

# BSM1 - Welcome back Bayes

---

Lecturer: Riccardo Corradin

# Introduction

Welcome to Bayesian Statistical Models!

Before starting, I gratefully acknowledge Alessandra Guglielmi and Tommaso Rigon. Part of the material presented in this module is inspired by their lecture notes and examples.

---

- This module is about **models**. Models are one of the fundamental tools of a statistician toolbox.
- Ideally, a model is nothing but a mechanism for reasoning about the world, an (almost) objective way to describe scientific, economic, environmental, social, astronomical, etc, phenomena.
- In this module we will explore the **construction, properties, inferential procedures** and **summaries** of data analysis with **Bayesian models**.
- The material is mainly composed by three components:
  - slides, containing the methodological part;
  - code, usually with synthetic examples;
  - case studies, presenting real data analysis.

# Introduction

- The declination of a model itself depends on specific context we are working on, but we mainly distinguish among two classes: **deterministic** and **probabilistic** models.
  - The first is approximating the whole reality without any uncertainty or stochastic error.
  - the second class of models involves stochastic terms which introduce uncertainty and randomness.
- Probabilistic models are nothing but **distributional assumptions** combined with a **structural part**.
- For example, our dear linear model with Gaussian distributed error term has **structural part** (linear predictor) and a **distributional assumption** (Gaussian error).
- Usually the structural part of the model is parametrized, for example by a parameter  $\theta \in \Theta \subseteq \mathbb{R}^p$ , determining the behavior of our model.
- In a frequentist approach, our probabilistic model is expressed by setting a distribution on our data, say  $\mathbf{Y} \in \mathbb{Y} \subseteq \mathbb{R}^d$ , like

$$\mathbf{Y} \sim f(\mathbf{y} \mid \theta)$$

where  $f$  denotes a probability mass or density function.

- In a frequentist setting, the source of randomness is entirely driven by the distribution of our data. Hence, once we observe a sample, most of the inferential procedures look for value of  $\theta$  that better describe the observed data.

# Introduction

- Differently from the frequentist approach, where the probabilistic model is set only on the data, a Bayesian model is nothing but a distributional assumption **jointly for data and parameter**<sup>1</sup>

$$(\mathbf{Y}_1, \dots, \mathbf{Y}_n, \boldsymbol{\theta}) \sim \mathcal{L}(\mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\theta})$$

- We can exploit the chain rule, rewriting the previous distribution as

$$\mathcal{L}(\mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\theta}) = \mathcal{L}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- A Bayesian model is composed by a (prior) **distributional assumption for the parameter**, here denoted by  $\pi(\boldsymbol{\theta})$ , and another **distributional assumption for the data**  $\mathcal{L}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\theta})$ .
- We also assume conditional independence of the data, i.e.

$$\mathcal{L}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i \mid \boldsymbol{\theta}),$$

which says that the shared information among distinct observations is fully driven by the parameter  $\boldsymbol{\theta}$ .

---

<sup>1</sup> $\mathcal{L}$  denotes a generic distributional law and will be used consistently in the module

**Should I be Bayesian?**

—

# Should I be Bayesian?

---

Here some reasons to be Bayesian from Professors of the MSc in SSE.

- **Uncertainty quantification**, the Bayesian approach is genuinely tailored to quantify the **uncertainty of our estimates**.
- **Sequential update**, once new data are coming, we can **update our posterior belief** in force of the new information (in tractable cases it is easy).
- **More intuitive and interpretable**, as the parameters themselves follow a distribution.
- **Different assumptions**, the Bayesian paradigm assumes **exchangeable observations** instead of independent and identically distributed, which is a weaker assumption (and more realistic in many scenarios).
- **With complex models**, doing MCMC is **easier** (and **funnier**) rather than doing optimization.
- **You should not be**, don't ruin your life.

## Bayes' Theorem

---

# Bayes' Theorem

A first version of Bayes' Theorem.

## Theorem

*Let  $E$  be an event contained in  $F_1 \cup \dots \cup F_t$ , where the generic  $F_j$ ,  $j = 1, \dots, t$ , is a measurable event,  $F_i \cap F_j = \emptyset$  for any  $i \neq j$ , and  $P(E) > 0$ . Then, for the generic  $F_j$  the following holds*

$$P(F_j | E) = \frac{P(E | F_j)P(F_j)}{\sum_{j=1}^t P(E | F_j)P(F_j)}. \quad (1)$$

Ideally, we have a prior opinion on a set of possible events  $\{F_1, \dots, F_t\}$ . Then, suppose we observe an event  $E$ , with non-null probability, for which we know the probability of that event conditionally on each  $F_j$ ,  $j = 1, \dots, t$ .

Thanks to the Bayesian rule described in Equation (1), we can update our belief conditioned on the information driven by the event  $E$ .



# Bayes' Theorem

---

Proof.

# Bayes' Theorem

We recall and define the following quantities

- $\pi(\theta)$  is the **prior** distribution, which express our prior belief on the parameter space  $\Theta \subseteq \mathbb{R}^p$ .
- $L(\mathbf{y}_{1:n} \mid \theta)$  is the **likelihood function**, where  $\mathbf{y}_{1:n} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ ,  $\mathbf{y}_i \in \mathbb{Y} \subseteq \mathbb{R}^d$ , which measures how likely is a specific value of  $\theta$  given the observations  $\mathbf{y}_{1:n}$ .

Similarly to Equation (1), modern Bayesian approaches found on expressing our posterior belief over  $\Theta$  by updating our prior belief  $\pi(\theta)$  conditioning on the information coming through the observed sample  $\mathbf{y}_{1:n}$ .

## Theorem

*Let  $\mathbf{y}_{1:n}$  an observed sample and  $\theta$  a parameter of interest. Let  $\pi(\theta)$  be a distribution expressing our prior guess over  $\Theta$  and  $L(\theta \mid \mathbf{y}_{1:n})$  the likelihood function. Then*

$$\pi(\theta \mid \mathbf{y}_{1:n}) = \frac{L(\mathbf{y}_{1:n} \mid \theta)\pi(\theta)}{m(\mathbf{y}_{1:n})}, \quad (2)$$

*where  $m(\mathbf{y}_{1:n}) = \int_{\Theta} L(\mathbf{y}_{1:n} \mid \theta)\pi(\theta)d\theta$  is the marginal distribution of  $\mathbf{y}_{1:n}$ .*

**Important remark:**  $\pi(\theta \mid \mathbf{y}_{1:n}) \propto L(\mathbf{y}_{1:n} \mid \theta)\pi(\theta)$ .

# Bayes' Theorem

---

## Example

Suppose that we have a room full of pets. We know that 30% of them are dogs and 70% are cats, and hopefully they are friendly with each other. We further know that they are just of two colors, gray or brown. In particular, among the dogs 80% are brown, while among the cats 30% are brown. What is the probability of being a dog given that the color is brown?

## Example

We are proud citizen of Fantasytown. In the next month we will have the election of the major of our dear city. Two parties are running for the election, A and B. Let  $\theta$  be the probability that an elector votes for the party A. Suppose that we don't have any prior opinion on a specific value for such a probability, and we set  $\theta \sim Unif(0, 1)$ . Assuming each vote distributed as a Bernoulli distribution, i.e. the generic  $y_i \sim Be(\theta)$ , what is the posterior distribution  $\pi(\theta \mid y_1, \dots, y_n)$ ?

## A glimpse on exchangeability

One of the key differences between the frequentist and Bayesian approaches lies in their underlying assumptions about the observed data.

- In a frequentist setting, data are assumed to be sampled **independent and identically distributed** from a common distribution.
- In a Bayesian setting, data are assumed to be **exchangeable**.

Exchangeability is a weaker assumption on the data. In practice, we are assuming that the joint distribution of a sample is symmetric, i.e. the distribution of the data is invariant with respect to permutation

$$\mathcal{L}(y_1, \dots, y_n) \stackrel{d}{=} \mathcal{L}(y_{\lambda(1)}, \dots, y_{\lambda(n)}),$$

where  $\lambda : \mathbb{N}_n \rightarrow \mathbb{N}_n$  is a permutation of  $\{1, \dots, n\}$ .

- In practice, the order we observe our sample does not matter on the inferential procedure we are doing.
- Thanks to the **De Finetti representation theorem**, exchangeability implies conditional independence and justify the existence of a prior distribution.
- Such a condition can be further relaxed, e.g. partial exchangeability in mixed model (see slide block 3).

## Predicting the future

---

## Predicting the future

- Producing inference about unknown observable quantities, i.e. **predictive inference**, is quite natural in a Bayesian setting.
- The distribution of a generic unknown (but observable)  $y$  is given by

$$\mathcal{L}(y) = \int_{\Theta} \mathcal{L}(y, \theta) d\theta = \int_{\Theta} f(y | \theta) \pi(\theta) d\theta.$$

- After we collect  $n$  observations, we might be interested to study the distribution of  $y_{n+1}$  given our prior guess updated by  $y_{1:n}$ . Hence, we have

$$\begin{aligned}\mathcal{L}(y_{n+1} | y_{1:n}) &= \int_{\Theta} \mathcal{L}(y, \theta | y_{1:n}) d\theta \\ &= \int_{\Theta} f(y | \theta, y_{1:n}) \pi(\theta | y_{1:n}) d\theta \\ &= \int_{\Theta} f(y | \theta) \pi(\theta | y_{1:n}) d\theta,\end{aligned}$$

where the last step comes from the conditional independence.

- We can easily perform predictive inference averaging with respect to the posterior distribution.
- $\mathcal{L}(y_{n+1} | y_{1:n})$  is called **predictive distribution** or, more precisely, posterior predictive distribution.

## Choosing the prior distribution

---



# Choosing the prior distribution

---

The prior distribution expresses our **belief** on the parameter space  $\Theta$ .

Clearly, different distributional assumption resemble different belief. For example the followings.

- Within the same distributional family, we might have a prior belief more or less dispersed over  $\Theta$ .
- We can have a truncated distribution because we know that some values are not plausible.

Ideally, there are two main properties that play a crucial role in the prior specification.

- **Conjugacy.**
- **Informativeness.**

We will see many example during the module of priors satisfying one or both the previous.

# Conjugate prior distributions

We consider two distributional families:

- the class of **sampling distribution**  $\mathcal{F}$ , with  $f(\mathbf{y} \mid \boldsymbol{\theta}) \in \mathcal{F}$ ;
- the class of **prior distribution**  $\mathcal{P}$ , with  $\pi(\boldsymbol{\theta}) \in \mathcal{P}$ .

We can define the **conjugacy** as follow.

## Definition

We say that a class of prior distribution  $\mathcal{P}$  is conjugate for a class of sampling distribution  $\mathcal{F}$  if

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \in \mathcal{P}, \quad \text{for all } f(\mathbf{y} \mid \boldsymbol{\theta}) \in \mathcal{F} \text{ and } \pi(\boldsymbol{\theta}) \in \mathcal{P}.$$

- Usually we restrict our attention to distributional families  $\mathcal{P}$ , e.g.  $\mathcal{P}$  is the family of univariate Gaussian distribution, etc.
- Note that

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n f(\mathbf{y}_i \mid \boldsymbol{\theta}) = \{\pi(\boldsymbol{\theta}) f(\mathbf{y}_1 \mid \boldsymbol{\theta})\} \prod_{i=2}^n f(\mathbf{y}_i \mid \boldsymbol{\theta})$$

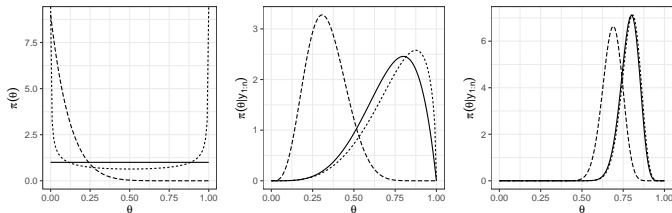
if  $\mathcal{P}$  is conjugate for a single  $f(\mathbf{y} \mid \boldsymbol{\theta})$ , then is conjugate for the likelihood.

- Conjugacy has a trivial interpretation, a posteriori we just update the parameters.

# Informative priors or not?

Informativeness **strongly impact** the way we can specify our prior guess. When specifying a prior, we can be in the following cases.

- **Informative**, we trust our prior belief. Ideally, we center our guess around a value with a small prior dispersion. We need a strong empirical information to move such a belief to other region of the support.
- **Weakly informative**, we center our guess, but we are not particularly confident about that, so we specify the prior distribution with a large dispersion (dangerous if not symmetric and/or in specific experimental settings).
- **Noninformative**. The prior plays a minimal role in the posterior distribution. Ideally, inference is unaffected by the prior setting.



**Figure 1:** From prior to posterior, with three different prior distributions. Dashed line, informative. Full line, weakly informative. Dotted line, noninformative. Left plot: prior distribution. Middle plot: posterior with 5 observations. Right plot: posterior with 25 observations.

## Jeffreys' invariance principle

One approach to be noninformative relies in the **principle of invariance** with respect to reparametrizations.

Ideally, Jeffreys' principle state that if we have a one-to-one transformation of our parameter  $\lambda = q(\theta)$ , such that the prior can be expressed with a change of variable as

$$\pi(\lambda) = \pi(\theta) \left| \frac{d\theta}{d\lambda} \right| = \pi(\theta) \left| \frac{d}{d\theta} q(\theta) \right|^{-1},$$

our prior guess should be invariant with respect of such transformation.

Jeffrey suggests to specify a prior starting from the **Fisher information** of  $\theta$

$$I(\theta) = \mathbb{E} \left[ \left( \frac{d \log f(\mathbf{y} | \theta)}{d\theta} \right) \left( \frac{d \log f(\mathbf{y} | \theta)}{d\theta} \right)^{\top} \middle| \theta \right] = \mathbb{E} \left[ \frac{d^2 \log f(\mathbf{y} | \theta)}{d\theta^2} \middle| \theta \right].$$

Then, a prior distribution invariant with respect to one-to-one reparametrizations can be constructed as

$$\pi_J(\theta) \propto [I(\theta)]^{1/2}.$$

**Important remark:** some remarkable examples of Jeffreys' prior give an improper distribution, i.e.  $\int_{\Theta} \pi(\theta) d\theta = +\infty$ . Nevertheless, also in this cases the posterior can still be a proper distribution.

## Jeffreys' invariance principle

---

We can easily check that the Jeffreys' prior is invariant with respect to reparametrization, i.e. if  $\lambda = q(\theta)$  is a one-to-one transformation, then

$$\pi_J(\lambda) = \pi_J(\theta) \left| \frac{d\theta}{d\lambda} \right|.$$

### Example

Suppose we have  $Y \sim N(0, \sigma^2)$ , a Gaussian distribution with known mean and unknown variance, with

$$f(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{y^2}{2\sigma^2} \right\}$$

Hence, we want to find the Jeffreys' prior for  $\sigma^2$ .

## Point estimates, credible regions and tests

---

# Inference and Bayes

---

We set up our prior specification. We observe some data. We update our prior belief. Now, we want to perform some **inference** with our **posterior belief**, expressed as a distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}).$$

As usual, we want to tackle three main tasks with our inference.

- **Point estimates**, we want to summarize our posterior guess with a single value which is representative of our updated belief.
- **Intervals**, we want to provide a range of values which are plausible given our updated belief.
- **Tests**, we want to use our updated belief to answer specific inferential questions, expressed in terms of hypotheses.



# Point estimates

A natural viewpoint to present point estimate in Bayesian framework is through **decision theory**.

Suppose we have a **loss function**

$$R(\mathbf{a}, \boldsymbol{\theta}) : \mathcal{A} \times \Theta \rightarrow \mathbb{R}_+$$

where  $\mathcal{A}$  is a set of possible actions and  $\Theta$  is the parameter space.

- Such a function **quantifies the loss** we are committing by choosing an action  $\mathbf{a}$  when the parameter is  $\boldsymbol{\theta}$ .
- Ideally, our optimal action  $\mathbf{a}$  is **minimizing the loss** we are committing.
- Further, such a loss should be minimized **for any value** of  $\boldsymbol{\theta}$ .

Specifically, the loss function can be averaged either **a priori**

$$\mathbb{E}[R(\mathbf{a}, \boldsymbol{\theta})] = \int_{\Theta} R(\mathbf{a}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

or **a posteriori**

$$\mathbb{E}[R(\mathbf{a}, \boldsymbol{\theta}) \mid \mathbf{y}_1, \dots, \mathbf{y}_n] = \int_{\Theta} R(\mathbf{a}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}_1, \dots, \mathbf{y}_n) d\boldsymbol{\theta}.$$

# Point estimates

We can extend this theory to derive a point estimate strategy. We set the action space equal to the parameter space,  $\mathcal{A} \equiv \Theta$ .

## Definition

The estimate is the **value minimizing the loss** while **is averaged** with respect to all the possible parameter choices, i.e. a **priori** we have

$$\hat{\theta} = \arg \min_{\theta^* \in \Theta} \left\{ E[R(\theta^*, \theta)] \right\},$$

while a **posteriori** we have

$$\hat{\theta} = \arg \min_{\theta^* \in \Theta} \left\{ E[R(\theta^*, \theta) \mid y_1, \dots, y_n] \right\}.$$

- Different loss functions lead to different point estimates, e.g. the **quadratic loss** function leads to the prior or posterior **mean**, while the **linear loss** function leads to the prior or posterior **median**.
- Point estimates usually are aggressive way to summarize our belief, as they synthesize a whole distribution on a single atom.
  - However, they're quite intuitive and easy to communicate.

### Example

Let  $R(\theta^*, \theta)$  be a quadratic loss function, i.e.

$$R(\theta^*, \theta) = (\theta^* - \theta)^\top (\theta^* - \theta).$$

Then, we can show that the point estimate with such a loss function is the (prior or posterior) expectation.

## Credible intervals

---

- With interval estimation we aim to produce a **set of reasonable values** for the parameter of interest in our analysis, **incorporating some uncertainty** quantification in our estimation processes.
- Given that the posterior information is represented by **an entire distribution**  $\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$ , the definition of interval estimates within the Bayesian framework is **quite natural**.
- The region  $C_\alpha$  is a  $100(1 - \alpha)\%$  credible interval (or Bayesian credible region) for  $\boldsymbol{\theta}$  if

$$P(\boldsymbol{\theta} \in C_\alpha \mid \mathbf{y}_{1:n}) = 1 - \alpha.$$

- However such definition is not unique, as it is possible to define different strategies to derive  $C_\alpha$ , which lead to different regions.
  - We can cut the support of  $\boldsymbol{\theta}$  in different way, but preserving the same amount of mass in the subset  $C_\alpha$ .
  - Some way of producing subsets are more justified and reasonable.
- The most commonly used strategies are **highest (posterior) density intervals** and **equally tailed intervals**.

**Highest (posterior) density intervals** are credible regions where we consider parameter values with the highest density function  $\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$ .

## Definition

The region  $C_\alpha$  is a  $100(1 - \alpha)\%$  highest (posterior) density interval for  $\boldsymbol{\theta}$  if

$$C_\alpha = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) \geq \gamma\},$$

where  $\gamma$  is chosen such that  $P(\boldsymbol{\theta} \in C_\alpha \mid \mathbf{y}_{1:n}) = \gamma$ .

- Given its construction, the highest posterior density interval produces the **smallest region** with respect to some measure.
- When the distribution is not symmetric, the probability mass left outside the region can be divided into **unequal part** on the tails.
- When the posterior distribution has a complex behaviour, such as multimodality, is not symmetric, etc., computing such region may not be easy.
- To compute the HPD we should know the exact values of the posterior density function, i.e. it is not possible to use the proportionality relation  $\propto$  in the computation, and we need to evaluate the normalization constant of the posterior distribution.
  - For complex problems the evaluation of such constant is not a trivial task.

## Credible intervals

**Equally tailed intervals** are credible regions constructed to leave equal probability mass on the tails outside of the posterior density function  $\pi(\theta \mid \mathbf{y}_{1:n})$  outside the region of interest.

### Definition

The region  $C_\alpha$  is a  $100(1 - \alpha)\%$  **equally tailed interval** for  $\theta$  if

$$C_\alpha = [c_{\alpha/2}, c_{1-\alpha/2}] = \left\{ \theta : P(\theta < c_{\alpha/2} \mid \mathbf{y}_{1:n}) = P(\theta > c_{\alpha/2} \mid \mathbf{y}_{1:n}) = \frac{\alpha}{2} \right\}.$$

- In practice,  $c_{\alpha/2}$  and  $c_{1-\alpha/2}$  are the quantiles of order  $\alpha/2$  and  $1 - \alpha/2$  of the posterior distribution.
- Depending on specific problems, the derivation of such quantiles can be an easy or a difficult task.
- The construction of an equally tailed interval guarantees that both tails have **the same probability mass**, but when the posterior distribution is not symmetric it means that we **eventually include parameters with low values of the posterior density function**.
- Not so easy with multivariate distribution.

**Testing hypotheses** is broadly used to support or refute some opinion on a phenomena of interest, specified as a partition of the parameter space  $\Theta$  or as model settings. Here we focus on the first case, the latter will be discussed later in the module.

Our starting point is a system of **hypotheses**

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1,$$

such that  $\Theta_0 \cup \Theta_1 = \Theta$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ .

In a **frequentist setting**, a test statistics is measuring on an empirical level if the observed data **support or not** the null hypothesis. In the case that there is a strong empirical evidence against  $H_0$  we reject the null hypothesis.

- The two hypotheses in a frequentist framework are **not symmetric**, in the sense that we need to assume an hypothesis true (null) to derive a test statistics.  
→ Testing  $H_0$  against  $H_1$  gives us no information about testing  $H_1$  against  $H_0$ .

In a Bayesian setting, we are measuring how **the empirical information is supporting** either  $H_0$  or  $H_1$ . There is no need to assume one of the hypotheses true.

- The two hypotheses in a Bayesian framework are **symmetric and can be exchanged**. We already have a distribution to use, the posterior distribution.  
→ Testing  $H_0$  against  $H_1$  gives is equivalent to  $H_1$  against  $H_0$ .

There are **different strategies** to perform hypothesis test in a **Bayesian framework**. In this module we resort to the **Bayes factor** to test our hypotheses.

## Definition

The Bayes factor can be defined as

$$\text{BF}_{01} = \frac{\text{posterior odds}_{01}}{\text{prior odds}_{01}} = \frac{\frac{P(\Theta_0 | \mathbf{y}_{1:n})}{P(\Theta_1 | \mathbf{y}_{1:n})}}{\frac{P(\Theta_0)}{P(\Theta_1)}} = \frac{P(\Theta_0 | \mathbf{y}_{1:n}) P(\Theta_1)}{P(\Theta_1 | \mathbf{y}_{1:n}) P(\Theta_0)},$$

i.e. the ratio of the posterior odds and the prior odds.

- The subscript 01 denotes the numerator and denominator quantities used for the odds calculation.
- The prior guess is incorporated in the testing procedure.
- In practice,  $\text{BF}_{01}$  is measuring how much **the empirical information** is shifting our guess toward  $H_0$ , but **adjusting by the prior guess**.
- Note that

$$\text{BF}_{10} = \frac{P(\Theta_1 | \mathbf{y}_{1:n}) P(\Theta_0)}{P(\Theta_0 | \mathbf{y}_{1:n}) P(\Theta_1)} = \frac{1}{\frac{P(\Theta_0 | \mathbf{y}_{1:n}) P(\Theta_1)}{P(\Theta_1 | \mathbf{y}_{1:n}) P(\Theta_0)}} = \frac{1}{\text{BF}_{01}}.$$



About the interpretation of Bayes factors, there are some **general guidelines** based on the observed value or its log-transformation.

BF <sub>01</sub>	log BF <sub>01</sub>	evidence
<1	<0	negative
1–3	0–2	weakly positive
3–12	2–5	positive
12–150	5–10	strongly positive
>150	>10	very strongly positive

- The Bayes factor can be rewritten as

$$\text{BF}_{01} = \frac{P(\Theta_0 \mid \mathbf{y}_{1:n}) P(\Theta_1)}{P(\Theta_1 \mid \mathbf{y}_{1:n}) P(\Theta_0)} = \frac{P(\Theta_0, \mathbf{y}_{1:n}) P(\mathbf{y}_{1:n}) P(\Theta_1)}{P(\Theta_1, \mathbf{y}_{1:n}) P(\mathbf{y}_{1:n}) P(\Theta_0)} = \frac{P(\mathbf{y}_{1:n} \mid \Theta_0)}{P(\mathbf{y}_{1:n} \mid \Theta_1)}.$$

- The Bayes factor can be generalized to the model comparison case. Suppose we have two models  $M_0$  and  $M_1$ . Then we can perform a comparison with

$$\text{BF}_{01} = \frac{P(\mathbf{y}_{1:n} \mid M_0)}{P(\mathbf{y}_{1:n} \mid M_1)}.$$

We will see more in details the model comparison case later in the module.

A peculiar case is given by testing an **atomic** and a **diffuse hypothesis**

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0.$$

Suppose now we have a diffuse prior distribution  $\pi(\theta)$ .

We have a problem. Under such a prior assumption,

$$P(\theta = \theta_0) = 0,$$

and we **cannot construct the Bayes factor**. Indeed, we can define a new prior starting from  $\pi(\theta)$  to solve this issue.

## Proposition

*Let  $\pi(\theta)$  be a diffuse prior over  $\Theta$ . Suppose we want to test an atomic versus a diffuse hypothesis,  $H_0 : \theta = \theta_0$  and  $H_1 : \theta \neq \theta_0$ . Define a new prior*

$$\pi_1(\theta) = \beta_0 \delta_{\theta_0}(\theta) + (1 - \beta_0) \pi(\theta),$$

*where  $\beta_0$  is the prior probability associated with  $H_0$ . Then*

$$\text{BF}_{01} = \frac{L(\mathbf{y}_{1:n} \mid \theta_0)}{m(\mathbf{y}_{1:n})}$$

*with  $m(\mathbf{y}_{1:n})$  marginal distribution of  $\mathbf{y}_{1:n}$ .*





## Example

Let us consider a model

$$f(y \mid \theta) = 2\theta y e^{-\theta y^2}, \quad y > 0, \theta > 0.$$

Suppose we observed a sample of size  $n = 4$ , with  $y_1^2 + \cdots + y_4^2 = 5.71$ .

- i) Find the family of prior distribution conjugate to the previous model and compute the posterior distribution.
- ii) Write down the expression of the Bayes factor to test

$$H_0 : \theta = 1 \quad \text{vs} \quad H_1 : \theta \neq 1.$$

- iii) Choose the parameters of the prior distribution that guarantee

$$E[\theta] = 1, \quad \text{var}(\theta) = 10.$$

Perform the test. Do you prefer  $H_0$  or  $H_1$ ?









## Sampling from posterior distributions

---

# Sampling from posterior distributions

---

- In an **ideal** world, once we have a posterior distribution everything is **analytically tractable**.
- In a more **realistic** world, even if we do not have analytical tractability, we can **easily sample** from the posterior distribution of interest, and use that sample for inferential purposes.
- In the **real** world, we cannot. We need tailored strategies to produce a sample even when the posterior distribution is **hardly tractable**.

There are many strategies to work with hardly tractable distributions. Among these, in the module we will consider three alternatives:

- Metropolis-Hastings;
- Gibbs sampler;
- Hamiltonian Monte Carlo.

These strategies can be used to produce a Markov chain whose **ergodic distribution** is the posterior **distribution of interest**.

# The Metropolis-Hastings

---

One of the **early approaches** to produce a Markov chain with a specific distribution as ergodic is the **Metropolis-Hastings**.

The idea is to generate a chain with Markovian dependence  $\{\theta^{(t)}\}_{t \geq 1}$  such that is behaving like the target distribution.

Let  $\pi(\theta \mid \mathbf{y}_{1:n})$  be the distribution of interest that we want to sample from. The core of Metropolis-Hastings is an **auxiliary proposal (instrumental) distribution**  $q(\theta^N \mid \theta)$ , such that the support of  $\pi(\theta \mid \mathbf{y}_{1:n})$  is a subset of the proposal support.

- $q(\theta^N \mid \theta) = q(\theta^N)$ , independent proposal.
- $q(\theta^N \mid \theta) = q(\theta^N - \theta)$ , random walk.

Suppose we propose a value  $\theta^{(N)} \sim q(\theta^N \mid \theta)$ . We then **accept** the proposed value with probability

- $\alpha(\theta^N, \theta) = \min \left\{ 1, \frac{\pi(\theta^N \mid \mathbf{y}_{1:n})}{\pi(\theta \mid \mathbf{y}_{1:n})} \right\}$ , Metropolis algorithm.
- $\alpha(\theta^N, \theta) = \min \left\{ 1, \frac{\pi(\theta^N \mid \mathbf{y}_{1:n})q(\theta \mid \theta^N)}{\pi(\theta \mid \mathbf{y}_{1:n})q(\theta^N \mid \theta)} \right\}$ , Metropolis-Hastings algorithm.

The latter produce a chain **with ergodic distribution**  $\pi(\theta^N \mid \mathbf{y}_{1:n})$ .

# The Metropolis-Hastings

---

- We remark that
  - $\alpha(\theta^N, \theta)$  is the probability to **move** from  $\theta$  to  $\theta^N$ ;
  - $[1 - \alpha(\theta^N, \theta)]$  is the probability to **stay** on  $\theta$ .
- The algorithm works not only for posterior distribution, but for **general probability distributions** that satisfy its **assumptions**.
- The algorithm works also up to a normalization constant. For example, let  $\pi(\theta \mid \mathbf{y}_{1:n}) = C \times \pi^*(\theta \mid \mathbf{y}_{1:n})$ . Then is easy to see that

$$\alpha(\theta^N, \theta) = \min \left\{ 1, \frac{\pi(\theta^N \mid \mathbf{y}_{1:n})q(\theta \mid \theta^N)}{\pi(\theta \mid \mathbf{y}_{1:n})q(\theta^N \mid \theta)} \right\} = \min \left\{ 1, \frac{\cancel{C}\pi^*(\theta^N \mid \mathbf{y}_{1:n})q(\theta \mid \theta^N)}{\cancel{C}\pi^*(\theta \mid \mathbf{y}_{1:n})q(\theta^N \mid \theta)} \right\}$$

An idea of **implementation** is the following.

---

**Algorithm 1** Pseudocode for the Metropolis-Hastings algorithm, sample of size  $T$ .

---

- 1: Set initial values for  $\theta^{(0)}$ .
  - 2: **for**  $t = 1$  to  $T$  **do**
  - 3:   Sample  $\theta^N \sim q(\theta^N \mid \theta^{(t-1)})$ .
  - 4:   Set  $\alpha(\theta^N, \theta^{(t-1)}) = \min \left\{ 1, \frac{\pi(\theta^N \mid \mathbf{y}_{1:n})q(\theta^{(t-1)} \mid \theta^N)}{\pi(\theta^{(t-1)} \mid \mathbf{y}_{1:n})q(\theta^N \mid \theta^{(t-1)})} \right\}$ .
  - 5:   Sample  $U \sim \text{Unif}(0, 1)$ .
  - 6:   If  $U < \alpha(\theta^N, \theta^{(t-1)})$ , set  $\theta^{(t)} = \theta^N$ . Else, set  $\theta^{(t)} = \theta^{(t-1)}$ .
  - 7: **end for**
-

# The Metropolis-Hastings

Recall that a Markov chain satisfies the **detailed balance condition** if there exists a function  $\pi(\theta \mid \mathbf{y}_{1:n})$  such that

$$\mathcal{K}(\theta^N, \theta^{(t-1)})\pi(\theta^N \mid \mathbf{y}_{1:n}) = \mathcal{K}(\theta^{(t-1)}, \theta^N)\pi(\theta^{(t-1)} \mid \mathbf{y}_{1:n}),$$

where  $\mathcal{K}(\theta^N, \theta^{(t-1)})$  is the transition kernel, and that such condition implies (i)  $\pi(\theta^{(t-1)} \mid \mathbf{y}_{1:n})$  is the **invariant distribution** of the chain and (ii) the chain is **reversible**.

Then, it is possible to prove the following result.

## Theorem

Let  $\{\theta^{(t)}\}_{t \geq 1}$  be the chain produced by a Metropolis-Hastings algorithm, such that the support of the target distribution  $\pi(\theta \mid \mathbf{y}_{1:n})$  is covered by the support of the proposal  $q(\cdot \mid \theta^{(t-1)})$ . Then,

- (a) the kernel of the chain satisfies the detailed balance condition with  $\pi(\theta \mid \mathbf{y}_{1:n})$ ;
- (b)  $\pi(\theta \mid \mathbf{y}_{1:n})$  is the stationary distribution of the chain.

Note that (a)  $\Rightarrow$  (b).



Quite commonly, even if we are unable to sample directly a distribution for the whole parameter  $\theta$ , we can express one (or more) of its dimension conditioning on the others.

Intuitively, this is the idea beyond **Gibbs samplers**.

Suppose, for simplicity, that we can **express our parameter of interest** as  $\theta = \{\theta_1, \theta_2\}$ .

If the conditional distributions

$$\pi(\theta_1 \mid \theta_2, \mathbf{y}_{1:n}),$$

$$\pi(\theta_2 \mid \theta_1, \mathbf{y}_{1:n}),$$

can be sampled easily, then we can iteratively update both subsets of parameters.

The previous concept can be extended to  $d > 2$  subsets. An idea of **implementation** is the following.

---

**Algorithm 2** Pseudocode for the Metropolis-Hastings algorithm, sample of size  $T$ .

---

- 1: Set initial values for  $\theta_1^{(0)}, \dots, \theta_d^{(0)}$ .
  - 2: **for**  $t = 1$  to  $T$  **do**
  - 3:   **for**  $j = 1$  to  $d$  **do**
  - 4:     Sample  $\theta_j^{(r)} \sim \pi(\theta_j^{(r)} \mid \theta_1^{(r)}, \dots, \theta_{j-1}^{(r)}, \theta_{j+1}^{(r-1)}, \dots, \theta_d^{(r-1)} \mathbf{y}_{1:n})$
  - 5:   **end for**
  - 6: **end for**
-

# Gibbs sampler

The **Gibbs sampler** is a special case of **Metropolis-Hastings**, where we accept all the proposed values. Note that if we propose a move from  $\theta_1^{(r-1)}$  to  $\theta_1^N$ , then we have that the proposal distribution is  $\pi(\theta_1^N | \theta_2^{(r-1)}, \mathbf{y}_{1:n})$ . The corresponding acceptance rate is then

$$\begin{aligned}\alpha(\theta_1^N, \theta_1^{(r-1)}) &= \frac{\pi(\theta_1^N, \theta_2^{(r-1)} | \mathbf{y}_{1:n})\pi(\theta_1^{(r-1)} | \theta_2^{(r-1)}, \mathbf{y}_{1:n})}{\pi(\theta_1^{(r-1)}, \theta_2^{(r-1)} | \mathbf{y}_{1:n})\pi(\theta_1^N | \theta_2^{(r-1)}, \mathbf{y}_{1:n})} \\ &= \frac{\pi(\theta_1^N | \theta_2^{(r-1)}, \mathbf{y}_{1:n})\pi(\theta_2^{(r-1)} | \mathbf{y}_{1:n})\pi(\theta_1^{(r-1)} | \theta_2^{(r-1)}, \mathbf{y}_{1:n})}{\pi(\theta_1^{(r-1)} | \theta_2^{(r-1)}, \mathbf{y}_{1:n})\pi(\theta_2^{(r-1)} | \mathbf{y}_{1:n})\pi(\theta_1^N | \theta_2^{(r-1)}, \mathbf{y}_{1:n})} = 1.\end{aligned}$$

All the **properties** of **Metropolis-Hastings** hold, for example that the invariant distribution of the chain is  $\pi(\theta^{N(1)}, \theta_{r-1}^{(2)} | \mathbf{y}_{1:n})$

- **Full Gibbs sampler**, where we have  $d$  dimension and we express each dimension conditionally on the others, e.g.

$$\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d.$$

- **Blocked Gibbs sampler**, we express some of the dimension conditioned on the others, e.g.

$$\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d.$$

- **Collapsed Gibbs sampler**, we marginalize some of the dimensions, e.g.

$$\theta_1 | \theta_3, \dots, \theta_d.$$



# Hamiltonian Monte Carlo

---

**Hamiltonian Monte Carlo** is a particular class of samplers which defines a kind of **Metropolis-Hastings** where the proposal is informed by the target distribution shape.

- Suppose we want to **sample**  $\theta \sim \pi(\theta \mid \mathbf{y}_{1:n})$ .
- We can think of  $\theta$  as the **position** of a dynamical system.
- We augment the problem by an auxiliary variable, the **momentum** of our system, for which we assume typically an **independent Gaussian distribution**.

The Hamiltonian Monte Carlo intuitively works by iterating the following steps

- (a) update position and momentum, according to a trajectory based on Hamiltonian dynamics;
- (b) perform a Metropolis-Hastings step to accept the proposed value.

Some remarks.

- The Hamiltonian Monte Carlo is exploring nicely the target distribution support.
- Works for continuous target distributions.

In short, we have a position  $\boldsymbol{\theta} \in \mathbb{R}^d$  and a momentum  $\mathbf{p} \in \mathbb{R}^d$ . The system is described by a function of  $(\boldsymbol{\theta}, \mathbf{p})$ , say  $H(\boldsymbol{\theta}, \mathbf{p})$ , known as **Hamiltonian function**.

→ Such a function describe the **evolution** of the system over time, specifically by looking at its **partial derivatives**

$$\begin{cases} \frac{d\theta_j}{dt} = \frac{\partial H}{\partial p_j}, & j = 1, \dots, d, \\ \frac{dp_j}{dt} = -\frac{\partial H}{\partial \theta_j}, & j = 1, \dots, d. \end{cases}$$

We consider function  $H(\boldsymbol{\theta}, \mathbf{p})$  of the type

$$\underbrace{H(\boldsymbol{\theta}, \mathbf{p})}_{\text{ENERGY}} = \underbrace{E(\boldsymbol{\theta})}_{\text{POTENTIAL ENERGY}} + \underbrace{K(\mathbf{p})}_{\text{KINETIC ENERGY}},$$

with  $K(\mathbf{p}) = \frac{1}{2}(\mathbf{p}M^{-1}\mathbf{p})$ . The latter leads to an independent  $\mathbf{0}$ -mean Gaussian distribution for the momentum.

Assuming the previous function  $K(\mathbf{p})$ , we have

$$\begin{cases} \frac{d\theta_j}{dt} = [M^{-1}\mathbf{p}]_j, & j = 1, \dots, d, \\ \frac{d\theta_j}{dt} = -\frac{\partial E}{\partial \theta}. \end{cases}$$

We can solve the previous numerically, resorting to a discretization. For example, using a **leapfrog integrator**.

We start from  $t = 0$ , with an initial value for  $\boldsymbol{\theta}$  and  $\mathbf{p}$ . We iterate the following to get a trajectory for position and momentum.

$$\begin{aligned} p_j(t + \epsilon/2) &= p_j(t) - \frac{\epsilon}{2} \left[ \frac{\partial E}{\partial \theta_j}(\boldsymbol{\theta}(t)) \right], \\ \theta_j(t + \epsilon) &= \theta_j(t) + \epsilon \left\{ [M^{-1}\mathbf{p}(t + \epsilon/2)]_j \right\}, \\ p_i(t + \epsilon/2) &= p_i(t) - \frac{\epsilon}{2} \left[ \frac{\partial E}{\partial \theta_i}(\boldsymbol{\theta}(t)) \right]. \end{aligned}$$

Ideally, with the previous for  $t = 1, \dots, L$  we can produce  $L$  distinct values of position and momentum, approximating locally the system trajectory.

In our environment, we have

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) = C_{\boldsymbol{\theta}} e^{-E(\boldsymbol{\theta})}, \quad \mathcal{L}(\mathbf{p}) = C_{\mathbf{p}} e^{-K(\mathbf{p})},$$

and we can simulate the joint distribution  $\mathcal{L}(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y}_{1:n}) \propto e^{-\{E(\boldsymbol{\theta})+K(\mathbf{p})\}}$ . An idea of **implementation** is the following.

---

**Algorithm 3** Pseudocode for the Hamiltonian Monte Carlo algorithm, sample of size  $T$ .

---

- 1: Set initial values for  $\boldsymbol{\theta}^{(0)}$ ,  $\epsilon$ ,  $L$ .
- 2: **for**  $t = 1$  to  $T$  **do**
- 3:   Sample the momentum  $\mathbf{p}_0 \sim C_{\mathbf{p}} e^{-K(\mathbf{p})}$  and set  $\boldsymbol{\theta}_0^N = \boldsymbol{\theta}^{(r-1)}$
- 4:   **for**  $t = 1$  to  $L$  **do**
- 5:      $\mathbf{p}_t = \mathbf{p}_{t-1} - \frac{\epsilon}{2} \nabla E(\boldsymbol{\theta}_{t-1}^N)$
- 6:      $\boldsymbol{\theta}_t^N = \boldsymbol{\theta}_{t-1}^N + \epsilon M^{-1} \mathbf{p}_t$
- 7:      $\mathbf{p}_t = \mathbf{p}_t - \frac{\epsilon}{2} \nabla E(\boldsymbol{\theta}_t^N)$
- 8:   **end for**
- 9:   Negate the momentum  $\mathbf{p}_L = -\mathbf{p}_L$ .
- 10:   Perform a Metropolis-Hastings step with acceptance rate

$$\alpha = \min \left\{ 1, \frac{\exp \{E(\boldsymbol{\theta}_L^N) + K(\mathbf{p}_L)\}}{\exp \{E(\boldsymbol{\theta}_0^N) + K(\mathbf{p}_0)\}} \right\}.$$

- 11: **end for**

## Example

Let us consider the case with

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}.$$

Write explicitly the quantities required to implement an Hamiltonian Monte Carlo to sample from the distribution of  $\boldsymbol{\theta}$ . Try to implement the algorithm in R.

**STAN**

---

- STAN is a probabilistic programming language implementing **full Bayesian statistical inference with MCMC sampling** (NUTS, HMC) and penalized maximum likelihood estimation with Optimization (BFGS).
- STAN is **coded in C++** and runs on all major platforms (Linux, Mac, Windows).
- STAN is an **open-source** software developed mainly at Columbia University.
- STAN can be accessed through **several interfaces**: RStan (R), PyStan (Python), MatlabStan (Matlab) and also CmdStan (shell),  
<http://mc-stan.org/users/interfaces/cmdstan>

Multiple Markov chain Monte Carlo (MCMC) algorithms and optimization algorithms are implemented for the inference:

(1) MCMC algorithms:

- Hamiltonian Monte Carlo (HMC) (default)
- No-U-Turn sampler

(2) Optimization algorithms:

- BFGS algorithm (default), Nesterov's accelerated gradient descent algorithm, Newton's method

## Typical workflow of using RStan

---

- Represent a **statistical model** by writing its log-posterior density (up to a normalization constant independent from the parameters)
  - this can be done using STAN modeling language.
- Translate the model coded in **STAN to C++** code using the function `stanc`.
- **Compile the C++ code** for the model using a C++ compiler (such as `g++`) to create a Dynamic Shared Object that can be loaded by R-
- Run the DSO to **sample** from the posterior.
- **Diagnose convergence** of the MCMC chains of sample.
- Conduct model **inference**.

[Don't worry! A single `rstan` call performs implicitly steps 2, 3 and 4. ]

- All the built-in C++ functions and operators are available.
- For the matrices functions of basic arithmetics, solvers, decompositions, .. are available (check the manual).
- Each distribution of the library has:
  - pseudo random number generator;
  - log density or mass function;
  - cumulative distribution.



# STAN has specific blocks...

---

- **DATA:**
  - Given input data
  - Executed first and load
- **TRANSFORMED DATA:**
  - Transform variables for convenience
- **PARAMETERS:**
  - Result output parameters
  - Updated at each iteration
- **TRANSFORM. PARAMETERS:**
  - Transform parameters for convenience
- **MODEL:**
  - Describe the model
- **GENERATED QUANTITIES:**
  - Generate quantities for monitoring convergence

... the order must be kept,  
the blocks are optional (except model block)

## Variable and expression types

---

- **Primitive:** `int` and `real`
- **Matrix:** `vector[n]`, `row_vector[n]`, `matrix[m, n]`
- **Bounded:** primitive or matrix type, with  
 $\langle lower = L \rangle$ ,  $\langle upper = U \rangle$ ,  $\langle lower = L, upper = U \rangle$
- **Constrained:** `unit_vector` for unit-length vectors, `simplex` for unit simplexes, `ordered` for ordered vectors, `corr_matrix` and `cov_matrix` for symmetric and positive definite matrices.
- **Arrays:** collection of object of any type
- **Sampling:**  $y \sim normal(mu, sigma)$
- **Increments log-probability:** `increment_log_prob(lp)`  
add the value `lp` to the log-density
- **For/while loop:** `for(n in 1 : N), while(cond)`
- **Block:** `{.....}` (allows local variables)
- **Print:** `print(" TH = ", theta)`

## Printing, plotting, and summarizing:

- **show** Print the default summary for the model.
- **print** Print a customizable summary for the model. See `print.stanfit`.
- **plot** Create various plots summarizing the fitted model. See `plot,stanfit-method`.
- **summary** Summarize the distributions of estimated parameters and derived quantities using the posterior draws. See `summary,stanfit-method`.
- **get\_posterior\_mean** Get the posterior mean for parameters of interest (using `pars` to specify a subset of parameters). Returned is a matrix with one column per chain and an additional column for all chains combined.
- ....

see <https://mc-stan.org/rstan/reference/stanfit-class.html> for a summary.

Let's play a bit with STAN!