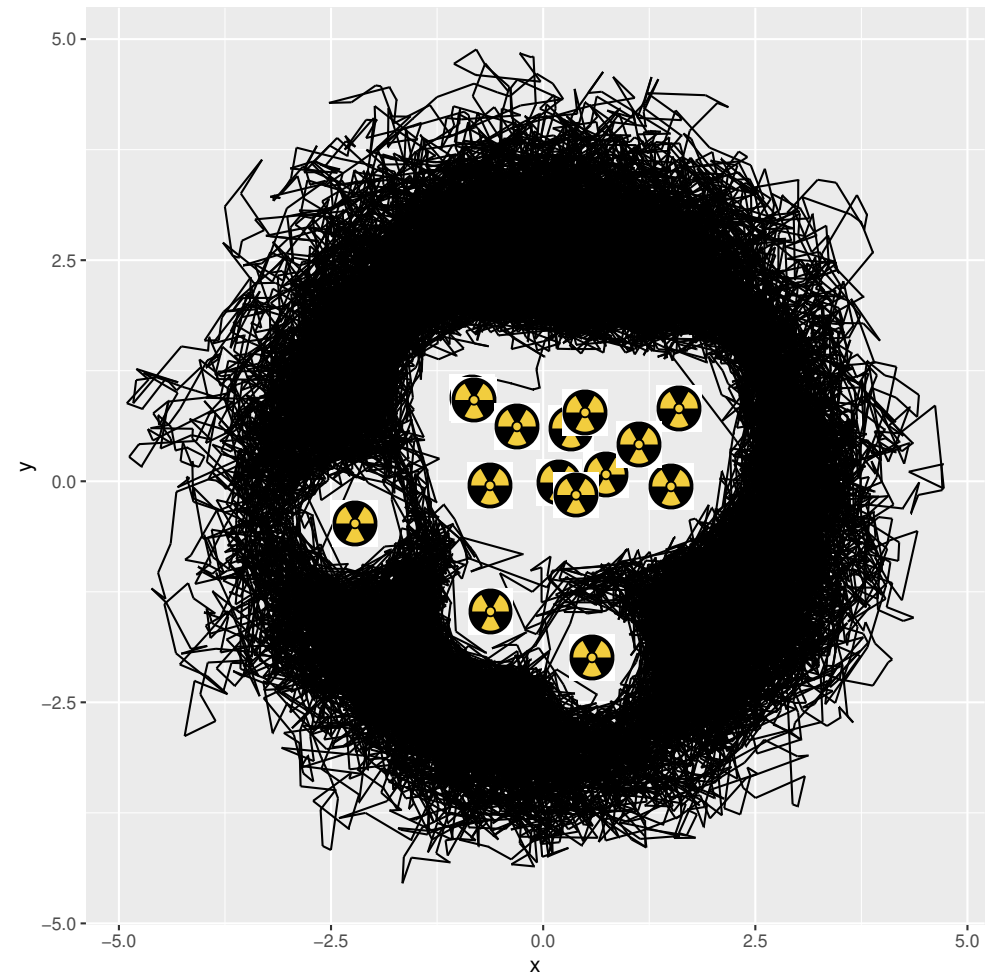
Write an R / Python code to sample from this distribution using a gaussian random walk M.-H. algorithm with innovations  $N(0, \sigma \text{Id}_2)$ .


## A more interesting application

**Exercise 4** (C. P., Robert (2007)). Let  $\mu_1, \dots, \mu_p \in \mathbb{R}^2$  be  $p$  fixed repulsive points. We aim at sampling from

$$g(\theta) \propto \exp\left(-\frac{\|\theta\|_2^2}{2}\right) \prod_{j=1}^p \exp\left(-\frac{1}{\|\theta - \mu_j\|_2^2}\right).$$

Write an R / Python code to sample from this distribution using a gaussian random walk M.-H. algorithm with innovations  $N(0, \sigma \text{Id}_2)$ .



**Figure 9:** Sample path of the Markov chain. The repulsive points are represented as . Settings:  $p = 15$ ,  $\theta_0 = (-1, 1)^\top$ ,  $\sigma = 0.1$

# The curse of dimensionality is everywhere

---

**Exercise 5.** Suppose we wish to simulate from  $U(\mathcal{S}_d)$  where  $\mathcal{S}_d = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$ . To do so we use a random walk M.-H. sampler, i.e.,  $X_* = X_t + \varepsilon_t$  where  $\varepsilon_t = (\varepsilon_{t,1}, \dots, \varepsilon_{t,d})$  with  $\varepsilon_{t,i} \stackrel{\text{iid}}{\sim} U(-L, L)$ ,  $L > 1$ . Given  $X_t = \mathbf{0}$ , show that the acceptance probability satisfies

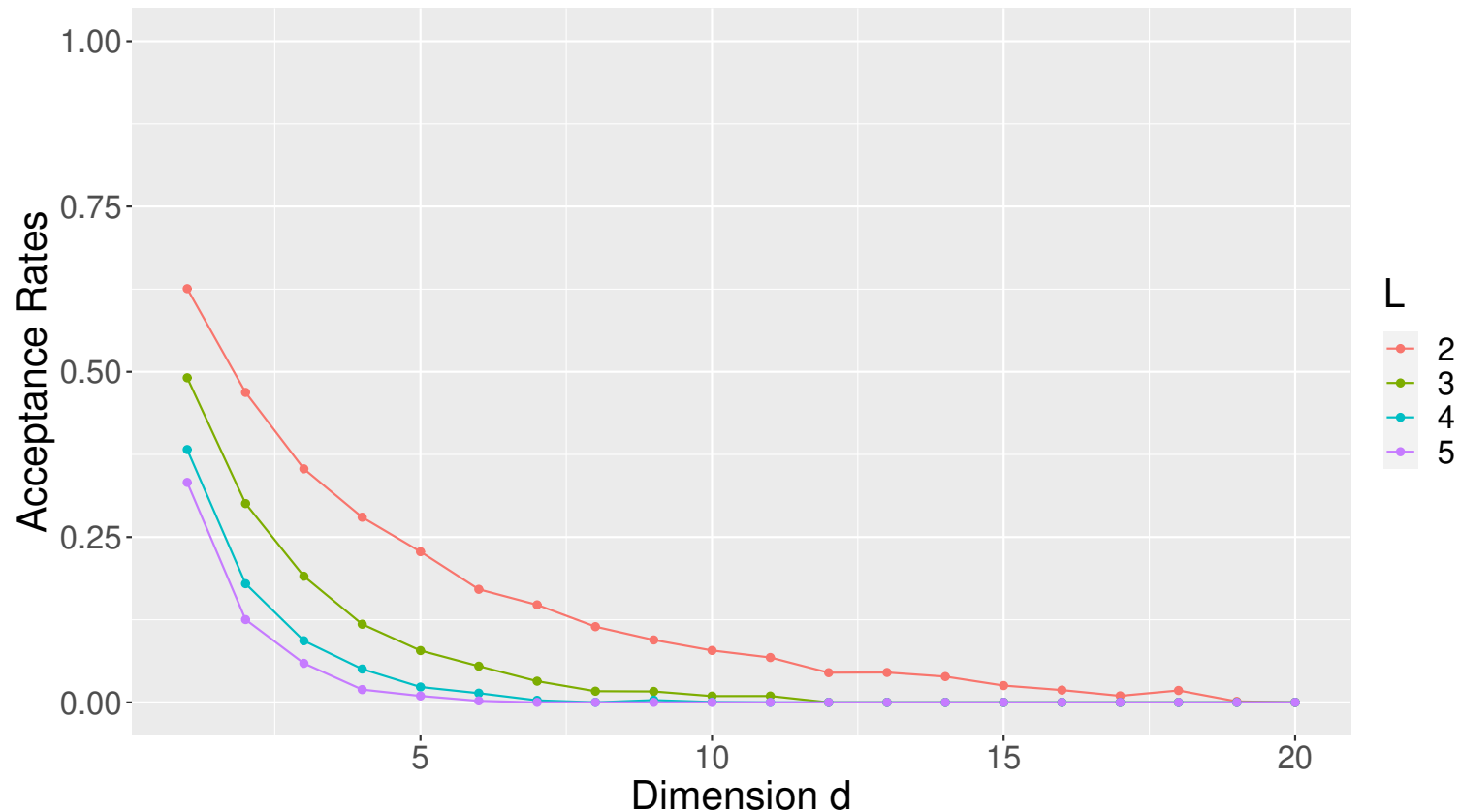
$$\mathbb{E}\{\alpha(X_t, X_*)\} \longrightarrow 0, \quad d \rightarrow \infty.$$

How would you interpret this result?

*Proof.* Start by simplifying the expression of  $\alpha(X_t, X_*)$  and compute  $\Pr(X_* \in \mathcal{S}_d \mid X_t = x_t)$ . Conclude. □

# Numerical illustration of this curse

To investigate this issue a bit further we simulate from a  $d$ -variate standard Normal distribution using a random walk M.-H. with  $U\{[-L, L]^d\}$  innovations.



**Figure 10:** *The curse of dimensionality applies to the M.-H. updating scheme.*

# Multivariate dimensional problems

---

- Loosely speaking the issue just stated is connected to the fact that random walks on  $\mathbb{R}^d$  get “lost” when  $d \geq 3$ .<sup>2</sup>
- As a consequence, when  $d \geq 2$ , it is common practice to use more specialized sampler such as the Gibbs sampler.
- Interestingly the Gibbs sampler corresponds to the M.-H. algorithm with a very specific proposal kernel  $K$ .

---

<sup>2</sup>This is “loosely speaking” since the Markov chain  $\{X_t : t \geq 0\}$  is actually not a random walk (because of the acceptance / rejection stage) and so won't get lost. . .



# The (random scan) Gibbs sampler

---

---

## Algorithm 2: Random scan Gibbs sampler.

---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ ,  $p > 1$ , initial state  $X_0 \in \mathbb{R}^p$ ,  $N \in \mathbb{N}_*$ .

**output**: A Markov chain whose stationary distribution is  $g$ .

*/\* Notation: for  $x \in \mathbb{R}^p$  and  $I \subset \{1, \dots, p\}$ ,  $x_{-I} = \{x_j : j \in \{1, \dots, p\} \setminus I\}$  \*/*

- 1 **for**  $t \leftarrow 1$  **to**  $N$  **do**
  - 2     Set  $X_{t+1} \leftarrow X_t$ ;
  - 3     Draw a coordinate  $I \sim U\{1, \dots, p\}$ —or any dist. on  $\{1, \dots, p\}$ ;
  - 4     Draw  $X_* \sim g(\cdot \mid X_{t,-I})$ , i.e., from the **full conditional distribution**;
  - 5     Let  $X_{t+1,I} \leftarrow X_*$ ;
  - 6 **Return** the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;
-

# The (random scan) Gibbs sampler

---

## Algorithm 2: Random scan Gibbs sampler.

---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ ,  $p > 1$ , initial state  $X_0 \in \mathbb{R}^p$ ,  $N \in \mathbb{N}_*$ .

**output**: A Markov chain whose stationary distribution is  $g$ .

*/\* Notation: for  $x \in \mathbb{R}^p$  and  $I \subset \{1, \dots, p\}$ ,  $x_{-I} = \{x_j : j \in \{1, \dots, p\} \setminus I\}$  \*/*

- 1 **for**  $t \leftarrow 1$  **to**  $N$  **do**
  - 2     Set  $X_{t+1} \leftarrow X_t$ ;
  - 3     Draw a coordinate  $I \sim U\{1, \dots, p\}$ —or any dist. on  $\{1, \dots, p\}$ ;
  - 4     Draw  $X_* \sim g(\cdot \mid X_{t,-I})$ , i.e., from the **full conditional distribution**;
  - 5     Let  $X_{t+1,I} \leftarrow X_*$ ;
  - 6 **Return** the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;
- 

 The proposal kernel is thus  $K(x_t, x_*) = \frac{1}{p}g(x_{*,i} \mid x_{t,-i})\delta_{x_{t,-i}}(x_{*,,-i})$ , hence

$$\alpha(x_t, x_*) = \min \left\{ 1, \frac{g(x_*)g(x_{t,i} \mid x_{*,,-i})}{g(x_t)g(x_{*,i} \mid x_{t,-i})} \right\} = \min \left\{ 1, \frac{g(x_*)g(x_{t,i} \mid x_{*,,-i})g(x_{*,,-i})}{g(x_t)g(x_{*,i} \mid x_{t,-i})g(x_{t,-i})} \right\} = 1.$$

## The systematic scan Gibbs sampler

---

Rather than selecting at random a coordinate to update, we cycle through each coordinate.

---

**Algorithm 3:** Systematic scan Gibbs sampler.

---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ ,  $p > 1$ , initial state  $X_0 \in \mathbb{R}^p$ ,  $N \in \mathbb{N}_*$ .

**output:** A Markov chain whose stationary distribution is  $g$ .

```
1 for  $t \leftarrow 1$  to  $N$  do
2   |   Set  $X_{t+1} \leftarrow X_t$ ;
3   |   for  $j \leftarrow 1$  to  $p$  do
4   |   |   Draw  $X_* \sim g(\cdot \mid X_{t+1,-j})$ ;
5   |   |   Let  $X_{t+1,j} \leftarrow X_*$ ;
6 Return the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;
```

---

# The systematic scan Gibbs sampler

Rather than selecting at random a coordinate to update, we cycle through each coordinate.

---

**Algorithm 3:** Systematic scan Gibbs sampler.


---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ ,  $p > 1$ , initial state  $X_0 \in \mathbb{R}^p$ ,  $N \in \mathbb{N}_*$ .

**output:** A Markov chain whose stationary distribution is  $g$ .

```
1 for  $t \leftarrow 1$  to  $N$  do
2   Set  $X_{t+1} \leftarrow X_t$ ;
3   for  $j \leftarrow 1$  to  $p$  do
4     Draw  $X_* \sim g(\cdot \mid X_{t+1,-j})$ ;
5     Let  $X_{t+1,j} \leftarrow X_*$ ;
6 Return the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;
```

---

 Provided the chain is long enough, in practice there is little difference between systematic and random scan scheme. To do theoretical work, random scan is easier to work with; while in practice we often (if not always) use systematic scan.

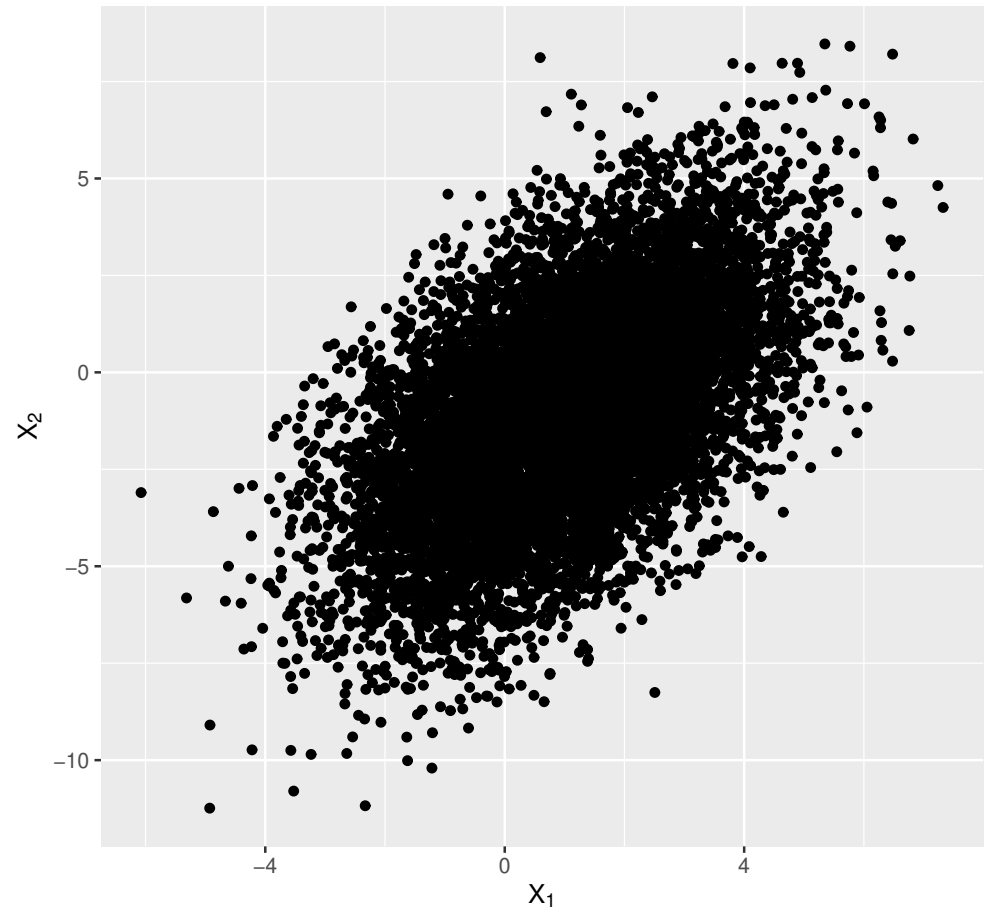
## Another toy (but still dumbass) example

---

**Exercise 6.** We aim at sampling from a bivariate Normal distribution with mean  $\mu = c(1, -1)$  and covariance matrix  $\Sigma = \begin{bmatrix} 3 & 2.5 \\ 2.5 & 7 \end{bmatrix}$ . Write a pseudo-code and then an R / Python code to simulate from this model using a Gibbs sampler.

## Another toy (but still dumbass) example

**Exercise 6.** We aim at sampling from a bivariate Normal distribution with mean  $\mu = c(1, -1)$  and covariance matrix  $\Sigma = \begin{bmatrix} 3 & 2.5 \\ 2.5 & 7 \end{bmatrix}$ . Write a pseudo-code and then an R / Python code to simulate from this model using a Gibbs sampler.



**Figure 11:** *Sample path of the Markov chain.*

# The M.-H. within Gibbs sampler

Sampling from the full conditional distributions is not always possible, if so, you can use a M.-H. updating scheme.

---

**Algorithm 4:** M.-H. within Gibbs sampler (with random scan).

---

**input** : Target distribution  $g$  on  $\mathbb{R}^p$ ,  $p > 1$ , initial state  $X_0 \in \mathbb{R}^p$ ,  $N \in \mathbb{N}_*$ , proposal kernels  $K_j(\cdot, \cdot)$ ,  $j = 1, \dots, p$ .

**output:** A Markov chain whose stationary distribution is  $g$ .

```
1 for  $t \leftarrow 1$  to  $N$  do
2   Draw a coordinate  $I \sim U\{1, \dots, p\}$ —or another discrete distribution on  $\{1, \dots, p\}$ ;
3   Draw a proposal  $X_{*,I} \sim K(X_{t-1}, \cdot)$ ;
4   Let  $X_* = (X_{*,1}, \dots, X_{*,p})^\top$  with
      
$$X_{*,j} = \begin{cases} X_{t-1,j}, & \text{if } j \neq I \\ X_{*,I}, & \text{otherwise.} \end{cases}$$

5   Set  $X_t$  according to the M.-H. updating scheme;
6 Return the Markov chain  $\{X_t : t = 0, \dots, N\}$ ;
```

---

---

**Exercise 7.** Redo Exercise 6 but using a M.-H. within Gibbs to work on an even more dumbass example.



---

**Exercise 7.** Redo Exercise 6 but using a M.-H. within Gibbs to work on an even more dumbass example.

 Unless you're using a M.-H. within Gibbs sampler, targeting the 50% or 25% acceptance rate is irrelevant for Gibbs sampling. However, thinning and removing the burnin period should be considered!



[BAYESIAN MODE ON]

(The target distribution will now be the posterior distribution)

# FC Nantes scoring abilities

## Exercise 8.

We are interesting in modelling the number of goals scored by FC Nantes—or your favourite football team. To this aim we consider the following Bayesian model

$$\begin{aligned} N_i \mid \lambda &\stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda), & i = 1, \dots, n, \\ \lambda &\sim \text{Gamma}(\alpha, \beta), & \alpha, \beta \text{ known.} \end{aligned}$$



1. Give the (explicit) posterior distribution.
2. Write a MCMC sampler to sample from this distribution.
3. Retrieve the data for this year, e.g., from [here](#).
4. Put some sensible value for the hyper parameters  $\alpha, \beta$ , run your code and check if it matches the theoretical results of question 1.
5. Give an estimate and a (symmetric) credible interval for the expected number of goals scored by FC Nantes.
6. Comment about the Ligue 1.

# Bioassay study

**Exercise 9.** To model the dose–response relation, i.e., how the probability of death is related to the dose  $x_i$ , we consider the Bayesian model:

$$Y_i \mid \theta_i \stackrel{\text{ind}}{\sim} \text{Binomial}(n_i, \theta_i),$$
$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)},$$

$$\theta_i = \Pr(\text{death} \mid x_i), \pi(\beta_0, \beta_1) \propto 1.$$

**Table 1:** *Bioassay data from Racine et al. (1986)*

Dose ( $x_i$ ) in log g / ml	Number ( $n_i$ ) of animals	Number ( $y_i$ ) of deaths
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

1. Write an MCMC sampler to sample from the posterior distribution of the above model and generate a (long enough) Markov chain.
2. In bioassay studies, a parameter of interest is the LD50, the dose level at which the probability of death is 50%. Based on your previous simulation, plot the posterior distribution of LD50.

# Survival analysis

**Table 2:** Motorette failure time. Right censored observations are marked with a +.

$(^{\circ}F)$	Failure time (hours)									
150	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+
170	1764	2772	3444	3542	3780	4860	5196	5448+	5448+	5448+
190	408	408	1344	1344	1440	1680+	1680+	1680+	1680+	1680+
220	408	408	504	504	504	528+	528+	528+	528+	528+

**Exercise 10.** Table 2 contains failure times  $y_{ij}$  from an accelerated life trial in which ten motorettes were tested at each of four temperatures, with the objective of predicting lifetime at  $130^{\circ}F$ . We analyse these data using a Weibull model with

$$\Pr(Y_{ij} \leq y \mid X = x_i) = 1 - \exp \left\{ - \left( \frac{y}{\theta_i} \right)^{\gamma} \right\}, \quad \theta_i = \exp(\beta_0 + \beta_1 x_i), \quad i = 1, \dots, 4, \quad j = 1, \dots, 10,$$

where failure time are in units of hundreds of hours and  $x_i = \ln(\text{temperature}/100)$ . We take independent priors on the parameters,  $N(0, 100)$  on  $\beta_0$  and  $\beta_1$  and exponential with mean 2 on  $\gamma$ .

## Motorette (following)

---

1. Write a Gibbs sampler for this model.
2. Analyze the generated Markov chain and comment any potential issues.
3. How would you predict the failure time when  $X = 130^\circ F$ ?

0. Introduction

1. Bayesian Refresher

1.5 Bayesian asymptotics

2. Intractable posterior

▷ 3. Hierarchical models

4. Finite mixture models

5. Approximate Bayesian Computation

## 3. Hierarchical models

# Motivations

---

- Data often depict different layers of variation, that one has to modelled:
  - success of surgical interventions may depend on patients (age/state of health) within surgeons (different experience/skill) within hospitals (different environments/skill of nursing staff)
  - student's marks may depend on the classroom, which depend on school, which depend on school districts. . .
- For each layer we actually observed draws from their respective population, e.g., patients/doctors drawn from a given hospital, schools drawn from a given school district.
- This suggest having different layer of randomness.
- Often motivated by the so-called concept of **borrowing strength**



# Bayesian hierarchical model

---

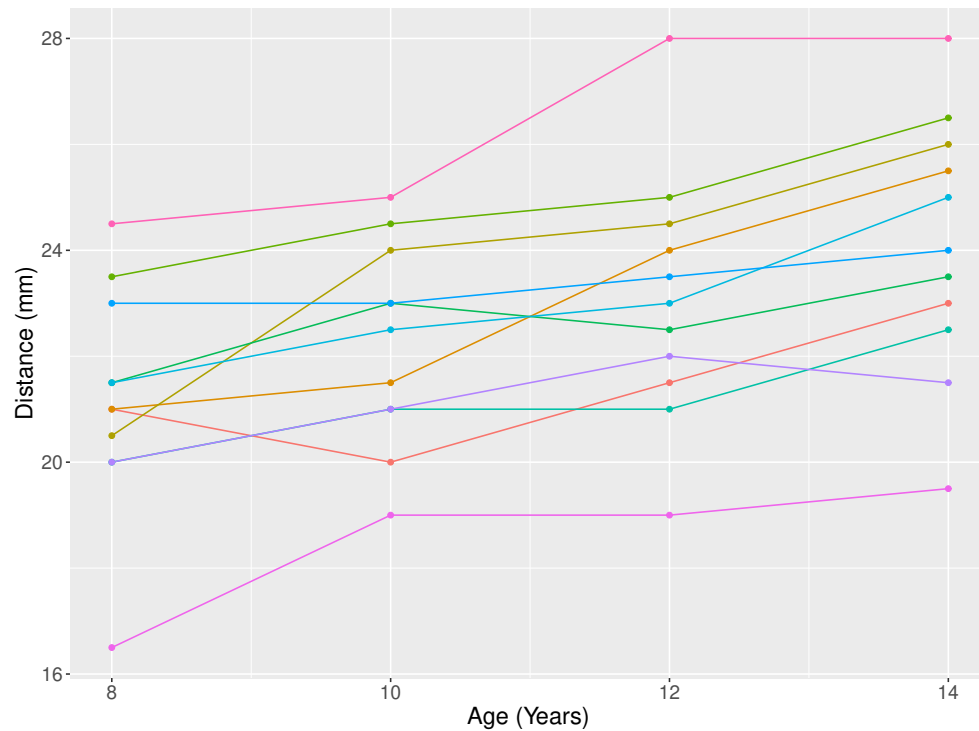
**Definition 11.** A statistical model  $\{f(x; \theta) : x \in \mathbb{R}^p, \theta \in \Theta\}$  is a **hierarchical model** if we have

$$f(x; \theta) = \int f_1(x | z_1) f_2(z_1 | z_2) \cdots f_d(z_{d-1} | z_d) f(z_d) \mathrm{d}z_1 \cdots \mathrm{d}z_d.$$

In the above expression, the  $z_j$ 's are called **latent variables**.

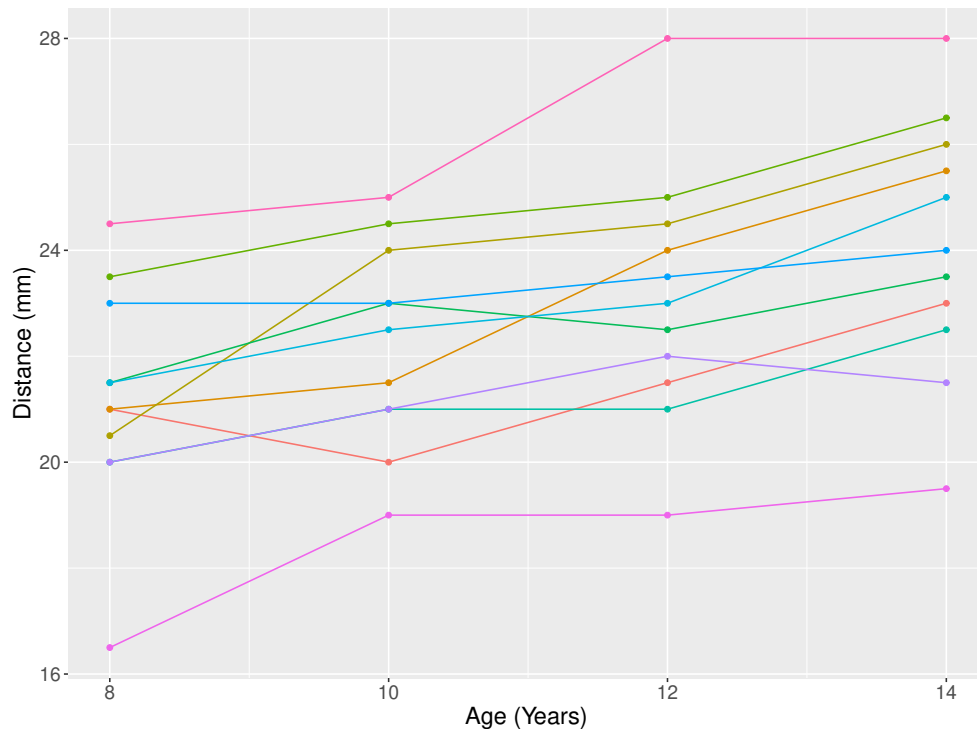
If in addition we put a prior distribution on  $\theta$  then we have a **Bayesian hierarchical model**.

**Example 2.** X-rays of the children's skulls were shot by orthodontists to measure the distance from the hypophysis to the pterygomaxillary fissure. Shots were taken every two years from 8 years of age until 14 years of age.



**Figure 12:** *The data collected by the orthodontists.*

**Example 2.** X-rays of the children's skulls were shot by orthodontists to measure the distance from the hypophysis to the pterygomaxillary fissure. Shots were taken every two years from 8 years of age until 14 years of age.



$$Y_{ij} \mid b_j \stackrel{\text{ind}}{\sim} N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$

$$b_j \sim N(0, \sigma_b^2),$$

- $Y_{ij}$ : distance
- $x_{ij}$ : age of subject  $j$  at index  $i$
- Bayesian: priors on  $\beta_1, \sigma_b^2, \sigma^2$ .

**Figure 12:** *The data collected by the orthodontists.*

# Graphs

---

**Definition 12.** A graph is a pair  $\mathcal{G} = (V, E)$  where:

- $V$  is a set whose elements are called **vertices**;
- $E$  is a subset of  $V \times V$  whose elements are called **edges**.

A graph  $G = (V, E)$  is said to be **directed** when edges are replaced by **arrows**<sup>3</sup>

# Graphs

---

**Definition 12.** A graph is a pair  $\mathcal{G} = (V, E)$  where:

- $V$  is a set whose elements are called **vertices**;
- $E$  is a subset of  $V \times V$  whose elements are called **edges**.

A graph  $G = (V, E)$  is said to be **directed** when edges are replaced by **arrows**<sup>3</sup>

- Why am I talking about graph in this lecture?

# Graphs

---

**Definition 12.** A graph is a pair  $\mathcal{G} = (V, E)$  where:

- $V$  is a set whose elements are called **vertices**;
- $E$  is a subset of  $V \times V$  whose elements are called **edges**.

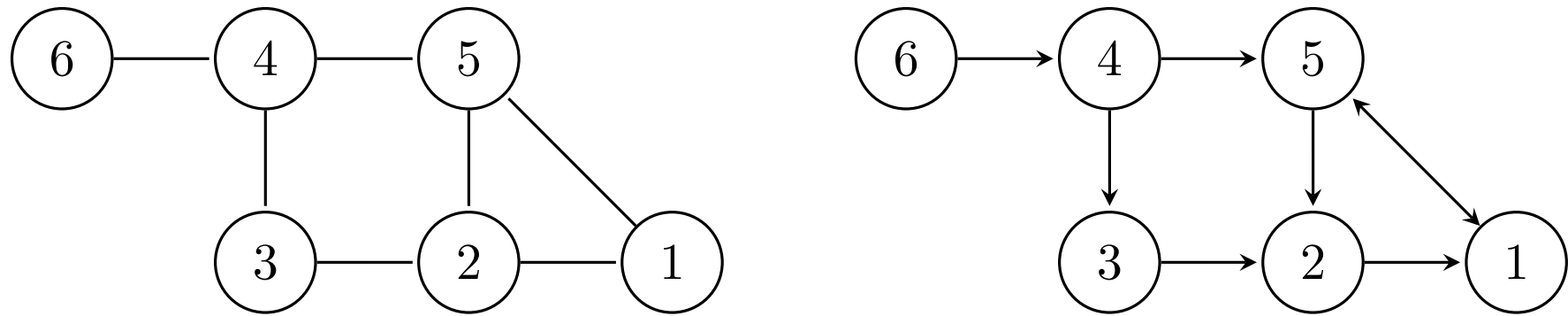
A graph  $G = (V, E)$  is said to be **directed** when edges are replaced by **arrows**<sup>3</sup>

- Why am I talking about graph in this lecture?
- Because you can represent statistical models as graphs: each node correspond to a random variable.
- Such a representation is called **(probabilistic) graphical model**.
- In this lecture we will mainly focus on models based on directed acyclic graphs.

---

<sup>3</sup>some people add the additional condition that you cannot have arrows on yourself, i.e., no loop.

# Example of graphs



**Figure 13:** *Example of two graphs. Left : (undirected) graph. Right: Directed graph.*

## Some vocabulary

---

**Definition 13.** Let  $G = (V, E)$  be a directed graph. For any  $i \in V$ , we define

- the **parents** of  $i$  as the set

$$\{j \in V : \text{there is an arrow from } j \text{ to } i\};$$

- the **child** of  $i$  as the set

$$\{j \in V : \text{there is an arrow from } i \text{ to } j\};$$

- the **descendants** of  $i$  as the set

$$\{j \in V : \text{there is a path of arrows from } i \text{ to } j\};$$

- the **non descendants** of  $i$  as the set

$$V \setminus \{\{i\} \cup \{\text{descendants of } i\}\}.$$



# Conditional independence

---

**Definition 14.** Let  $X, Y, Z$  be random variables . We say that  $X$  and  $Y$  are **conditionally independent** given  $Z$ , denoted  $X \perp Y \mid Z$ , if for all  $x, y, z$  we have

$$f(x, y \mid z) = f(x \mid z)f(y \mid z),$$

where  $f(\cdot \mid z)$  denotes the conditional density.

**Proposition 5.** *If  $X$  and  $Y$  are conditionally independent given  $Z$ , then  $f(x \mid y, z) = f(x \mid z)$ .*

*Proof.* Easy. Just write the definition of conditional density and simplify. □

# Directed acyclic graph (DAG)

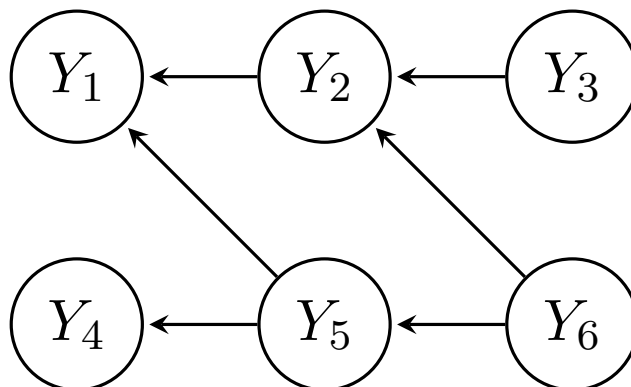
**Definition 15.** A **directed acyclic graph (DAG)** is a graphical model that represents a hierarchical dependence structure, i.e., for all  $i \in V$

$$Y_i \perp \text{non descendants of } Y_i \mid \text{parents of } Y_i.$$

It is **directed** because it is a directed graph and **acyclic** because it is impossible to start from a node and get back to it using a path of arrows.

**Example 3.** The hierarchical dependence structure

$f(y) = f(y_1 \mid y_2, y_5)f(y_2 \mid y_3, y_6)f(y_3)f(y_4 \mid y_5)f(y_5 \mid y_6)f(y_6)$  gives:



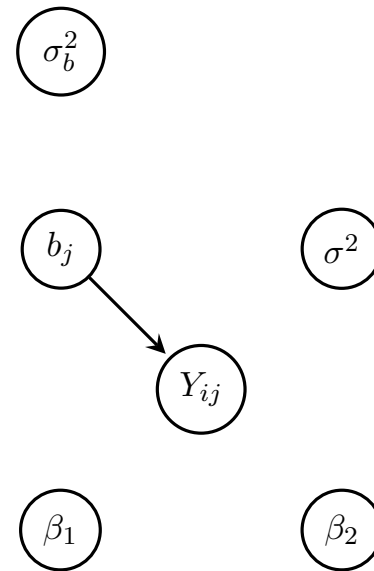
---

**Example 4.** Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid b_j, \sim N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$
$$b_j \sim N(0, \sigma_b^2),$$

**Example 4.** Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid b_j, \sim N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$
$$b_j \sim N(0, \sigma_b^2),$$

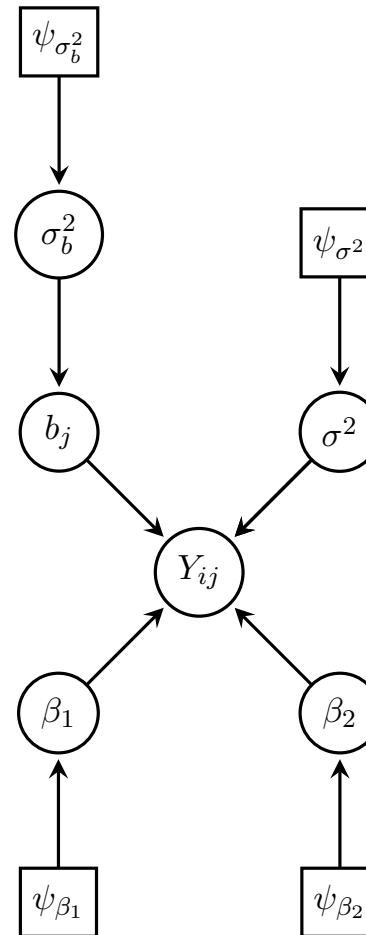


**Example 4.** Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid b_j, \sim N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$

$$b_j \sim N(0, \sigma_b^2),$$

And if we go Bayesian... (prior hyperparameters are denoted by squares)



# Factorization of a DAG and full conditional distributions

- Since, by definition, for any DAG  $G = (V, E)$  we have

$$f(y) = \prod_{j \in V} f(y_j \mid \text{parents of } y_j).$$

- Hence the full conditional distributions write

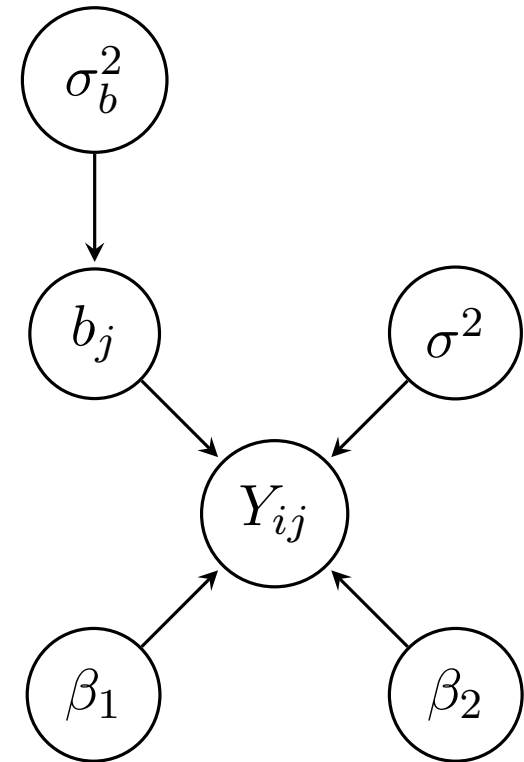
$$\begin{aligned} f(y_j \mid y_{-j}) &\propto f(y) \\ &\propto \prod_{i \in V} f(y_i \mid \text{parents of } y_i) \\ &\propto f(y_j \mid \text{parents of } y_j) \prod_{\substack{i \in V: \\ y_i \text{ child of } y_j}} f(y_i \mid \text{parents of } y_i). \end{aligned}$$

**Exercise 11.** Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid \beta_1, \beta_2, b_j, \sigma^2 \stackrel{\text{ind}}{\sim} N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$
$$b_j \sim N(0, \sigma_b^2),$$

with prior distribution

$$\pi(\theta) = \pi(\beta_1)\pi(\beta_2)\pi(\sigma_b^2)\pi(\sigma^2).$$



Derive the full conditional distributions required for a Gibbs sampler.

# Latent Dirichlet Allocation

The **Latent Dirichlet Allocation (LDA)** is a stochastic model on the structure of text documents. Let  $Y_{d,n}$  be the  $n$ -th word in the  $d$ -th document. The model writes

$$\begin{aligned} Y_{d,n} &| \Phi, Z_{d,n} \stackrel{\text{ind}}{\sim} \text{Discrete}(\Phi_{Z_{d,n}}), & d = 1, \dots, D, \quad n = 1, \dots, N_d \\ Z_{d,n} &| \theta_d \stackrel{\text{ind}}{\sim} \text{Discrete}(\theta_d), & d = 1, \dots, D, \quad n = 1, \dots, N_d \\ \theta_d &| \alpha \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\alpha), & d = 1, \dots, D \\ \Phi_t &| \beta \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\beta), & t = 1, \dots, T. \end{aligned}$$

In the above model,

- $Z_{.,.}$  are latent variables identifying the **theme** of each word;
- $\theta_d$  is the document signature, i.e., a discrete distribution on possible themes, for document  $d$ ;
- $\Phi_t$  is the theme signature, i.e., a discrete distribution on the **vocabulary**, for theme  $t$ .



## Latent Dirichlet Allocation (following)

- Exercise 12.** 1. Give the full conditional distributions for this model.
2. If you were to write a Gibbs sampler based on your previous result, this wouldn't scale well for big data. Hence a **collapsed Gibbs sampler**, i.e., marginalizing the posterior w.r.t.  $\theta$  and  $\Phi$ , is often used. One can show that

$$\pi(\mathbf{Z} \mid \mathbf{y}) \propto \prod_{d=1}^D B(\alpha + n_{d,\cdot}) \prod_{t=1}^T B(\beta + n_{\cdot,t,\cdot}),$$

where  $B$  is the multivariate Beta function.

Give the full conditional distribution for this collapsed Gibbs sampler.

*Remark.* You might first want to show that for  $e_j$  the  $j$ -th vector of the canonical basis of  $\mathbb{R}^d$ , we have for all  $x \in \mathbb{R}^d$ ,

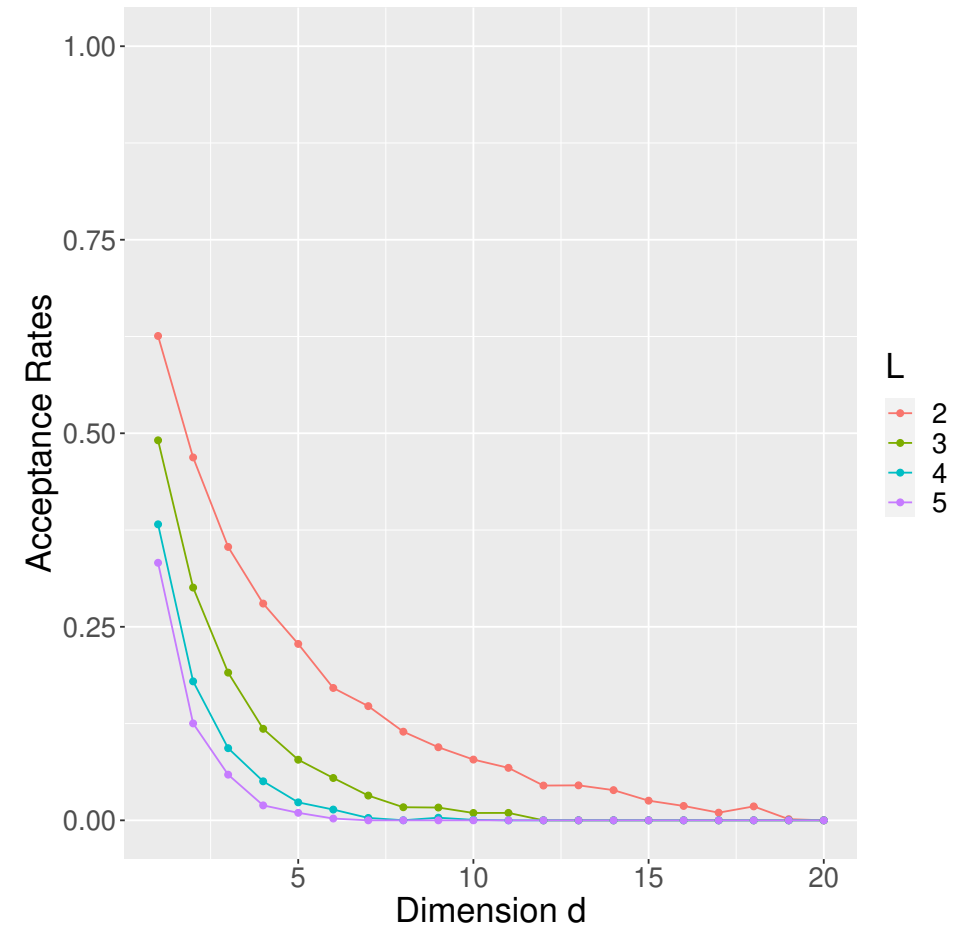
$$B(x + e_j) = \frac{x_j}{\sum_{i=1}^d x_i} B(x),$$

and then use this result to simplify those full conditional distributions.

# Coagulation time

**Table 3:** *Coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets. Data were rounded but we ignore this problem here.*

Diet	Measurements
A	62, 60, 63, 59
B	63, 67, 71, 64, 65, 66
C	68, 66, 71, 67, 68, 68
D	56, 62, 60, 61, 63, 64, 63, 59



**Figure 14:** *The coagulation time data set.*

## Coagulation time (following)

---

**Exercise 13.** A simple model for the blood data is a **one-way layout**, where we suppose there are two levels of variation. First, each individual has a mean  $\theta_t$  which is measured with error on each occasion, so that

$$Y_{ij} \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j, \quad j = 1, \dots, J.$$

Secondly, we suppose that each mean  $\theta_j$  is drawn from a distribution of means, corresponding to the members of the population from which the six individuals were drawn, so that  $\theta_j \stackrel{\text{iid}}{\sim} N(\mu, \sigma_\theta^2)$ .

For Bayesian modelling we need prior densities for  $\mu$ ,  $\sigma^2$  and  $\sigma_\theta^2$  and we use

$$\mu \sim N(\mu_0, \tau^2), \quad \sigma^2 \sim \text{InverseGamma}(\alpha, \beta), \quad \sigma_\theta^2 \sim \text{InverseGamma}(\alpha_\theta, \beta_\theta).$$

1. Find the corresponding DAG.
2. Give an MCMC algorithm to sample from the posterior distribution.

# What is the likelihood in Bayesian hierarchical models?

---

- Consider the (one layer) Bayesian hierarchical model where

$$f(y, \psi, \theta) = f(y | \psi)f(\psi | \theta)\pi(\theta)$$

- Depending on the aim of the study the likelihood might be

$f(y | \psi)$  if focus is on  $\psi$ .

$f(y | \theta)$  if focus is on  $\theta$ .

 Hence the meaning of predictions has to be set before doing inference!

# Model selection for hierarchical models

- Information criteria are numerical quantities that help us in identifying the “best” model from a bunch of candidates, say  $\mathcal{M}_1, \dots, \mathcal{M}_k$ .
- Most often the lower, the better.
- You probably (I hope!) already now two of them

$$AIC(\mathcal{M}) = D(\hat{\theta}) + 2k, \quad k = \text{number of parameters in } \theta$$

$$BIC(\mathcal{M}) = D(\hat{\theta}) + k \log n, \quad n = \text{sample size,}$$

where  $D(\theta) = -2 \log f(y | \theta)$  and is known as the **deviance**.

👉 Both AIC / BIC put emphasis on predicting from  $f(y | \theta)$ , i.e., from the top layer. BIC differs to AIC and aims to identify the “true model” as  $n \rightarrow \infty$ .

# Deviance information criterion

**Definition 16.** The **Deviance Information Criterion (DIC)** of a Bayesian hierarchical model is

$$DIC(\mathcal{M}) = D(\theta) + 2p_{\text{eff}}, \quad p_{\text{eff}} = \mathbb{E}_{\pi}[D(\theta) | Y] - D(\hat{\theta}), \quad \hat{\theta} = \mathbb{E}_{\pi}[\theta | Y].$$

The quantity  $p_{\text{eff}}$  is known as the **effective number of parameters**.

- Given a Markov chain  $\{\theta_t : t = 1, \dots, T\}$  drawn from the posterior, we can estimate

$$p_{\text{eff}} = \frac{1}{T} \sum_{t=1}^T D(\theta_t) - D(\hat{\theta}) = \bar{D}(\theta) - D(\hat{\theta}),$$

and the DIC as

$$DIC(\mathcal{M}) = D(\hat{\theta}) + 2p_{\text{eff}}.$$

-  DIC puts emphasis on predicting from  $f(y | \psi)$ , i.e., from the bottom layer.

# Bayes Factor

---

TALK ABOUT BAYES FACTOR

# Illustration

---

- Recall our coagulation time model

$$Y_{ij} \mid \theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j, \quad j = 1, \dots, J$$
$$\theta_j \sim N(\mu, \sigma_\theta^2)$$

with prior distribution  $\pi(\mu, \sigma^2, \sigma_\theta^2) = \pi(\mu)\pi(\sigma^2)\pi(\sigma_\theta^2)$ .

- If interest is in predicting the coagulation time for
  - future animals in those diets then use DIC
  - a future animal following a random diet then use AIC / BIC.



0. Introduction

1. Bayesian Refresher

1.5 Bayesian asymptotics

2. Intractable posterior

3. Hierarchical models

▷ 4. Finite mixture models

5. Approximate Bayesian Computation

## 4. Finite mixture models

# Finite mixture models

---

**Definition 17.** A continuous random variable  $X$  is said to follow a **finite mixture model** if  $X$  has density

$$f(x; \psi) = \sum_{k=1}^K \omega_k f_k(x; \theta_k),$$

where  $\omega_k \geq 0$ ,  $\sum_{k=1}^K \omega_k = 1$ ,  $f_k$  are p.d.f. (typically within the same family) and  $\theta = (\omega, \boldsymbol{\theta})$ ,  $\omega = (\omega_1, \dots, \omega_K)$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ .

**Example 5.** A (two component) **Gaussian mixture** is given by

$$f(x; \theta) = \omega_1 \varphi(x; \mu_1, \Sigma_1) + (1 - \omega_1) \varphi(x; \mu_2, \Sigma_2).$$

# Likelihood for finite mixture models

---

- Suppose we have  $n$  independent observations  $x_1, \dots, x_n$  from the mixture model

$$f(x; \psi) = \sum_{k=1}^K \omega_k f_k(x; \theta_k).$$

- The likelihood is thus

$$L(\psi) = \prod_{i=1}^n \sum_{k=1}^K \omega_k f_k(x_i; \theta_k),$$

which shows  $K^n$  terms and yield to intractable likelihood (too CPU demanding).

# Likelihood for finite mixture models

- Suppose we have  $n$  independent observations  $x_1, \dots, x_n$  from the mixture model

$$f(x; \psi) = \sum_{k=1}^K \omega_k f_k(x; \theta_k).$$

- The likelihood is thus

$$L(\psi) = \prod_{i=1}^n \sum_{k=1}^K \omega_k f_k(x_i; \theta_k),$$

which shows  $K^n$  terms and yield to intractable likelihood (too CPU demanding).



We need an alternative to be able to estimate mixture models.

# Incomplete–data formalism

---

- A common practice with finite mixture model is to adopt the **incomplete–data** point of view.
- For each observation  $X_i$ , we associate a **latent variable**  $Z_i \in \{1, \dots, K\}$  specifying the **class** of  $X_i$ .
- The mixture model thus writes

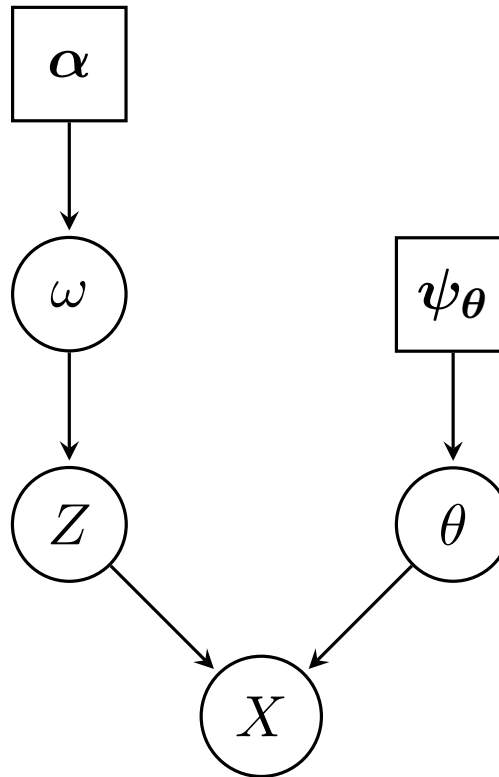
$$\begin{aligned} X_i \mid Z_i, \theta &\sim f_{Z_i}(\cdot \mid \theta_{Z_i}) \\ Z_i \mid \omega &\sim \text{Discrete}(\omega). \end{aligned}$$

- The **completed likelihood** based on  $(\mathbf{x}, \mathbf{z})$  thus writes

$$L(\psi) = \prod_{i=1}^n \omega_{z_i} f_{z_i}(x_i; \theta_{z_i}),$$

which now shows only  $n$  terms to compute (as usual).

# DAG of the incomplete-data formalism



---

**Exercise 14.** Write an R / Python code to estimate the posterior distribution from the following Gaussian mixture model:

$$f(x) = \sum_{k=1}^K \omega_k \varphi(x; \mu_k, \sigma_k^2), \quad K \text{ known,}$$

where  $\varphi(\cdot; \mu, \sigma^2)$  denotes the p.d.f. of the Gaussian random variable with mean  $\mu$  and variance  $\sigma^2$ .

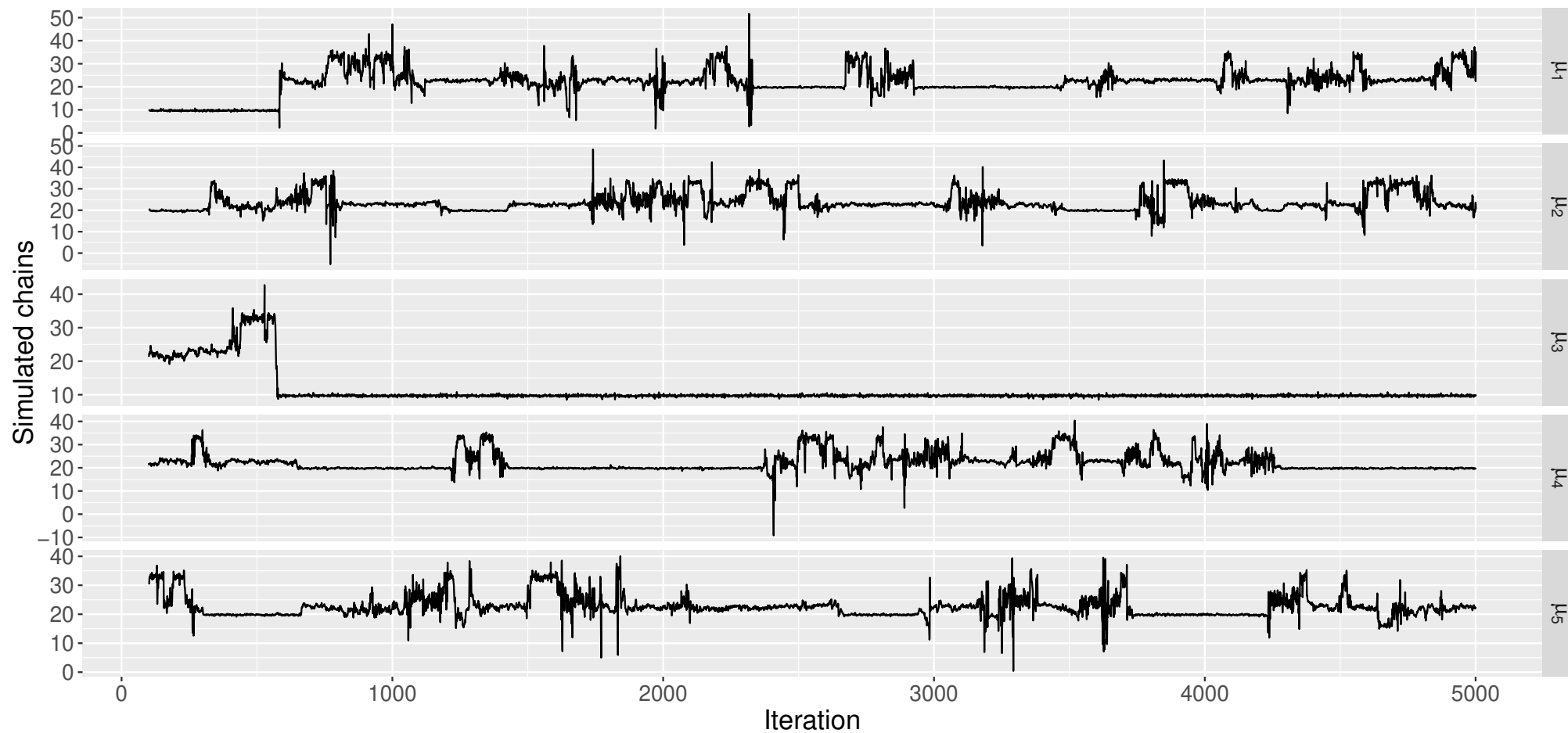
We will assume independent priors for  $\omega$ ,  $\mu_k$  and  $\sigma_k^2$ , i.e.,

$$\omega \sim \text{Dirichlet}(1, \dots, 1)$$

$$\mu_k \sim N(20, 100)$$

$$\sigma_k^2 \sim \text{InverseGamma}(0.1, 0.1).$$

Test your code on the galaxies dataset (available from the MASS R package) with  $K = 6$  and comment.



**Figure 15:** *A typical trace plot of the output of a Gibbs sampler on a mixture model.*



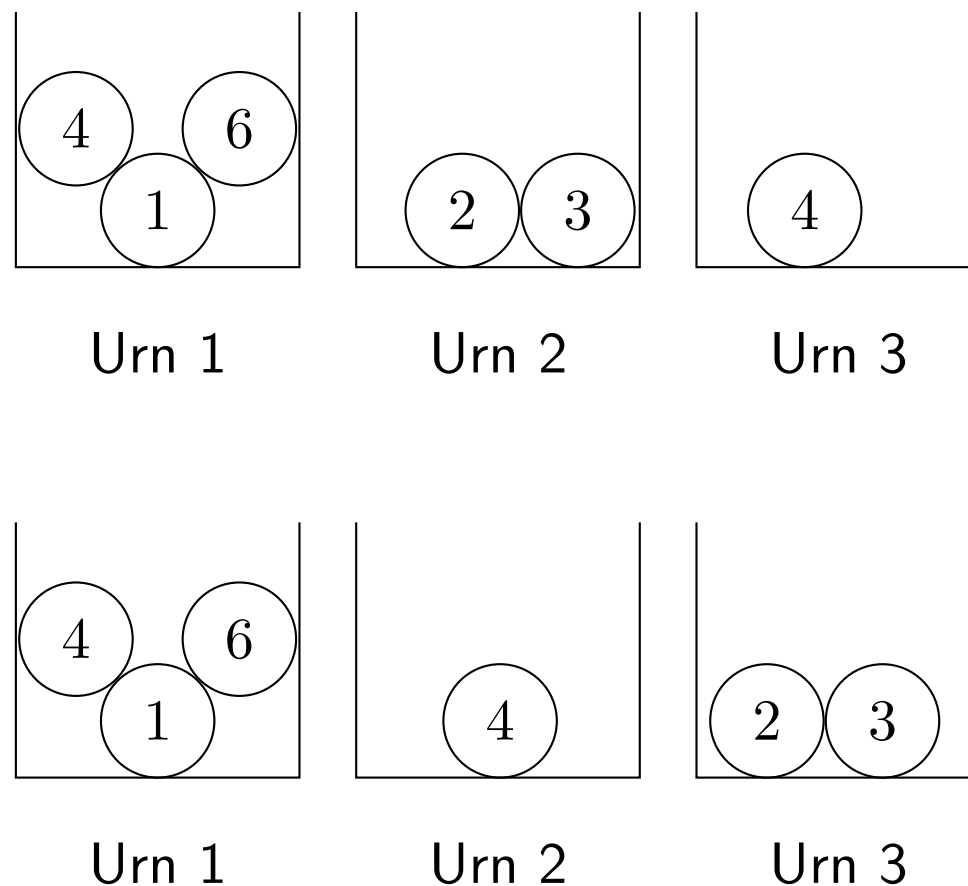
# Label switching

---

- What we just experienced is called **label switching**.

**Definition 18.** Consider a parametric statistical model  $\{f(x; \theta) : x \in E, \theta \in \Theta\}$ . We say that the model is **identifiable** if, for all  $x \in E$ , the mapping  $\theta \mapsto f(x; \theta)$  is one–one.

- Every mixture model is by essence non identifiable, e.g., think about balls that we put in urns or wait for the next slide.
- For a  $K$  component mixture, there are (at least)  $K!$  points where the likelihood is the same.



**Figure 16:** *Label switching for mixture models. Each urn corresponds to a given class and each ball correspond to latent variable associated to an observation  $X_i$  (for instance if ball 1 is in urn 2  $\Leftrightarrow Z_1 = 2$ ). Hence the two rows gives the same mixture.*

# Dealing with label switching

---

- Although controversial, one common way to bypass this hurdle is to add some *dependence across parameters*
- One widely used option is to further assume that

$$\mu_1 \leq \mu_2 \leq \dots \leq \mu_K.$$

- Obviously one can do exactly the same with the class probabilities  $\omega_k$  or variances  $\sigma_k^2$ .
- While coding such an ordering correspond to add an extra step at each iteration of your MCMC sampler where you *reorganize your data to meet your additional constraint*

---

## Algorithm 5: Gibbs sampler for mixture model (with weight ordering).

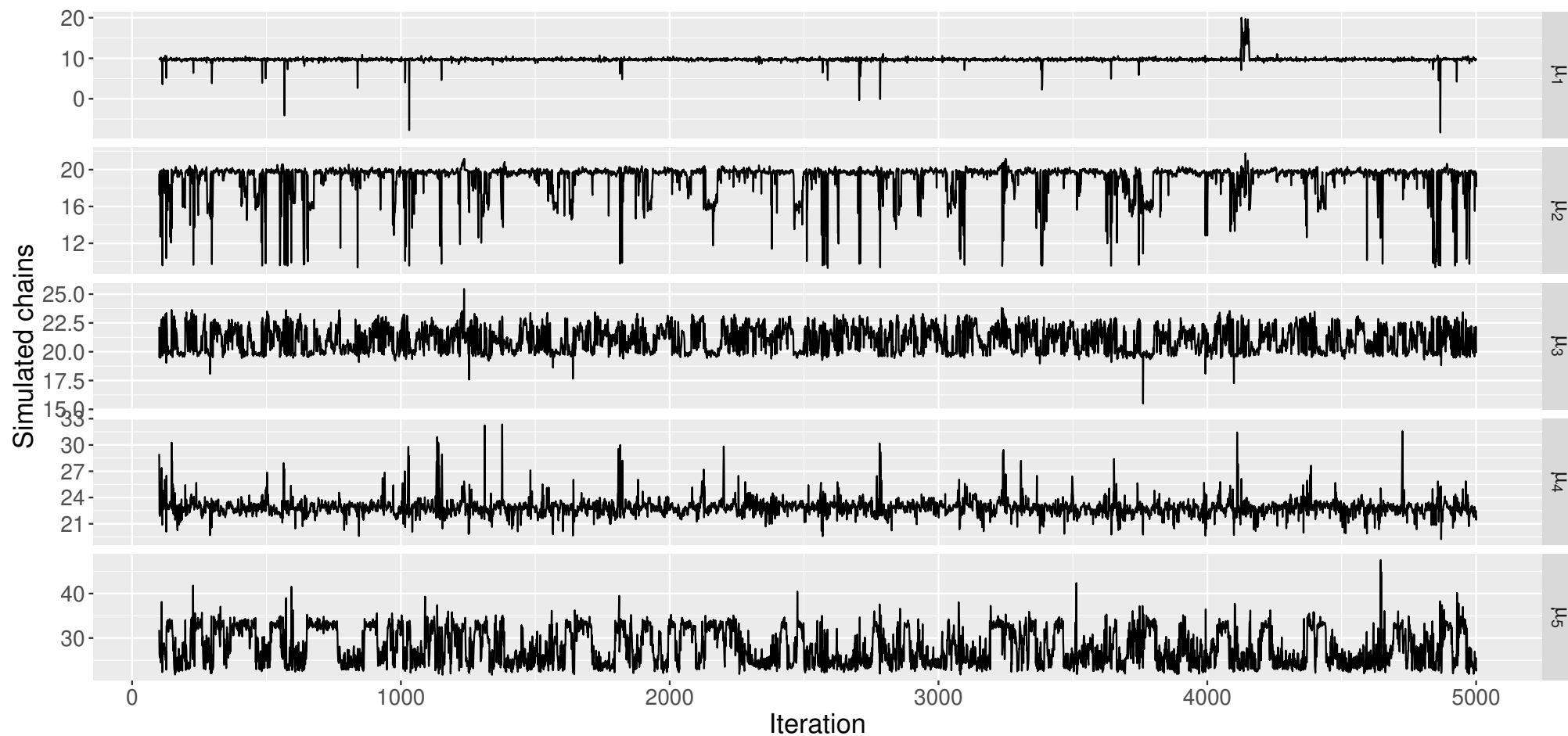
---

**input** : A finite mixture model  $f(x; \psi) = \sum_{k=1}^K \omega_k f_k(x; \theta_k)$ , initial state  $\theta_0$ , some data  $x_1, \dots, x_n$ .

**output**: A Markov chain whose stationary distribution is  $\pi(\theta | x)$ .

```
1 for  $t \leftarrow 1$  to  $N$  do
  /* Update the latent variable */
2   for  $i \leftarrow 1$  to  $n$  do
3      $z_{t,i} \sim \pi(z_i | \dots)$ ;
  /* Update the mixture weights  $\omega_k$  */
4   for  $k \leftarrow 1$  to  $K$  do
5      $\omega_{t,k} \sim \pi(\omega_k | \dots)$ ;
  /* Update the  $\theta_k$  parameters */
6   for  $k \leftarrow 1$  to  $K$  do
7      $\theta_{t,k} \sim \pi(\theta_k | \dots)$ ;
  /* Reordering to (hopefully) mitigate the label switching issue */
8 for  $t \leftarrow 0$  to  $N$  do
9   Reorder  $\mu_t$  into increasing order and  $\theta_t$  accordingly;
10 Return the Markov chain  $\{(\omega_t, \theta_t) : t = 0, \dots, N\}$ ;
```

---



**Figure 17:** *Applying reordering to our Gibbs sampler.*

- 
- Note that other techniques were developed to bypass the label switching problem.
  - The above reordering procedure has the advantage of being simple but has severe limitations.
  - Talking about these different approaches is beyond the scope of this lecture but should be preferred!

0. Introduction

---

1. Bayesian Refresher

---

1.5 Bayesian  
asymptotics

---

2. Intractable  
posterior

---

3. Hierarchical  
models

---

4. Finite mixture  
models

---

5. Approximate  
Bayesian  
▷ Computation

---

# 5. Approximate Bayesian Computation

# Motivation

---

- The posterior distribution is given by

$$\pi(\theta | y) \propto f(y | \theta)\pi(\theta)$$

- As a consequence all the MCMC samplers introduced so far require that we are able to compute the likelihood  $f(y | \theta)$ .
- Suppose it is not possible because the likelihood
  - is too CPU demanding
  - has no closed form
- Usual MCMC algorithm are thus useless
- Is it possible to still derive useful algorithms?
- The answer is **yes** and belongs on **likelihood free approaches**
- Recall that our goal is to sample from the posterior distribution



# Motivation

---

- The posterior distribution is given by

$$\pi(\theta | y) \propto f(y | \theta)\pi(\theta)$$

- As a consequence all the MCMC samplers introduced so far require that we are able to compute the likelihood  $f(y | \theta)$ .
- Suppose it is not possible because the likelihood
  - is too CPU demanding
  - has no closed form
- Usual MCMC algorithm are thus useless
- Is it possible to still derive useful algorithms?
- The answer is **yes** and belongs on **likelihood free approaches**
- Recall that our goal is to sample from the posterior distribution

 Likelihood free approaches substitute the evaluation of the likelihood for simulation from the model.

# A first algorithm

---

**Algorithm 6:** Likelihood free rejection sampling.

---

**input** : A Bayesian statistical model  $\{f(\cdot | \theta), \pi(\cdot)\}$  and sample size  $N$ .

**output:** A (independent) sample from the posterior  $\pi(\theta | y)$ .

```
1  $t \leftarrow 1$ ;  
2 while  $t < N$  do  
3   Draw a proposal parameter  $\theta_*$  from the prior distribution  $\pi(\theta)$ ;  
4   Simulate synthetic data  $y_*$  from  $f(\cdot | \theta_*)$ ;  
5   if  $y_* = y$  then  
6     Set  $\theta_t \leftarrow \theta_*$ ;  
7      $t \leftarrow t + 1$ ;  
8 Return  $\{\theta_t : t = 0, \dots, N\}$ ;
```

---

# A first algorithm

---

**Algorithm 6:** Likelihood free rejection sampling.

---

**input** : A Bayesian statistical model  $\{f(\cdot | \theta), \pi(\cdot)\}$  and sample size  $N$ .

**output:** A (independent) sample from the posterior  $\pi(\theta | y)$ .

```
1  $t \leftarrow 1$ ;  
2 while  $t < N$  do  
3   Draw a proposal parameter  $\theta_*$  from the prior distribution  $\pi(\theta)$ ;  
4   Simulate synthetic data  $y_*$  from  $f(\cdot | \theta_*)$ ;  
5   if  $y_* = y$  then  
6     Set  $\theta_t \leftarrow \theta_*$ ;  
7      $t \leftarrow t + 1$ ;  
8 Return  $\{\theta_t : t = 0, \dots, N\}$ ;
```

---

 The above algorithm is indeed likelihood-free since it does not require to compute the likelihood.

---

**Proposition 6.** *The generated sample is i.i.d. from  $\pi(\theta \mid y)$ .*

*Proof.* Sketch: To get the distribution of  $\theta_t$  start with the joint distribution of  $(\theta_t, y_*)$  and marginalize w.r.t.  $y_*$ . □

# Curse of dimensionality again and again

---

- In the above algorithm we accept proposal  $\theta_*$  iff  $y_* = y$
- Clearly if  $f(y | \theta)$  is a **continuous model** it occurs with probability 0
- For discrete model the probability is positive but decreases quickly to 0 as:
  - the number of levels gets bigger
  - the sample size  $n$ ,  $y = (y_1, \dots, y_n)$ , increases.

# Curse of dimensionality again and again

---

- In the above algorithm we accept proposal  $\theta_*$  iff  $y_* = y$
- Clearly if  $f(y | \theta)$  is a **continuous model** it occurs with probability 0
- For discrete model the probability is positive but decreases quickly to 0 as:
  - the number of levels gets bigger
  - the sample size  $n$ ,  $y = (y_1, \dots, y_n)$ , increases.

 We need to relax this constraint...

# Mitigate the curse of dimensionality

---

**Exercise 15.** Consider a Bernoulli model with  $p = 1/4$  with prior distribution  $U(0, 1)$ . Ignoring that  $p = 1/4$ , we currently have  $p_* = 3/4$ .

1. Having observed  $y = 1$ , what is the probability to accept  $p_*$ ?
2. Having observed  $y = (1, 0, 0, 1, 0, 0, 0)$ , what is the probability to accept  $p_*$ ?

# Mitigate the curse of dimensionality

---

**Exercise 15.** Consider a Bernoulli model with  $p = 1/4$  with prior distribution  $U(0, 1)$ . Ignoring that  $p = 1/4$ , we currently have  $p_* = 3/4$ .

1. Having observed  $y = 1$ , what is the probability to accept  $p_*$ ?
2. Having observed  $y = (1, 0, 0, 1, 0, 0, 0)$ , what is the probability to accept  $p_*$ ?

- To mitigate the curse of dimensionality, we substitute  $y_* = y$  for  $\|T(y_*) - T(y)\| < \epsilon$  where
- $T(\cdot)$  is a summary statistic (possibly multivariate);
  - $\|\cdot\|$  is any (pseudo) norm.



---

**Algorithm 7:** ABC rejection sampling.

---

**input** : A Bayesian statistical model  $\{f(\cdot | \theta), \pi(\cdot)\}$ , sample size  $N$ , tolerance value  $\varepsilon$ , summary statistic  $T$  and (pseudo) norm  $\|\cdot\|$

**output:** A (independent) sample approximately drawn from  $\pi(\theta | y)$ .

```
1  $t \leftarrow 1$ ;  
2 while  $t < N$  do  
3   Draw a proposal parameter  $\theta_*$  from the prior distribution  $\pi(\theta)$ ;  
4   Simulate synthetic data  $y_*$  from  $f(\cdot | \theta_*)$ ;  
5   if  $\|T(y_*) - T(y)\| < \varepsilon$  then  
6     Set  $\theta_t \leftarrow \theta_*$ ;  
7      $t \leftarrow t + 1$ ;  
8 Return  $\{\theta_t : t = 1, \dots, N\}$ ;
```

---

**Proposition 7.** *The distribution of  $\theta_t$  is*

$$\pi_{T,\varepsilon}(\theta_* \mid y) \propto \int f(y_* \mid \theta_*) \pi(\theta_*) 1_{\{\|T(y_*) - T(y)\| < \varepsilon\}} dy_*.$$

*Proof.* As the previous one!

□

**Proposition 7.** *The distribution of  $\theta_t$  is*

$$\pi_{T,\varepsilon}(\theta_* | y) \propto \int f(y_* | \theta_*) \pi(\theta_*) 1_{\{\|T(y_*) - T(y)\| < \varepsilon\}} dy_*.$$

*Proof.* As the previous one! □

□ Clearly we have two limiting cases:

- as  $\varepsilon \rightarrow 0$ ,  $\pi_{T,\varepsilon}(\theta | y) \rightarrow \pi(\theta | y)$  if  $T$  is such that  $T(x) = T(y) \Rightarrow x = y$
- as  $\varepsilon \rightarrow \infty$ ,  $\pi_{T,\varepsilon}(\theta | y) \rightarrow \pi(\theta)$

**Proposition 7.** *The distribution of  $\theta_t$  is*

$$\pi_{T,\varepsilon}(\theta_* | y) \propto \int f(y_* | \theta_*) \pi(\theta_*) 1_{\{\|T(y_*) - T(y)\| < \varepsilon\}} dy_*.$$

*Proof.* As the previous one! □

□ Clearly we have two limiting cases:

- as  $\varepsilon \rightarrow 0$ ,  $\pi_{T,\varepsilon}(\theta | y) \rightarrow \pi(\theta | y)$  if  $T$  is such that  $T(x) = T(y) \Rightarrow x = y$
- as  $\varepsilon \rightarrow \infty$ ,  $\pi_{T,\varepsilon}(\theta | y) \rightarrow \pi(\theta)$

 Taking  $\varepsilon$  too large is useless!

# Application

---

Consider the following Bayesian statistical model

$$\begin{aligned} Y \mid \mu &\sim N(\mu, 1) \\ \mu &\sim N(0, 4) \end{aligned}$$

Note that, due to the use of conjugate prior, the posterior is known exactly and is  $N(\tilde{\mu}, \tilde{\sigma}^2)$  where

$$\tilde{\sigma}^2 = \left( \frac{1}{4} + \frac{n}{1} \right)^{-1}, \quad \tilde{\mu} = \tilde{\sigma}^2 \frac{\sum_{i=1}^n Y_i}{1}.$$

For the numerical application we set  $n = 50$ ,  $\mu = 1$  and  $\varepsilon = 0.025$ .

**Figure 18:** Comparison of the posterior distribution obtained using ABC rejection sampling with the true posterior distribution.

---

**Exercise 16.** Implement an ABC rejection sampling for the above Bayesian model where  $\|\cdot\| = \|\cdot\|_2$  and the summary statistics is

- the sample mean
- the sample median
- the sample standard deviation
- the bivariate vector (sample mean, sample vector)

Comment your results. What is expected?

# Why moving to Markov chains?

---

- The above algorithm generates an **independent sample** from  $\pi(\theta | y)$
- But it is **inefficient** since it does not use information of **accepted proposal**  $\theta_*$
- Clearly one can benefit from accepted proposal  $\theta_*$  using random perturbations around  $\theta_*$

# Why moving to Markov chains?

---

- The above algorithm generates an **independent sample** from  $\pi(\theta | y)$
- But it is **inefficient** since it does not use information of **accepted proposal**  $\theta_*$
- Clearly one can benefit from accepted proposal  $\theta_*$  using random perturbations around  $\theta_*$

 MCMC algorithms were exactly defined for this purpose!



---

**Algorithm 8:** ABC–MCMC algorithm.

---

**input** : A Bayesian statistical model  $\{f(\cdot | \theta), \pi(\cdot)\}$ , sample size  $N$ , tolerance value  $\varepsilon$ , summary statistic  $T$  and (pseudo) norm  $\|\cdot\|$ , a proposal kernel  $K(\cdot, \cdot)$  and initial simulated data  $\tilde{y}_0$  such that  $\|T(\tilde{y}_0) - T(y)\| < \varepsilon$ .

**output:** A Markov chain whose stationary distribution is  $\pi_\varepsilon(\theta | y)$ .

1 **for**  $t \leftarrow 1$  **to**  $N$  **do**

2     Draw a **proposal parameter**  $\theta_*$  from the proposal kernel  $K(\theta_{t-1}, \cdot)$ ;

3     Simulate synthetic data  $\tilde{y}_*$  from  $f(\cdot | \theta_*)$ ;

4     Compute the acceptance probability

$$\alpha(\theta_{t-1}, \theta_*) = \min \left\{ 1, \frac{\pi(\theta_*) f(\tilde{y}_* | \theta_*) 1_{\{\|T(\tilde{y}_*) - T(y)\| < \varepsilon\}} K(\theta_*, \theta_{t-1}) f(\tilde{y}_{t-1} | \theta_{t-1})}{\pi(\theta_{t-1}) f(\tilde{y}_{t-1} | \theta_{t-1,*}) K(\theta_*, \theta_{t-1}) f(\tilde{y}_* | \theta_*)} \right\}$$

5     Set

$$(\theta_t, \tilde{y}_t) = \begin{cases} (\theta_*, \tilde{y}_*), & \text{with probability } \alpha(\theta_{t-1}, \theta_*) \\ (\theta_{t-1}, \tilde{y}_{t-1}), & \text{with probability } 1 - \alpha(\theta_{t-1}, \theta_*) \end{cases}$$

6 **Return**  $\{\theta_t : t = 0, \dots, N\}$ ;


## Degree of approximation

---

- Recall that we are actually sampling from  $\pi_{T,\varepsilon}(\theta \mid y)$  rather than  $\pi(\theta \mid y)$  and the amount of observation depends on
  - the summary statistics  $T$
  - the tolerance value  $\varepsilon$
- Choosing relevant  $T$  is application dependent
- One can easily play with  $\varepsilon$ .
- One strategy is to use **adaptive threshold values**  $\varepsilon$ , i.e.,  $\varepsilon$  is now  $\varepsilon_t \downarrow 0$ .
- However keep in mind that if  $\varepsilon_t$  decreases too
  - slowly** the sampler is inefficient as we mainly sample from  $\pi(\theta)$
  - quickly** we may get stuck in some specific region.
- A rule of thumb is to store as well the divergences  $\|T(y_*) - T(y)\|$  and compute an empirical quantile of fixed order and update  $\varepsilon$  each  $K$  iterations.

## Degree of approximation

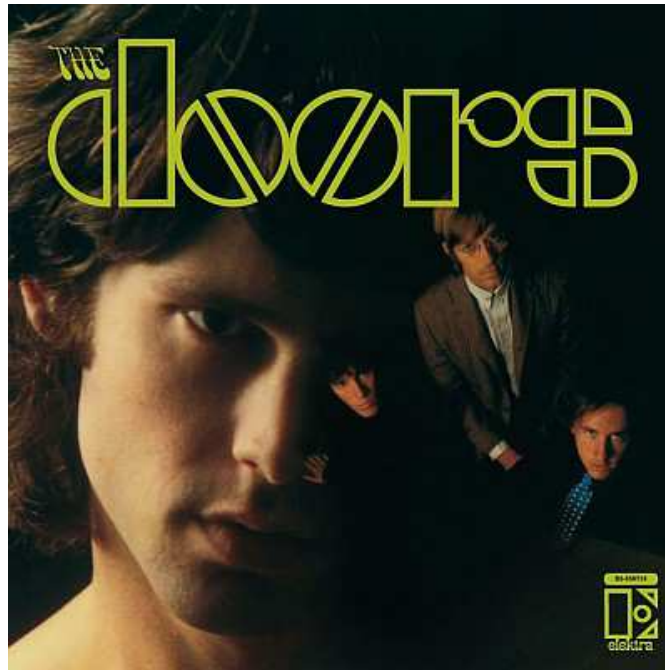
- Recall that we are actually sampling from  $\pi_{T,\varepsilon}(\theta | y)$  rather than  $\pi(\theta | y)$  and the amount of observation depends on
  - the summary statistics  $T$
  - the tolerance value  $\varepsilon$
- Choosing relevant  $T$  is application dependent
- One can easily play with  $\varepsilon$ .
- One strategy is to use **adaptive threshold values**  $\varepsilon$ , i.e.,  $\varepsilon$  is now  $\varepsilon_t \downarrow 0$ .
- However keep in mind that if  $\varepsilon_t$  decreases too
  - slowly** the sampler is inefficient as we mainly sample from  $\pi(\theta)$
  - quickly** we may get stuck in some specific region.
- A rule of thumb is to store as well the divergences  $\|T(y_*) - T(y)\|$  and compute an empirical quantile of fixed order and update  $\varepsilon$  each  $K$  iterations.

 Contrary to conventional MCMC algorithms, we usually target an acceptance probability around 1%. (Recall high acceptance rates induce sampling from the prior)

---

**Exercise 17.** Implement an ABC sampler on the FC Nantes scoring abilities and compare your results with those already obtained.

**Exercise 18.** Implement an ABC sampler on the Bioassay study and compare your results with those already obtained.



THIS IS THE END...