CHAPITRE 1

Introduction

Dans ce chapitre, nous introduisons les notions de base de statistiques utiles pour la suite, parmi lesquelles les notions d'expérience, de modèle, d'estimateur et de régions de confiance. Nous illustrons ces notions par des exemples et faisons quelques rappels de probabilité. Enfin nous définissons la notion de loi conditionnelle qui joue un rôle central dans la suite.

1.1 Modèles statistiques

L'objet de départ en statistique est une suite d'observations, appelée *données*, typiquement sous la forme d'une suite numérique $x_1, \dots, x_n$.

La modélisation statistique consiste à écrire $x_i = X_i(\omega)$: les données sont vues comme des réalisations de *variables aléatoires* $X_1, \dots, X_n$, voir un cours basique de probabilités pour les notions élémentaires d'aléatoire.

Definition 1. Une expérience statistique est la donnée de

- un objet aléatoire X à valeurs dans un espace E muni d'une tribu d'événements $\mathcal{E}$.
- une famille de mesures de probabilité sur $(E, \mathcal{E})$ appelée *modèle*

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\},$$

où Θ est un ensemble appelé *espace des paramètres*.

Loi. Une 'mesure de probabilité' s'appelle aussi 'loi'.

Souvent, X consiste en un n-uplet $X = X^{(n)} = (X_1, \dots, X_n)$. Dans ce cas, les quantités E et $\mathcal{P}$ de la définition précédente dépendent de n.

Modèle du n-échantillon. Lorsque $X = X^{(n)} = (X_1, \dots, X_n)$, on prendra souvent $P_\theta^{(n)} = P_\theta \otimes \cdots \otimes P_\theta = P_\theta^{\otimes n}$. Si un n-uplet $(Y_1, \dots, Y_n)$ est de loi $P_\theta^{\otimes n}$, on dira que les variables $Y_1, \dots, Y_n$ sont indépendantes et identiquement distribuées (en abbrégé iid).

Definition 2. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est identifiable si, pour tous $\theta, \theta' \in \Theta$,

$$P_\theta = P_{\theta'} \quad \Rightarrow \quad \theta = \theta'.$$

L'identifiabilité d'un modèle implique que pour une loi donnée Q dans $\mathcal{P}$, il y a un *unique* paramètre tel que $Q = P_\theta$. C'est une propriété très importante, qui assure que le modèle est *bien* paramétré. Nous donnons quelques exemples ci-dessous.

Definition 3. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est dominé s'il existe une mesure positive μ sur E telle que, pour tous $\theta \in \Theta$, P_θ admet une densité p_θ par rapport à μ, soit

$$dP_\theta(x) = p_\theta(x) d\mu(x).$$

Notons qu'il faut que la mesure μ soit la même pour tous les $\theta \in \Theta$. On parle de mesure *dominante*. Dans la suite, nous travaillerons toujours avec des modèles dominés.

Notations. Si X est une variable aléatoire de loi Q, on note $X \sim Q$. Cela signifie que pour toute g intégrable par rapport à Q, soit $g \in L^1(Q)$,

$$E_{X \sim Q}[g(X)] = E_Q[g(X)] = \int_E g(x) dQ(x).$$

Si $Y \sim P_\theta$, on abrège souvent $E_{Y \sim P_\theta}$ en E_θ. Egalement, dans le cas du modèle du n-échantillon ci-dessus, on note simplement E_θ en lieu et place de $E_{P_\theta^{\otimes n}}$.

Exemples primordiaux

1 Le modèle fondamental est

$$\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}.$$

C'est un modèle dominé, pour μ la mesure de Lebesgue sur $\mathbb{R}$,

$$dP_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} dx.$$

Il s'agit aussi d'un modèle identifiable. En effet, ici $P_\theta = \mathcal{N}(\theta, 1)$, et si $P_\theta = P_{\theta'}$,

(a) *méthode 1.* Si $\mathcal{N}(\theta, 1) = \mathcal{N}(\theta', 1)$, l'espérance pour les deux lois est la même : $E_{\mathcal{N}(\theta,1)}[X] = E_{\mathcal{N}(\theta',1)}[X]$. Or $E_\theta[X] = E_{\mathcal{N}(\theta,1)}[X] = \int x e^{-(x-\theta)^2/2} dx/\sqrt{2\pi} = \theta$. Donc $\theta = \theta'$.

(b) *méthode 2.* Si deux lois à densité par rapport à μ sont égales, alors leurs densité sont égales μ-presque partout. Or $\theta \neq \theta'$ implique que $p_\theta(x) \neq p_{\theta'}(x)$ pour tout $x \in \mathbb{R}$. Ainsi $P_\theta \neq P_{\theta'}$, donc le modèle est identifiable (c'est la contraposée de la définition).

2 Le modèle de tirage de pile ou face est

$$\mathcal{P} = \{\mathrm{Be}(\theta),\ \theta \in [0,1]\},$$

où $\mathrm{Be}(\theta)$ est la loi de Bernoulli. C'est la loi discrète définie par : si $X \sim \mathrm{Be}(\theta)$,

$$P[X = 1] = \theta, \qquad P[X = 0] = 1 - \theta,$$

ce que l'on note aussi $\mathrm{Be}(\theta) = (1-\theta)\delta_0 + \theta\delta_1$. C'est un modèle dominé par $\mu = \delta_0 + \delta_1$. Le modèle est identifiable, par exemple parce que $E_\theta X = \theta$ donc $P_\theta = P_{\theta'}$ implique que $\theta = E_\theta X = E_{\theta'} X = \theta'$.

Definition 4. Un estimateur ponctuel $\hat{\theta}(X)$ dans une expérience statistique $(X, \mathcal{P})$ est une fonction mesurable de X, à valeurs dans l'espace des paramètres Θ.

Une statistique $S(X)$ est une fonction mesurable quelconque de X.

Exemple. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, si l'on dispose d'observations $X = (X_1, \ldots, X_n)$, alors $\hat{\theta}_1(X) = 1$, $\hat{\theta}_2(X) = \overline{X} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ sont des estimateurs (ponctuels) de θ, et aussi des statistiques.

1.2 Approches statistiques

Nous introduisons deux points de vue principaux, l'approche *fréquentiste* et l'approche *bayésienne*. Ces deux approches ont le même point de départ : l'expérience statistique définie plus haut, et en particulier le modèle $\mathcal{P}$. La principale différence réside dans l'hypothèse que l'on fait sur la loi suivie par les données X.

1.2.1 Approche fréquentiste

Dans l'approche *fréquentiste*, on suppose

$$\exists\, \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

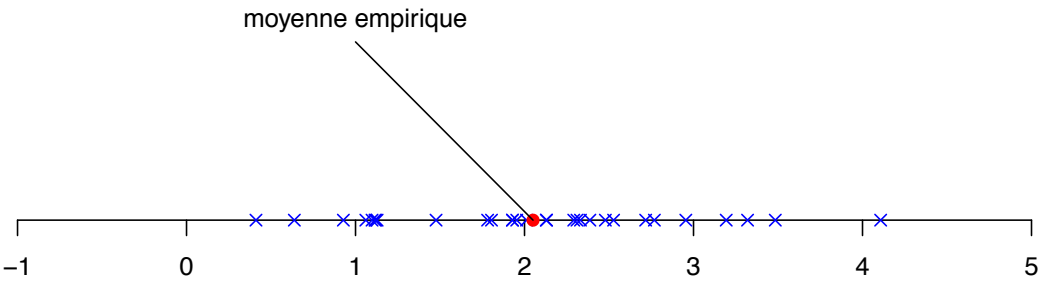
Dans ce cadre, θ_0 s'appelle vraie valeur du paramètre. Typiquement, θ_0 est inconnu et on cherche à l'"estimer" (à l'approcher), à l'aide des données X.

Exemple (modèle fondamental). Supposons l'expérience statistique donnée par $X = (X_1, \dots, X_n)$ et $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$. L'approche fréquentiste consiste à supposer qu'il existe $\theta_0 \in \mathbb{R}$ tel que

$$(X_1, \dots, X_n) \sim \mathcal{N}(\theta_0, 1)^{\otimes n},$$

c'est-à-dire que les données sont i.i.d. de loi commune $\mathcal{N}(\theta_0, 1)$. La Figure 1.1 représente $n = 30$ points tirés aléatoirement de façon indépendante suivant une loi $\mathcal{N}(\theta_0, 1)$. La vraie valeur de θ a été prise égale à $\theta_0 = 2$. On constate que l'échantillon est assez concentré autour de 2, et que la moyenne empirique $\sum_{i=1}^n X_i/n$ est proche de 2.

FIGURE 1.1 – Echantillon de taille $n = 30$ d'une loi $\mathcal{N}(\theta_0, 1)$



Notation. Souvent, lorsque $X = (X_1, \dots, X_n)$ et que l'on travaille avec un modèle d'échantillonage, on notera simplement $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$ au lieu de $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$.

Grandes questions dans le cadre fréquentiste (on peut aussi les poser dans le cadre bayésien)

1. *Estimation.* Il s'agit de construire un estimateur $T(X_1, \dots, X_n)$ qui soit proche, en un sens à préciser, de la vraie valeur θ_0 du paramètre θ.
2. *Intervalles/régions de confiance.* On cherche à construire $\mathcal{C} = \mathcal{C}(X_1, \dots, X_n)$ sous-ensemble (aléatoire) de Θ tel que $\theta_0 \in \mathcal{C}(X_1, \dots, X_n)$ avec grande probabilité.
3. *Tests.* On veut répondre par "vrai" ou "faux" à une propriété donnée de P_θ en construisant $\varphi(X_1, \dots, X_n)$ à valeurs dans $\{0, 1\}$.

1.2.2 Approche bayésienne (intuition)

Thomas Bayes (1702-1761) et Laplace (1749-1827) ont été des pionniers de la méthodologie bayésienne. Dans cette approche, on modélise toutes les quantités inconnues par des variables aléatoires. Ainsi "tout est aléatoire".

Une intuition possible derrière cette approche est que plutôt que de modéliser des quantités par des nombres, il peut être intéressant de les modéliser plutôt par des lois de probabilité.

Prenons comme exemple la température à un endroit précis à un instant du temps. Par exemple, la température dans la pièce où se trouve le lecteur à l'instant présent. On peut répondre en utilisant un thermomètre et en donnant la valeur lue, mettons $T = 18$ degrés. On peut aussi penser qu'il est peut-être raisonnable de tenir compte d'une petite erreur de mesure possible, et donc de répondre plutôt, $\Pi_T = \mathcal{N}(18, v)$, une loi gaussienne centrée en 18, de petite variance, par exemple $v = 0.5$. Si l'on s'intéresse ensuite à la température demain au même endroit, sachant que soit le temps est resté le même, soit la température a augmenté autour de 3 degrés, au lieu de répondre $T' = 18$ ou $T' = 21$, on peut aussi proposer $\Pi_{T'} = \frac{1}{2}\mathcal{N}(18, v) + \frac{1}{2}\mathcal{N}(21, v)$.

La façon dont on formalise l'approche *bayésienne* est de supposer aléatoire le paramètre inconnu θ du modèle, avec la loi de ce paramètre appelée *loi a priori*. Cette loi reflète notre connaissance a priori (éventuelle) du paramètre. Ainsi, par exemple dans le modèle $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, si l'on sait à l'avance que le paramètre θ est positif, il est assez naturel de prendre une loi a priori sur θ qui porte sur $\mathbb{R}^+$, par exemple la loi exponentielle $\mathcal{E}(1)$. Ensuite, une fois des données $X_1, \ldots, X_n$ observées, on va mettre à jour la loi a priori en utilisant l'"information" contenue dans les données. Formellement, cette mise à jour se fait par une opération de conditionnement, ce que nous verrons au Chapitre 2. On obtient alors une nouvelle loi, la loi a posteriori, qui est la 'mise à jour de la loi a priori' une fois les données observées. Notons que si l'on n'a pas de connaissance préalable comme la positivité ci-dessus, on pourra choisir plutôt une loi 'qui met un peu de masse partout', comme la loi $\mathcal{N}(0, 1)$, ce qui reflète le fait que potentiellement le paramètre pourrait être partout sur $\mathbb{R}$.

Illustrons les idées ci-dessus dans le cadre du modèle fondamental $\{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$, avec pour loi a priori sur θ la loi $\mathcal{N}(0, 1)$. Nous verrons au Chapitre 2 qu'après avoir observé n données $X_1, \ldots, X_n$, la loi a posteriori est $\Pi_n = \mathcal{N}\left(\frac{n}{n+1}\overline{X}, \frac{1}{n+1}\right)$. La Figure 1.2 représente la densité de la loi a priori, et de la loi a posteriori Π_3, obtenue après observation des données X_1, X_2, X_3.

1.3 Exemples de modèles

Voici quelques modèles statistiques classiques, décrits par les lois P_θ correspondantes.

→ modèle fondamental

$$P_\theta = \mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}.$$

→ modèle gaussien à moyenne *et* variance inconnues.
Le paramètre du modèle est $\theta = (\mu, \sigma^2)$ et

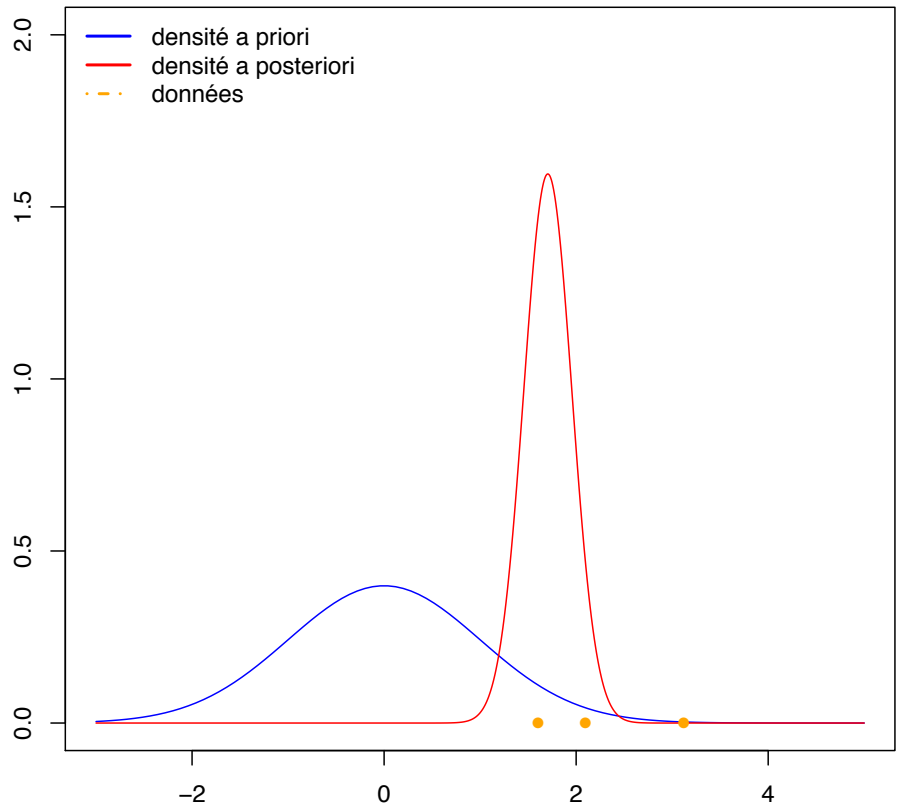
$$P_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2),$$

avec ici $\Theta = \{(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$.

→ modèle gaussien en dimension 2.
Il s'agit de l'ensemble des lois

$$\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \Sigma\right),$$

FIGURE 1.2 – Densités a priori et a posteriori



avec μ_1, μ_2 réels et Σ une matrice 2×2 définie positive, et $\mathcal{N}(v, M)$ désigne la loi d'un vecteur gaussien centré en v et de matrice de variance-covariance M, voir la définition 5 ci-dessous.

→ modèles de translation et changement d'échelle
Il s'agit de la famille de lois de

$$X = \sigma Y + \mu, \quad \text{avec } \sigma > 0,\ \mu \in \mathbb{R},$$

pour Y un variable aléatoire réelle de densité f.
Exercice. Montrer que la densité d'une telle variable X est $\sigma^{-1} f(\frac{\cdot - \mu}{\sigma})$.

→ lois gamma $\Gamma(t, \lambda)$
Il s'agit de lois de densité $f_{t,\lambda}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$,

$$f_{t,\lambda}(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x} \mathbb{1}_{x>0},$$

pour $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u}$ la fonction Gamma. La loi exponentielle $\mathcal{E}(\lambda)$ est la loi $\Gamma(1, \lambda)$.

→ modèle "non-lisse"

$$P_\theta = \text{Unif}[0, \theta],$$

avec pour densité $f_\theta(x) = \theta^{-1} \mathbb{1}_{[0,\theta]}(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

→ nous verrons d'autres exemples dans la suite du cours et en TD.

1.4 Outils de probabilité

Dans le cadre de ce cours, nous travaillerons essentiellement avec des variables aléatoires à valeurs dans $\mathbb{R}$, $\mathbb{R}^d$, $d \geq 1$. Un cas particulier est celui de variables à valeurs discrètes dans $\mathbb{N}$ ou $\mathbb{Z}$. Dans tout ce qui suit, on pourra prendre les espaces d'arrivée égaux à $\mathbb{R}$ pour fixer les idées.

Types de lois de probabilité sur $\mathbb{R}$, lois discrètes et lois à densité.

La loi de probabilité d'une variable aléatoire X à valeurs dans $\mathbb{R}$ est complètement déterminée par sa fonction de répartition, définie pour t réel par

$$F_X(t) = P[X \leq t].$$

On peut classifier des types de lois suivant la nature de la fonction de répartition F_X. Les deux exemples essentiels sont

→ S'il existe une fonction mesurable positive f telle que

$$F_X(t) = \int_{-\infty}^{t} f(u)du,$$

alors X est dite *à densité*, et f est sa densité par rapport à la mesure de Lebesgue.

→ Si la fonction de répartition F_X est constante par morceaux,

$$F_X(t) = \sum_{i \in \mathcal{D}} p_i \mathbb{1}_{[a_i,\infty)}(t),$$

avec $\mathcal{D}$ un ensemble fini ou dénombrable, $(a_i)_{i\in\mathcal{D}}$ une suite de réels, et $\sum_{i\in\mathcal{D}} p_i = 1$, alors la loi de X est dite *discrète*.

Exemples

→ La loi normale $\mathcal{N}(\mu, \sigma^2)$ est, par définition, une loi à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, avec pour densité

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(u-\mu)^2\right\}.$$

→ La loi de Bernoulli $\text{Be}(\theta)$ est la loi discrète de fonction de répartition donnée par $F(t) = p\mathbb{1}_{[0,\infty)} + (1-p)\mathbb{1}_{[1,\infty)}$.

Lois à densité générales.

Soit $(E, \mathcal{E})$ un espace muni d'une tribu $\mathcal{E}$ et soit μ une mesure positive σ-finie sur E.

On rappelle qu'une mesure μ est σ-finie sur E s'il existe une partition dénombrable de E, $E = \cup_{i\geq 1} E_i$, telle que $\mu(E_i) < \infty$. Une telle hypothèse d'ordre technique permet d'invoquer des théorèmes comme celui de Fubini et sera toujours supposée dans la suite.

Par exemple pour $E = \mathbb{R}$ on prend la plupart du temps μ égale à la mesure de Lebesgue sur $\mathbb{R}$. Si la loi P sur E vérifie que pour tout $A \in \mathcal{E}$,

$$P[A] = \int_A p(x) d\mu(x),$$

ce que l'on note aussi $dP(x) = p(x)d\mu(x)$ ou $dP = pd\mu$, on dit que P est à densité p par rapport à μ.

Exemple. On rappelle que δ_x, la masse de Dirac en x, est la mesure positive définie, pour tout A mesurable, par $\delta_x[A] = \mathbb{1}_{x\in A}$.

→ Sur $E = \{0, 1\}$ (ou $E = \mathbb{N}$), la loi de Bernoulli P_θ =Be(θ) admet une densité par rapport à la mesure $\mu = \delta_0 + \delta_1$. En effet, on peut écrire,

$$\begin{aligned} P_\theta[\{0\}] &= 1 - \theta = (1 - \theta)\mu[\{0\}] \\ P_\theta[\{1\}] &= \theta = \theta\mu[\{1\}]. \end{aligned}$$

et donc $dP_\theta = p_\theta d\mu$ avec $p_\theta(x) = (1 - \theta)\mathbb{1}_{x=0} + \theta\mathbb{1}_{x=1}$. On peut par ailleurs remarquer que $P_\theta = (1 - \theta)\delta_0 + \theta\delta_1$.

Lois produits.

Soit P une mesure de probabilité sur $(E, \mathcal{E})$ et Q une mesure de probabilité sur $(F, \mathcal{F})$. Alors la loi produit $P \otimes Q$ est la loi sur l'espace produit $E \times F$ muni de la tribu produit qui vérifie

$$(P \otimes Q)(A \times B) = P(A) \times Q(B),$$

pour tout $A \in \mathcal{E}$ et $B \in \mathcal{F}$. De plus, si P a une densité p par rapport à une mesure dominante μ sur E et Q une densité q par rapport à une mesure dominante ν sur F, alors $P \otimes Q$ a pour densité $p \times q$ par rapport à $\mu \otimes \nu$

$$\begin{aligned} d(P \otimes Q)(x, y) &= p(x)q(y)d(\mu \otimes \nu)(x, y) \\ &= p(x)q(y)d\mu(x)d\nu(y). \end{aligned}$$

Lois produits et indépendance. Deux variables aléatoires X et Y sont indépendantes si et seulement si la loi du couple $P_{(X,Y)}$ est le produit de la loi P_X de X et P_Y de Y, soit $P_{(X,Y)} = P_X \otimes P_Y$.

Exemple. La loi sur $\mathbb{R}^2$ dont la densité par rapport à la densité produit Leb($\mathbb{R}$) $\otimes$ Leb($\mathbb{R}$) est

$$\frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

est une loi produit. En effet, on reconnait le produit de deux lois normales standard $\mathcal{N}(0,1)$. Donc cette loi est $\mathcal{N}(0,1) \otimes \mathcal{N}(0,1)$.

Plus généralement, on peut faire des produits de plusieurs lois, ou de n fois la même loi. Ainsi, $Q = P^{\otimes n}$ est une mesure de probabilité sur l'espace produit E^n. Si P a une densité p par rapport à une mesure dominante μ sur E, soit $dP(x) = p(x)d\mu(x)$, alors $P^{\otimes n}$ a une densité sur E^n par rapport à $\mu^{\otimes n}$, égale à $q(x_1, \ldots, x_n) = \prod_{i=1}^n p(x_i)$.

Vecteurs gaussiens

Pour $d \geq 1$, soit $\mu \in \mathbb{R}^d$ et V une matrice définie positive, c'est-à-dire telle que $y^T V y > 0$ pour tout y non nul de $\mathbb{R}^d$, où y^T désigne la transposée.

Definition 5. Un vecteur aléatoire X de $\mathbb{R}^d$ suit une loi $\mathcal{N}(\mu, V)$ si sa densité par rapport à la mesure de Lebesgue dans $\mathbb{R}^d$ est, pour $|V| = \det(V)$,

$$x \to \frac{1}{\sqrt{(2\pi)^d |V|}} \exp\left\{-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right\}.$$

Notons en particulier que si V est une matrice diagonale (et donc également V^{-1}), la densité de la loi $\mathcal{N}(\mu, V)$ s'exprime comme un produit de densités coordonnée par coordonnée. Cela signifie donc d'après ce qui précède que les coordonnées X_i de X sont alors indépendantes. Si en revanche V n'est pas diagonale, V^{-1} non plus et la densité ne s'écrit pas comme un produit : les coordonnées X_i ne sont alors pas indépendantes.

Convergences.

Pour $x \in \mathbb{R}^d$, $d \geq 1$, on note $\|x\|^2 = \sum_{i=1}^d x_i^2$.

Definition 6. Soit $X_1, \ldots, X_n, \ldots$ et X des variables aléatoires à valeurs dans $\mathbb{R}^d$, $d \geq 1$, définies sur un même espace de probabilité. La suite (X_n) converge en probabilité vers X, ce que l'on note $X_n \xrightarrow{P} X$, si

$$\forall \varepsilon > 0, \qquad P\left[\|X_n - X\| > \varepsilon\right] \to 0 \quad (n \to \infty).$$

Proposition 1. [loi des grands nombres] Soit $(X_n)_{n\geq 1}$ une suite de variables iid à valeurs dans $\mathbb{R}^d$, $d \geq 1$, avec $E[\|X_1\|] < \infty$. Alors (la convergence a aussi lieu presque-sûrement)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} EX_1.$$

Definition 7. Dans une expérience statistique X, $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, l'estimateur $\hat{\theta}(X)$ est consistant si, pour tout $\theta \in \Theta$,

$$\hat{\theta}(X) \overset{P_\theta}{\to} \theta.$$

Definition 8. Soit $(X_n)_{n\geq 1}$ et X des variables aléatoires quelconques à valeurs dans $\mathbb{R}^d$. On dit que X_n converge en loi vers X ce que l'on note $X_n \overset{\mathcal{L}}{\to} X$ si pour toute fonction $f : \mathbb{R}^k \to \mathbb{R}$ continue bornée,

$$E[f(X_n)] \to E[f(X)] \qquad (n \to \infty).$$

De même, on dira que (X_n) converge en loi vers une loi P si $E[f(X_n)] \to E[f(X)]$ pour $X \sim P$, pour toute fontion f continue bornée.

On note la propriété importante suivante de la convergence en loi. On rappelle que pour $A \subset \mathbb{R}^d$, la frontière de A est $\partial A = \overline{A} \setminus \overset{\circ}{A}$.

Proposition 2. Si $X_n \overset{\mathcal{L}}{\to} X$ dans $\mathbb{R}^d$, alors pour tout borélien A de $\mathbb{R}^d$ pour lequel $P[X \in \partial A] = 0$, on a

$$P[X_n \in A] \to P[X \in A] \qquad (n \to \infty).$$

Application. Si $Z_n \overset{\mathcal{L}}{\to} \mathcal{N}(0,1)$, alors pour tout intervalle I de $\mathbb{R}$,

$$P[Z_n \in I] \to P[Z \in I] \qquad (n \to \infty).$$

Proposition 3. [TCL dans $\mathbb{R}^d$] Soit (X_n) une suite de variables aléatoires iid dans $\mathbb{R}^d$, avec $E[\|X_1\|^2] < \infty$. Soit $\mu = EX_1$ et $V = E[(X_1 - E(X_1))(X_1 - E(X_1))^T]$. Alors

$$\sqrt{n}(\overline{X} - \mu) \overset{\mathcal{L}}{\to} \mathcal{N}(0, V),$$

où la $\mathcal{N}(0, V)$ est la loi gaussienne centrée sur $\mathbb{R}^d$ de covariance V.

Definition 9. Un estimateur $\hat{\theta} = \hat{\theta}(X)$ de θ est dit asymptotiquement normal si, pour X de loi $P_\theta^{(n)}$, et pour Σ_θ une matrice positive, quand $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta).$$

Proposition 4. [théorème de l'image continue] Soient X_n, X des variables aléatoires à valeurs dans $\mathbb{R}^d$. Soit $g : \mathbb{R}^d \to \mathcal{Y}$ une fonction continue, pour $\mathcal{Y}$ un espace métrique quelconque. Alors $X_n \xrightarrow{\mathcal{L}} X$ implique $g(X_n) \xrightarrow{\mathcal{L}} g(X)$. Egalement, $X_n \xrightarrow{P} X$ implique $g(X_n) \xrightarrow{P} g(X)$.

Proposition 5. [Lemme de Slutsky] Soient X_n, Y_n, Z_n des suites de variables aléatoires, X une variable aléatoire fixée, et a, b des constantes. On suppose

$$X_n \xrightarrow{\mathcal{L}} X, \quad Y_n \xrightarrow{\mathcal{L}} b, \quad Z_n \xrightarrow{\mathcal{L}} a.$$

Alors $Z_n X_n + Y_n \xrightarrow{\mathcal{L}} aX + b$.

Remarques. Si $a = 0$, alors $aX = 0X = 0$.
Si a est une constante, alors on peut vérifier que convergence que convergence en loi ou en proba vers a sont équivalentes : $Z_n \xrightarrow{\mathcal{L}} a$ si et seulement si $Z_n \xrightarrow{P} a$.

Exercice. Montrer que si X_n est asymptotiquement normal, alors X_n est consistant. *[On pourra écrire $X_n - \theta = \sqrt{n}(X_n - \theta)/\sqrt{n}$]*

1.5 Outils de statistique

Nous introduisons d'abord une notion de risque qui sera précisée dans la suite du cours. Dans la suite, $\hat{\theta} = \hat{\theta}(X)$ est un estimateur dans une expérience statistique $(X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\})$.

Point important : estimateurs et statistiques sont des fonctions mesurables de X seulement ; elles ne sont pas autorisées pas à dépendre du paramètre inconnu θ (sinon, on pourrait prendre $\hat{\theta} = \theta$).

Definition 10.

- Cas où $\Theta \subset \mathbb{R}$. Le risque quadratique d'un estimateur $\hat{\theta}(X)$ au point θ est la fonction $\theta \to R(\theta, \hat{\theta})$ définie par

$$R(\theta, \hat{\theta}) = E_\theta \left[(\hat{\theta} - \theta)^2 \right] = \int (\hat{\theta}(x) - \theta)^2 dP_\theta(x).$$

- Cas où (Θ, d) est un espace métrique. Le risque quadratique est

$$R(\theta, \hat{\theta}) = E_\theta \left[d(\hat{\theta}, \theta)^2 \right].$$

Typiquement, on souhaitera contrôler le risque en donnant des bornes pour celui-ci.

Voyons maintenant quelques inégalités pour contrôler la probabilité de déviation $P_\theta[|\hat{\theta} - \theta| \geq t]$.

Inégalité de Markov. Soit Y une variable aléatoire réelle et $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ une fonction croissante. Alors

$$P[|Y| \geq t] \leq \frac{1}{\psi(t)} E[\psi(|Y|)].$$

En particulier, pour la fonction $x \to x^p$, avec p entier, on obtient

$$P[|Y| \geq t] \leq t^{-p} E[|Y|^p].$$

L'*inégalité de Tchébychev* est un cas particulier, avec $\mathrm{Var}[Y] = E[(Y - EY)^2]$,

$$P[|Y - EY| \geq t] \leq \frac{1}{t^2} \mathrm{Var}[Y].$$

Preuve.

L'inégalité de Markov découle de, en utilisant la croissance de ψ,

$$E\psi(|Y|) \geq E\left[\psi(|Y|) \mathbb{1}_{|Y| \geq t}\right] \geq \psi(t) E[\mathbb{1}_{|Y| \geq t}] = \psi(t) P[|Y| \geq t].$$

On en déduit l'inégalité de Tchébychev en l'appliquant à $Z = E - EY$ et $\psi(x) = x^2$.

Conséquence de l'inégalité de Tchébychev.

$$P_\theta[|\hat{\theta} - \theta| \geq t] \leq \frac{1}{t^2} E_\theta[(\hat{\theta} - \theta)^2] = \frac{R(\theta, \hat{\theta})}{t^2}.$$

Ainsi, un risque quadratique petit implique qu'avec grande probabilité, $|\hat{\theta} - \theta|$ est petit.

Décomposition biais-variance, cas où $\Theta \subset \mathbb{R}$. On peut toujours décomposer

$$\hat{\theta} - \theta = \hat{\theta} - E_\theta \hat{\theta} + E_\theta \hat{\theta} - \theta.$$

En prenant le carré puis l'espérance, et en utilisant la linéarité de l'espérance, qui donne $E_\theta(\{\hat\theta - E_\theta\hat\theta\}\{E_\theta\hat\theta - \theta\}) = \{E_\theta\hat\theta - \theta\}E_\theta(\hat\theta - E_\theta\hat\theta) = 0$, on obtient

$$\begin{aligned} E_\theta(\hat\theta - \theta)^2 &= E\left[(\hat\theta - E_\theta\hat\theta)^2\right] + \left[E_\theta\hat\theta - \theta\right]^2 + 2E_\theta\left[\{\hat\theta - E_\theta\hat\theta\}\{E_\theta\hat\theta - \theta\}\right] \\ &= \quad \left[E_\theta\hat\theta - \theta\right]^2 \quad + \quad E\left[(\hat\theta - E_\theta\hat\theta)^2\right] \\ &= \text{Biais au carré} \quad + \quad \text{Variance} \end{aligned}$$

Pour rendre le risque petit, on cherchera donc à rendre à la fois le biais et la variance petits.

Exemples de calculs de risques. Soient $X_1, \dots, X_n$ iid $\mathcal{N}(\theta, 1)$

→ estimateur constant $\hat\theta = 1$.

> $R(\theta, \hat\theta) = E_\theta(1-\theta)^2 = (\theta-1)^2$. Le risque est imbattable si $\theta = 1$ puisqu'il est nul, mais si $\theta \neq 1$ il est strictement positif et ne tend pas vers 0.

→ estimateur $\hat\theta(X) = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$

> Effectuons le calcul explicite pas à pas
>
> $$\begin{aligned} R(\theta, \hat\theta) = E_\theta(\overline{X} - \theta)^2 &= \frac{1}{n^2}E_\theta\Big[\sum_{i=1}^n (X_i - \theta)\Big]^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n E_\theta(X_i - \theta)^2 + \frac{1}{n^2}\sum_{i\neq j} E_\theta\left[(X_i - \theta)(X_j - \theta)\right] \end{aligned}$$
>
> Pour $i \neq j$, les variables X_i et X_j sont indépendantes donc
>
> $$E_\theta[(X_i - \theta)(X_j - \theta)] = E_\theta[X_i - \theta]E_\theta[X_j - \theta] = 0.$$
>
> Les X_i sont de même loi, donc $E_\theta[(X_i - \theta)^2] = E_\theta[(X_1 - \theta)^2] = \mathrm{Var}_\theta[X_1]$ pour tout i, soit
>
> $$R(\theta, \hat\theta) = \frac{1}{n}\mathrm{Var}_\theta[X_1] = \frac{1}{n}.$$
>
> D'un point de vue global, le risque est bien meilleur que celui de l'estimateur constant. On peut aussi aller plus vite dans le calcul ci-dessus en écrivant
>
> $$R(\theta, \hat\theta) = \frac{1}{n^2}\mathrm{Var}_\theta\Big[\sum_{i=1}^n X_i\Big] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\theta(X_i) = \frac{1}{n}\mathrm{Var}_\theta(X_1) = \frac{1}{n},$$
>
> en utilisant que pour des variables indépendantes, la variance de la somme est la somme des variances.

Exercice. Soit $X \sim \mathrm{Bin}(n, p)$, la loi binomiale de paramètres n, p. On rappelle qu'il s'agit de la loi de $\sum_{i=1}^n \varepsilon_i$ où ε_i sont iid de loi de Bernoulli $\mathrm{Be}(p)$. Pour $\hat\theta = X/n$,

1. écrire la décomposition biais-variance.
2. montrer que $R(\theta, \hat\theta) \leq 1/(4n)$.

Intervalles et régions de confiance.

On note $[a, b]$ un intervalle d'extrémités a et b.

Definition 11. Soit $\alpha > 0$.

- Cas $\Theta \subset \mathbb{R}$. Un intervalle de confiance de niveau (au moins) $1-\alpha$ est un intervalle aléatoire $I(X) = [a(X), b(X)]$ où $a(X), b(X)$ sont des statistiques à valeurs dans $\mathbb{R}$ vérifiant
$$P_\theta[\,\theta \in I(X)\,] \geq 1-\alpha \qquad \forall\, \theta \in \Theta$$
- Cas général. Une région de confiance de niveau (au moins) $1-\alpha$ est $\mathcal{R}(X) \subset \Theta$ avec
$$P_\theta[\,\theta \in \mathcal{R}(X)\,] \geq 1-\alpha \qquad \forall\, \theta \in \Theta$$

On remarquera que Θ lui-même est toujours une région de confiance, de niveau de confiance égal à 1. Cependant, on souhaite en général trouver une région la plus petite possible (ou proche de la plus petite), telle que le niveau de confiance reste au moins de $1-\alpha$.

Construction d'intervalles de confiance.

1ère technique [utilisation de bornes en probabilité]. On part d'une inégalité comme celle de Tchébychev qui contrôle la probabilité de déviation de $\hat{\theta}$ à θ. Pour $t > 0$,
$$P_\theta[|\hat{\theta} - \theta| > t] \leq t^{-2} R(\hat{\theta}, \theta).$$

→ *Exemple : expérience binomiale.* On observe $X \sim \mathrm{Bin}(n, \theta)$. On pose $\hat{\theta} = X/n$. D'après l'exercice ci-dessus, $R(\hat{\theta}, \theta) \leq 1/(4n)$, donc pour tout $t > 0$
$$P_\theta[|\hat{\theta} - \theta| > t] \leq \frac{1}{4nt^2}$$
soit aussi, en prenant l'événement complémentaire,
$$P_\theta[\theta \in [\hat{\theta} - t, \hat{\theta} + t]] \geq 1 - \frac{1}{4nt^2}.$$
Ainsi pour que $[\hat{\theta} - t, \hat{\theta} + t]$ soit un intervalle de confiance de niveau $1-\alpha$ il suffit de choisir t de sorte que $\alpha = 1/(4nt^2)$ soit $t = 1/\sqrt{4n\alpha}$. On a donc obtenu que
$$I(X) = \left[\hat{\theta} - \frac{1}{\sqrt{4n\alpha}}, \hat{\theta} + \frac{1}{\sqrt{4n\alpha}}\right] = \left[\hat{\theta} \pm \frac{1}{\sqrt{4n\alpha}}\right]$$
est un intervalle de confiance de niveau au moins $1-\alpha$.

Remarque importante : $I(X)$ ne doit pas dépendre de θ ! Or en général $R(\hat{\theta}, \theta)$ dépend de θ. Par exemple dans l'exemple binomial, il vaut $\theta(1-\theta)/n$. C'est pourquoi on peut dans ce cas

le majorer pour obtenir une quantité indépendante de θ.

On peut aussi utiliser d'autres inégalités à la place de celle de Tchébychev, par exemple celles de Markov, Hoeffding ... (voir TDs). On peut parfois aussi utiliser la loi de $\hat{\theta}$ si elle est connue, ce qui n'est pas très fréquent, pour construire l'intervalle de confiance.

Exercice. Dans le modèle fondamental avec observations $X_1, \dots, X_n$ iid de loi $\mathcal{N}(\theta, 1)$, construire un intervalle de confiance pour θ de niveau (au moins) $1-\alpha$ à partir de $\hat{\theta} = \overline{X}$. On donne l'inégalité $P[|\mathcal{N}(0,1)| \geq t] \leq e^{-t^2/2}$ pour tout $t > 0$.

2ème technique [intervalles de confiance asymptotiques]. On peut utiliser une convergence en loi quand $n \to \infty$ pour construire un intervalle de confiance asymptotique. Supposons, pour $\Theta \subset \mathbb{R}$, que l'on dispose d'un estimateur $\hat{\theta} = \hat{\theta}_n(X)$ asymptotiquement normal, soit

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)).$$

et que $\theta \to \sigma^2(\theta)$ est continue. Soit z_α tel que $P[|\mathcal{N}(0,1)| \leq z_\alpha] = 1-\alpha$. Alors

$$I(X) = \left[\hat{\theta}_n(X) - z_\alpha \frac{\sigma(\hat{\theta}_n(X))}{\sqrt{n}} \, , \, \hat{\theta}_n(X) + z_\alpha \frac{\sigma(\hat{\theta}_n(X))}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$, c'est-à-dire un intervalle tel que

$$\liminf_{n \to \infty} P_\theta[\theta \in I_n(X)] = 1 - \alpha.$$

Preuve.

On constate que

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} = \frac{\sigma(\hat{\theta}_n)}{\sigma(\theta)} \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)}.$$

Comme $\hat{\theta}_n$ est asymptotiquement normal, il est consistant, voir exercice en 1.4, donc $\hat{\theta}_n \xrightarrow{P} \theta$. Par image continue (Proposition 4), on en déduit $\sigma(\hat{\theta}_n) \xrightarrow{P} \sigma(\theta)$. Par ailleurs,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Grâce au lemme de Slutsky (Proposition 5), on en déduit

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

La proposition 2 permet d'en déduire

$$P_\theta \left[-z_\alpha \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \leq z_\alpha \right] \to P\left[|\mathcal{N}(0,1)| \leq z_\alpha\right] = 1 - \alpha,$$

ce qui s'écrit exactement $P_\theta[\theta \in I(X)] \to 1-\alpha$, ce qu'il fallait démontrer.

VRAISEMBLANCE.

Soit $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ un modèle statistique et $X_1, \dots, X_n$ des observations iid de loi P_θ. Supposons le modèle dominé par rapport à une mesure dominante μ, soit $dP_\theta = p_\theta d\mu$. La densité du n-uplet $(X_1, \dots, X_n)$ par rapport à $\mu^{\otimes n}$ est donc $p_\theta(x_1) \dots p_\theta(x_n)$. Cette densité prise calculée en θ et aux points d'observation s'appelle *vraisemblance*.

Definition 12. La fonction vraisemblance est la fonction

$$\mathcal{V} : \theta \to \prod_{i=1}^{n} p_\theta(X_i).$$

1.6 Lois conditionnelles

On commence par rappeler, pour A, B des événements avec $P(B) > 0$, la définition de la probabilité de 'A sachant B'. Celle-ci est définie par

$$P[A \,|\, B] = \frac{P(A \cap B)}{P(B)}.$$

1.6.1 Le cas discret

Cadre. Soit E un ensemble dénombrable, on peut penser à $\mathbb{N}$ pour fixer les idées. Soient X et Y deux variables aléatoires à valeurs dans E.

On souhaite définir la loi conditionnelle de Y sachant X.

Notons que, s'agissant de variables discrètes, les lois de X et Y sont complètement définies par les données de $P[X = e]$ et $P[Y = e]$ pour tous les éléments possibles e de E. Si Q est la loi $\mathcal{L}(Y \,|\, X = x)$ que l'on cherche à définir, il suffit donc aussi de se donner $Q(\{e\})$ pour tout $e \in E$. On définit tout simplement ces quantités à l'aide de la formule ci-dessus pour la probabilité de A sachant B.

Definition 13. Soit $x \in E$ fixé. La loi conditionnelle de $Y \,|\, X = x$, parfois aussi notée $\mathcal{L}(Y \,|\, X = x)$ est définie par, pour tout $e \in E$, et $x \in E$ tel que $P[X = x] > 0$,

$$P[Y = e \,|\, X = x] = \frac{P[Y = e, X = x]}{P[X = x]}.$$

Exemple. Soient Y, Z deux variables aléatoire *indépendantes* de lois $Y \sim \mathrm{Be}(1/2)$ et $Z \sim \mathrm{Be}(1/2)$. On pose $X = Y + Z$. Quelle est la loi conditionnelle $\mathcal{L}(Y \,|\, X = 1)$?

Notons déjà que $X = 1$ si et seulement $Y = 1$ et $Z = 0$, ou bien $Y = 0$ et $Z = 1$. En utilisant la définition de la loi conditionnelle ainsi que l'indépendance de Y et Z,

$$P[Y = 1 \,|\, X = 1] = \frac{P[X = 1, Y = 1]}{P[X = 1]} = \frac{P[Z = 0, Y = 1]}{P[Y = 1, Z = 0] + P[Y = 0, Z = 1]}$$
$$= \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}.$$

Par ailleurs, comme Y ne prend que les valeurs 0 ou 1, on en déduit que $P[Y = 0 \,|\, X = 1] = 1 - P[Y = 1 \,|\, X = 1] = 1 - \frac{1}{2} = \frac{1}{2}$. On en conclut que

$$\mathcal{L}(Y \,|\, X = 1) = \mathrm{Be}\Big(\frac{1}{2}\Big) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

En procédant de la même manière, on obtient (exercice)

$$\mathcal{L}(Y \,|\, X = 0) = \delta_0 \quad \text{et} \quad \mathcal{L}(Y \,|\, X = 2) = \delta_1.$$

Par extension, on définit la loi conditionnelle de Y sachant X, notée $\mathcal{L}(Y \,|\, X)$, comme la loi égale à $\mathcal{L}(Y \,|\, X = x)$ si $X = x$. Dans l'exemple ci-dessus,

$$\mathcal{L}(Y \,|\, X) = \begin{cases} \delta_0 & \text{si } X = 0 \\ \mathrm{Be}\left(\frac{1}{2}\right) & \text{si } X = 1 \\ \delta_1 & \text{si } X = 2, \end{cases}$$

ce qu'on peut aussi écrire de manière un peu plus compacte comme

$$\mathcal{L}(Y \,|\, X) = \left(1 - \frac{X}{2}\right)\delta_0 + \frac{X}{2}\delta_1.$$

1.6.2 Le cas à densité

Cadre.

On se donne

- un espace E muni d'une tribu $\mathcal{E}$ et un espace F muni d'une tribu $\mathcal{F}$
- une mesure α positive σ-finie sur E et une mesure β positive σ-finie sur F
- une variable aléatoire X sur E et une variable aléatoire Y sur F.

On *suppose* que le couple (X, Y) admet une densité notée $f(x, y)$ par rapport à $\alpha \otimes \beta$, ce que l'on écrit aussi, si $P_{X,Y}$ désigne la loi du couple,

$$dP_{X,Y}(x, y) = f(x, y) d\alpha(x) d\beta(y).$$

Loi et densité marginales.

Proposition 6. Dans le cadre ci-dessus, la loi de X seule, appelée loi marginale de X, est la loi P_X de densité f_X donnée par

$$f_X(x) = \int f(x,y)d\beta(y).$$

Preuve.

Pour toute fonction g mesurable bornée, en utilisant le théorème de Fubini,

$$\begin{aligned} E[g(X)] &= \int\int g(x)f(x,y)d\alpha(x)d\beta(y) \\ &= \int g(x)\left[\int f(x,y)d\beta(y)\right]d\alpha(x) = \int g(x)f_X(x)d\alpha(x). \end{aligned}$$

De même, la loi marginale de Y est la loi P_Y dont la densité sur F par rapport à β est donnée par $f_Y(y) = \int f(x,y)d\alpha(x)$.

✍ A partir de la loi du couple (X,Y), on a facilement déduit les lois individuelles de X et Y. Il est important de noter que l'opération inverse n'est pas possible en général sans hypothèse supplémentaire. Il y a en général beaucoup de lois jointes possibles correspondant à deux lois marginales données, voir TDs.

Loi conditionnelle.

Definition 14. La loi conditionnelle de Y sachant $X = x$ est la loi de densité, sur F par rapport à β, donnée par, pour $f_X(x) > 0$,

$$f_{Y\,|\,X=x}(y) = \frac{f(x,y)}{\int f(x,y)d\beta(y)} = \frac{f(x,y)}{f_X(x)}.$$

On notera parfois $f(y\,|\,x)$ au lieu de $f_{Y\,|\,X=x}(y)$ s'il n'y a pas de risque de confusion. Notons que par définition, $y \to f(y\,|\,x)$ est une densité par rapport à β, soit $\int f(y\,|\,x)d\beta(y) = 1$.

Notons que pour avoir une quantité définie pour tous les x de E, on peut étendre la définition de $f_{Y\,|\,X=x}(y)$ au cas où $f_X(x) = 0$ en posant le quotient ci-dessus égal à une valeur quelconque (par exemple 0) lorsque $f_X(x) = 0$. Ces points x n'auront typiquement pas d'incidence dans les calculs car l'ensemble des x tels que $f_X(x) = 0$ est un ensemble de P_X-mesure nulle.

Exercice. Vérifier que le cas discret est un cas particulier de la formule ci-dessous, pour lequel

E et F sont dénombrables, et α, β sont les mesures de comptage sur E et F respectivement, $\alpha = \sum_{e\in E} \delta_e$, $\beta = \sum_{f\in F} \delta_f$.

✍ A partir de la densité conditionnelle de $Y \mid X$ et de la densité marginale de X, on retrouve la densité jointe du couple (X, Y), puisque par définition $f(x, y) = f_{Y\mid X=x}(y) f_X(x)$.

Exemple. Soit un couple (X, Y) de variables aléatoires sur $E = \mathbb{R}^+ \times \mathbb{R}^+$ de densité

$$f(x, y) = xe^{-x(y+1)}$$

par rapport à la mesure de Lebesgue restreinte à $\mathbb{R}^+ \times \mathbb{R}^+$. Déterminons la loi conditionnelle de Y sachant X. Il suffit de diviser la densité jointe $f(x, y)$ par la densité marginale $f_X(x) = \int_0^\infty xe^{-x(y+1)}dy = e^{-x}$. Ainsi

$$f_{Y\mid X=x}(y) = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy}$$

On reconnait la densité d'une loi exponentielle de paramètre x. Ainsi, $\mathcal{L}(Y \mid X = x) = \mathcal{E}(x)$. On écrit aussi $\mathcal{L}(Y \mid X) = \mathcal{E}(X)$. Notons que la loi marginale de X a pour densité e^{-x}, ainsi $\mathcal{L}(X) = \mathcal{E}(1)$.

Utilisation du symbole $\propto$='proportionnel à'. Une autre façon de faire pour déterminer la densité conditionnelle est de remarquer qu'il s'agit de reconnaître dans l'expression $f(x, y)/f_X(x)$ une densité en y. En ce sens $f_X(x)$ est simplement une constante de normalisation. De même, tout facteur dépendant seulement de x dans $f(x, y)$ peut se mettre en facteur et intervient seulement dans la normalisation. On écrit ceci à l'aide du symbole proportionnel à '$\propto$'

$$xe^{-x(y+1)} = xe^{-x}e^{-xy} \propto e^{-xy}.$$

La loi dont la densité en y est proportionnelle à e^{-xy} est bien la loi $\mathcal{E}(x)$. Cette méthode évite de devoir calculer la densité marginale $f_X(x)$. Dans cet exemple, ce calcul était immédiat mais ce n'est pas toujours le cas, nous verrons d'autres exemples au prochain chapitre.

Exercice. Déterminer la densité de la loi marginale de Y et montrer que la loi conditionnelle de $X \mid Y$ est une loi Gamma $\Gamma(2, Y + 1)$.

Notion d'espérance conditionnelle

On rappelle l'abbréviation $f(y \mid x) = f_{Y\mid X=x}(y)$.

Definition 15. Si $E[|Y|] < \infty$, on définit l'espérance conditionnelle $E[Y \mid X]$ par

$$E[Y \mid X] = \int y f(y \mid X) d\beta(y).$$

Plus généralement, pour ϕ mesurable avec $\phi(Y)$ intégrable,

$$E[\phi(Y) \mid X] = \int \phi(y) f(y \mid X) d\beta(y).$$

Proposition 7. Pour toute $h : E \times F \to \mathbb{R}$ mesurable, à condition que la variable $h(X,Y)$ soit intégrable,

$$E[h(X,Y)] = E[E[h(X,Y) \,|\, X]] = \int\int h(x,y) dP_{Y \,|\, X=x}(y) dP_X(x).$$

En particulier, sous les mêmes conditions, si $h(X,Y) = \varphi(X)\psi(Y)$, pour φ, ψ mesurables,

$$E[\psi(Y)\varphi(X)] = E[E[\psi(Y) \,|\, X]\varphi(X)].$$

Preuve.

$$\begin{aligned} E[h(X,Y)] &= \int\int h(x,y) f(x,y) d\alpha(x) d\beta(y) \\ &= \int\int h(x,y) \frac{f(x,y)}{f_X(x)} f_X(x) d\alpha(x) d\beta(y) \\ &= \int \left[\int h(x,y) dP_{Y \,|\, X=x}(y) \right] f_X(x) d\alpha(x), \end{aligned}$$

où on a utilisé le théorème de Fubini pour la dernière égalité.

Proposition 8. Dans le cadre précédent, soit (X,Y) un couple de variables aléatoires de densité $f(x,y)$ par rapport à $\alpha \otimes \beta$. Supposons Y de carré intégrable : $E[Y^2] < \infty$. Alors

$$\inf \left\{ E[\,(Y - h(X))^2\,], \quad Eh(X)^2 < \infty \right\} = E[\,(Y - g(X))^2\,],$$

où $g(u) = E[Y \,|\, X = u]$.

Preuve.

On note que pour toute h telle que $E[h(X)^2] < \infty$,

$$E[(Y - h(X))^2] = E[(Y - g(X))^2] + E[(g(X) - h(X))^2].$$

En effet, le double produit est nul puisque, comme $g(X) = E[Y \,|\, X]$,

$$\begin{aligned} E[(Y - g(X))(g(X) - h(X))] &= E[E[Y - g(X) \,|\, X](g(X) - h(X))] \\ &= E[(g(X) - g(X))(g(X) - h(X))] = 0. \end{aligned}$$

On déduit de la première identité ci-dessus que $E[(Y - h(X))^2] \geq E[(Y - g(X))^2]$ pour toute h telle que $E[h(X)^2] < \infty$. Pour conclure il suffit de montrer que $E[g(X)^2] < \infty$. Or

$$Eg(X)^2 = E\left[\int yf(y \mid X)d\beta(y)\right]^2 = \int \left[\int yf(y \mid x)d\beta(y)\right]^2 f_X(x)d\alpha(x)$$
$$\leq \int\int y^2 f(y \mid x)d\beta(y)f_X(x)d\alpha(x) = \int\int y^2 f(x,y)d\beta(y)d\alpha(x),$$

où la dernière ligne résulte de l'inégalité de Cauchy-Schwarz (ou Jensen pour le carré)

$$\left[\int yf(y \mid x)d\beta(y)\right]^2 \leq \int y^2 f(y \mid x)d\beta(y) \int f(y \mid x)d\beta(y) = \int y^2 f(y \mid x)d\beta(y),$$

puisque par définition $y \to f(y \mid x)$ est une densité. En utilisant le théorème de Fubini, nous constatons que $\int\int y^2 f(x,y)d\beta(y)d\alpha(x) = \int\int y^2 f(x,y)d\alpha(x)d\beta(y) = E[Y^2] < \infty$ par hypothèse, ce qui montre $E[g(X)^2] < \infty$.

Interprétation en termes de projection. Soit $\mathcal{H}$ l'espace vectoriel (fermé) de toutes les fonctions $h(X)$ avec h mesurable et $E[h(X)^2] < \infty$. La fonction g, l'espérance conditionnelle de Y sachant $X = \cdot$, est simplement la projection orthogonales de Y sur $\mathcal{H}$.

Remarque culturelle. Le cadre à densité n'est pas le seul cadre où l'on puisse définir des lois conditionnelles. On peut plus généralement proposer une définition de la loi conditionnelle comme opérateur de 'désintégration', dans l'esprit de l'identité de la Proposition 7. Cette notion plus générale est utile notamment en statistique bayésienne pour des modèles complexes, lorsque θ n'est plus un paramètre d'un espace de dimension finie mais par exemple une fonction, mais nous ne la considèrerons pas dans le cadre de ce cours.

1.7 Plan du cours

1. Introduction
2. L'approche bayésienne
3. Bayésien et théorie de la décision
4. Critères de choix de lois a priori
5. Convergences de lois a priori
6. Tests bayésiens
7. Algorithmes de simulation

CHAPITRE 2

L'approche bayésienne

Nous définissons le cadre bayésien, avec les notions de lois a priori et a posteriori. Nous expliquons comment calculer les densités a posteriori grâce à la formule de Bayes. Nous définissons les notions de lois conjuguées, de lois impropres et de régions de crédibilité, et présentons deux façons de construire ces dernières.

2.1 Définitions

Le point de départ est toujours une expérience statistique : on se donne X objet aléatoire et $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ un modèle statistique. On supposera ici $\Theta \subset \mathbb{R}^d$, pour $d \geq 1$ fixé.

Definition 1. (1ère partie) Le *cadre bayésien* consiste dans un premier temps à munir l'espace des paramètres Θ d'une mesure de probabilité Π, appelée loi a priori.

On ne connait pas θ, on lui attribue une loi de probabilité. Dans le cadre bayésien, θ a pour loi Π, que l'on *choisit*.

Cadre dominé.

On suppose toujours dans la suite que

→ les lois P_θ ont toutes une densité p_θ par rapport à une mesure positive σ-finie μ sur E

$$dP_\theta = p_\theta d\mu$$

→ la loi Π a une densité π par rapport à une mesure positive σ-finie ν sur Θ

$$d\Pi = \pi d\nu$$

L'étape suivante consiste à dire comment intervient X. Plus précisément, nous allons spécifier la loi du couple (X, θ). Pour cela, nous supposerons que l'application

$$\begin{array}{ccc} (x,\theta) & \to & p_\theta(x) \\ E \times \Theta & \to & \mathbb{R}^+ \end{array} \tag{1}$$

est mesurable, où $E \times \Theta$ est muni de la tribu produit. Cette hypothèse est presque toujours satisfaite, elle sert simplement à ce que les quantités qui suivent soient bien définies.

Proposition 1. Supposons l'application (1) mesurable. Alors la fonction

$$(x, \theta) \to p_\theta(x)\pi(\theta)$$

est une densité de probabilité par rapport à $\mu \otimes \nu$.

Preuve.

Grâce à (1), l'application $(x, \theta) \to p_\theta(x)\pi(\theta)$ est mesurable comme produit de fonctions mesurables, et positive par définition. Le théorème de Fubini donne alors que

$$\begin{aligned} \int \int p_\theta(x)\pi(\theta)d\mu(x)d\nu(\theta) &= \int \left[\int p_\theta(x)d\mu(x) \right] \pi(\theta)d\nu(\theta) \\ &= \int 1 \cdot \pi(\theta)d\nu(\theta) = 1. \end{aligned}$$

Definition 1. (2ème partie) Dans le *cadre bayésien*, on suppose (1) et on définit

$$\mathcal{L}(\,(X,\theta)\,) = \text{ loi de densité } p_\theta(x)\pi(\theta)$$

par rapport à $\mu \otimes \nu$. La loi de θ et la loi conditionnelle $\mathcal{L}(X \,|\, \theta)$ sont alors données par

$$\begin{array}{c} \theta \sim \Pi \\ X \,|\, \theta \sim P_\theta. \end{array} \tag{2}$$

Vérifions que les lois de θ et de $X \,|\, \theta$ sont bien celles données dans la définition. La densité de θ s'obtient en intégrant la densité jointe

$$\int p_\theta(x)\pi(\theta)d\mu(x) = \pi(\theta),$$

donc $\theta \sim \Pi$. Ceci est cohérent avec la première partie de la Définition 1.

La densité de $X \,|\, \theta$ s'obtient par la formule de la densité conditionnelle

$$f_{X\,|\,\theta}(x) = \frac{p_\theta(x)\pi(\theta)}{\int p_\theta(x)\pi(\theta)d\mu(x)} = \frac{p_\theta(x)\pi(\theta)}{\pi(\theta)} = p_\theta(x),$$

donc $\mathcal{L}(X \mid \theta) = P_\theta$ comme annoncé.

✍ La loi marginale de X s'obtient également par intégration de la densité jointe. C'est la loi de densité $x \to \int p_\theta(x)\pi(\theta)d\theta$ par rapport à μ.

✍ Attention ! Dans le cadre bayésien, la loi de X *n'est donc pas* P_θ, qui est la loi de $X \mid \theta$.

Une fois défini le cadre, la façon bayésienne de construire un 'estimateur' est de conditionner l'information de départ, contenue dans la loi a priori, par l'observations, c'est-à-dire X. On obtient ainsi la définition suivante.

Definition 2. La loi a posteriori est la loi conditionnelle $\mathcal{L}(\theta \mid X)$ dans le cadre bayésien de la définition 1. C'est une loi sur Θ, qui est notée $\Pi[\cdot \mid X]$.

Notons que sous l'hypothèse (1) que nous supposerons vérifiée dans la suite, il est équivalent de se donner la loi jointe de (X, θ) ou les deux lois de θ et de $X \mid \theta$ suivant (2). Nous fairons donc simplement référence à (2) quand nous parlerons de formalisme ou de cadre bayésien.

2.2 Formule de Bayes

Théorème 1. [formule de Bayes] La loi a posteriori a une densité par rapport à ν égale à

$$f_{\theta \mid X=x}(\theta) = \frac{p_\theta(x)\pi(\theta)}{\int p_\theta(x)\pi(\theta)d\nu(\theta)}$$

Preuve.
Il suffit de combiner la définition 1 et la formule de la densité conditionnelle.

Cas du modèle d'échantillonage.

Soit une expérience statistique d'échantillonage où $X = (X_1, \dots, X_n)$ et $P_\theta = P_\theta^{(n)} = \otimes_{i=1}^n P_\theta$. Le formalisme bayésien s'écrit

$$\theta \sim \Pi$$
$$X_1, \dots, X_n \mid \theta \sim P_\theta^{(n)} = \bigotimes_{i=1}^n P_\theta$$

La densité jointe de (X, θ) par rapport à $\mu^{\otimes n} \otimes \nu$ est donc la fonction

$$(x_1, \dots, x_n, \theta) \to p_\theta(x_1) \times \cdots \times p_\theta(x_n) \times \pi(\theta) = \prod_{i=1}^n p_\theta(x_i)\pi(\theta).$$

La loi marginale de $X = (X_1, \ldots, X_n)$ a elle pour densité

$$(x_1, \ldots, x_n) \to \int \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)d\nu(\theta).$$

La formule de Bayes donne donc pour densité conditionnelle de θ sachant X

$$f_{\theta \,|\, X_1=x_1,\ldots,X_n=x_n}(\theta) = \frac{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)}{\int \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)d\nu(\theta)}.$$

Dans la pratique, les données observées sont $X_1, \ldots, X_n$, donc on écrira cette expression directement aux points observés.

Théorème 2. [Bayes et échantillonage] La loi a posteriori dans le modèle d'échantillonage a une densité par rapport à ν égale à

$$f_{\theta \,|\, X_1,\ldots,X_n}(\theta) = \frac{\prod_{i=1}^{n} p_\theta(X_i)\pi(\theta)}{\int \prod_{i=1}^{n} p_\theta(X_i)\pi(\theta)d\nu(\theta)}$$

Preuve.

C'est la même que pour le Théorème 1 avec la densité de P_θ remplacée par celle de $P_\theta^{\otimes n}$.

Interprétation. La densité a posteriori en tant que fonction de θ est proportionnelle à

$$\left[\prod_{i=1}^{n} f_\theta(X_i)\right] \pi(\theta).$$

Cette quantité est le produit de la vraisemblance, cf. Chapitre 1, et de la densité a priori. La loi a posteriori peut donc s'interpréter comme une *mise à jour* de la loi a priori à l'aide des données. C'est l'opération de conditionnement qui permet cette mise à jour.

Remarque. La plupart du temps, μ et ν sont prises égales à la mesure de Lebesgue sur $\mathbb{R}$, ou à une mesure discrète, typiquement la mesure de comptage sur les entiers.

Exemples.

→ L'exemple historique de Bayes.

Thomas Bayes (1763) considère le problème suivant. Une boule de billard roule sur une ligne de longueur 1, avec une probabilité uniforme de s'arrêter en un point. Supposons qu'elle s'arrête en p. Une deuxième boule roule n fois dans les mêmes conditions, et on note X le nombre de fois où elle s'est arrêtée *à gauche* de la première boule. Bayes se demande : connaissant X, quelle inférence peut-on mener sur p ?

Exercice. Ecrire cette expérience dans un formalisme bayésien (c'est le cas de le dire !), où Θ est l'intervalle $[0,1]$. Quelle est le paramètre, la loi a priori ? Répondre à la question de Bayes en calculant la densité a posteriori.

$\rightarrow$ Le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1)\ \theta \in \mathbb{R}\}$

a) *Cas d'une observation* $X = X_1$. Le cadre bayésien s'écrit

$$\begin{aligned} X \,|\, \theta &\sim \mathcal{N}(\theta, 1) \\ \theta &\sim \Pi \end{aligned}$$

Choisissons une loi a priori, prenons $\Pi = \mathcal{N}(0,1)$. Nous avons donc ici

$$\begin{aligned} dP_\theta(x) = p_\theta(x)dx, \qquad & p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \\ d\Pi(\theta) = \pi(\theta)d\theta, \qquad & \pi(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \end{aligned}$$

La loi a posteriori $\Pi[\cdot \,|\, X]$ est une loi sur $\Theta = \mathbb{R}$, de densité par rapport à la mesure de Lebesgue donnée par

$$f_{\theta\,|\,X}(\theta) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}}{\int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} d\theta}.$$

Il s'agit maintenant de déterminer la loi dont la densité en θ est donnée par cette expression.

Méthode 1 – 'on écrit tout'

$$\begin{aligned} f_{\theta\,|\,X}(\theta) &= \frac{e^{-\theta^2 + \theta X - \frac{X^2}{2}}}{\int e^{-\theta^2 + \theta X - \frac{X^2}{2}} d\theta} = \frac{e^{-\left(\theta - \frac{X}{2}\right)^2 - \frac{X^2}{4}}}{\int e^{-\left(\theta - \frac{X}{2}\right)^2 - \frac{X^2}{4}} d\theta} \\ &= \frac{e^{-\left(\theta - \frac{X}{2}\right)^2}}{\int e^{-\left(\theta - \frac{X}{2}\right)^2} d\theta}. \end{aligned}$$

L'intégrale au dénominateur est aussi égale à $\int e^{-u^2} du$, qui vaut $\sqrt{\pi}$, puisque la densité d'une $\mathcal{N}(0, 1/2)$ est $u \to e^{-u^2}/\sqrt{\pi}$ et par définition intègre à 1. Ainsi

$$f_{\theta\,|\,X}(\theta) = \frac{1}{\sqrt{\pi}} e^{-\left(\theta - \frac{X}{2}\right)^2}.$$

On reconnait la densité d'une loi $\mathcal{N}\left(\frac{X}{2}, \frac{1}{2}\right)$.

Méthode 2 – 'proportionnel à'. Le symbole $\propto$ ci-dessous signifie 'à constante de proportionnalité près', où cette constante peut dépendre de tout sauf de θ,

$$\begin{aligned} f_{\theta\,|\,X}(\theta) &\propto e^{-\theta^2 + \theta X - \frac{X^2}{2}} \propto e^{-\theta^2 + \theta X} \\ &\propto e^{-\left(\theta - \frac{X}{2}\right)^2 + \frac{X^2}{4}} \propto e^{-\left(\theta - \frac{X}{2}\right)^2}. \end{aligned}$$

La loi (unique) dont la densité est proportionnelle à cette expression est une loi $\mathcal{N}\left(\frac{X}{2}, \frac{1}{2}\right)$.

On constate qu'il n'est pas utile de garder l'intégrale au dénominateur dans les calculs, puisque c'est une expression qui dépend de X seulement (dans l'exemple c'est même une constante) et pas de θ, et intervient donc seulement en termes de constante de normalisation. Dans la pratique, on utilise donc quasi-systématiquement la méthode du 'proportionnel à'.

b) *Cas de n observations* $X_1, \dots, X_n$. Le cadre bayésien s'écrit

$$\begin{aligned} X = (X_1, \dots, X_n) \,|\, \theta &\sim \mathcal{N}(\theta, 1)^{\otimes n} \\ \theta &\sim \Pi \end{aligned}$$

La loi a posteriori $\Pi[\cdot \,|\, X]$ est une loi sur $\Theta = \mathbb{R}$, de densité par rapport à la mesure de Lebesgue donnée par

$$f_{\theta \,|\, X_1,\dots,X_n}(\theta) = \frac{\left\{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}}\right\} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}}{\int \left\{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}}\right\} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} d\theta}.$$

Déterminons de quelle loi il s'agit avec la méthode du 'proportionnel à'

$$\begin{aligned} f_{\theta \,|\, X_1,\dots,X_n}(\theta) &\propto \exp\left(-\sum_{i=1}^{n} \frac{1}{2}(X_i - \theta)^2 - \frac{\theta^2}{2}\right) \\ &\propto \exp\left(-\frac{n+1}{2}\theta^2 + n\overline{X}\theta - n\overline{X^2}\right) \\ &\propto \exp\left(-\frac{n+1}{2}\left(\theta - \frac{n\overline{X}}{n+1}\right)^2\right). \end{aligned}$$

On en conclut

$$\Pi[\cdot \,|\, X_1, \dots, X_n] = \mathcal{N}\left(\frac{n\overline{X}}{n+1}, \frac{1}{n+1}\right).$$

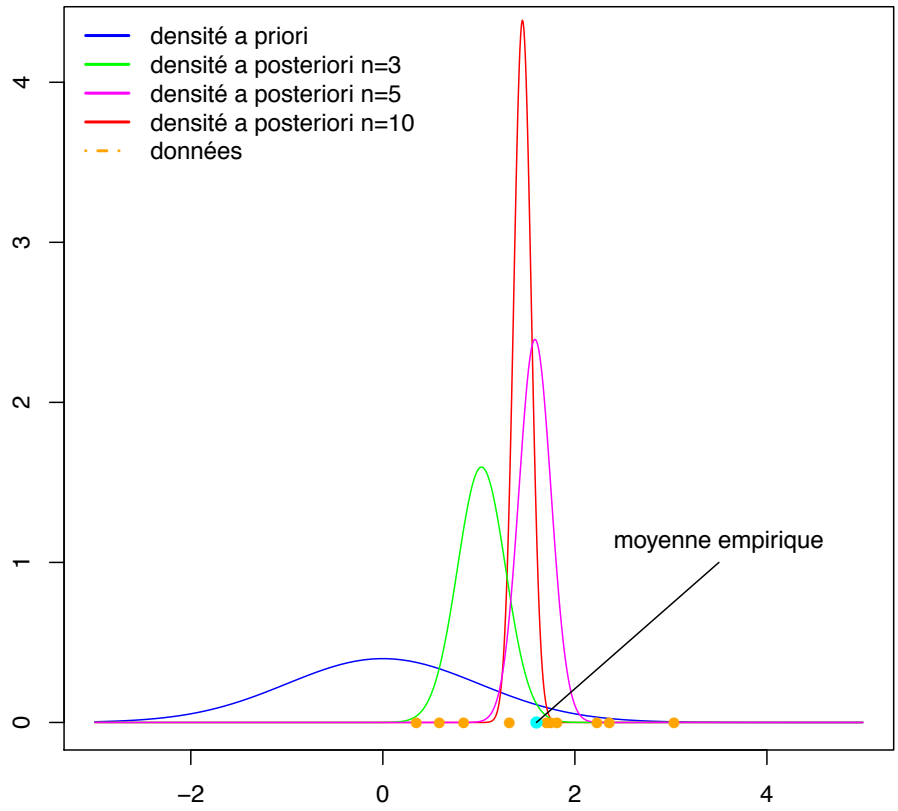
La figure 2.1 trace la loi a priori, les données, et les loi a posteriori correspondantes à $n = 3, 5, 10$ observations. On constate que la loi a posteriori se concentre 'près de $\overline{X}$' et que l''incertitude' – que l'on peut décrire comme l'écart-type de la loi a posteriori – décroit, comme $1/\sqrt{n}$ quand n augmente.

Exercice. Dans le modèle fondamental avec n observations, si la loi a priori Π sur θ est une $\mathcal{N}(a, v)$, montrer que

$$\Pi[\cdot \,|\, X_1, \dots, X_n] = \mathcal{N}\left(\frac{av^{-1} + n\overline{X}}{v^{-1} + n}, \frac{1}{v^{-1} + n}\right).$$

Vérifier que la moyenne de la loi a posteriori est une moyenne pondérée de la moyenne de la loi a priori et de la moyenne des données.

FIGURE 2.1 – Densités a priori et a posteriori



On note que dans le modèle fondamental, en choisissant un a priori gaussien, la loi a posteriori est elle-même gaussienne.

2.3 Aspects de la loi a posteriori

Dans l'exemple du modèle fondamental ci-dessus, nous constatons que la moyenne de la loi a posteriori (sachant X) vaut

$$\int \theta d\Pi(\theta \mid X) = E\left[\mathcal{N}\left(\frac{n\overline{X}}{n+1}, \frac{1}{n+1}\right) \mid X\right] = \frac{n\overline{X}}{n+1}.$$

Typiquement, plusieurs aspects de la loi a posteriori pourront nous intéresser.

Definition 3. Soit une expérience statistique $X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, soit Π une loi a priori sur θ, et $\Pi[\cdot \mid X]$ l'a posteriori correspondant. On définit, si ces quantités existent,

- la moyenne a posteriori

$$\bar{\theta} = \bar{\theta}(X) = \int \theta d\Pi(\theta \mid X).$$

- le mode a posteriori : c'est un point $\hat{\theta}_m(X)$ où le maximum de la densité a posteriori $\theta \to f_{\theta\,|\,X}(\theta)$ est atteint. On le note

$$\hat{\theta}_m(X) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ f_{\theta\,|\,X}(\theta).$$

- la variance a posteriori est la variance de la loi a posteriori. Si $\Theta \subset \mathbb{R}$,

$$v(X) = \int (\theta - \bar{\theta}(X))^2 d\Pi(\theta\,|\,X).$$

Si $\Theta \subset \mathbb{R}^d,\ d \geq 2$, il s'agit de la matrice de variance-covariance a posteriori

$$v(X) = \int (\theta - \bar{\theta}(X))(\theta - \bar{\theta}(X))^T d\Pi(\theta\,|\,X).$$

On note que ces quantités peuvent parfois ne pas être définies, par exemple si la loi a posteriori n'a pas d'espérance ou de moment d'ordre 2, ou si elle n'a pas de mode.

Definition 4. Dans le cadre précédent, si $\Theta \subset \mathbb{R}$, soit $F_{\theta\,|\,X}(\cdot)$ la fonction de répartition de la loi a posteriori $\Pi[\cdot\,|\,X]$. Supposons que $F_{\theta\,|\,X}$ admette une application réciproque $F^{-1}_{\theta\,|\,X}$. On définit alors les quantiles a posteriori comme, pout tout $t \in [0,1]$,

$$q_X(t) = F^{-1}_{\theta\,|\,X}(t).$$

Le quantile $q_X(1/2)$ s'appelle médiane a posteriori $\hat{\theta}^{med}(X)$.

Si la fonction $F_{\theta\,|\,X}$ est continue strictement croissante, ce qui est le cas en particulier si la loi a posteriori a une densité strictement positive par rapport à la mesure de Lebesgue, alors $F^{-1}_{\theta\,|\,X}$ est bien définie. Plus généralement, on peut toujours poser

$$q_X(t) = F^{-}_{\theta\,|\,X}(t),$$

où, pour une fonction de répartition G quelconque, G^- est la fonction quantile généralisés

$$G^-(u) = \inf\{y \in [0,1],\ G(y) \geq u\}.$$

Dans l'exemple du modèle fondamental avec a priori $\mathcal{N}(0,1)$ sur θ, on a

$$\bar{\theta}(X) = \hat{\theta}^m(X) = \hat{\theta}^{med}(X) = \frac{n\overline{X}}{n+1} \quad \text{et} \quad v(X) = \frac{1}{n+1}.$$

Notons que les statistiques $\bar{\theta}(X) = \hat{\theta}^m(X) = \hat{\theta}^{med}(X)$ sont des estimateurs ponctuels au sens usuel du terme. Dans l'exemple du modèle fondamental, ils sont même très proches de $\overline{X}$. Nous dirons plus sur ce sujet aux chapitres 3 et 5.

2.4 Conjugaison

Definition 5. Une famille $\mathcal{F}$ de lois a priori est dite conjuguée par rapport au modèle $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ si, pour toute loi a priori $\Pi \in \mathcal{F}$, si Π est prise comme loi a priori dans le cadre bayésien de ce modèle, la loi a posteriori $\Pi[\cdot \,|\, X]$ associée appartient aussi à $\mathcal{F}$.

Parfois, plutôt que de parler de conjugaison par rapport à un modèle, on parle aussi de conjugaison par rapport à un type de vraisemblance donnée. En effet, on déduit du modèle une vraisemblance, et le modèle n'intervient dans la loi a posteriori qu'au travers de la vraisemblance.

Exemples de familles de lois a priori conjuguées

→ la famille des lois gaussiennes $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$ est conjuguée par rapport au modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$.

→ la famille des lois Beta est conjuguée pour des vraisemblances binomiales, voir TDs.

→ la famille des lois de Dirichlet est conjuguée pour des vraisemblances multinomiales, voir TDs.

D'autres exemples seront vus au Chapitre 4.

Intérêts des familles de lois conjuguées

1. on dispose d'une expression explicite de la loi a posteriori comme élément de la classe de départ $\mathcal{F}$. Par exemple, dans le modèle fondamental, si $\Pi = \mathcal{N}(a, v)$, on a vu plus haut que $\Pi[\cdot \,|\, X] = \mathcal{N}\left(\frac{av^{-1}+n\overline{X}}{v^{-1}+n}, \frac{1}{v^{-1}+n}\right)$. Les paramètres de la loi a priori sont simplement 'mis à jour' à l'aide des données.
2. si l'on sait générer facilement des variables aléatoires simulées suivant un élément quelconque de la famille de lois $\mathcal{F}$ conjuguée, alors il est donc aisé de simuler des variables qui suivent la loi a posteriori, car $\Pi[\cdot \,|\, X]$ appartient à $\mathcal{F}$.
3. c'est un critère de choix possible de lois a priori, cf. Chapitre 4.

✍ Dans l'exemple du modèle fondamental, la famille de lois $\{\mathcal{N}(\mu, 1),\ \mu \in \mathbb{R}\}$ *n'est pas* conjuguée. En effet, la variance des lois a posteriori $\Pi[\cdot \,|\, X]$ n'est en général pas égale à 1.

2.5 Lois a priori impropres

Definition 6. Un a priori Π est dit impropre si Π est une mesure *positive* sur Θ, de masse infinie, soit

$$\Pi(\Theta) = +\infty.$$

Un a priori impropre Π n'est pas une mesure de probabilité sur Θ, puisque la masse totale ne vaut pas 1, c'est donc par abus de langage qu'on parle de *loi* a priori impropre.

Definition 7. Dans le cadre d'une expérience $X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, et $dP_\theta = p_\theta d\mu$, si l'on met un a priori impropre Π sur θ, la loi a posteriori correspondante $\Pi[\cdot \mid X]$ est la loi sur Θ de densité par rapport à Π égale à

$$\theta \to \frac{p_\theta(X)}{\int p_\theta(X) d\Pi(\theta)},$$

à condition que $\int p_\theta(X) d\Pi(\theta)$ soit finie μ-presque partout.

Exemple. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, prenons comme loi a priori la mesure de Lebesgue $\Pi = \mathrm{Leb}_{\mathbb{R}}$ sur $\mathbb{R}$.

Notons que $\int p_\theta(x) d\Pi(\theta) = \int \frac{e^{-\frac{-(x-\theta)^2}{2}}}{\sqrt{2\pi}} d\theta = 1 < \infty$. Pour n observations, la densité a posteriori est

$$\theta \to \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}}}{\int \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} d\theta} \ \propto\ e^{-\frac{n}{2}(\theta - \overline{X})^2}.$$

On en déduit $\Pi[\cdot \mid X] = \mathcal{N}\left(\overline{X}, \frac{1}{n}\right)$.

Un intérêt des a priori impropres est qu'ils permettent parfois d'avoir des calculs plus simples. Ils interviennent aussi souvent comme a priori 'non-informatifs', une notion que nous verrons au Chapitre 4.

2.6 Régions de crédibilité

Faisons un premier bilan rapide de ce que nous avons obtenu jusqu'ici. Partant d'une expérience statistique $X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ et d'une loi a priori Π sur Θ, nous avons construit une mesure de probabilité, la loi a posteriori $\Pi[\cdot \mid X_1, \ldots, X_n]$, qui dépend des données.

Par rapport à l'approche fréquentiste où l'on considère typiquement un estimateur $\hat{\theta}(X_1, \ldots, X_n)$ à valeurs dans Θ, on obtient ici un objet, $\Pi[\cdot \mid X_1, \ldots, X_n]$, à valeurs dans $\mathcal{M}_1(\Theta)$, l'ensemble des mesures de probabilité sur Θ.

Nous avons vu à la définition 3 que l'on pouvait à partir de la loi a posteriori construire des estimateurs ponctuels comme la moyenne ou le mode a posteriori. Mais peut-être pourrait-on également tirer profit du fait que la loi a posteriori donne non seulement une information sur une 'localisation', via par exemple la moyenne a posteriori, mais aussi une information sur la 'dispersion', par exemple via la variance a posteriori et les quantiles a posteriori. Ainsi, une loi a posteriori dont la variance est très petite sera très concentrée autour de sa moyenne et on peut penser qu'elle donnera plus d'informations sur le paramètre θ qu'une loi a posteriori

à variance plus grande.

Question. Ne pourrait-on pas utiliser $\Pi[\cdot \,|\, X_1, \dots, X_n]$ pour obtenir des intervalles ou des régions de 'confiance' ? Cette question motive la définition suivante.

Definition 8. Une région de crédibilité $A \subset \Theta$ de niveau (au moins) $1-\alpha$ pour $\Pi[\cdot \,|\, X_1, \dots, X_n]$ est un ensemble mesurable $A = A(X_1, \dots, X_n)$ tel que

$$\Pi[A \,|\, X_1, \dots, X_n] \geq 1 - \alpha.$$

✍ Si l'on ne fait pas d'hypothèse spécifique, il n'y a aucune raison pour qu'une région de crédibilité soit une région de confiance. Cela n'a en principe même pas de sens de parler de région de confiance dans un cadre bayésien où il n'y a pas de 'vrai' θ_0 comme dans le cadre fréquentiste. Nous verrons cependant au Chapitre 5 qu'il est possible de faire une analyse fréquentiste des lois a posteriori, et que sous certaines conditions une région de crédibilité peut être une région de confiance, éventuellement asymptotiquement.

Il y a en général de nombreux choix possibles pour construire une région de crédibilité. Par exemple, Θ est toujours une région de crédibilité 1. Bien sûr, en pratique on cherchera à construire une région 'la plus petite possible' ou proche de celle-ci. Ci-dessous nous voyons en détails deux constructions classiques.

1 Construction via des quantiles a posteriori

On suppose ici pour simplifier que

- $\Theta \subset \mathbb{R}$. Il s'agit donc de construire un intervalle de crédibilité.
- la fonction de répartition a posteriori, pour $X = (X_1, \dots, X_n)$,

$$t \to F_{\theta \,|\, X}(t) = \Pi[(-\infty, t] \,|\, X]$$

  est continue strictement croissante sur $\mathbb{R}$, et admet donc une réciproque $F^{-1}_{\theta \,|\, X} = q_X$.

Dans ce cadre, on pose alors

$$a_n(X) = q_X\left(\frac{\alpha}{2}\right),$$
$$b_n(X) = q_X\left(1 - \frac{\alpha}{2}\right).$$

Par construction $\Pi[(-\infty, a_n(X)] \,|\, X] = \alpha/2$. Puisque la fonction $t \to F_{\theta \,|\, X}(t)$ est continue et donc sa réciproque également, on a aussi $\Pi[(-\infty, a_n(X)) \,|\, X] = \alpha/2$. Par ailleurs, par construction $\Pi[(b_n(X), +\infty] \,|\, X] = \alpha/2$. On en déduit

$$\Pi[\,[a_n(X), b_n(X)]\,|\, X] = 1 - \alpha.$$

Sous les hypothèses précédentes, nous avons donc construit un intervalle de crédibilité (exactement) $1-\alpha$. Ce choix est 'bilatère', dans le sens où on prend des quantiles à gauche et à

droite. On pourrait aussi - mais ce choix est moins courant - prendre un quantile unilatère et poser $J(X) = (-\infty, \beta(X)]$ avec $\beta(X) = q_X(1-\alpha)$.

2 Régions 'Highest Probability Density' "HPD"

Soit Q une loi de probabilité sur Θ de densité g par rapport à une mesure ν. On commence par définir un 'ensemble de niveau' pour Q. Pour tout $y \geq 0$, on définit

$$\mathcal{L}(y) = \{\theta \in \Theta, \quad g(\theta) \geq y\}.$$

La région $\mathcal{L}(y)$ consiste en l'ensemble des paramètres pour lesquels la densité g en ce paramètre dépasse le niveau y.

Definition 9. Une région HD ('highest density') au niveau $1-\alpha$ pour une loi Q de densité g est $\mathcal{H} \subset \Theta$ de la forme

$$\mathcal{H} = \mathcal{L}(y_\alpha),$$

avec $\mathcal{L}(y)$ défini ci-dessus et y_α définit par

$$y_\alpha = \sup \left\{ y \in \mathbb{R}^+, \quad Q[\mathcal{L}(y)] \geq 1-\alpha \right\}.$$

Remarquons que cette définition implique que

$$Q[\mathcal{L}(y_\alpha)] \geq 1-\alpha.$$

En effet : [l'argument ci-dessous relève plus de la théorie de la mesure et peut être admis]

> Si y_n est une suite telle que $y_n \uparrow y_\alpha$ alors $\mathcal{L}(y_n) \downarrow \Lambda := \cap_{n\geq 1}\mathcal{L}(y_n)$ puisque les $\mathcal{L}(y)$ sont emboîtés. Ceci implique que $Q[\mathcal{L}(y_n)] \downarrow Q[\Lambda]$ (propriété classique des mesures). Par ailleurs par définition de y_α, on a $\Lambda = \mathcal{L}(y_\alpha)$ ce qui s'obtient en vérifiant la double inclusion. En passant à la limite dans $Q[\mathcal{L}(y_n)] \geq 1-\alpha$, on en conclut $Q[\mathcal{L}(y_\alpha)] \geq 1-\alpha$.

Une région HD est donc par construction le plus petit parmi les ensembles de niveau $\mathcal{L}(y)$ qui ont une probabilité au moins $1-\alpha$ sous Q. La figure 2.2 illustre la définition précédente.

Definition 10. Dans une expérience statistique $X, \mathcal{P}$ avec une loi a priori Π sur $\Theta \subset \mathbb{R}^d$, soit $\Pi[\cdot \,|\, X]$ la loi a posteriori. Une région HPD ('highest posterior density') au niveau $1-\alpha$ est une région HD au niveau $1-\alpha$ pour la loi $\Pi[\cdot \,|\, X]$.

Dans l'énoncé ci-dessous, le *volume* d'un ensemble mesurable A est un synonyme pour $\nu(A) = \int_A d\nu(\theta)$. Si ν est la mesure de Lebesgue (ce qui pour nous est le cas la plupart du temps), alors $\nu(A)$ est le volume usuel dans $\mathbb{R}^d$.

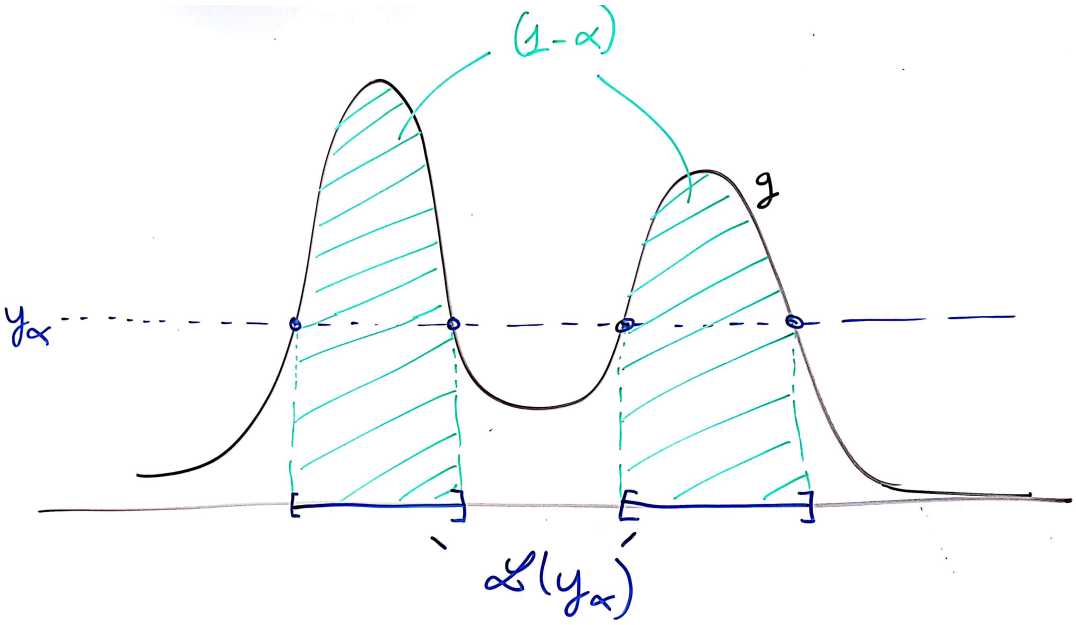


FIGURE 2.2 – La réunion des deux intervalles en bleu sur l'axe des abscisses est la région HPD au niveau $1-\alpha$ pour la densité g dessinée. La région hachurée en vert a une aire égale à $(1-\alpha)\,\%$.

Théorème 3. Dans le cadre de la définition 10, une région HPD au niveau $1-\alpha$ est de volume minimal parmi les régions de même niveau de crédibilité pour $\Pi[\cdot \,|\, X]$.

Preuve.

Soit D une région HPD de niveau $1-\alpha$ et notons pour simplifier $g(\theta)$ la densité a posteriori $f_{\theta\,|\,X}(\theta)$. Par définition, D est de la forme

$$D = \{\theta \in \Theta, \quad g(\theta) \geq y_\alpha\} = \mathcal{L}(y_\alpha).$$

Il suffit de montrer que si une région $C \subset \Theta$ a un crédibilité au moins aussi grande que D, soit $\Pi[C \,|\, X] \geq \Pi[D \,|\, X]$, alors $\mathrm{Vol}(C) \geq \mathrm{Vol}(D)$, où Vol désigne le volume dans $\mathbb{R}^d$. Il est équivalent de démontrer la contraposée, soit que $\mathrm{Vol}(C) < \mathrm{Vol}(D)$ implique $\Pi[C \,|\, X] < \Pi[D \,|\, X]$. Notons que

$$\begin{aligned}\mathrm{Vol}(C) &= \mathrm{Vol}(C \cap D) + \mathrm{Vol}(C \cap D^c)\\ \mathrm{Vol}(D) &= \mathrm{Vol}(C \cap D) + \mathrm{Vol}(D \cap C^c).\end{aligned}$$

Si $\mathrm{Vol}(C) < \mathrm{Vol}(D)$, on a donc $\mathrm{Vol}(C \cap D^c) < \mathrm{Vol}(D \cap C^c)$. Or

$$\begin{aligned}
\Pi[D \cap C^c \,|\, X] &= \int_{D \cap C^c} g(\theta) d\nu(\theta) \\
&\geq y_\alpha \mathrm{Vol}(D \cap C^c) && \text{par définition de } D \\
&> y_\alpha \mathrm{Vol}(C \cap D^c) && \text{via l'inégalité ci-dessus} \\
&\geq \int_{C \cap D^c} g(\theta) d\nu(\theta) && \text{par définition de } D^c \\
&\geq \Pi[C \cap D^c \,|\, X]
\end{aligned}$$

[remarque : la quatrième inégalité peut être une égalité si jamais $\mathrm{Vol}(C \cap D^c) = 0$]
En ajoutant de part et d'autre de cette inégalité la quantité $\Pi[C \cap D \,|\, X]$, on obtient $\Pi[C \,|\, X] < \Pi[D \,|\, X]$, ce qu'il fallait démontrer.

✍ En général, les deux constructions $\boxed{1}$ et $\boxed{2}$ donnent des régions différentes. Un exemple est donné par la figure 2.2, où la région HPD est une union de deux intervalles disjoints, donc est nécessairement différente d'une région obtenue par quantiles comme pour $\boxed{1}$, où l'on obtient un seul intervalle. En revanche, les constructions coïncident si la densité a posteriori est continue, unimodale et symétrique sur $\mathbb{R}$, voir TDs.

✍ Du point de vue pratique, la méthode $\boxed{1}$ est souvent plus facile a mettre en oeuvre, car elle nécessite seulement de connaître deux des quantiles a posteriori, tandis que $\boxed{2}$ nécessite de travailler avec les ensembles de niveau de la densité a posteriori.

Dans ce chapitre, nous avons vu la construction de l'objet central bayésien, la loi a posteriori. Cette loi et/ou certains de ses apects comme la moyenne a posteriori constituent des estimateurs de θ (en un sens généralisé pour $\Pi[\cdot \,|\, X]$ puisque $\Pi[\cdot \,|\, X]$ est une probabilité et non un point de Θ). Certains de ces estimateurs sont-ils optimaux en un certain sens ? Nous examinons plusieurs notions d'optimalité au Chapitre 3.