

# Bayesian Statistics

Luiz Max de Carvalho[lmax.fgv@gmail.com]

PhD-level course  
School of Applied Mathematics (EMAp/FGV), Rio de Janeiro.

March 30, 2021

- This is a 60-hour, PhD-level course on Bayesian inference.
- We have 11 planned weeks. Reading material is posted at <https://github.com/maxbiostat/BayesianStatisticsCourse/>
- Assessment will be done via a written exam (70%) and an assignment (30%);
- Tenets:
  - ◊ Respect the instructor and your classmates;
  - ◊ Read before class;
  - ◊ Engage in the discussion;
  - ◊ Don't be afraid to ask/disagree.
- Books are
  - ◊ Robert (2007);
  - ◊ Hoff (2009);
  - ◊ Bernardo and Smith (2000).

## Bayes's Theorem

What do

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)}, \quad (1)$$

and

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\sum_{i=1}^n \Pr(B \mid A_i) \Pr(A_i)}, \quad (2)$$

and

$$p(\theta \mid \mathbf{y}) = \frac{l(\mathbf{y} \mid \theta) \pi(\theta)}{\int_{\Theta} l(\mathbf{y} \mid t) \pi(t) dt}, \quad (3)$$

and

$$p(\theta \mid \mathbf{y}) = \frac{l(\mathbf{y} \mid \theta) \pi(\theta)}{m(\mathbf{y})}, \quad (4)$$

all have in common? In this course, we will find out how to use Bayes's rule in order to draw statistical inferences in a coherent and mathematically sound way.

## Bayesian Statistics is a complete approach

Our whole paradigm revolves around the posterior:

$$p(\theta | \mathbf{x}) \propto l(\theta | \mathbf{x})\pi(\theta).$$

Within the Bayesian paradigm, you are able to

- Perform point and interval inference about unknown quantities;

$$\delta(\mathbf{x}) = E_p[\theta] := \int_{\Theta} t p(t | \mathbf{x}) dt,$$

$$\Pr(a \leq \theta \leq b) = 0.95 = \int_a^b p(t | \mathbf{x}) dt;$$

- Compare models:

$$\text{BF}_{12} = \frac{\Pr(M_1 | \mathbf{x})}{\Pr(M_2 | \mathbf{x})} = \frac{\Pr(\mathbf{x} | M_1) \Pr(M_1)}{\Pr(\mathbf{x} | M_2) \Pr(M_2)};$$

- Make predictions:  $g(\tilde{x} | \mathbf{x}) := \int_{\Theta} f(\tilde{x} | t) p(t | \mathbf{x}) dt;$
- Make decisions:  $E_p[U(r)].$

## Statistical model: informal definition

Stuff you say at the bar:

### Definition 1 (Statistical model: informal)

*DeGroot, def 7.1.1, pp. 377 A statistical model consists in identifying the random variables of interest (observable and potentially observable), the specification of the joint distribution of these variables and the identification of parameters ( $\theta$ ) that index this joint distribution. Sometimes it is also convenient to assume that the parameters are themselves random variables, but then one needs to specify a joint distribution for  $\theta$  also.*

## Statistical model: formal definition

Stuff you say in a Lecture:

### Definition 2 (Statistical model: formal)

*McCullagh, 2002. Let  $\mathcal{X}$  be an arbitrary sample space,  $\Theta$  a non-empty set and  $\mathcal{P}(\mathcal{X})$  the set of all probability distributions on  $\mathcal{X}$ , i.e.  $P : \Theta \rightarrow [0, \infty)$ ,  $P \in \mathcal{P}$ . A parametric statistical model is a function  $P : \Theta \rightarrow \mathcal{P}(\mathcal{X})$ , that associates each point  $\theta \in \Theta$  to a probability distribution  $P_\theta$  over  $\mathcal{X}$ .*

### Examples:

- Put  $\mathcal{X} = \mathbb{R}$  and  $\Theta = (-\infty, \infty) \times (0, \infty)$ . We say  $P$  is a *normal* (or *Gaussian*) statistical model<sup>1</sup> if for every  $\theta = \{\mu, \sigma^2\} \in \Theta$ ,

$$P_\theta(x) \equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

- Put  $\mathcal{X} = \mathbb{N} \cup \{0\}$  and  $\Theta = (0, \infty)$ .  $P$  is a Poisson statistical model if, for  $\lambda \in \Theta$ ,

$$P_\lambda(k) \equiv \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, \dots$$

<sup>1</sup>Note the abuse of notation: strictly speaking,  $P_\theta$  is a probability **measure** and not a *density* as we have presented it here.

## Principle I: the sufficiency principle

Sufficiency plays a central role in all of Statistics.

### Definition 3 (Sufficient statistic)

Let  $x \sim f(x | \theta)$ . We say  $T : \mathcal{X} \rightarrow \mathbb{R}$  is a **sufficient statistic** for the parameter  $\theta$  if  $\Pr(X = x | T(x), \theta)$  is independent of  $\theta$ .

This is the basis for a cornerstone of Statistics,

### Theorem 1 (Factorisation theorem)

Under mild regularity conditions, we can write:

$$f(x | \theta) = g(T(x) | \theta)h(x | T(x)).$$

We can now state

### Idea 1 (Sufficiency principle (SP))

For  $x, y \in \mathcal{X}$ , if  $T$  is sufficient for  $\theta$  and  $T(x) = T(y)$ , then  $x$  and  $y$  should lead to the same inferences about  $\theta$ .

## Principle II: the Likelihood principle

The Likelihood Principle (LP) is a key concept in Statistics, of particular Bayesian Statistics.

### Idea 2 (Likelihood Principle)

*The information brought by an observation  $x \in \mathcal{X}$  about a parameter  $\theta \in \Theta$  is **completely** contained in the likelihood function  $l(\theta | x) \propto f(x | \theta)$ .*

### Example 1 (Uma vez Flamengo...)



## Principle II: the Likelihood principle

Suppose a pollster is interested in estimating the fraction  $\theta$  of football fans that cheer for Clube de Regatas do Flamengo (CRF). They survey  $n = 12$  people and get  $x = 9$  supporters and  $y = 3$  “antis”. Consider the following two designs:

- i) Survey 12 people and record the number of supporters;
- ii) Survey until they get  $y = 3$ .

The likelihoods for both surveys are, respectively,

$$x \sim \text{Binomial}(n, \theta) \implies l_1(\theta | x, n) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},$$

$$n \sim \text{NegativeBinomial}(y, 1 - \theta) \implies l_2(\theta | n, y) = \binom{n-1}{y-1} \theta^y (1 - \theta)^{n-y},$$

hence

$$l_1(\theta) \propto l_2(\theta) \propto \theta^3 (1 - \theta)^9.$$

Therefore, we say that these two experiments bring exactly the same information about  $\theta$ .

A generalised version of the LP can be stated as follows:

### Theorem 2 (Likelihood Proportionality Theorem (Gonçalves and Franklin, 2019))

Let  $\Theta$  be a nonempty set and  $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$  be a family of probability measures on  $(\Omega, \mathcal{A})$  and  $\nu_1$  and  $\nu_2$  be  $\sigma$ -finite measures on  $(\Omega, \mathcal{A})$ . Suppose  $P \ll \nu_1$  and  $P \ll \nu_2$  for all  $P \in \mathcal{P}$ . Then there exists a measurable set  $A \in \mathcal{A}$  such that  $P_\theta(A) = 1$  for all  $\theta \in \Theta$  and there exist  $f_{1,\theta} \in \left[ \frac{dP_\theta}{d\nu_1} \right]$  and  $f_{2,\theta} \in \left[ \frac{dP_\theta}{d\nu_2} \right]$  and a measurable function  $h$  such that

$$f_{1,\theta}(\omega) = h(\omega)f_{2,\theta}(\omega), \forall \theta \in \Theta \forall \omega \in A.$$

## Principle III: stopping rule principle

A subject of contention between inference paradigms is the role of stopping rules in the inferences drawn.

### Idea 3 (Stopping rule principle (SRP))

*Let  $\tau$  be a stopping rule directing a series of experiments  $\mathcal{E}_1, \mathcal{E}_2, \dots$ , which generates data  $\mathbf{x} = (x_1, x_2, \dots)$ . Inferences about  $\theta$  should depend on  $\tau$  only through  $\mathbf{x}$ .*

### Example 3 (Finite stopping rules)

Suppose experiment  $\mathcal{E}_i$  leads to the observation of  $x_i \sim f(x_i \mid \theta)$  and let  $\mathcal{A}_i \subset \mathcal{X}_1 \times \dots \times \mathcal{X}_i$  be a sequence of events. Define

$$\tau := \inf \{n : (x_1, \dots, x_n) \in \mathcal{A}_n\}.$$

It can be shown that  $\Pr(\tau < \infty) = 1$  (exercise 1.20 BC).

## Principle IV: the conditionality principle

We will now state one of the main ingredients of the derivation of the LP. The Conditionality Principle (CP) is a statement about the permissible inferences from randomised experiments.

### Idea 4 (Conditionality Principle)

*Let  $\mathcal{E}_1$  and  $\mathcal{E}_2$  be two experiments about  $\theta$ . Let  $Z \sim \text{Bernoulli}(p)$  and*

- If  $Z = 1$ , perform  $\mathcal{E}_1$  to generate  $x_1 \sim f_1(x_1 \mid \theta)$ ;*
- If  $Z = 0$  perform  $\mathcal{E}_2$  to generate  $x_2 \sim f_2(x_2 \mid \theta)$ .*

*Inferences about  $\theta$  should depend **only** on the selected experiment,  $\mathcal{E}_i$ .*

## Deriving the Likelihood Principle

Birnbaum (1962) showed that the simpler and mostly uncontroversial Sufficiency and Conditionality principles lead to the Likelihood Principle.

Theorem 2 (Birnbaum's theorem (Birnbaum, 1962))

$$SP + CP \implies LP. \quad (5)$$

Proof.

Sketch:

- Define a function  $EV(\mathcal{E}, x)$  to quantify the evidence about  $\theta$  brought by data  $x$  from experiment  $\mathcal{E}$  and consider a randomised experiment  $\mathcal{E}^*$  in which  $\mathcal{E}_1$  and  $\mathcal{E}_2$  are performed with probability  $p$ ;
- Show that CP implies  $EV(\mathcal{E}^*, (j, x_j)) = EV(\mathcal{E}_j, x_j), j = 1, 2$ ;
- Show that SP implies  $EV(\mathcal{E}^*, (1, x_1)) = EV(\mathcal{E}^*, (2, x_2))$  when

$$l(\theta \mid x_1) = cl(\theta \mid x_2).$$

□

See Robert (2007), pg.18 for a complete proof.

## Recommended reading

---

 Robert (2007) Ch. 1;

▶▶ Next lecture: Robert (2007) Ch. 2 and \* Schervish (2012) Ch.3;

## Belief functions

Let  $F, G$  and  $H \in \mathcal{S}$  be three (possibly overlapping) statements about the world. For example, consider the following statements about a person:

$F = \{\text{votes for a left-wing candidate}\};$

$G = \{\text{is in the 10\% lower income bracket}\};$

$H = \{\text{lives in a large}\};$

### Definition 4 (Belief function)

*For  $A, B \in \mathcal{S}$ , a belief function  $\text{Be} : \mathcal{S} \rightarrow \mathbb{R}$  assigns numbers to statements such that  $\text{Be}(A) < \text{Be}(B)$  implies one is more confident in  $B$  than in  $A$ .*

## Belief functions: properties

---

It is useful to think of  $\text{Be}$  as **preferences over bets**:

- $\text{Be}(F) > \text{Be}(G)$  means we would bet on  $F$  being true over  $G$  being true;
- $\text{Be}(F \mid H) > \text{Be}(G \mid H)$  means that, **conditional** on knowing  $H$  to be true, we would bet on  $F$  over  $G$ ;
- $\text{Be}(F \mid G) > \text{Be}(F \mid H)$  means that if we were forced to bet on  $F$ , we would be prefer doing so if  $G$  were true than  $H$ .



## Belief functions: axioms

---

In order for  $\text{Be}$  to be **coherent**, it must adhere to a certain set of properties/axioms. A self-sufficient collection is:

A1 (boundedness of complete [dis]belief):

$$\text{Be}(\neg H \mid H) \leq \text{Be}(F \mid H) \leq \text{Be}(H \mid H), \forall F \in \mathcal{S};$$

A2 (monotonicity):

$$\text{Be}(F \text{ or } G \mid H) \leq \max \{ \text{Be}(F \mid H), \text{Be}(G \mid H) \};$$

A3 (sequentiality): There exists  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  such that

$$\text{Be}(F \text{ and } G \mid H) = f(\text{Be}(G \mid H), \text{Be}(F \mid G \text{ and } H)).$$

### Exercise 1 (Probabilities and beliefs)

*Show that the axioms of belief functions map one-to-one to the axioms of probability:*

*P1.  $0 \leq \Pr(E), \forall E \in \mathcal{S}$ ;*

*P2.  $\Pr(\mathcal{S}) = 1$ ;*

*P3. For any countable sequence of disjoint statements  $E_1, E_2, \dots \in \mathcal{S}$  we have*

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i).$$

Hint: derive the consequences (e.g. monotonicity) of these axioms and compare them with the axioms of belief functions.

## Useful probability laws

### Definition 5 (Partition)

If  $H = \{H_1, H_2, \dots, H_K\}$ ,  $H_i \in \mathcal{S}$ , such that  $H_i \cap H_j = \emptyset$  for all  $i \neq j$  and  $\bigcup_{k=1}^K H_k = H$ , we say  $H$  is a partition of  $\mathcal{S}$ .

For any  $H \in \mathcal{D}(\mathcal{S})$ :

- **Total probability:**  $\sum_{k=1}^K \Pr(H_k) = 1$ ;
- **Marginal probability:**

$$\Pr(E) = \sum_{k=1}^K \Pr(E \cap H_k) = \sum_{k=1}^K \Pr(E | H_k) \Pr(H_k),$$

for all  $E \in \mathcal{S}$ ;

- Consequence  $\implies$  Bayes's rule:

$$\Pr(H_j | E) = \frac{\Pr(E | H_j) \Pr(H_j)}{\sum_{k=1}^K \Pr(E | H_k) \Pr(H_k)}.$$

# Independence

We will now state a central concept in probability theory and Statistics.

## Definition 6 ( (Conditional) Independence)

For any  $F, G \in \mathcal{S}$ , we say  $F$  and  $G$  are **conditionally independent** given  $A$  if

$$\Pr(F \cap G \mid A) = \Pr(F \mid A) \Pr(G \mid A).$$

## Remark 1

If  $F$  and  $G$  are conditionally independent given  $A$ , then

$$\Pr(F \mid A \cap G) = \Pr(F \mid A).$$

## Proof.

First, notice that the axioms P1-P3 imply  $\Pr(F \cap G \mid A) = \Pr(G \mid A) \Pr(F \mid A \cap G)$ . Now use conditional independence to write

$$\Pr(G \mid A) \Pr(F \mid A \cap G) = \Pr(F \cap G \mid A) = \Pr(F \mid A) \Pr(G \mid A),$$

$$\Pr(G \mid A) \Pr(F \mid A \cap G) = \Pr(F \mid A) \Pr(G \mid A).$$

### Definition 7 (Exchangeable)

We say a sequence of random variables  $Y = \{Y_1, Y_2, \dots, Y_n\}$  are **exchangeable** if

$$\Pr(Y_1, Y_2, \dots, Y_n) = \Pr(Y_{\xi_1}, Y_{\xi_2}, \dots, Y_{\xi_n}),$$

for all **permutations**  $\xi$  of the labels of  $Y$ .

### Example 4 (Uma vez Flamengo... continued)

Suppose we survey 12 people and record whether they cheer for Flamengo  $Y_i = 1$  or not  $Y_i = 0$ ,  $i = 1, 2, \dots, 12$ . What value should we assign to :

- $p_1 := \Pr(1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1)$ ;
- $p_2 := \Pr(1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1)$ ;
- $p_3 := \Pr(1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0)$ ?

If your answer is  $p_1 = p_2 = p_3$  then you are saying the  $Y_i$  are (at least partially) exchangeable!

## An application of conditional independence

For  $\theta \in (0, 1)$ , consider the following sequence of probability statements:

$$\begin{aligned}\Pr(Y_{12} = 1 \mid \theta) &= \theta, \\ \Pr(Y_{12} = 1 \mid Y_1, \dots, Y_{11}, \theta) &= \theta, \\ \Pr(Y_{11} = 1 \mid Y_1, \dots, Y_{10}, Y_{12}, \theta) &= \theta.\end{aligned}$$

These imply that the  $Y_i$  are conditionally independent and identically distributed (iid), and in particular:

$$\begin{aligned}\Pr(Y_1 = y_1, \dots, Y_{12} = y_{12} \mid \theta) &= \prod_{i=1}^{12} \theta^{y_i} (1 - \theta)^{1-y_i}, \\ &= \theta^S (1 - \theta)^{12-S},\end{aligned}$$

with  $S := \sum_{i=1}^{12} y_i$ . Also, under a uniform prior,

$$\Pr(Y_1, \dots, Y_{12}) = \int_0^1 t^S (1 - t)^{12-S} \pi(t) dt = \frac{(S+1)!(12-S+1)!}{13!} = \binom{13}{S+1}^{-1}.$$

## Relaxing exchangeability (a bit)

Sometimes total symmetry can be a burden. We can relax this slightly by introducing the concept of **partial exchangeability**:

### Definition 8 (Partially exchangeable)

Let  $\mathbf{X} = \{X_1, \dots, X_n\}$  and  $\mathbf{Y} = \{Y_1, \dots, Y_m\}$  be two sets of random variables. We say  $\mathbf{X}$  and  $\mathbf{Y}$  are **partially exchangeable** if

$$\Pr(X_1, \dots, X_n; Y_1, \dots, Y_m) = \Pr(X_{\xi_1}, \dots, X_{\xi_n}; Y_{\sigma_1}, \dots, Y_{\sigma_m}),$$

for any two permutations  $\xi$  and  $\sigma$  of  $1, \dots, n$  and  $1, \dots, m$ , respectively.

### Example 5 (Uma vez Flamengo...continued)

To see how exchangeability can be relaxed into partial exchangeability, consider  $\mathbf{X}$  and  $\mathbf{Y}$  as observations coming from populations from Rio de Janeiro and Ceará, respectively. If the covariate “state” were deemed to not matter, then we would have complete exchangeability.

## A statistically useful remark

### Remark 2 (Exchangeability from conditional independence)

Take  $\theta \sim \pi(\theta)$ , i.e., represent uncertainty about  $\theta$  using a probability distribution. If  $\Pr(Y_1 = y_1, \dots, Y_n = y_n \mid \theta) = \prod_{i=1}^n \Pr(Y_i = y_i \mid \theta)$ , then  $Y_1, \dots, Y_n$  are exchangeable.

### Proof.

Sketch: Use

- Marginalisation;
- Conditional independence;
- Commutativity of products in  $\mathbb{R}$ ;
- Definition of exchangeability.





## A fabulous theorem!

### Theorem 3 (De Finetti's theorem<sup>2</sup>)

If  $\Pr(Y_1, \dots, Y_n) = \Pr(Y_{\xi_1}, \dots, Y_{\xi_n})$  for all permutations  $\xi$  of  $1, \dots, n$ , then

$$\Pr(Y_1, \dots, Y_n) = \Pr(Y_{\xi_1}, \dots, Y_{\xi_n}) = \int_{\Theta} \Pr(Y_1, \dots, Y_n \mid t) \pi(t) dt, \quad (6)$$

for some choice of triplet  $\{\theta, \pi(\theta), f(y_i \mid \theta)\}$ , i.e., a parameter, a prior and a sampling model.

See Proposition 4.3 in [Bernardo and Smith \(2000\)](#) for a proof outline. Here we shall prove the version from [De Finetti \(1931\)](#).

---

<sup>2</sup>Technically, the theorem stated here is more general than the representation theorem proven by De Finetti in his seminal memoir, which concerned binary variables only.

## Consequences

This theorem has a few important implications, namely:

- $\pi(\theta)$  represents our beliefs about  $\lim_{n \rightarrow \infty} \sum_i (Y_i \leq c)/n$  for all  $c \in \mathcal{Y}$ ;
- $\{Y_1, \dots, Y_n \mid \theta \text{ are i.i.d}\} + \{\theta \sim \pi(\theta)\} \implies \{Y_1, \dots, Y_n \text{ are exchangeable for all } n\}$ ;
- If  $Y_i \in \{0, 1\}$ , we can also claim that:
  - ◊ If the  $Y_i$  are assumed to be independent, then they are distributed Bernoulli conditional on a random quantity  $\theta$ ;
  - ◊  $\theta$  has a prior measure  $\Pi \in \mathcal{P}((0, 1))$ ;
  - ◊ By the strong law of large numbers (SLLN),  $\theta = \lim_{n \rightarrow \infty} (\frac{1}{n} \sum_{i=1}^n Y_i)$ , so  $\Pi$  can be interpreted as a “belief about the limiting relative frequency of 1’s”.

## The soul of Statistics

---

As the exchangeability results above clearly demonstrate, being able to use conditional independence is a handy tool. More specifically, knowing on what to condition so as to make things exchangeable is key to statistical analysis.

### Idea 5 (Conditioning is the soul of Statistics<sup>3</sup>)


*Knowing on what to condition can be the difference between an unsolvable problem and a trivial one. When confronted with a statistical problem, always ask yourself “What do I know for sure?” and then “How can I create a conditional structure to include this information?”.*

---

<sup>3</sup>This idea is due to Joe Blitzstein, who did his PhD under no other than the great Persi Diaconis.

## Recommended reading

---

 Hoff (2009) Ch. 2 and \*Schervish (2012) Ch.1;

- \*Paper: Diaconis and Freedman (1980) explains why if  $n$  samples are taken from an exchangeable population of size  $N \gg n$  without replacement, then the sample  $Y_1, \dots, Y_n$  can be modelled as approximately exchangeable;

►► Next lecture: Robert (2007) Ch. 3.

## Priors: a curse and a blessing

- Priors are the main point of contention between Bayesians and non-Bayesians;
- As we shall see, there is usually no unique way of constructing a prior measure;
- Moreover, in many situations the choice of prior is not inconsequential.
- There is always a question of when to stop adding uncertainty...



## Determination of priors: existence

It is usually quite hard to determine a (unique) prior even when substantial knowledge. Why? One reason is that a prior measure is guaranteed to exist only when there is a **coherent ordering** of the Borel sigma-algebra  $\mathcal{B}(\Theta)$ . This entails that the following axioms hold:

(A1) Total ordering: For all measurable  $A, B \in \mathcal{B}(\Theta)$  one and only one of these can hold:

$$A < B, B < A \text{ or } A \sim B.$$

(A2) Transitivity: For measurable  $A_1, A_2, B_1, B_2 \in \mathcal{B}(\Theta)$  such that  $A_1 \cap A_2 = \emptyset = B_1 \cap B_2$  and  $A_i \leq B_i, i = 1, 2$  then the following holds:

$$\diamond A_1 \cup A_2 \leq B_1 \cup B_2;$$

$$\diamond \text{ If } A_1 < B_1 \text{ then } A_1 \cup A_2 < B_1 \cup B_2;$$

(A3) For any measurable  $A, \emptyset \leq A$  and also  $\emptyset < \Theta$ ;

(A4) Continuity: If  $E_1 \supset E_2 \dots$  is a decreasing sequence of measurable sets and  $B$  is such that  $B \leq E_i$  for all  $i$ , then

$$B \leq \bigcap_{i=1}^{\infty} E_i.$$

## Approximation I: marginalisation

---

One way to approach the problem of determining a prior measure is to consider the marginal distribution of the data:

$$m(x) = \int_{\Theta} f(x | \theta) \pi(\theta) d\theta. \quad (7)$$

In other words we are trying to solve an inverse problem in the form of an integral equation by placing restrictions on  $m(x)$  and calibrating  $\pi$  to satisfy them.

## Approximation II: moments

Another variation on the integral-equation-inverse-problem theme is to consider expectations of measurable functions. Suppose

$$E_{\pi}[g_k] := \int_{\Theta} g_k(t) \pi(t) dt = w_k. \quad (8)$$

For instance, if the analyst knows that  $E_{\pi}[\theta] = \mu$  and  $\text{Var}_{\pi}(\theta) = \sigma^2$ , then this restricts the class of functions in  $\mathcal{L}_1(\Theta)$  that can be considered as prior density<sup>4</sup>. One can also consider *order statistics* by taking  $g_k(x) = \mathbb{I}_{(-\infty, a_k]}(x)$ .

---

<sup>4</sup>As we shall see in the coming lectures,  $\pi$  needs not be in  $\mathcal{L}_1(\Theta)$ , i.e., needs not be **proper**. But this “method-of-moments” approach is then complicated by lack of integrability.



## Maximum entropy priors

The moments-based approach is not complete in the sense that it does not lead to a unique prior measure  $\pi$ .

### Definition 9 (Entropy)

*The entropy of a probability distribution  $P$  is defined as*

$$H(P) := E_P[-\log p] = - \int_{\mathcal{X}} \log p(x) dP(x). \quad (9)$$

When  $\theta$  has finite support, we get the familiar

$$H(P) = - \sum_i p(\theta_i) \log(p(\theta_i)).$$

We can leverage this concept in order to pick  $\pi$ .

### Definition 10 (Maximum entropy prior)

*Let  $\mathcal{P}_r$  be a class of probability measures on  $\mathcal{B}(\Theta)$ . A maximum entropy prior in  $\mathcal{P}_r$  is a distribution that satisfies*

$$\arg \max_{P \in \mathcal{P}_r} H(P).$$

When  $\Theta$  is finite, we can write

$$\pi^*(\theta_i) = \frac{\exp \{ \sum_{k=1} \lambda_k g_k(\theta_i) \}}{\sum_j \exp \{ \sum_{k=1} \lambda_k g_k(\theta_j) \}},$$

where the  $\lambda_k$  are Lagrange multipliers. In the uncountable case things are significantly more delicate, but under regularity conditions there exists a reference measure  $\Pi_0$  such that

$$\begin{aligned} H_\Pi &= E_{\pi_0} \left[ \log \left( \frac{\pi(\theta)}{\pi_0(\theta)} \right) \right], \\ &= \int_{\Theta} \log \left( \frac{\pi(\theta)}{\pi_0(\theta)} \right) \Pi_0(d\theta). \end{aligned}$$

### Exercise 2 (Maximum entropy Beta prior)

*Find the maximum entropy Beta distribution under the following constraints:*

- $E[\theta] = 1/2$ ;
- $E[\theta] = 9/10$ .

**Hint:** If  $P$  is a Beta distribution with parameters  $\alpha$  and  $\beta$ , then

$$H_P = \log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta),$$

where  $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$  is the Beta function and  $\psi(x) = \frac{d}{dx} \log(\Gamma(x))$  is the digamma function.

## Parametric approximations: easy-peasy

In some situations, the “right” parametric family presents itself naturally.

### Example 6 (Eliciting Beta distributions)

Let  $x_i \sim \text{Binomial}(n_i, p_i)$  be the number of Flamengo supporters out of  $n_i$  people surveyed. Over the years, the average of  $p_i$  has been 0.70 with variance 0.1. If we assume  $p_i \sim \text{Beta}(\alpha, \beta)$  we can elicit an informative distribution based on historical data by solving the system of equations

$$E[\theta] = \frac{\alpha}{\alpha + \beta} = 0.7,$$

$$\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.1.$$

## Parametric approximations: difficulties

Other times we may have a hard time narrowing down the prior to a specific parametric family. Consider the following example.

### Example 7 (Normal or Cauchy?)

Suppose  $x_i \sim \text{Normal}(\theta, 1)$  and we are informed that  $\Pr(\theta \leq -1) = 1/4$ ,  $\Pr(\theta \leq 0) = 1/2$  and  $\Pr(\theta \leq 1) = 3/4$ . Seems like plenty of information. It can be shown that

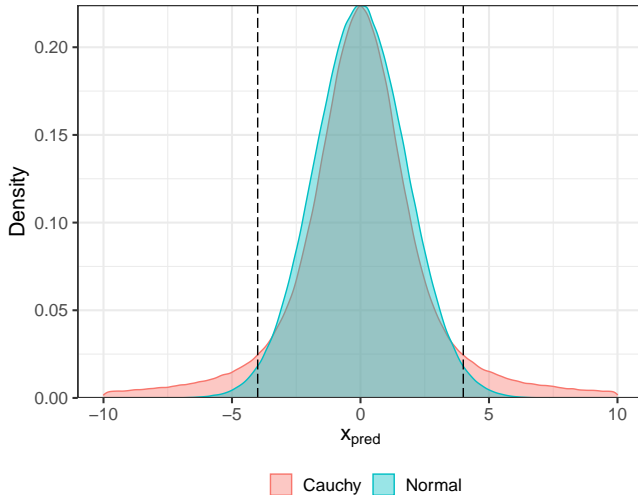
$$\pi_1(\theta) = \frac{1}{\sqrt{2\pi}2.19} \exp\left(-\frac{\theta^2}{2 \times 2.19}\right) \text{ (Normal),}$$

$$\pi_2(\theta) = \frac{1}{\pi(1 + \theta^2)} \text{ (Cauchy),}$$

both satisfy the requirements. Unfortunately, under quadratic loss we get  $\delta_1(4) = 2.75$  and  $\delta_2(4) = 3.76$  and differences are exacerbated for  $|x| \geq 4$ .

## Why, though?

Remember the marginal approach? It is illuminating in this case. Heres  $m(x)$ :



Prior predictive distributions of  $x$  under Normal and Cauchy priors.

# Conjugacy

Conjugacy is a central concept in Bayesian statistics. It provides a functional view of the prior-posterior mechanic that emphasises tractability over coherence.

## Definition 11 (Conjugate)

*A family  $\mathcal{F}$  of distributions on  $\Theta$  is called **conjugate** or closed under sampling for a likelihood  $f(x | \theta)$  if, for every  $\pi \in \mathcal{F}$ ,  $p(\theta | x) \in \mathcal{F}$ .*

## Arguments for using conjugate priors

- “Form-preservation”: in a limited-information setting it makes sense that  $p(\theta | x)$  and  $\pi(\theta)$  lie on the same family, since the information in  $x$  might not be enough to change the structure of the model, just its parameters;
- Simplicity: when you do not know a whole lot, it makes sense to KISS<sup>5</sup>;
- Sequential learning: since  $\mathcal{F}$  is closed under sampling, one can update a sequence of posteriors  $p_i(\theta | x_1, \dots, x_i)$  as data comes in.

---

<sup>5</sup>Keep it simple, stupid!

## Exponential families

The exponential family of distributions is a cornerstone of statistical practice, underlying many often-used models. Here are a few useful definitions.

### Definition 12 ((Natural) Exponential family)

Let  $\mu$  be a  $\sigma$ -finite measure on  $\mathcal{X}$  and let  $\Theta$  be a non-empty set serving as the parameter space. Let  $C : \Theta \rightarrow (0, \infty)$  and  $h : \mathcal{X} \rightarrow (0, \infty)$  and let  $R : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^k$  and  $T : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^k$ . The family of distributions with density

$$f(x \mid \theta) = C(\theta)h(x) \exp(R(\theta) \cdot T(x))$$

w.r.t.  $\mu$  is called an **exponential family**. Moreover, if  $R(\theta) = \theta$ , the family is said to be **natural**.

### Definition 13 (Regular exponential family)

We say a natural exponential family  $f(x \mid \theta)$  is **regular** if the natural parameter space

$$N := \left\{ \theta : \int_{\mathcal{X}} \exp(\theta \cdot x) h(x) d\mu(x) < \infty \right\}, \quad (10)$$

is an open set of the same dimension as the closure of the convex hull of  $\text{supp}(\mu)$ .



## Conjugacy and sufficiency

There is an intimate link between sufficiency (i.e. the existence of sufficient statistics) and conjugacy. The following is a staple of Bayesian theory.

### Theorem 4 (Pitman-Koopman-Darmois)

*If a family of distributions  $f(\cdot | \theta)$  whose support does not depend on  $\theta$  is such that, for a sample size large enough, there exists a sufficient statistic of fixed dimension, then  $f(\cdot | \theta)$  is an exponential family.*

The support condition is not a complete deal breaker, however:

### Remark 3 (Quasi-exponential)

*The  $\text{Uniform}(-\theta, \theta)$  and  $\text{Pareto}(\theta, \alpha)$  families are called quasi-exponential due to the fact that there do exist sufficient statistics of fixed dimension for these families, even though their supports depend on  $\theta$ .*

## Conjugacy in the exponential family

I hope you are convinced of the utility of the exponential family by now. It would be nice to have an automated way to deduce a conjugate prior for  $f(x | \theta)$  when it is in the exponential family. This is exactly what the next result gives us.

### Remark 4 (Conjugate prior for the exponential family)

*A conjugate family for  $f(x | \theta)$  is given by*

$$\pi(\theta | \mu, \lambda) = K(\mu, \lambda) \exp(\theta \cdot \mu - \lambda g(\theta)), \quad (11)$$

*such that the posterior is given by  $p(\theta | \mu + x, \lambda + 1)$ .*

Please do note that (11) is only a valid density when  $\lambda > 0$  and  $\mu/\lambda$  belongs to the interior of the natural space parameter. Then, it is a  $\sigma$ -finite measure. See [Diaconis and Ylvisaker \(1979\)](#) for more details.

## Conjugacy: common families

Table 3.3.1. *Natural conjugate priors for some common exponential families*

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative Binomial $\mathcal{N}eg(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \dots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}a(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Taken from Robert (2007), page 121.

## Conjugacy: drawbacks

---

Conjugate modelling is certainly useful, but has its fair share of pitfalls.

### Arguments against using conjugate priors

- Conjugate priors are restrictive *a priori*: in many settings, specially in high dimensions, the set of conjugate priors that retain tractability is so limited so as to not be able to encode all prior information available;
- Conjugate priors are not truly subjective: they limit the analyst's input to picking values for the hyperparameters;
- Conjugate priors are restrictive *a posteriori*: you are stuck with a given structure forever, no matter how much data you run into.

## Recommended reading

---

 Robert (2007) Ch. 3;

►► Next lecture: Robert (2007) Ch. 3.6, Seaman III et al. (2012), Gelman et al. (2017) and Simpson et al. (2017).

## References

---

- Bernardo, J. M. and Smith, A. F. (2000). *Bayesian Theory*. John Wiley & Sons.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.
- De Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. In *Atti della R Accademia Nazionale dei Lincei*, volume 4, pages 251–299.
- Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.
- Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, pages 269–281.
- Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555.
- Gonçalves, F. B. and Franklin, P. (2019). On the definition of likelihood function. *arXiv preprint arXiv:1906.10733*.
- Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

## References

---

- Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.
- Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.
- Seaman III, J. W., Seaman Jr, J. W., and Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, 66(2):77–84.
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28.