BAYESIAN STATISTICAL MODELS

Chapter 1 - Introduction

Lecturer: Riccardo Corradin University of Milano-Bicocca

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.

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 idependent 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.

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:

$a)$ notes, containing the methodological part;

$b)$ code, usually with synthetic examples;

$c)$ case studies, presenting real data analysis.

1 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. Hence, is exactly describing what happens, without any uncertainty. The second class of models involves stochastic terms which introduce uncertainty and randomness, and require specific mathematical tools.
Probabilistic models are nothing but distributional assumptions combined with a structural part. For example, our dear linear model with Gaussian distributed error term has

- a structural part (linear predictor);
- a distributional assumption (Gaussian error).

Usually the structural part of the model is parametrized, for example by a parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, determining the behavior of our model. Specific choices of the parameters support depends on which model we are considering. For example, if we are interested into model the location of a Gaussian distribution, $\Theta \equiv \mathbb{R}$, while if we want to model its variance $\Theta \equiv \mathbb{R}_+$.

Regarding how to proceed to build up our models, there are two main approaches that we can set. These equal to make specific assumption on our observations but also on the model structure, ad define the two macro areas of statistics: frequentist and Bayesian approaches.

- In a frequentist approach, our probabilistic model is expressed by setting a distribution on our data, say $\boldsymbol{Y} \in \mathbb{Y} \subseteq \mathbb{R}^d$, like
$$\boldsymbol{Y} \sim f(\boldsymbol{y} \mid \boldsymbol{\theta}),$$
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 $\boldsymbol{\theta}$ that better describe the observed data.
- 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[1]
$$(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n, \boldsymbol{\theta}) \sim \mathcal{L}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n, \boldsymbol{\theta}).$$
We can exploit the chain rule, rewriting the previous distribution as
$$\mathcal{L}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n, \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$
A Bayesian model is composed by a (prior), i.e., distributional assumption for the parameter, here denoted by $\pi(\boldsymbol{\theta})$, and another distributional assumption for the data $\mathcal{L}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \mid \boldsymbol{\theta})$. We also assume conditional independence of the data, i.e., given the parameter $\boldsymbol{\theta}$, data factorize as
$$\mathcal{L}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i \mid \boldsymbol{\theta}),$$
which says that the shared information among distinct observations is fully driven by the parameter $\boldsymbol{\theta}$.

2 Bayes Theorem

The Bayes Theorem is a fundamental result in probability and mathematical statistics, which can be exploited to update our prior knowledge conditioning on empirical information. In its simple version, it can be written as follows.

Theorem 2.1. *Let E be an event contained in $F_1 \cup \cdots \cup F_t$, 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*
$$\Pr(F_j \mid E) = \frac{\Pr(E \mid F_j)\Pr(F_j)}{\sum_{j=1}^{t} \Pr(E \mid F_j)\Pr(F_j)}. \tag{1}$$

[1] $\mathcal{L}$ denotes a generic distributional law and will be used consistently in the module

Ideally, we have a prior opinion on a set of possible events $\{F_1, \ldots, F_t\}$. 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, $j = 1, \ldots, t$. Thanks to the Bayesian rule described in Equation (1), we can update our belief conditioned on the information driven by the event E.

Proof. We first note that

$$\sum_{j=1}^{t} P(E \mid F_j)P(F_j) = \sum_{j=1}^{t} P(E \cap F_j) = P(E),$$

given that $F_i \cap F_j = \emptyset$, for every $i \neq j$. Then we have

$$P(F_j \cap E) = P(F_j \mid E)P(E)$$

and

$$P(F_j \cap E) = P(E \mid F_j)P(F_j).$$

Therefore,

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

□

We recall and define the following quantities, which will be used consistently during the notes.

- $\pi(\boldsymbol{\theta})$ is the prior distribution, i.e. a probability mass function or a density function which express our prior belief on the parameter space $\Theta \subseteq \mathbb{R}^p$. It measures and describes our knowledge on the parameter θ without having observed any data.
- $\mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta})$ is the likelihood function, where $\boldsymbol{y}_{1:n} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \in \mathbb{Y} \subseteq \mathbb{R}^d$, which measures how likely is a specific value of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}_{1:n}$. This is the empirical measure introduced in the mode. For any value of $\boldsymbol{\theta}$, it provides a quantification of how likely is the sample we observed.

Similarly to Equation (1), modern Bayesian approaches found on expressing our posterior belief over Θ by updating our prior belief $\pi(\boldsymbol{\theta})$, conditioning on the information coming through the observed sample $\boldsymbol{y}_{1:n}$. The latter is measured resorting to the likelihood function. Hence, the Bayes theorem can be rewritten in terms of pmf/df as in the following result.

Theorem 2.2. *Let $\boldsymbol{y}_{1:n}$ an observed sample and $\boldsymbol{\theta}$ a parameter of interest. Let $\pi(\boldsymbol{\theta})$ be a distribution expressing our prior guess over Θ and $\mathrm{L}(\boldsymbol{\theta} \mid \boldsymbol{y}_{1:n})$ the likelihood function. Then*

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

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

Remark 2.3. The posterior distribution is proportional to the likelihood times the prior

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

Quite often, posterior inference on $\boldsymbol{\theta}$ can be done just looking at our prior guess and the empirical information. Hence, we can omit the normalization constant when not needed.

Exercise 2.4. 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?

Exercise 2.5. 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 θ 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 $\pi(\theta \mid y_1, \dots, y_n)$?

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

One of the most celebrated theorem in mathematical statistics and probability is the de Finetti representation theorem. Suppose that we have a set of observations $Y_{1:n}^{\intercal} = (Y_1, \dots, Y_n)$. Furthermore, the observations are assumed to be exchangeable. Then the following holds true.

Theorem 3.1 (De Finetti, 1937)**.** *The sequence $Y_{1:n}$ is exchangeable if and only if there exists a probability measure Q such that, for $A = A_1 \times \cdots \times A_n$, we have*

$$\Pr(Y_{1:n} \in A) = \int_{\Theta} \prod_{i=1}^{n} \Pr(y_i \in A_i \mid \theta) Q(\mathrm{d}\theta).$$

In practice, exchangeability assumption means that the order we observe our sample does not affect our inference. 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.

Exercise 3.2. We consider the case with Bernoulli distributed observations. Think about the following sampling scheme.

- We start from an urn with balls of two colors, R red balls and B blue balls.
- We sample a ball from the urn. We look at its color, and we replace the ball in the urn with

one more ball of the same color.

- We repeat the previous process.

The previous scheme correspond to the marginal distribution of a Bernoulli sample with a suitable prior on the success probability. Show the followings.

- The sequence is exchangeable.
- The probability measure Q is a $Beta(R, B)$ distribution.
- The representation of the previous theorem holds.

4 PREDICTING THE FUTURE

Producing inference about unknown observable quantities, i.e. predictive inference, is one of the fundamental tasks a statistician should do while doing inference. Luckily, it is quite natural in a Bayesian setting. The distribution of a generic unknown (but observable) y is given by

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

After we collect n observations, we might be interested to study the distribution of $\boldsymbol{y}_{n+1}$ given our prior guess updated by $\boldsymbol{y}_{1:n}$. Hence, once conditioning on the previous sample, at the sampling step $n+1$ we have

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

where the last step comes from the conditional independence. We can easily perform predictive inference by simply averaging with respect to the posterior distribution. Furthermore,

$$\mathcal{L}(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}_{1:n})$$

is called predictive distribution or, more precisely, posterior predictive distribution.

5 CHOOSING THE PRIOR DISTRIBUTION

The prior distribution expresses our belief on the parameter space Θ. Clearly, different distributional assumption resemble different belief. For example, within the same distributional family, we might have a prior belief more or less dispersed over Θ. Similarly, we can have a truncated distribution because we know that some values are not plausible. In general, if we have a specific belief on our parameter space, it can be formalized mathematically as a distributional choice. On the counterpart, there are some choices that are more vague and reduce the impact of our prior guess on our posterior inference.

Ideally, there are two main properties that play a crucial role in the prior specification, and we want to choose distributions that relate to them.

- Conjugacy.
- Informativeness.

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

5.1 CONJUGATE PRIOR DISTRIBUTIONS

Conjugate distributions are prior choices known for their tractability. Hence, doing posterior inference with our model becomes quite simple. We consider two distributional families that plays a crucial role.

- The class of sampling distribution $\mathcal{F}$, with $f(\boldsymbol{y} \mid \boldsymbol{\theta}) \in \mathcal{F}$, which are the possible distribution describing a data generating process, indexed by a parameter $\boldsymbol{\theta}$.
- The class of prior distribution $\mathcal{P}$, with $\pi(\boldsymbol{\theta}) \in \mathcal{P}$, which are the possible distribution a priori for the parameter indexing the data generating process.

We can define conjugacy as in the following.

Definition 5.1. We say that a class of prior distribution $\mathcal{P}$ is conjugate for a class of sampling distribution $\mathcal{F}$ if

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

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

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

Hence, if $\mathcal{P}$ is conjugate for a single $f(\boldsymbol{y} \mid \boldsymbol{\theta})$, then is conjugate for the likelihood. Conjugacy has a trivial interpretation. Since we stay in the same distributional family, to perform posterior inference we simply have to update the parameters of π.

For the exponential families something nice happens. Recall that a distribution belongs to the exponential family if can be expressed as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\{-\phi(\boldsymbol{\theta})^{\intercal} t(\boldsymbol{y})\},$$

where $t(\boldsymbol{y})$ is a vector of sufficient statistics. Suppose also that the prior distribution is an exponential family as well, of the form

$$\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta})^{\nu} \exp\{-\phi(\boldsymbol{\theta})^{\intercal} \boldsymbol{\nu}\}.$$

Hence, the posterior distribution can be expressed as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{1:n}) \propto g(\boldsymbol{\theta})^{\eta+n} \exp\left\{-\phi(\boldsymbol{\theta})^{\intercal} \left(\boldsymbol{\nu} + \sum_{j=1}^{n} t(\boldsymbol{y}_i)\right)\right\}.$$

In practice, for exponential families we can find a conjugate prior distribution, which gives us a simple expression.

Exercise 5.2. Let Y be distributed as an Exponential distribution with

$$f(y \mid \theta) = \theta e^{\theta y}.$$

Using the previous relation of for the exponential families, show that the Gamma distribution is a conjugate prior for the exponential family.

5.2 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.

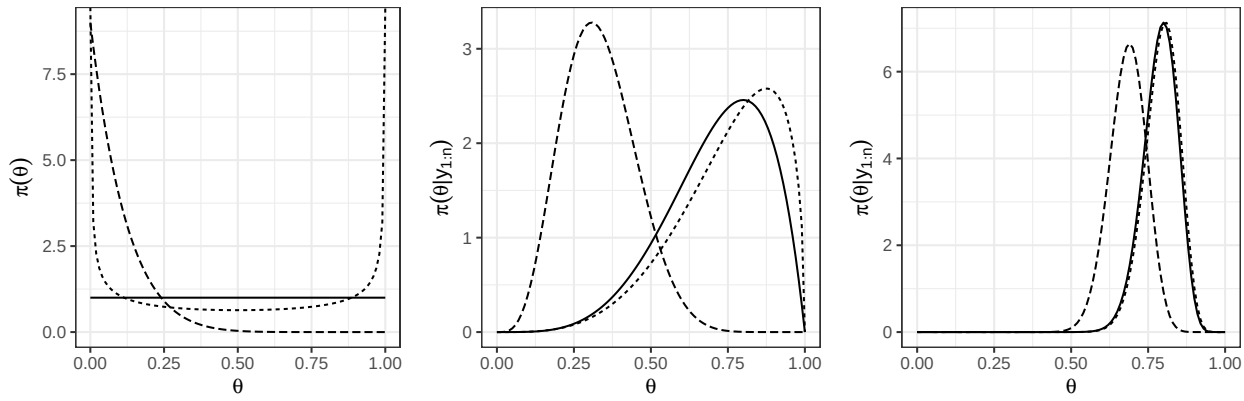

Figure 1: From prior to posterior, with three different prior distributions. Dashed line, informative. Full line, weakly informative. Dotted line, noninformative. Left plot: prior distribution. Middle plot: posterior with 5 observations. Right plot: posterior with 25 observations.

5.3 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 $\boldsymbol{\lambda} = q(\boldsymbol{\theta})$, such that the prior can be expressed with a change of variable as

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

our prior guess should be invariant with respect of such transformation. Jeffrey suggests to specify a prior starting from the Fisher information of $\boldsymbol{\theta}$

$$I(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\mathrm{d}\log f(\boldsymbol{y} \mid \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right)\left(\frac{\mathrm{d}\log f(\boldsymbol{y} \mid \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right)^{\intercal}\Big|\boldsymbol{\theta}\right] = \mathbb{E}\left[\frac{\mathrm{d}^2\log f(\boldsymbol{y} \mid \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|\boldsymbol{\theta}\right].$$

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

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

Remark 5.3. some remarkable examples of Jeffreys prior give an improper distribution, i.e.

$$\int_{\Theta} \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = +\infty.$$

Nevertheless, also in this cases the posterior can still be a proper distribution.

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

$$\pi_J(\boldsymbol{\lambda}) = \pi_J(\boldsymbol{\theta})\Big|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\lambda}}\Big|.$$

Here, $\left|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\lambda}}\right|$ denotes the determinant of the transformation Jacobian $A_{\boldsymbol{\lambda}}$. Hence, for the Fisher information matrix, we have

$$I(\boldsymbol{\lambda}) = A_{\boldsymbol{\lambda}}^{\intercal}\, I(\boldsymbol{\theta})\, A_{\boldsymbol{\lambda}},$$

and

$$\begin{aligned}\pi_J(\boldsymbol{\lambda}) \propto \sqrt{|I(\boldsymbol{\lambda})|} &= \sqrt{|A_{\boldsymbol{\lambda}}^{\intercal}\, I(\boldsymbol{\theta})\, A_{\boldsymbol{\lambda}}|} \\ &= \sqrt{|A_{\boldsymbol{\lambda}}|^2\, |I(\boldsymbol{\theta})|} = |A_{\boldsymbol{\lambda}}|\, \sqrt{|I(\boldsymbol{\theta})|} \propto \pi_J(\boldsymbol{\theta}).\end{aligned}$$

Exercise 5.4. 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, find the Jeffreys prior for σ^2.

6 Point estimates, credible regions and tests

Suppose we set up our prior specification. We observe some data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. We update our prior belief. Now we want to perform some inference, and summarize somehow our posterior belief, expressed here as a distribution

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

Commonly, in statistical analysis we want to tackle three main tasks with our inference.

- Point estimates, we want to summarize our posterior guess with a single value belonging to the parameters support which is representative of our updated belief. This value correspond to the best estimate of the parameter, with respect to some criteria.

- Intervals, we want to provide a range of values which are plausible given our updated belief. As summarizing our posterior belief in a single value can be insufficient, we want to identify a subset of the parameter space Θ of plausible values.
- Tests, we want to use our updated belief to answer specific inferential questions, expressed in terms of hypotheses.

6.1 POINT ESTIMATES

A natural viewpoint to present point estimate in Bayesian framework is through decision theory. Suppose we have a loss function

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

where $\mathcal{A}$ is a set of possible actions and Θ is the parameter space. Such a function quantifies the loss we are committing by choosing an action $\boldsymbol{a}$ when the parameter value is $\boldsymbol{\theta}$. Ideally, our optimal action $\boldsymbol{a}$ is minimizing the loss we are committing, under the parameter θ. Further, such a loss should be minimized for any value of $\boldsymbol{\theta}$.

Since we do not have a single value of $\boldsymbol{\theta}$, but a distribution over its support, we can take suitable functional of the previous loss function. Specifically, the loss function can be averaged either a priori

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

or a posteriori

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

hence resulting in the expected loss we are committing by choosing a specific action $\boldsymbol{a}$, either a priori or a posteriori.

We can extend this idea to derive a point estimate strategy. Ideally, the possible actions we can take are choosing specific values for the parameter $\boldsymbol{\theta}$. Hence, we set the action space equal to the parameter space, $\mathcal{A} \equiv \Theta$. Ideally, we can find the optimal point of $\mathcal{A}$ with respect to the previous functionals.

Proposition 6.1. *The point estimate under the loss function R is the value minimizing the loss while is averaged with respect to all the possible parameter choices, i.e. a priori we have*

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}^* \in \Theta}{\operatorname{argmin}} \Big\{ \mathbb{E}\left[R(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \right] \Big\},$$

while a posteriori we have

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}^* \in \Theta}{\operatorname{argmin}} \Big\{ \mathbb{E}\left[R(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_n \right] \Big\}.$$

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

Exercise 6.2. Let $R(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ be a quadratic loss function, i.e.

$$R(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = (\boldsymbol{\theta}^* - \boldsymbol{\theta})^\intercal(\boldsymbol{\theta}^* - \boldsymbol{\theta}).$$

Show that the point estimate with such a loss function is the (prior or posterior) expectation.

6.2 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 $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{1:n})$, the definition of interval estimates within the Bayesian framework is quite natural.

Definition 6.3. The region C_α is a $100(1-\alpha)\%$ credible interval (or Bayesian credible region) for $\boldsymbol{\theta}$ if

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

However such definition is not unique, as it is possible to define different strategies to derive C_α, 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_α. 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 $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{1:n})$.

Definition 6.4. The region C_α is a $100(1-\alpha)\%$ highest (posterior) density interval for $\boldsymbol{\theta}$ if

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

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

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.

Remark 6.5. 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.

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

Definition 6.6. The region C_α is a $100(1-\alpha)\%$ equally tailed interval for $\boldsymbol{\theta}$ if

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

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.

6.3 TESTS

Testing hypotheses is broadly used to support or refute some opinion on a phenomena of interest, specified as a partition of the parameter space Θ 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

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \qquad vs \qquad H_1 : \boldsymbol{\theta} \in \Theta_1,$$

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 6.7. The Bayes factor can be defined as

$$\mathrm{BF}_{01} = \frac{\text{posterior odds}_{01}}{\text{prior odds}_{01}} = \frac{\frac{\Pr(\Theta_0|\boldsymbol{y}_{1:n})}{\Pr(\Theta_1|\boldsymbol{y}_{1:n})}}{\frac{\Pr(\Theta_0)}{\Pr(\Theta_1)}} = \frac{\Pr(\Theta_0 \mid \boldsymbol{y}_{1:n})}{\Pr(\Theta_1 \mid \boldsymbol{y}_{1:n})}\frac{\Pr(\Theta_1)}{\Pr(\Theta_0)},$$

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 incorporasted in the testing procedure. In practice, BF_{01} is measuring how much the empirical information is shifting our guess toward H_0, but adjusting by the prior guess. Note that

$$\mathrm{BF}_{10} = \frac{\Pr(\Theta_1 \mid \boldsymbol{y}_{1:n})}{\Pr(\Theta_0 \mid \boldsymbol{y}_{1:n})}\frac{\Pr(\Theta_0)}{\Pr(\Theta_1)} = \frac{1}{\frac{\Pr(\Theta_0|\boldsymbol{y}_{1:n})}{\Pr(\Theta_1|\boldsymbol{y}_{1:n})}\frac{\Pr(\Theta_1)}{\Pr(\Theta_0)}} = \frac{1}{\mathrm{BF}_{01}}.$$

Remark 6.8. About the interpretation of Bayes factors, there are some general guidelines based on the observed value or its log-transformation.

BF_{01}	$\log \mathrm{BF}_{01}$	evidence
<1	<0	negative
1–3	0–2	weakly positive
3–12	2–5	positive
12–150	5–10	strongly positive
>150	>10	very stronly positive

Further, the Bayes factor can be rewritten as

$$\mathrm{BF}_{01} = \frac{\Pr(\Theta_0 \mid \boldsymbol{y}_{1:n})}{\Pr(\Theta_1 \mid \boldsymbol{y}_{1:n})} \frac{\Pr(\Theta_1)}{\Pr(\Theta_0)} = \frac{\Pr(\Theta_0, \boldsymbol{y}_{1:n}) \Pr(\boldsymbol{y}_{1:n})}{\Pr(\Theta_1, \boldsymbol{y}_{1:n}) \Pr(\boldsymbol{y}_{1:n})} \frac{\Pr(\Theta_1)}{\Pr(\Theta_0)} = \frac{\Pr(\boldsymbol{y}_{1:n} \mid \Theta_0)}{\Pr(\boldsymbol{y}_{1:n} \mid \Theta_1)},$$

hence, we can work with the marginal distribution of the data under different hypotheses. The Bayes factor can be generalize to the model comparison case. Suppose we have two models M_0 and M_1. Then we can perform a comparison with

$$\mathrm{BF}_{01} = \frac{\Pr(\boldsymbol{y}_{1:n} \mid M_0)}{\Pr(\boldsymbol{y}_{1:n} \mid M_1)}.$$

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

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad vs \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Suppose now we have a diffuse prior distribution $\pi(\boldsymbol{\theta})$. We have a problem. Under such a prior assumption,

$$P(\boldsymbol{\theta} = \boldsymbol{\theta}_0) = 0,$$

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

Proposition 6.9. *Let $\pi(\boldsymbol{\theta})$ be a diffuse prior over Θ. Suppose we want to test an atomic versus a diffuse hypothesis, $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Define a new prior*

$$\pi_1(\boldsymbol{\theta}) = \beta_0 \delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) + (1 - \beta_0)\pi(\boldsymbol{\theta}),$$

where β_0 is the prior probability associated with H_0. Then

$$\mathrm{BF}_{01} = \frac{\mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)}{m(\boldsymbol{y}_{1:n})}$$

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

Proof. With the prior $\pi_1(\boldsymbol{\theta})$ we have

$$\begin{aligned} m_1(\boldsymbol{y}_{1:n}) &= \int_\Theta \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}) \pi_1(\mathrm{d}\boldsymbol{\theta}) \\ &= \beta_0 \int_\Theta \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}) \delta_{\boldsymbol{\theta}_0}(\mathrm{d}\boldsymbol{\theta}) + (1 - \beta_0) \int_\Theta \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}) \pi(\mathrm{d}\boldsymbol{\theta}) \\ &= \beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0) + (1 - \beta_0) m(\boldsymbol{y}_{1:n}), \end{aligned}$$

where $\mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)$ corresponds to the likelihood function evaluated at $\boldsymbol{\theta}_0$, i.e., the marginal under the prior $\delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, and $m(\boldsymbol{y}_{1:n})$ to the marginal under $\pi(\boldsymbol{\theta})$. A posteriori, we have

$$P(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \boldsymbol{y}_{1:n}) = \frac{\int_{\boldsymbol{\theta}_0} \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta})\pi_1(\mathrm{d}\boldsymbol{\theta})}{m_1(\boldsymbol{y}_{1:n})} = \frac{\beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)}{\beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0) + (1-\beta_0) m(\boldsymbol{y}_{1:n})}.$$

Therefore, the Bayes factor equals

$$\begin{aligned}
\mathrm{BF}_{01} &= \frac{\text{posterior odds}_{01}}{\text{prior odds}_{01}} \\
&= \left(\frac{1-\beta_0}{\beta_0}\right) \frac{P(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \mid \boldsymbol{y}_{1:n})}{P(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \mid \boldsymbol{y}_{1:n})} \\
&= \left(\frac{1-\beta_0}{\beta_0}\right) \frac{\frac{\beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} | \boldsymbol{\theta}_0)}{m_1(\boldsymbol{y}_{1:n})}}{1 - \frac{\beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} | \boldsymbol{\theta}_0)}{m_1(\boldsymbol{y}_{1:n})}} \\
&= \left(\frac{1-\beta_0}{\beta_0}\right) \frac{\beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)}{m_1(\boldsymbol{y}_{1:n}) - \beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)} \\
&= (1-\beta_0) \frac{\mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)}{(1-\beta_0) m(\boldsymbol{y}_{1:n}) + \beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0) - \beta_0 \mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)} \\
&= \frac{\mathrm{L}(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}_0)}{m(\boldsymbol{y}_{1:n})}.
\end{aligned}$$

□

Exercise 6.10. Let us consider a model

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

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

$$H_0 : \theta = 1 \qquad vs \qquad H_1 : \theta \neq 1.$$

iii) Choose the parameters of the prior distribution that guarantee

$$\mathbb{E}[\theta] = 1, \qquad \mathrm{var}(\theta) = 10.$$

Perform the test. Do you prefer H_0 or H_1?