

Statistics and Learning

Lecturers (alphabetic order):

Julien Bect, Gilles Faÿ, Ziad Kobeissi, Laurent Le Brusquet,
Vincent Lescarret, Arshak Minasyan, Arthur Tenenhaus[†] & Xujia Zhu

[†] Course coordinator

1/42

Lecture 5/9 Bayesian estimation

Course objectives

- ▶ Introduce the concept of prior information.
- ▶ Present the basics of the Bayesian approach.
- ▶ Demonstrate how to construct estimators using prior information.

2/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercices (with solutions)
- 6 – Appendices

3/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercices (with solutions)
- 6 – Appendices

Recap: comparing estimators

Quadratic risk: $R_\theta(\hat{\eta}) = \mathbb{E}_\theta (\|\hat{\eta} - g(\theta)\|^2)$.

Definition

We say that $\hat{\eta}'$ is (weakly) **preferable** to $\hat{\eta}$ if

- ▶ $\forall \theta \in \Theta, R_\theta(\hat{\eta}') \leq R_\theta(\hat{\eta})$,

We say that it is **strictly preferable** to $\hat{\eta}$ if, in addition,

- ▶ $\exists \theta \in \Theta, R_\theta(\hat{\eta}') < R_\theta(\hat{\eta})$,

Remarks

- ▶ The relation “is preferable to” is a **partial order** on risk functions.
- ▶ **In general there is no optimal estimator**, i.e., no estimator that is preferable to all the others (unless we restrict the class of estimators that is considered).

4/42

Comparing (all) estimators: two approaches

Two approaches make it possible to refine the comparison for the cases where the risk functions cannot be compared:

- 1 the **minimax** (or “worst case”) approach:

$$R_{\max}(\hat{\eta}) = \sup_{\theta \in \Theta} R_\theta(\hat{\eta}),$$

▮ not discussed in this class;

- 2 the **Bayesian** (or “average case”) approach:

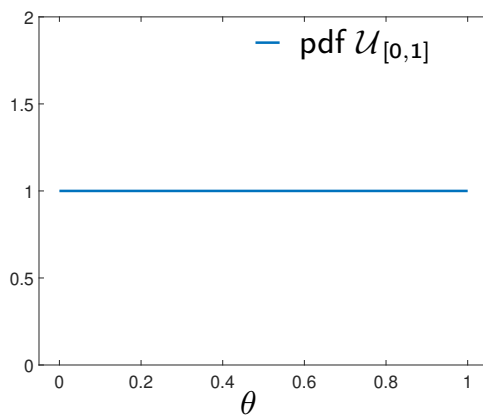
$$R_{\text{Bayes}, \pi}(\hat{\eta}) = \int_{\Theta} R_\theta(\hat{\eta}) \pi(d\theta),$$

where π is a probability measure on Θ , to be chosen.

▮ this is the topic of this lecture.

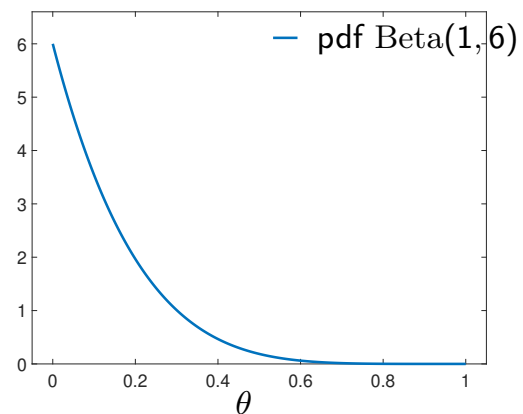
5/42

Example: white balls / red balls (see lecture #1)



Measure π : uniform over $[0, 1]$

$$\hat{\theta}_a = \frac{\sum_{i=1}^n X_i + 1}{n + 2}$$



Measure π : Beta(1, 6)

beta distrib.

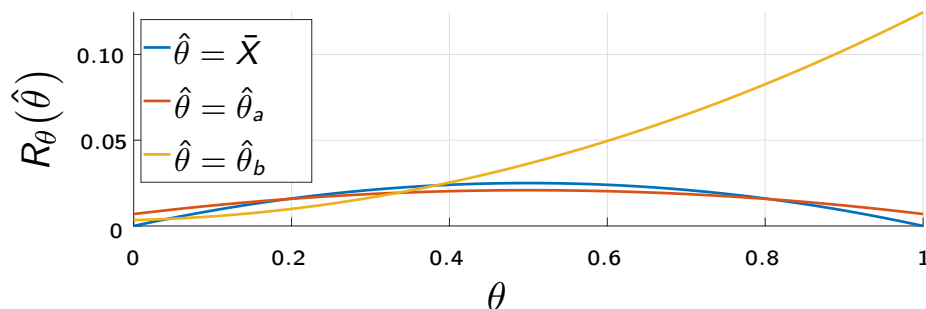
$$\hat{\theta}_b = \frac{\sum_{i=1}^n X_i + 1}{n + 7}$$

Observation: $\hat{\theta}_b < \hat{\theta}_a$,

the second estimator provides smaller estimates

6/42

Example: white balls / red balls (with $n = 10$)



	$\hat{\theta} = \bar{X}$	$\hat{\theta} = \hat{\theta}_a$	$\hat{\theta} = \hat{\theta}_b$
$R_{\max}(\hat{\theta})$	0.025 $\frac{1}{4n}$	≈ 0.0208 $\frac{1}{4(n+2)}$	≈ 0.1246 $\frac{36}{(n+7)^2}$ ✱
$R_{\text{Bayes}, \pi}(\hat{\theta})$ with $\pi \sim \mathcal{U}_{[0,1]}$	≈ 0.0167 $\frac{1}{6n}$	≈ 0.0162 $\frac{n+4}{6(n+2)^2}$	≈ 0.0456 $\frac{n+69}{6(n+7)^2}$
$R_{\text{Bayes}, \pi}(\hat{\theta})$ with $\pi \sim \text{Beta}(1, 6)$	≈ 0.0107 $\frac{3}{28n}$	≈ 0.0129 $\frac{3n+22}{28(n+2)^2}$	≈ 0.0089 $\frac{3n+42}{28(n+7)^2}$

exercise 2

Establish the expressions of R_{\max} and $R_{\text{Bayes}, \pi}$ for $\hat{\theta} = \bar{X}$.

✱ valid for $n \leq 77$

7/42

Unknown parameter \rightarrow random variables

We will assume from now on a dominated model: pdf $f_\theta(\underline{x})$.

Consider the Bayesian risk (quadratic, in this case)

$$\begin{aligned} R_{\text{Bayes}, \pi}(\hat{\eta}) &= \int_{\Theta} R_{\theta}(\hat{\eta}) \pi(d\theta) \\ &= \int_{\Theta} \mathbb{E}_{\theta} (\|\hat{\eta} - g(\theta)\|^2) \pi(d\theta). \end{aligned}$$

It can be re-written as :

$$R_{\text{Bayes}, \pi}(\hat{\eta}) = \iint_{\underline{\mathcal{X}} \times \Theta} \|\hat{\eta}(\underline{x}) - g(\theta)\|^2 \underbrace{f_{\theta}(\underline{x}) \nu(d\underline{x}) \pi(d\theta)}_{\text{Proba. measure on } \underline{\mathcal{X}} \times \Theta}.$$

8/42

Unknown parameter \rightarrow random variables (cont'd)

Let us introduce a new random variable ϑ , such that

$$(\underline{X}, \vartheta) \sim f_{\theta}(\underline{x}) \nu(d\underline{x}) \pi(d\theta). \quad (\star)$$

Then the Bayesian risk can be re-written more simply as:

$$R_{\text{Bayes}, \pi} = \mathbb{E} (\|\hat{\eta} - g(\vartheta)\|^2),$$

where the expectation is, this time, over both \underline{X} and ϑ .

Bayesian approach

In Bayesian statistics, the unknown parameter θ is (also) modeled as a random variable.

(Technical remark: the introduction of a new random variable ϑ such that (\star) holds is always possible, if we are willing to replace the underlying set Ω by $\tilde{\Omega} = \Omega \times \Theta$, provided that Θ is endowed with a σ -algebra \mathcal{F}_{Θ} such that $\theta \mapsto \mathbb{P}_{\theta}(E)$ is \mathcal{F}_{Θ} -measurable for all $E \in \mathcal{F}$.)

9/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
- 6 – Appendices

Bayesian statistical models

Technical assumptions: we assume from now on that

- ▶ Θ is endowed with a σ -algebra \mathcal{F}_Θ , e.g., if $\Theta \subset \mathbb{R}^p$, $\mathcal{F}_\Theta = \mathcal{B}(\Theta)$;
- ▶ $\theta \mapsto \mathbb{P}_\theta(E)$ is \mathcal{F}_Θ -measurable for all $E \in \mathcal{F}$ (σ -algebra on Ω).

Definition

A **Bayesian statistical model** consists of

- ▶ a statistical model as previously defined:

$$\left(\underline{\mathcal{X}}, \underline{\mathcal{A}}, \left\{ \mathbb{P}_\theta^{\mathcal{X}}, \theta \in \Theta \right\} \right),$$

- ▶ a probability distrib. π , called **prior distribution**, on $(\Theta, \mathcal{F}_\Theta)$.

Dominated model \rightarrow makes it possible to define a **likelihood**.

Joint, prior, and posterior distributions

Recall that we have introduced a new random variable ϑ , such that

$$(\underline{X}, \vartheta) \sim f_{\theta}(\underline{x}) \nu(d\underline{x}) \pi(d\theta). \quad (\star)$$

Bayesian vocabulary

We call:

- ▶ **joint distribution** the distribution of \underline{X} and ϑ , that is, (\star) ,
- ▶ **prior distribution** the marginal distribution \mathbb{P}^{ϑ} of ϑ , that is, π ,
- ▶ **posterior distribution** the distribution $\mathbb{P}^{\vartheta|\underline{X}}$ of ϑ given the data.

Interpretation (“subjective Bayes”)

- ▶ prior distribution \rightarrow **knowledge** about θ **before** data acquisition
- ▶ posteriori distribution \rightarrow ... **after** data acquisition

11/42

By the way... what is a conditional distribution ?

General definition: beyond the scope of this lecture!

(\Rightarrow uses the notion of kernel)

Let (U, V) be a pair of random variables (or vectors) that admits a density with respect to a product measure $\nu_1 \otimes \nu_2$.

We will *define* $\mathbb{P}^{V|U=u}$ as the measure with density

$$f^{V|U}(v | u) = \frac{f^{U,V}(u, v)}{f^U(u)}$$

with respect to ν_2 , for all u such that $f^U(u) > 0$.

Then we have, for any measurable function φ s.t. $\varphi(U, V) \in L^1$,

$$\mathbb{E}(\varphi(U, V) | U) \stackrel{\text{a.s.}}{=} \int_{\Theta} \varphi(U, v) f^{V|U}(v | U) \nu_2(dv).$$

12/42

Joint and marginal densities

We will assume[†] from now on that π admits a pdf

- ▶ wrt a measure ρ on $(\Theta, \mathcal{F}_\Theta)$, e.g., Lebesgue measure,
- ▶ we will write (abusively): $\pi(d\theta) = \pi(\theta) \rho(d\theta)$.

Proposition

The joint distribution admits the joint pdf

$$f^{(\underline{X}, \vartheta)}(\underline{x}, \theta) = f_\theta(\underline{x}) \pi(\theta),$$

and the corresponding marginal densities are

$$\begin{aligned} f^\vartheta(\theta) &= \pi(\theta), \\ f^{\underline{X}}(\underline{x}) &= \int f_\theta(\underline{x}) \pi(\theta) \rho(d\theta). \end{aligned}$$

[†]: This is not actually an assumption, since we can always use $\rho = \pi$ (with the pdf equal to 1).

Proof

Joint pdf (informal proof)

$$\begin{aligned} \mathbb{P}^{(\underline{X}, \vartheta)}(d\underline{x}, d\theta) &= f_\theta(\underline{x}) \nu(d\underline{x}) \pi(\theta) \rho(d\theta) \\ &= \underbrace{f_\theta(\underline{x}) \pi(\theta)}_{\text{joint pdf}} \nu(d\underline{x}) \rho(d\theta) \end{aligned}$$

Marginal densities \rightarrow we just need to integrate:

$$\begin{aligned} f^\vartheta(\theta) &= \int f_\theta(\underline{x}) \pi(\theta) \nu(d\underline{x}) = \pi(\theta), \\ f^{\underline{X}}(\underline{x}) &= \int f_\theta(\underline{x}) \pi(\theta) \rho(d\theta). \end{aligned}$$



Likelihood and Bayes' formula

Recall the **conditional density**:

$$f^{V|U}(v | u) = \frac{f^{(U,V)}(u, v)}{f^U(u)}, \quad \forall u \text{ s.t. } f^U(u) \neq 0. \quad (*)$$

Proposition

i) The conditional distribution of \underline{X} given ϑ admits the pdf

$$f^{\underline{X}|\vartheta}(\underline{x} | \theta) = f_{\theta}(\underline{x}) \quad (\text{"likelihood"}).$$

ii) The posterior distribution (ϑ given \underline{X}) admits the pdf :

$$f^{\vartheta|\underline{X}}(\theta | \underline{x}) = \frac{f_{\theta}(\underline{x}) \pi(\theta)}{f^{\underline{X}}(\underline{x})} \quad (\text{Bayes' formula}).$$

Proof. Simply apply (*) to the joint pdf. □

15/42

Remark: proportionality

The term $\frac{1}{f^{\underline{X}}(\underline{x})}$ plays the role of a **normalizing constant**:

$$f^{\vartheta|\underline{X}}(\theta | \underline{x}) = \frac{f_{\theta}(\underline{x}) \pi(\theta)}{f^{\underline{X}}(\underline{x})}.$$

Notation. The symbol " \propto " indicates **proportionality**. Thus,

$$f^{\vartheta|\underline{X}}(\theta | \underline{x}) \propto f_{\theta}(\underline{x}) \pi(\theta),$$

or, less formally,

$$\text{posterior pdf} \propto \text{likelihood} \times \text{prior pdf}.$$

The "constant" $f^{\underline{X}}(\underline{x})$ is often difficult to compute, but in some situations the computation can be avoided (MAP estimator, MCMC numerical methods...).

16/42

Example: white balls / red balls (cont'd)

Reminder: we want to estimate $\theta = \frac{W}{W+R}$ from $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$.

Density of the observations:

$$f_{\theta}(\underline{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{N(\underline{x})} (1 - \theta)^{n-N(\underline{x})}.$$

with $N(\underline{x}) = \sum_{i=1}^n x_i$.

We assume a prior distribution $\text{Beta}(a_0, b_0)$ for the parameter θ :

$$\pi(\theta) \propto \theta^{a_0-1} (1 - \theta)^{b_0-1},$$

and we denote, as before, ϑ the corresponding RV.

(The choice of the prior distribution will be discussed later.)

17/42

Example: white balls / red balls (cont'd)

Then we have:

$$\begin{aligned} f^{\vartheta|\underline{X}}(\theta | \underline{x}) &\propto f_{\theta}(\underline{x}) \pi(\theta) \\ &\propto \theta^{N(\underline{x})} (1 - \theta)^{n-N(\underline{x})} \cdot \theta^{a_0-1} (1 - \theta)^{b_0-1} \\ &= \theta^{a_0+N(\underline{x})-1} (1 - \theta)^{b_0+n-N(\underline{x})-1}. \end{aligned}$$

We recognize (up to a cst) the pdf of the $\text{Beta}(a_n, b_n)$ distrib., with

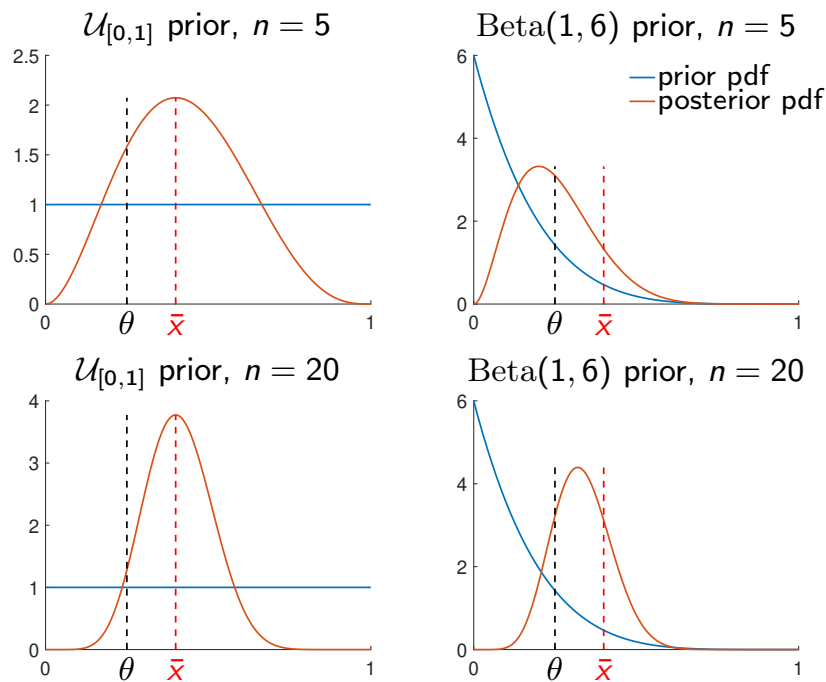
$$\begin{cases} a_n = a_0 + N, \\ b_n = b_0 + n - N. \end{cases}$$

beta distrib.

Conclusion. Posterior distribution: $\vartheta | \underline{X} \sim \text{Beta}(a_n, b_n)$.

18/42

Example: white balls / red balls (cont'd)



Remark: for $n \rightarrow \infty$, we have a $\mathbb{E}(\vartheta \mid \underline{X}_n) = \bar{X}_n + O(\frac{1}{n})$ with $\text{var}(\vartheta \mid \underline{X}_n) \simeq \frac{\theta(1-\theta)}{n}$.

19/42

Example: component reliability

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta) = \mathcal{E}(\frac{1}{\eta})$, hence the likelihood:

$$\begin{aligned} \mathcal{L}(\eta, \underline{x}_n) &= f(\underline{x}_n \mid \eta) = \prod_{i=1}^n \frac{1}{\eta} \exp\left(-\frac{1}{\eta} x_i\right) \\ &= \eta^{-n} \exp\left(-\frac{1}{\eta} \sum_{i=1}^n x_i\right). \end{aligned}$$

(Here, we directly use η as our unknown parameter.)

We choose (see below) a truncated $\mathcal{N}(\eta_0, \sigma_0^2)$ prior for η :

$$\pi(\eta) \propto \exp\left(-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}\right) \mathbf{1}_{\eta \geq 0}.$$

Abuse of notation: we simply denote f the conditional probability density, instead of $f^{\underline{X}_n \mid \eta}$, where η is the random variable associated with the parameter η .

20/42

Example: component reliability (cont'd)

Posterior distribution of η . From Bayes' formula we get:

$$f(\eta | \underline{x}_n) \propto \underbrace{\eta^{-n} \exp\left(-\frac{1}{\eta} \sum_{i=1}^n x_i\right)}_{\text{likelihood}} \cdot \underbrace{\exp\left(-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}\right)}_{\text{prior pdf}}.$$



This time we fail to recognize a “familiar” density

➡ numerical evaluation of the integrals

$$f(\underline{x}_n) = \int_0^{+\infty} \eta^{-n} e^{-\frac{1}{\eta} \sum_{i=1}^n x_i} e^{-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}} d\eta$$

$$\mathbb{E}(\eta | \underline{X}_n = \underline{x}_n) = \frac{1}{f(\underline{x}_n)} \int_0^{+\infty} \eta \cdot \eta^{-n} e^{-\frac{1}{\eta} \sum_{i=1}^n x_i} e^{-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}} d\eta$$

Abuse of notation (cont'd): we often use the same symbol (here, η) to represent both a point in the parameter space and the RV associated with the parameters.

21/42

Example: component reliability (cont'd)

Numerical application. $\eta_0 = 14.0$, $\sigma_0 = 1.0$ and the true value is $\eta = 11.4$.

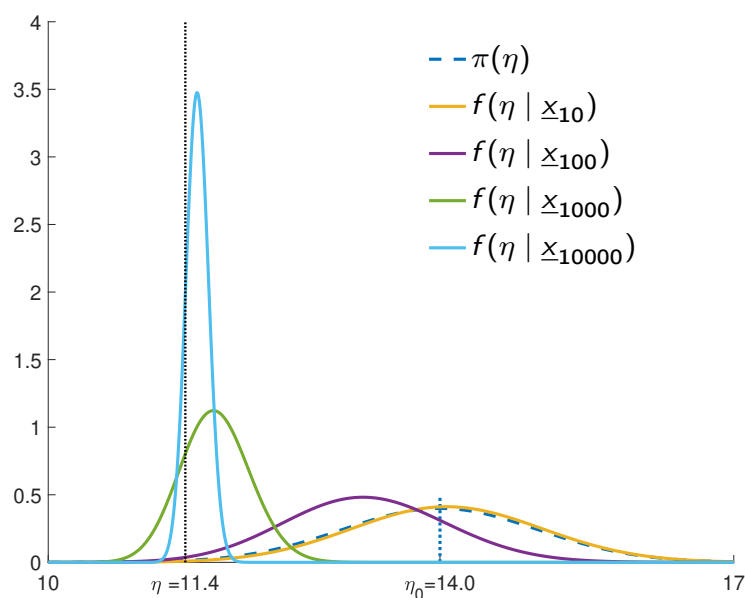


Figure – Prior and posterior densities of η , for four values of n .

22/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
- 6 – Appendices

Several approaches

Two kinds of sources of prior information:

- ▶ “historical” **data**,
- ▶ **experts**: subjective knowledge, field expertise, etc.

Advanced topics (not covered in this course):

- ▶ merging several sources of prior information,
- ▶ “weakly informative” or “objective” priors,
- ▶ least favorable priors (cf. minimax),
- ▶ ...

Example: white balls / red balls (cont'd)

Assume that we have data from a past experiment:

- ▶ sample of $n_0 = 20$ draws,
- ▶ $N_0 = 15$ white balls drawn.

Choice of a prior distribution

We can decide, e.g., to choose a $\text{Beta}(a_0, b_0)$ prior, with $a_0 = N_0 = 15$ and $b_0 = n_0 - N_0 = 5$.

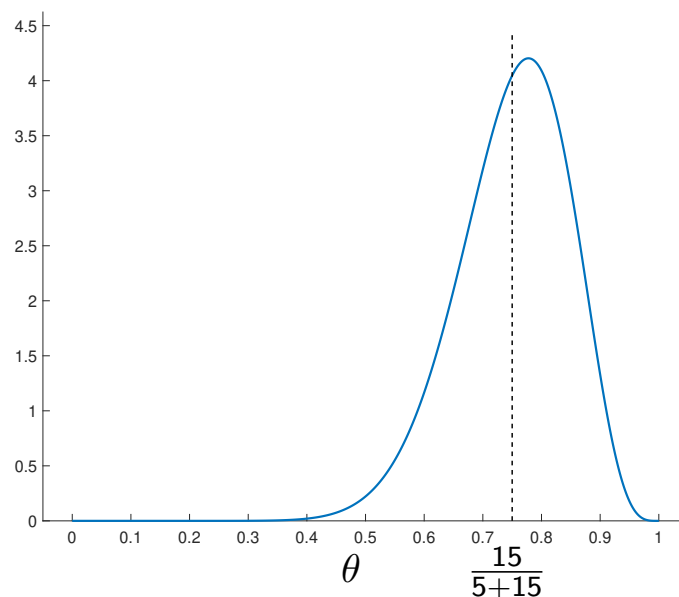
Arguments in favour of this choice:

- ▶ the shape of the distrib. makes computations easier (see below);
- ▶ **expectation** : $\frac{a_0}{a_0+b_0} = p_0$, with $p_0 = \frac{N_0}{n_0}$;
- ▶ **variance**: $\frac{a_0 b_0}{(a_0+b_0)^2(a_0+b_0+1)} \approx \frac{p_0(1-p_0)}{n_0} \Rightarrow$ variance of \bar{X}_{n_0} .

24/42

Example: white balls / red balls (cont'd)

Prior density $\text{Beta}(15, 5)$



25/42

Example: component reliability

We have the following pieces of information:

- ▶ The manufacturer claims that the lifetime of its components is approximately $\eta_0 = 6$ months.
- ▶ A field expert estimates that the accuracy of the manufacturer's data is roughly $\varepsilon_0 = 10\%$.

Choice of a prior distribution (elicitation)

We can decide, e.g., to choose a $\mathcal{N}(\eta_0, \sigma_0)$ prior, truncated to $[0, +\infty)$, with $\sigma_0 = \varepsilon_0 \eta_0 / 1.96$.

Arguments in favour of this choice:

- ▶ The prior is (approx.) centered on the manufacturer's value η_0 .
- ▶ $\approx 95\%$ of the prior probability is supported by the interval $[0.9\eta_0, 1.1\eta_0]$.
- ▶ The choice of a Gaussian shape and the value 95% are arbitrary.

26/42

Conjugate priors easier computations !

Families of conjugate prior distributions

A **family of distributions** (densities) is called **conjugate** for a given statistical model if, for any prior π in this family, the posterior $f^{\vartheta|\underline{X}}$ remains inside the family.

Examples.

- ▶ $\text{Ber}(\theta)$ sample + **bêta** prior,
- ▶ $\mathcal{N}(\mu, \sigma^2)$ sample with known σ^2 + \mathcal{N} prior on μ ,
- ▶ $\mathcal{N}(\mu, \sigma^2)$ sample with known μ + IG^\dagger prior on σ^2 ,
- ▶ $\mathcal{E}(\theta)$ sample + **gamma** prior,
- ▶ ...

 gamma distrib.

† : inverse gamma. $Z \sim \text{IG}$ if $1/Z$ has a gamma distribution.

27/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
- 6 – Appendices

Bayes estimators

Goal

We want to construct estimators of $\eta = g(\theta)$ taking into account

- ▶ the data \underline{x} ,
- ▶ and the prior distribution π .

Bayes estimators

Let $L : N \times N \rightarrow \mathbb{R}$ be a **loss function**.

► Reminder: we “lose” $L(\eta, \tilde{\eta})$ if we estimate $\tilde{\eta}$ when the true value is η .

Definition: Bayesian estimator

A **Bayesian estimator** is an estimator that minimizes the **posterior expected loss**:

$$\hat{\eta} = \arg \min_{\tilde{\eta} \in N} J(\tilde{\eta}, \underline{X})$$

with

$$\begin{aligned} J(\tilde{\eta}, \underline{x}) &= \mathbb{E} (L(g(\vartheta), \tilde{\eta}) \mid \underline{X} = \underline{x}) \\ &= \int_{\Theta} L(g(\theta), \tilde{\eta}) f^{\vartheta|\underline{X}}(\theta \mid \underline{x}) \rho(d\theta). \end{aligned}$$

($\Rightarrow J$ is well-defined for $\mathbb{P}^{\underline{X}}$ -almost all \underline{x} .)

Remark: equivalently, a Bayesian estimator minimizes the Bayes risk R_{π} .

29/42

Quadratic loss

Consider the quadratic loss function $L(\eta, \tilde{\eta}) = \|\eta - \tilde{\eta}\|^2$:

$$J(\tilde{\eta}, \underline{x}) = \int_{\Theta} \|g(\theta) - \tilde{\eta}\|^2 f^{\vartheta|\underline{X}}(\theta \mid \underline{x}) \rho(d\theta).$$

Proposition

In this case, the Bayesian estimator is

$$\hat{\eta} = \mathbb{E}(g(\vartheta) \mid \underline{X}) = \int_{\Theta} g(\theta) f^{\vartheta|\underline{X}}(\theta \mid \underline{X}) \rho(d\theta).$$

$\Rightarrow \hat{\eta}$ is the **posterior mean** of ϑ

Remark: it can also be written as

$$\hat{\eta}(\underline{x}) = \frac{\int_{\Theta} g(\theta) f_{\theta}(\underline{x}) \pi(\theta) \rho(d\theta)}{f^{\underline{X}}(\underline{x})} = \frac{\int_{\Theta} g(\theta) f_{\theta}(\underline{x}) \pi(\theta) \rho(d\theta)}{\int_{\Theta} f_{\theta}(\underline{x}) \pi(\theta) \rho(d\theta)}.$$

30/42

Example: white balls / red balls (cont'd)

With a $\text{Beta}(a_0, b_0)$ prior on ϑ , we have seen that:

$$\vartheta | \underline{X} \sim \text{Beta}(N + a_0, n - N + b_0)$$

with $N = \sum_{i=1}^n X_i$.

The expectation of the $\text{Beta}(a, b)$ distribution is $\frac{a}{a+b}$, thus:

$$\hat{\theta} = \mathbb{E}(\vartheta | \underline{X}) = \frac{N + a_0}{n + a_0 + b_0}.$$

Remark: we recover the expressions of $\hat{\theta}_a$ and $\hat{\theta}_b$ ([back to slide ??](#)).

31/42

Another example: Gaussian n -sample (with known σ^2)

It can be proved (see PC 5) that $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma_0^2)$

- ▶ with $\theta \in \mathbb{R}$ (unknown), $\sigma_0 > 0$ (known),
- ▶ and $\vartheta \sim \mathcal{N}(\mu, \tau^2)$,

then

$$\vartheta | \underline{X} \sim \mathcal{N}\left(\frac{\tau^2 \sum_{i=1}^n X_i + \sigma_0^2 \mu}{n\tau^2 + \sigma_0^2}, \frac{\tau^2 \sigma_0^2}{n\tau^2 + \sigma_0^2}\right)$$

Hence, the Bayesian estimator (for the quadratic loss):

$$\hat{\theta} = \lambda \bar{X} + (1 - \lambda) \mu \quad \text{with } \lambda = \frac{n\tau^2}{n\tau^2 + \sigma_0^2}$$

Interpretation

- ▶ when $n \rightarrow \infty$, $\hat{\theta} \approx \bar{X}$ (the prior no longer has influence)
- ▶ with finite n , when $\frac{\sigma_0}{\tau} \gg 1$, $\hat{\theta} \approx \mu_\theta$ (the data is ignored).

32/42

L^1 loss

Assume for simplicity that $\eta = \theta \in \mathbb{R}$.

Consider the loss function $L(\theta, \tilde{\theta}) = |\theta - \tilde{\theta}|$:

$$J(\tilde{\theta}, \underline{x}) = \int_{\Theta} |\theta - \tilde{\theta}| f^{\vartheta|\underline{X}}(\theta | \underline{x}) \rho(d\theta).$$

Proposition

In this case the Bayesian estimator $\hat{\theta}$ is such that

$$\int_{-\infty}^{\hat{\theta}} f^{\vartheta|\underline{X}}(\theta | \underline{X}) \rho(d\theta) = \int_{\hat{\theta}}^{\infty} f^{\vartheta|\underline{X}}(\theta | \underline{X}) \rho(d\theta) = \frac{1}{2} \quad \mathbb{P}^{\underline{X}}\text{-a.s.}$$

⇒ $\hat{\theta}$ is a **median** of the posterior density of ϑ

Remark: when ϑ has a symmetric posterior density, the two Bayesian estimators (L^1 and L^2 loss) coincide.

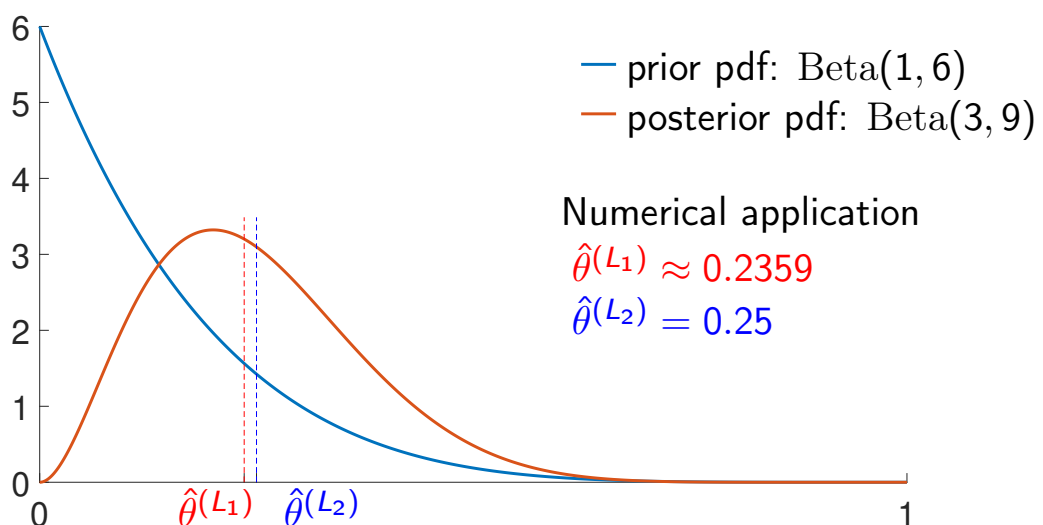
Example: mean of a Gaussian n -sample, with a Gaussian prior.

33/42

Example: white balls / red balls (cont'd)

Observed sample ($n = 5$): $\underline{x} = (W, R, R, W, R)$.

Prior on η : $\vartheta \sim \text{Beta}(1, 6)$, with $\theta = \mathbb{P}(X_1 = W)$.



34/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
 - 5.1 – Questions
 - 5.2 – Solutions
- 6 – Appendices

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
 - 5.1 – Questions
 - 5.2 – Solutions
- 6 – Appendices

Exercise 1 (exponential likelihood + gamma prior)

 solution

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$ with $\theta \in \Theta = (0, +\infty)$.

We endow θ with a Gamma(α_0, β_0) prior.

Questions


- i Show that the gamma priori is conjugate, and find the parameters α_n and β_n of the posterior distribution.
- ii Give the Bayesian estimator of θ , for the quadratic loss.
- iii prove that this estimator tends to the MLE when the parameters α_0 and β_0 tend to a certain limit to be specified.

35/42


Exercise 2 (maximal and Bayesian risks)

 solution

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$ with $\theta \in \Theta = (0, 1)$.

We want to estimate θ . The objective of this exercise is to establish the expressions of the maximal and Bayesian quadratic risks of $\hat{\theta} = \bar{X}$, announced on  slide 7.

Questions

- i Calculate the quadratic risk $R_\theta(\bar{X})$, and deduce the maximal risk $R_{\max}(\bar{X})$.
- ii Calculate the Bayesian risk $R_{\text{Bayes}, \pi}(\bar{X})$ when π is a  beta distribution with parameters $a > 0$ and $b > 0$.

36/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
 - 5.1 – Questions
 - 5.2 – Solutions
- 6 – Appendices

Solution of exercise 1

[back to questions](#)

Preliminary remark: in this solution we use the same notation, as often done in practice, for the “deterministic” parameter θ and the corresponding random variable, denoted by ϑ in the lecture.

i) First write the likelihood:

$$\mathcal{L}(\theta; \underline{x}) = f(\underline{x} | \theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i},$$

and the prior density:

$$\pi(\theta) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\beta_0 \theta} \propto \theta^{\alpha_0-1} e^{-\beta_0 \theta}.$$

The posterior density then follow from the Bayes formula:

$$f(\theta | \underline{x}) \propto \mathcal{L}(\theta; \underline{x}) \pi(\theta) \propto \theta^{\alpha_0+n} e^{-\theta(\beta_0 + \sum_{i=1}^n x_i)}$$

Solution of exercise 1

[back to questions](#)

The distribution of θ given \underline{X} , aka posterior distribution, is therefore a gamma distribution with parameters

- ▶ $\alpha_n = \alpha_0 + n$,
- ▶ $\beta_n = \beta_0 + \sum_{i=1}^n X_i$.

ii) The Bayesian estimator for the quadratic loss is given by the posterior expectation of θ given the data:

$$\mathbb{E}(\theta \mid \underline{X}) = \frac{\alpha_n}{\beta_n} = \frac{\alpha_0 + n}{\beta_0 + \sum_{i=1}^n X_i}.$$

iii) This estimator tends to the MLE $1/\bar{X}_n$ when both α_0 and β_0 tend to zero.

38/42

Solution of exercise 2

[back to questions](#)

i) \bar{X} is an unbiased estimator of $\theta = \mathbb{E}_\theta(X_1)$, therefore

$$R_\theta(\bar{X}) = \text{var}_\theta(\bar{X}) = \frac{1}{n} \text{var}_\theta(X_1) = \frac{\theta(1-\theta)}{n}.$$

The function $\theta \mapsto R_\theta(\hat{\theta})$ is a polynomial of degree two in θ , which attains its maximum at $\theta = \frac{1}{2}$, hence:

$$R_{\max}(\bar{X}) = \frac{1}{4n}.$$

ii) Let $B(a, b) = \Gamma(a)\Gamma(b) / \Gamma(a+b)$.

39/42

The Bayesian risk for $\pi = \text{Beta}(a, b)$ is:

$$\begin{aligned} R_{\text{Bayes}, \pi}(\bar{X}) &= \int R_{\theta}(\hat{\theta}) \pi(d\theta) \\ &= \int_0^1 \frac{\theta(1-\theta)}{n} \cdot \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\ &= \frac{1}{n B(a, b)} \int_0^1 \theta^a (1-\theta)^b d\theta \\ &= \frac{B(a+1, b+1)}{n B(a, b)} = \frac{1}{n} \frac{ab}{(a+b+1)(a+b)}. \end{aligned}$$

In particular,

- ▶ For $\pi = \mathcal{U}_{[0,1]} = \text{Beta}(1, 1)$, $R_{\text{Bayes}, \pi}(\bar{X}) = \frac{1}{6n}$.
- ▶ For $\pi = \text{Beta}(1, 6)$, $R_{\text{Bayes}, \pi}(\bar{X}) = \frac{3}{28n}$.

40/42

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Standard exercises (with solutions)
- 6 – Appendices

The beta family of distributions

Let $X \sim \text{Beta}(a, b)$ with $(a, b) = \theta \in (\mathbb{R}_+^*)^2$. Its pdf is :

$$f_\theta(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \mathbb{1}_{]0,1[}(x).$$

Moments

- ▶ expectation : $\mathbb{E}_\theta(X) = \frac{a}{a+b}$
- ▶ variance : $\text{var}_\theta(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Special case

- ▶ $\mathcal{U}_{[0,1]} = \text{Beta}(1, 1)$

Properties

- ▶ If $X \sim \text{Beta}(a, 1)$, then $-\log(X) \sim \mathcal{E}\left(\frac{1}{a}\right)$.
- ▶ If $X \sim \Gamma(a, \lambda)$, $Y \sim \Gamma(b, \lambda)$, and $X \perp Y$, then $\frac{X}{X+Y} \sim \text{Beta}(a, b)$.

⏮ back to slide ??

⏮ back to slide ??

⏮ back to slide 36

41/42

The gamma family of distributions

A random variable X follows the $\Gamma(p, \lambda)$ distribution, with parameters $p > 0$ and $\lambda > 0$, if it has the pdf

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^+}(x).$$

Moments

- ▶ expectation : $\mathbb{E}_\theta(X) = \frac{p}{\lambda}$
- ▶ variance : $\text{var}_\theta(X) = \frac{p}{\lambda^2}$

Special cases

- ▶ $\mathcal{E}(\lambda) = \Gamma(p=1, \lambda)$
- ▶ $\Gamma(p = \frac{n}{2}, \lambda = \frac{n}{2}) = \chi^2(n)$

Properties

- ▶ Let $a > 0$. If $X \sim \Gamma(p, \lambda)$, then $aX \sim \Gamma(p, \frac{\lambda}{a})$.
- ▶ If X and Y are independent, with $X \sim \Gamma(p, \lambda)$ and $Y \sim \Gamma(q, \lambda)$, then $X + Y \sim \Gamma(p + q, \lambda)$.

⏮ back to slide ??

42/42