CentraleSupélec

Statistics and Learning

Julien Bect

(julien.bect@centralesupelec.fr)

Teaching : CentraleSupélec / dept. of Statistics and Signal Processing

Research : Laboratory of Signals and Systems (L2S)

J. Bect & L. Le Brusquet — 1A — Statistics and Learning

Lecture 4/9

Bayesian estimation

In this lecture you will learn how to…

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

Lecture outline

1 – Introduction : the Bayes risk

2 – Bayesian statistics : prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

Lecture outline

1 – Introduction : the Bayes risk

2 – Bayesian statistics : prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

Recap : comparing estimators

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

Definition

We will 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 will 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).

Comparing (all) estimators : two approaches

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

1. the minimax (or « worst case ») approach :

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

➠ not discussed in this class ;

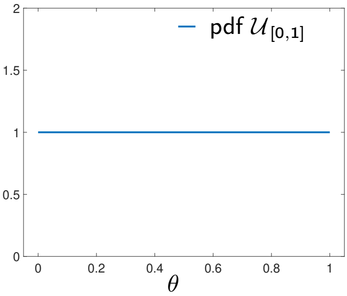
2. the **Bayesian** (or « average case ») approach :

$$R_{\mathrm{Bayes},\pi}\left(\hat{\eta}\right) \;=\; \int_\Theta R_\theta\left(\hat{\eta}\right) \pi(\mathrm{d}\theta),$$

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

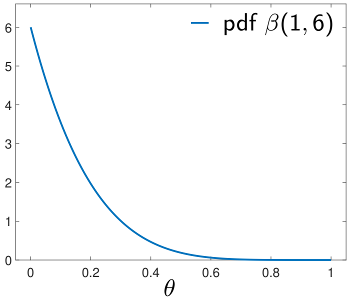
➠ this is the topic of this lecture.

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



Measure π : uniform over $[0,1]$

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



Measure π : $\beta\,(1,6)$

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

Observation : $\hat{\theta}_{\mathrm{b}} = \frac{n+2}{n+7}\,\hat{\theta}_{\mathrm{a}}$,

➠ the second estimator provides smaller estimates

The beta family of distributions

Let $X \sim \beta(a,b)$ with $(a,b) = \theta \in (\mathbb{R}_\star^+)^2$. 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 : $\mathbb{V}_\theta(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Special case

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

Properties

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

Unknown parameter → random variables

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

Consider the Bayesian risk (quadratic, in this case)

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

It can be re-written as :

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

Unknown parameter $\to$ random variables (cont'd)

Let us introduce a new random variable ϑ, such that

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

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

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

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 $\widetilde{\Omega} = \Omega \times \Theta$, provided that Θ is endowed with a σ-algebra $\mathscr{F}_\Theta$ such that $\theta \mapsto \mathbb{P}_\theta(E)$ is $\mathscr{F}_\Theta$-measurable for all $E \in \mathscr{F}$.)

Lecture outline

1 – Introduction : the Bayes risk

2 – Bayesian statistics : prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

Bayesian statistical models

Technical assumptions : we assume from now on that

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

Definition

A Bayesian statistical model consists of

- a statistical model as previously defined :

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

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

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

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

Bayesian vocabulary

We call :

- joint distribution the distribution of $\underline{X}$ and ϑ, that is, $(\star)$,
- prior distribution the marginal distribution $\mathbb{P}^{\vartheta}$ 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

Joint and marginal densities

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

- wrt a measure ν_Θ on $(\Theta, \mathscr{F}_\Theta)$, e.g., Lebesgue's measure,
- we will write (abusively) : $\pi(\mathrm{d}\theta) = \pi(\theta)\,\mathrm{d}\theta$.

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

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

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

Proof

Joint pdf (informal proof)

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

Marginal densities $\rightarrow$ we just need to integrate :

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

$\square$

Likelihood and Bayes' formula

Proba refresher : the conditional density of $Y \mid Z$ is equal to

$$f^{Y|Z}(y \mid z) \;=\; \frac{f^{(Y,Z)}(y,z)}{f^{Z}(z)}, \qquad \forall z \text{ s.t. } f^{Z}(z) \neq 0. \qquad (\star)$$

Proposition

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

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

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

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

Proof. Simply apply $(\star)$ to the joint pdf. $\square$

Remark : proportionality

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

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

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

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

or, less formally,

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

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. . .).

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

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

Density of the observations :

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

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

Let us choose a $\beta(a_0, b_0)$ prior :

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

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

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

Then we have :

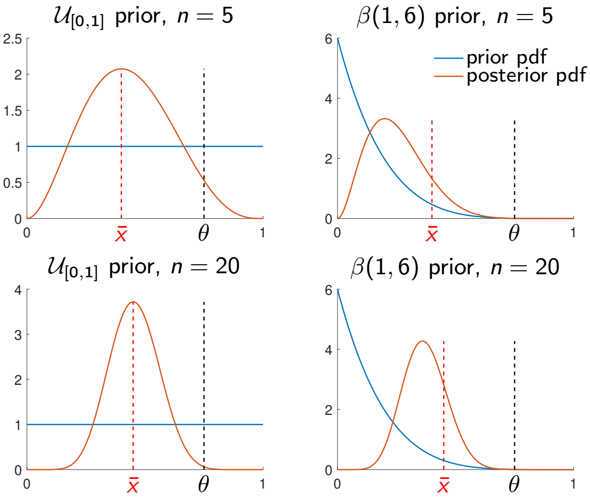
$$\begin{aligned} f^{\vartheta|\underline{X}}(\theta \mid \underline{x}) &\propto f_\theta(\underline{x})\, \pi(\theta) \\ &\propto \theta^{N(\underline{x})} (1-\theta)^{n-N(\underline{x})} \cdot \theta^{a_0-1} (1-\theta)^{b_0-1} \\ &= \theta^{a_0+N(\underline{x})-1} (1-\theta)^{b_0+n-N(\underline{x})-1}.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

We recognize (up to a cst) the pdf of the $\beta(a_n, b_n)$ distrib., with

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

Conclusion. Posterior distribution : $\vartheta \mid \underline{X} \sim \beta(a_n, b_n)$.

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



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

Example : component reliability

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

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

(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 :

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

Example : component reliability (cont'd)

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

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

⚠ This time we fail to recognize a "familiar" density

➠ numerical evaluation of the integral

$$f(\underline{x}) \;=\; \int \eta^{-n} \exp\left(-\frac{1}{\eta}\sum_{i=1}^{n} x_i\right) \exp\left(-\frac{(\eta-\eta_0)^2}{2\sigma_0^2}\right) \nu(\mathrm{d}\underline{x}).$$

Example : component reliability (cont'd)

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

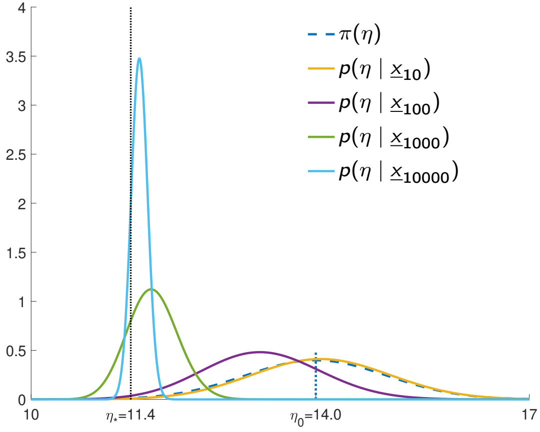


Figure – Prior and posterior densities of η, for four values of n.

Lecture outline

1 – Introduction : the Bayes risk

2 – Bayesian statistics : prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

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" 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.

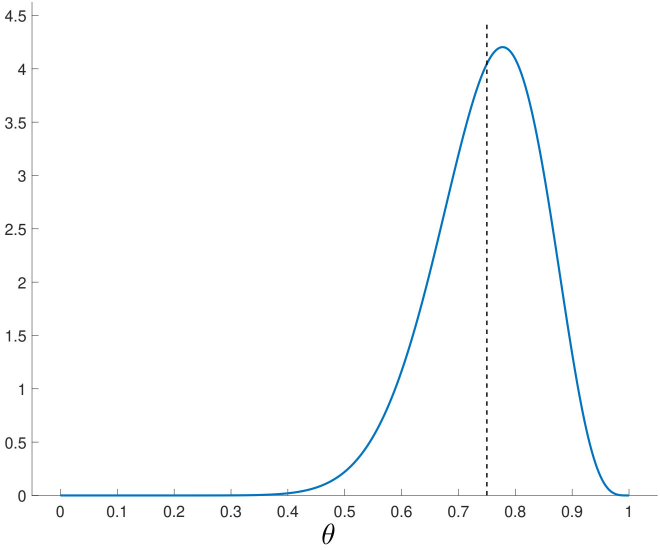
Choice of a prior distribution

We can decide, e.g., to choose a $\beta(a_0, b_0)$ 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 : $\frac{a_0 b_0}{(a_0+b_0)^2(a_0+b_0+1)} \approx \frac{p_0(1-p_0)}{n_0}$ ➠ variance of $\bar{X}_{n_0}$.

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



Example : component reliability

We have the following pieces of information :

- The manufacturer claims that tex 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 centered on the manufacturer's value η_0.
- $\approx$ 95% of the prior proba. is supported by the interval $[0.9\eta_0,\ 1.1\eta_0]$.
- The choice of the Gaussian form is arbitrary. . .

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.

- $\mathrm{Ber}(\theta)$ sample + β 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 + $\mathcal{IG}^\dagger$ prior on σ^2,
- $\mathcal{E}(\theta)$ sample + gamma prior,
- ...

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

Lecture outline

1 – Introduction : the Bayes risk

2 – Bayesian statistics : prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

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 $L : N \times N \to \mathbb{R}$ 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 :

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

with

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

Remark : equivalently, a Bayesian estimator minimizes the Bayes risk R_π.

Quadratic loss

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

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

Proposition

In this case, the Bayesian estimator is the posterior mean :

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

Remark : it can also be written as :

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

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

With a $\beta(a_0, b_0)$ prior on ϑ, we have seen that :

$$\vartheta|\underline{X} \sim \beta\left(N + a_0, n - N + b_0\right) \quad \text{with } N = \sum_{i=1}^{n} X_i.$$

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

$$\hat{\theta} = \frac{N + a_0}{n + a_0 + b_0}.$$

Remark : we recover the expressions of $\hat{\theta}_a$ and $\hat{\theta}_b$.

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

We have seen that, if

- $X_i, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma_0^2)$ with $\theta \in \mathbb{R}$ (unknown) and $\sigma_0 > 0$ (known),
- $\vartheta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$,

then

$$\vartheta | \underline{x} \sim \mathcal{N}\left(\frac{\sigma_\theta^2 \sum_{i=1}^n x_i + \sigma_0^2 \mu_\theta}{n\sigma_\theta^2 + \sigma_0^2}, \frac{\sigma_\theta^2 \sigma_0^2}{n\sigma_\theta^2 + \sigma_0^2} \right)$$

Hence the Bayesian estimator (for the quadratic loss) :

$$\hat{\theta} = \frac{\sigma_\theta^2 \sum_{i=1}^n X_i + \sigma_0^2 \mu_\theta}{n\sigma_\theta^2 + \sigma_0^2}$$

Interpretation

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

L^1 loss

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

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

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

Proposition

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

$$\int_{-\infty}^{\hat{\theta}} f_\theta(\underline{x})\pi(\theta)\mathrm{d}\theta = \int_{\hat{\theta}}^{\infty} f_\theta(\underline{x})\pi(\theta)\mathrm{d}\theta \left(= \frac{1}{2} \right)$$

➠ $\hat{\theta}$ is thus the median of the posterior density of ϑ

Remark : when $\vartheta|\underline{x}$ has a symmtric density, the two Bayesian estimators (L^1 and L^2 loss) coincide.

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

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

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

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