M1 – UPMC
2016 - 2017

Introduction aux statistiques bayésiennes

4M072

Enseignant
Ismaël Castillo

Table des matières

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

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

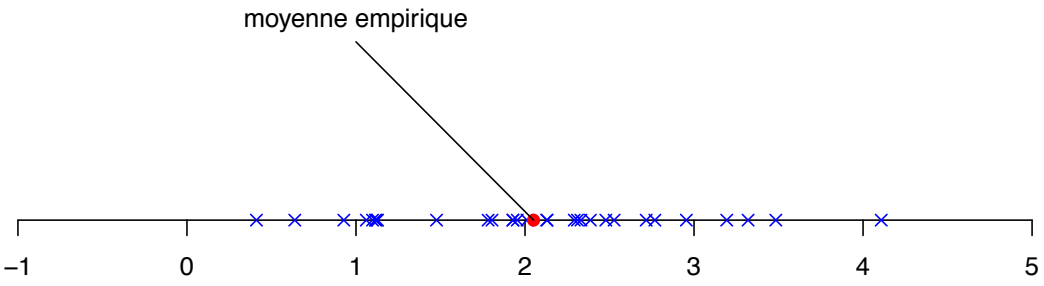
Dans ce cadre, θ_0 s'appelle vraie valeur du paramètre. Typiquement, θ_0 est inconnu et on cherche à l'"estimer" (à l'approcher), à l'aide des données X.

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

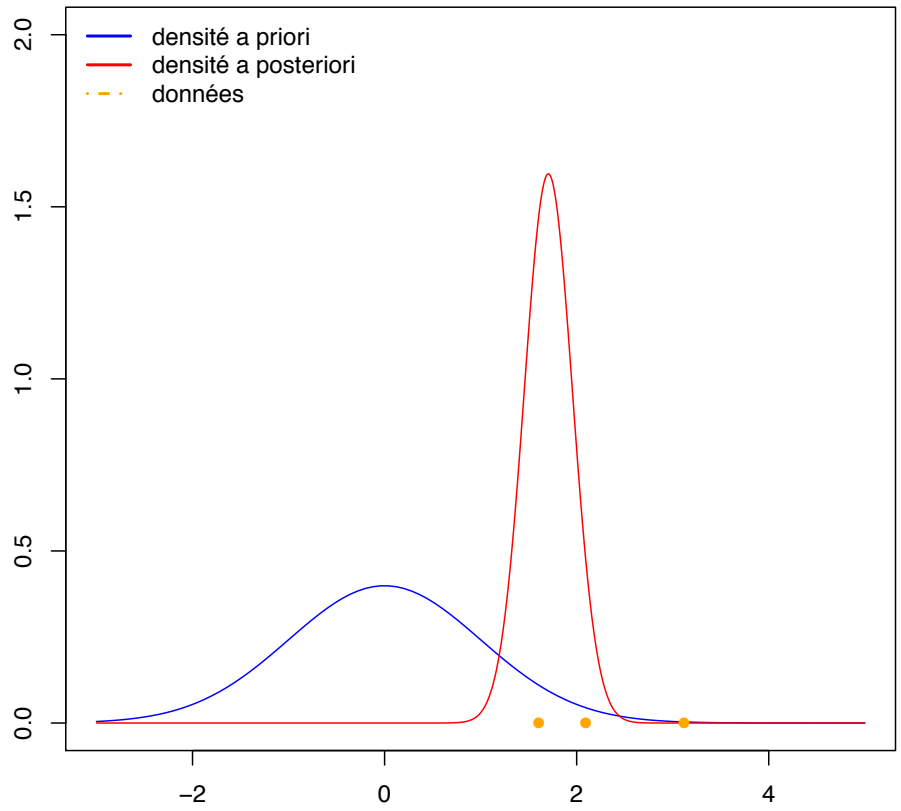
$$P_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2),$$

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 $\theta \in \Theta$,

$$\hat{\theta}(X) \xrightarrow{P_\theta} \theta.$$

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 $X_n \xrightarrow{\mathcal{L}} X$ 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 $X_n \xrightarrow{\mathcal{L}} X$ 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 $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, 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

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

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

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

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 $X_1, \ldots, X_n$ 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 $R(\theta,\hat\theta) \le 1/(4n)$.

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
$$P_\theta[\,\theta \in I(X)\,] \ge 1-\alpha \qquad \forall\,\theta \in \Theta$$
- Cas général. Une région de confiance de niveau (au moins) $1-\alpha$ est $\mathcal{R}(X) \subset \Theta$ avec
$$P_\theta[\,\theta \in \mathcal{R}(X)\,] \ge 1-\alpha \qquad \forall\,\theta \in \Theta$$

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 $t > 0$,
$$P_\theta[|\hat\theta - \theta| > t] \le t^{-2} R(\hat\theta,\theta).$$

→ *Exemple : expérience binomiale.* On observe $X \sim \mathrm{Bin}(n,\theta)$. On pose $\hat\theta = X/n$. D'après l'exercice ci-dessus, $R(\hat\theta,\theta) \le 1/(4n)$, donc pour tout $t > 0$
$$P_\theta[|\hat\theta - \theta| > t] \le \frac{1}{4nt^2}$$
soit aussi, en prenant l'événement complémentaire,
$$P_\theta[\theta \in [\hat\theta - t, \hat\theta + t]] \ge 1 - \frac{1}{4nt^2}.$$
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 $X_1, \ldots, X_n$ 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 $(X_1, \ldots, X_n)$ par rapport à $\mu^{\otimes n}$ est donc $p_\theta(x_1)\ldots p_\theta(x_n)$. 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 $Y \mid X = x$, parfois aussi notée $\mathcal{L}(Y \mid X = x)$ est définie par, pour tout $e \in E$, et $x \in E$ tel que $P[X = x] > 0$,

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

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 $\mathcal{L}(Y \mid X = 1)$?

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,

$$\begin{aligned} P[Y = 1 \mid X = 1] &= \frac{P[X = 1, Y = 1]}{P[X = 1]} = \frac{P[Z = 0, Y = 1]}{P[Y = 1, Z = 0] + P[Y = 0, Z = 1]} \\ &= \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}. \end{aligned}$$

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

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

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

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

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

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

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

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

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 $f_X(x) > 0$,

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

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

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

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

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

On reconnait la densité d'une loi exponentielle de paramètre x. Ainsi, $\mathcal{L}(Y \,|\, X = x) = \mathcal{E}(x)$. On écrit aussi $\mathcal{L}(Y \,|\, X) = \mathcal{E}(X)$. 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 $X \,|\, Y$ est une loi Gamma $\Gamma(2, Y + 1)$.

Notion d'espérance conditionnelle

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

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,

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

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

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

Preuve.

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

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 $g(u) = E[Y \mid X = u]$.

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 $g(X) = E[Y \mid X]$,

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

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

$$\begin{aligned} Eg(X)^2 = E\left[\int y f(y \mid X) d\beta(y)\right]^2 &= \int \left[\int y f(y \mid x) d\beta(y)\right]^2 f_X(x) d\alpha(x) \\ \leq \int \int y^2 f(y \mid x) d\beta(y) f_X(x) d\alpha(x) &= \int \int y^2 f(x, y) d\beta(y) d\alpha(x), \end{aligned}$$

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

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

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 $X = (X_1, \ldots, X_n)$ et $P_\theta = P_\theta^{(n)} = \otimes_{i=1}^n P_\theta$. Le formalisme bayésien s'écrit

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

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

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

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.

$\to$ 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* $X_1, \ldots, X_n$. Le cadre bayésien s'écrit

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

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

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

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

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

On en conclut

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

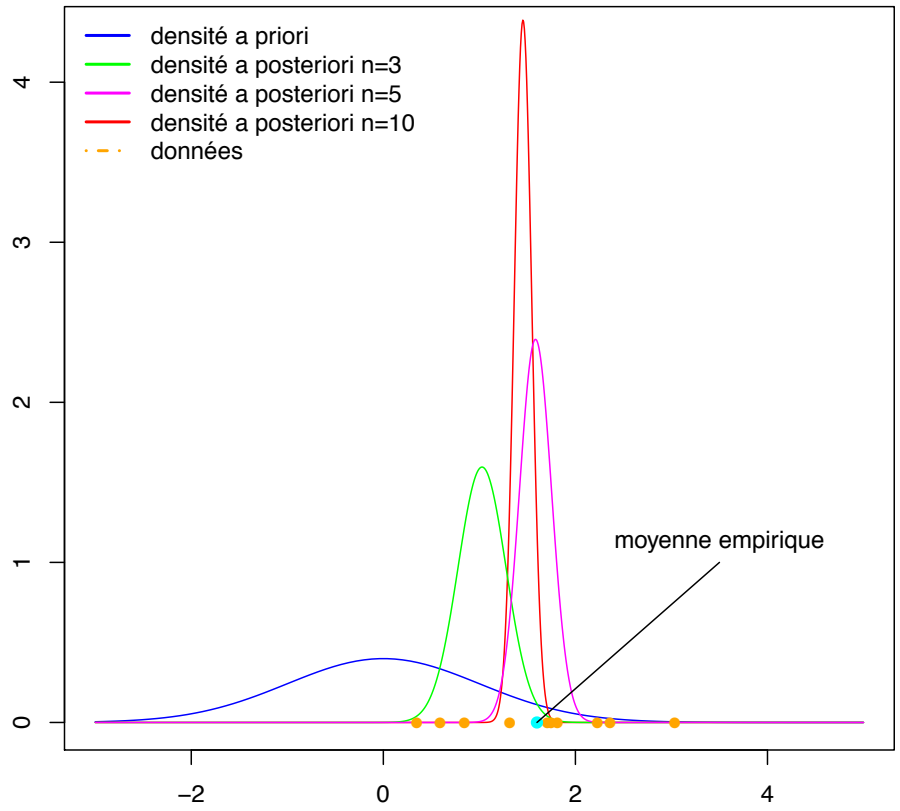
La figure 2.1 trace la loi a priori, les données, et les loi a posteriori correspondantes à $n = 3, 5, 10$ observations. On constate que la loi a posteriori se concentre 'près de $\overline{X}$' et que l''incertitude' – que l'on peut décrire comme l'écart-type de la loi a posteriori – décroit, comme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

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

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 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 $\mathbb{R}^d$.

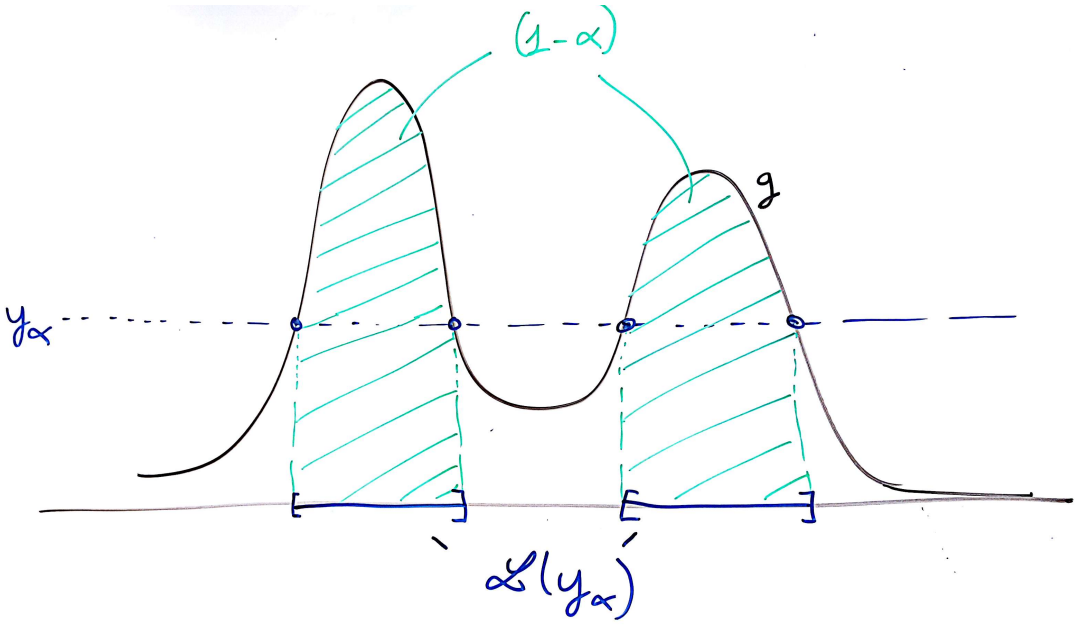


FIGURE 2.2 – La réunion des deux intervalles en bleu sur l'axe des abscisses est la région HPD au niveau $1-\alpha$ pour la densité g dessinée. La région hachurée en vert a une aire égale à $(1-\alpha)\,\%$.

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.

CHAPITRE 3

Bayésien et théorie de la décision

Dans ce chapitre, nous examinons des critères de choix d'estimateurs. Ceci exige au préalable de définir une notion de risque et de fonction de perte. Nous étudions trois critères classiques : l'admissibilité, le risque de Bayes et le risque minimax, ainsi que certaines relations existantes entre ces critères. Enfin, nous introduisons quelques outils pour minorer le risque minimax.

Dans une expérience statistique, à une loi a priori donnée correspond une loi a posteriori et de celle-ci on peut déduire plusieurs estimateurs tels que la moyenne, la médiane, le mode etc. Lequel choisir en pratique ? Quels critères de choix énoncer ? Plus généralement, y-a-t-il des estimateurs 'optimaux' parmi tous les estimateurs, pas nécessairement d'ailleurs des aspects d'une loi a posteriori ?

3.1 Risques, admissibilité

On se place dans le cadre d'une expérience $X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, le modèle étant dominé, $dP_\theta = p_\theta d\mu$ par une mesure μ sur un espace E. Dans ce cadre, un estimateur (ponctuel) T est une application mesurable $T : E \to \Theta$.

Definition 1. Une fonction de perte ℓ est une fonction $\ell : \Theta \times \Theta \to \mathbb{R}^+$ mesurable avec

$$\ell(\theta, \theta') = 0 \ \Leftrightarrow\ \theta = \theta', \qquad \forall\, \theta, \theta' \in \Theta.$$

Exemples.

1. Si $\Theta \subset \mathbb{R}$, la fonction $\ell(\theta, \theta') = (\theta - \theta')^2$ s'appelle perte quadratique.
2. Si $\Theta \subset \mathbb{R}$, la fonction $\ell(\theta, \theta') = |\theta - \theta'|$ s'appelle perte en valeur absolue.

3. Pour Θ quelconque, on définit la perte de Hellinger par $\ell(\theta, \theta') = h(P_\theta, P_{\theta'})$, où $h(P, Q)$ est la distance de Hellinger entre les lois P et Q définie par, pour $dP = pd\mu$ et $dQ = qd\mu$,

$$h(P, Q)^2 = \int (\sqrt{p} - \sqrt{q})^2 d\mu.$$

On note qu'il s'agit bien d'une fonction de perte au sens de la définition 1 dès que le modèle statistique considéré est identifiable, tandis que l'hypothèse (1) garantit la mesurabilité.

Definition 2. La fonction risque d'un estimateur $T(X)$ pour la fonction de perte ℓ est l'application

$$\begin{array}{rccl} R & : & \Theta & \to & \mathbb{R}^+ \\ & & \theta & \to & R(\theta, T) = E_\theta\left[\,\ell(\theta, T(X))\,\right] = \int \ell(\theta, T(x)) dP_\theta(x). \end{array}$$

Le risque au point θ de l'estimateur $T(X)$ est donc la perte moyenne de $T(X)$ en θ.

La fonction de perte, et le risque en résultant, peuvent être vus comme des 'coûts' associés aux estimateurs, et vont nous permettre de comparer ceux-ci entre eux. Cependant, définir une notion de *meilleur estimateur possible* est quelque chose de délicat, qui a mis longtemps à émerger historiquement.

On peut se convaincre de la difficulté intrinsèque du problème de définition de *meilleur estimateur possible* avec les deux exemples suivants.

→ l'estimateur constant

> Soit θ_0 fixé dans $\Theta \subset \mathbb{R}$. Posons $\hat{\theta} = \theta_0$, alors
>
> $$R(\hat{\theta}, \theta) = E_\theta(\theta - \theta_0)^2 = (\theta - \theta_0)^2.$$
>
> Le risque est nul si $\theta = \theta_0$, donc en ce point on ne peut faire mieux, mais le risque est strictement positif si $\theta \neq \theta_0$, donc beaucoup d' estimateurs feront mieux en ce points.

→ le phénomène de Hodges : voir TDs.

Dans la suite, on fixe une fonction de perte. Toutes les définitions et résultats qui suivent s'entendent donc à fonction de perte fixée. Les risques dépendent de cette fonction de perte. Par simplicité on ne rappelera pas tout le temps la fonction de perte.

Un premier critère de comparaison est l'admissibilité. On commence par définir l'inadmissibilité : un estimateur est inadmissible s'il existe un autre estimateur, strictement meilleur au sens suivant.

Definition 3. Un estimateur T est inadmissible s'il existe un estimateur T_1 tel que

$$\begin{aligned} &\forall\, \theta \in \Theta, && R(\theta, T_1) \leq R(\theta, T) && (i) \\ \text{et} \quad &\exists\, \theta_1 \in \Theta, && R(\theta_1, T_1) < R(\theta_1, T) && (ii) \end{aligned}$$

Un estimateur est admissible s'il n'est pas inadmissible.

L'admissibilité semble être une requête naturelle. Cependant, elle a ceci de peu satisfaisant que les estimateurs constants sont, souvent, admissibles.

Proposition 1. Soit $\theta_0 \in \Theta$. L'estimateur constant $T(X) = \theta_0$ est admissible si pour tout A mesurable, $P_{\theta_0}(A) = 0$ implique $P_\theta(A) = 0$ pour tout $\theta \in \Theta$.

Le modèle est dominé, donc $dP_\theta = p_\theta d\mu$. En particulier, si p_{θ_0} est strictement positive, alors $P_{\theta_0}(A) = 0$ implique $\mu(A) = 0$ qui à son tour implique $P_\theta(A) = 0$ pour tout $\theta \in \Theta$.

Preuve.

Raisonnons par l'absurde. Soit T l'estimateur constant égal à θ_0. Supposons que pour tout A mesurable, $P_{\theta_0}(A) = 0$ implique $P_\theta(A) = 0$ pour tout $\theta \in \Theta$, mais que T est inadmissible. Alors (i)-(ii) de la définition 3 sont vérifiés. En particulier, (i) pris en $\theta = \theta_0$ implique $R(\theta_0, T_1) = R(\theta_0, T) = 0$ soit

$$R(\theta_0, T_1) = \int \ell(\theta_0, T_1(x)) dP_{\theta_0}(x) = 0.$$

La fonction $x \to \ell(\theta_0, T_1(x))$ est mesurable positive, d'intégrale nulle contre P_{θ_0}. Elle est donc nulle P_{θ_0}-presque sûrement soit

$$P_{\theta_0}[\ell(\theta_0, T_1(X)) = 0] = 1.$$

En passant aux complémentaires, et en utilisant la propriété de l'énoncé, on en déduit que $P_\theta[\ell(\theta_0, T_1(X)) = 0] = 1$ pour tout θ, ce qui entraîne grâce à la propriété d'injectivité de la fonction de perte

$$P_\theta[T_1(X) = \theta_0] = 1.$$

D'après la partie (ii) de l'inadmissibilité, il existe $\theta_1 \in \Theta$ tel que $R(\theta_1, T_1) < R(\theta_1, T)$. Mais $R(\theta_1, T) = \int \ell(\theta_1, T_1(x)) dP_{\theta_0}(x) = \ell(\theta_1, \theta_0)$ car $T_1 = \theta_0$ P_θ-presque sûrement d'après ce qui précède. Mais $R(\theta_1, T) = \ell(\theta_1, \theta_0)$ également par définition de T. Contradiction.

Plusieurs estimateurs peuvent être admissibles simultanément. Par exemple, dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ avec n observations, nous verrons plus loin que $T_1 = \overline{X}$ est admissible. D'après la Proposition 1, l'estimateur constant $T_2(X) = \theta_0$, pour un $\theta_0 \in \mathbb{R}$ quelconque, est admissible, d'après la note sous la Proposition. Notons en particulier que,

pour $\theta_1 = \theta_0 + 2$, voir le Chapitre 1 pour le calcul du risque de $\overline{X}$,

$$R(\theta_0, T_2) = 0^2 = 0 < R(\theta_0, T_1) = \frac{1}{n}$$
$$R(\theta_1, T_2) = 2^2 = 4 > R(\theta_1, T_1) = \frac{1}{n}.$$

La notion d'admissibilité est un peu rigide et souvent peu pratique. D'autres notions sont plus utilisées dans la pratique, notamment celles de risque bayésien et de risque minimax.

Intuition. La notion de risque bayésien définit ci-dessous va nous donner une réponse complète possible à la question de trouver un estimateur de risque optimal. Cependant, cette notion dépendra de l'a priori choisit, ce qui n'en fait pas une réponse 'universelle'. Le risque minimax considéré ci-dessous est lui plus 'universel' mais correspond à une vision un peu 'pessimiste, du pire des cas'.

Soit Π une loi a priori *donnée* sur Θ. Rappelons que nous travaillons également à fonction de perte ℓ donnée. Ainsi les définitions ci-dessous dépendent implicitement de ℓ.

Definition 4. On appelle risque de Bayes ou parfois risque bayésien pour l'estimateur T et la loi a priori Π la quantité

$$R_B(\Pi, T) = \int_\Theta R(\theta, T) d\Pi(\theta).$$

Explicitement, en utilisant l'expression du risque on peut aussi écrire

$$R_B(\Pi, T) = \int \int \ell(\theta, T(x)) dP_\theta(x) d\Pi(\theta).$$

Interprétation. Le théorème de Fubini donne que $R_B(\Pi, T) = E[\ell(\theta, T(X))]$, où E désigne l'espérance par rapport à le loi du couple (X, θ) dans le cadre bayésien.

Definition 5. Un estimateur T_1 est dit de Bayes pour la loi a priori Π si

$$R_B(\Pi, T_1) = \inf_T R_B(\Pi, T),$$

où l'infimum porte sur tous les estimateurs T possibles. On note alors

$$R_B(\Pi) = \inf_T R_B(\Pi, T)$$

qui s'appelle risque de Bayes pour la loi a priori Π.

Un estimateur de Bayes pour Π a donc un risque qui minimise le risque bayésien pour Π, qui est une moyenne des risques ponctuels en θ suivant la loi a priori Π sur θ. Un tel estimateur

minimise donc un risque 'en moyenne selon Π'.

Exemple. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta,1)^{\otimes n},\ \theta \in \mathbb{R}\}$, considérons la loi a priori $\Pi \sim \mathcal{N}(0,1)$ et la fonction de perte quadratique $\ell(\theta,\theta') = (\theta-\theta')^2$. Calculons le risque de Bayes pour Π des estimateurs suivants

$$T_1(X) = 0, \quad T_2(X) = \overline{X}, \quad T_3(X) = \frac{n}{n+1}\overline{X}.$$

Auparavant notons que $\overline{X}$, sous la loi P_θ, peut s'écrire $\theta + \overline{\varepsilon}$, avec $\overline{\varepsilon}$ de loi $\mathcal{N}(0,1/n)$.

Avec la définition du risque bayésien et des estimateurs,

$$\begin{aligned} R_B(\Pi,T_1) &= \int R(\theta,T_1)d\Pi(\theta) \\ &= \int\int (\theta - T_1)^2 dP_\theta(x)d\Pi(\theta) \\ &= \int\int \theta^2 dP_\theta(x)d\Pi(\theta) = \int \theta^2 d\Pi(\theta) = 1 \end{aligned}$$

Pour T_2, en utilisant le fait vu précédemment que $R(\theta,T_2) = 1/n$,

$$R_B(\Pi,T_2) = \int \frac{1}{n} d\Pi(\theta) = \frac{1}{n}.$$

Pour T_3, nous calculons d'abord

$$\begin{aligned} E_\theta((T_3-\theta)^2) &= E_\theta\left[\left(\frac{n}{n+1}\overline{X} - \theta\right)^2\right] = E_\theta\left[\left(-\frac{1}{n+1}\theta + \frac{n}{n+1}\overline{\varepsilon}\right)^2\right] \\ &= \frac{\theta^2}{n+1} + \frac{n^2}{(n+1)^2} E\left[\mathcal{N}\left(0,\frac{1}{n}\right)^2\right] = \frac{\theta^2}{n+1} + \frac{n}{(n+1)^2}. \end{aligned}$$

Nous en déduisons

$$\begin{aligned} R_B(\Pi,T_3) &= \frac{1}{n+1}\int \theta^2 d\Pi(\theta) + \frac{n}{(n+1)^2} \\ &= \frac{1+n}{(n+1)^2} = \frac{1}{n+1}. \end{aligned}$$

On constate que $R_B(\Pi,T_3) < R_B(\Pi,T_2) < R_B(\Pi,T_1)$. Nous verrons dans la suite qu'en fait T_3 est un estimateur de Bayes pour Π et la fonction de perte quadratique.

Definition 6. Le risque maximal d'un estimateur T est

$$R_{max}(T) = \sup_{\theta\in\Theta} R(\theta,T).$$

De même que pour le risque de Bayes, il est alors naturel de chercher un estimateur qui est le meilleur du point de vue du risque maximal, ce qui amène à la définition suivante.

Definition 7. Le risque minimax R_M est défini comme

$$R_M = \inf_T R_{max}(T) = \inf_T \sup_{\theta\in\Theta} R(\theta, T),$$

où l'infimum porte sur tous les estimateurs possibles T. Un estimateur T_1 est minimax si

$$R_{max}(T_1) = R_M.$$

Puisque $R_{max}(T)$ peut être vu comme le 'pire risque' pour T sur l'ensemble des points $\theta \in \Theta$, un estimateur minimax s'interprète comme un estimateur 'le meilleur dans le pire des cas'. Le critère minimax est en un sens plus pessimiste que le critère du risque de Bayes ci-dessus, mais il a l'avantage d'être plus 'universel' en ce qu'il ne dépend pas de la loi a priori Π.

Exemple. Reprenons l'exemple précédent du modèle fondamental avec les estimateurs T_1, T_2 et T_3 et calculons le risque maximal de chacun.

$$\begin{aligned}
R_{max}(T_1) &= \sup_{\theta\in\mathbb{R}} E_\theta[(0-\theta)^2] = \sup_{\theta\in\mathbb{R}} \theta^2 = +\infty \\
R_{max}(T_2) &= \sup_{\theta\in\mathbb{R}} E_\theta\left[(\overline{X}-\theta)^2\right] = \frac{1}{n} \\
R_{max}(T_3) &= \sup_{\theta\in\mathbb{R}} E_\theta\left[\left(\frac{n}{n+1}\overline{X}-\theta\right)^2\right] = \sup_{\theta\in\mathbb{R}} \left[\frac{\theta^2}{n+1} + \frac{n}{(n+1)^2}\right] = +\infty.
\end{aligned}$$

Le fait que $\Theta = \mathbb{R}$ soit ici non borné fait que le risque maximal puisse être infini, ce qui advient même pour un estimateur 'raisonnable' comme T_3. On peut en fait montrer que T_2 est un estimateur minimax dans ce cadre.

Les différents risques vus plus haut ainsi que l'admissibilité sont des notions qui peuvent être reliées entre elles sous certaines hypothèses, comme nous le verrons dans la suite.

3.2 Risque bayésien et estimateurs de Bayes

Nous allons maintenant voir qu'il est souvent possible de proposer une construction spécifique d'un estimateur de Bayes pour une fonction de perte ℓ et un a priori Π donnés. Rappelons qu'un tel estimateur minimise en T le risque bayésien $R_B(\Pi, T) = \int R(\theta, T) d\Pi(\theta)$.

Definition 8. Soient ℓ une fonction de perte, Π une loi a priori et $T(X)$ un estimateur.

Le risque a posteriori $\rho(\Pi, T \mid X)$ est défini par, pour tout $x \in E$,

$$\rho(\Pi, T \mid X = x) = \int \ell(\theta, T(x)) d\Pi(\theta \mid X = x).$$

On notera aussi $\rho(\Pi, T \mid X) = \int \ell(\theta, T(X)) d\Pi(\theta \mid X)$.

Au lieu de prendre la moyenne du risque par rapport à la loi a priori comme pour le risque bayésien de la définition 5, la moyenne par rapport à la loi a posteriori donne le risque a posteriori.

Théorème 1. Une fonction de perte ℓ et une loi a priori Π étant donnés, l'estimateur $\hat{\theta}^{\Pi}(X)$ suivant, s'il existe, est un estimateur de Bayes pour Π

$$\hat{\theta}^{\Pi}(X) = \operatorname*{argmin}_{T} \rho(\Pi, T \mid X).$$

On suppose que la quantité ci-dessus existe, c'est-à-dire que l'on peut minimiser $\rho(\Pi, T \mid X)$ en T et que la quantité correspondante est mesurable.

On peut légitimement se demander en quoi le résultat du Théorème 1 est une simplification par rapport à la définition de l'estimateur de Bayes, qui introduit aussi un minimum. D'après le Théorème 1,

$$\begin{aligned}\hat{\theta}^{\Pi}(X) &= \operatorname*{argmin}_{T} R_B(\Pi, T) = \operatorname*{argmin}_{T} \rho(\Pi, T \mid X) \\ &= \int \int \ell(\theta, T) dP_\theta d\Pi(\theta) = \int \ell(\theta, T) d\Pi(\theta \mid X)\end{aligned}$$

Ainsi, la minimisation qui résulte de l'application du Théorème est celle d'une expression avec une seule intégrale, et non deux, ce qui simplifie souvent la recherche d'une expression explicite, comme nous le verrons plus bas.

Preuve.

On peut supposer qu'il existe un estimateur T tel que $R_B(\Pi, T)$ soit fini. Si ce n'est pas le cas, alors tout estimateur a un risque de Bayes pour Π infini, donc un estimateur quelconque est de Bayes. Pour tout T tel que $R_B(\Pi, T)$ est fini, on peut écrire, en utilisant le théorème

de Fubini, et en notant $m(x) = \int p_\theta(x) d\Pi(\theta)$ la densité marginale de X,

$$\begin{aligned} R_B(\Pi, T) &= \int R(\theta, T) d\Pi(\theta) \\ &= \int \int \ell(\theta, T(x)) dP_\theta(x) d\Pi(\theta) \\ &= \int \int \ell(\theta, T(x)) p_\theta(x) d\Pi(\theta) d\mu(x) \\ &= \int \int \ell(\theta, T(x)) \frac{p_\theta(x) d\Pi(\theta)}{m(x)} m(x) d\mu(x) \\ &= \int \rho(\Pi, T \,|\, X = x) m(x) d\mu(x). \end{aligned}$$

Par définition, $\rho(\Pi, T \,|\, X = x) \geq \rho(\Pi, \hat{\theta}^\Pi \,|\, X = x)$. On en déduit que

$$R_B(\Pi, T) \geq \int \rho(\Pi, \hat{\theta}^\Pi \,|\, X = x) m(x) d\mu(x),$$

puis en reprenant les égalités ci-dessus mais dans l'autre sens avec $\hat{\theta}^\Pi$ à la place de T, on conclut que

$$R_B(\Pi, T) \geq R_B(\Pi, \hat{\theta}^\Pi),$$

ce qu'il fallait démontrer.

Nous examinons maintenant des applications du Théorème 1 dans le cas de plusieurs fonctions de perte classiques.

1 Bayes et fonction de perte quadratique

Considérons, pour $\Theta \subset \mathbb{R}$, la fonction de perte quadratique

$$\ell(\theta, \theta') = (\theta - \theta')^2, \qquad \theta, \theta' \in \mathbb{R}.$$

Proposition 2. Soit ℓ la perte quadratique et soit Π une loi a priori sur $\Theta \subset \mathbb{R}$. On suppose $\int \theta^2 d\Pi(\theta) < \infty$. Un estimateur de Bayes pour ℓ et la loi Π est donné par

$$\hat{\theta}^\Pi(X) = \overline{\theta}(X) = \int \theta d\Pi(\theta \,|\, X),$$

la moyenne a posteriori pour la loi a priori Π.

Plus généralement, si $\Theta \subset \mathbb{R}^d$, $d \geq 1$, la fonction de perte quadratique est définie via $\ell(\theta, \theta') = \|\theta - \theta'\|^2$ et le résultat est identique avec la même preuve : l'estimateur de Bayes est la moyenne a posteriori.

Preuve.

Méthode 1. D'après le Théorème 1, il suffit de chercher l'estimateur de Bayes sous la forme

$$\hat{\theta}^{\Pi}(X) = \operatorname*{argmin}_{T} \int (T(X) - \theta)^2 d\Pi(\theta \,|\, X).$$

Pour une variable aléatoire Z de carré intégrable, la fonction $\psi : b \to E[(Z-b)^2]$ est minimale pour $b = E[Z]$ car $\psi(b) = E[(Z - EZ)^2] + (E(Z) - b)^2 \geq \psi(E[Z])$. Il suffit d'appliquer cette remarque à Z de loi $\mathcal{L}(\theta \,|\, X)$ pour conclure, en notant que $E[\theta^2 \,|\, X] < \infty$ – car son espérance est $E[\theta^2] < \infty$ – et $E[Z]$ est alors $E[\theta \,|\, X]$.

Méthode 2. On utilise la définition du risque de Bayes et la Proposition 8 du Chapitre 1. On applique celle-ci dans le cadre bayésien où la loi jointe de (X, θ) est spécifiée par $\theta \sim \Pi$ et $X \,|\, \theta \sim P_\theta$. On note en effet que le risque bayésien pour la fonction de perte quadratique d'un estimateur $h(X)$ est par définition $E[(\theta - h(X))^2]$, d'après la remarque sous la définition 4. Or cette expression est minimale pour $h(X) = E[\theta \,|\, X] = \int \theta d\Pi(\theta \,|\, X)$, ce qu'il fallait démontrer.

Exemple. Dans le modèle fondamental avec $\Pi = \mathcal{N}(0, 1)$, nous avons vu au chapitre 2 que $E[\theta \,|\, X] = n\overline{X}/(n+1)$. On en déduit avec le Théorème 1 que cet estimateur est de Bayes pour Π, comme annoncé plus haut.

2 Bayes et fonction de perte en valeur absolue

Considérons, pour $\Theta \subset \mathbb{R}$, la fonction de perte en valeur absolue

$$\ell(\theta, \theta') = |\theta - \theta'|, \qquad \theta, \theta' \in \mathbb{R}.$$

Proposition 3. Soit ℓ la perte en valeur absolue et soit Π une loi a priori sur $\Theta \subset \mathbb{R}$ admettant une densité π par rapport à la mesure de Lebesgue sur $\mathbb{R}$ avec $\int |\theta| d\Pi(\theta) < \infty$. Un estimateur de Bayes pour ℓ et la loi Π est donné par

$$\hat{\theta}^{\Pi}(X) = \hat{\theta}^{med}(X),$$

la médiane a posteriori pour la loi a priori Π.

Remarque. Modulo une preuve légèrement plus technique, on pourrait se passer de l'hypothèse de densité par rapport à la mesure de Lebesgue.

Preuve.

Pour abréger, nous noterons $Q(\cdot)$ la loi a posteriori $\Pi[\cdot \,|\, X]$. Notons que Q a une densité q par rapport à la mesure de Lebesgue $dQ(\theta) = q(\theta)d\theta$; en effet d'après la formule de Bayes $q(\theta) \propto p_\theta(x)\pi(\theta)$. Soit F_Q la fonction de répartition correspondante. D'après le Théorème 1, il suffit de chercher un minimum de la fonction,

$$y \to \int |y - \theta| dQ(\theta) =: I(y)$$

(le symbole $=:$ veut dire 'égal par définition à'). En séparant suivant le signe de $y-\theta$,

$$\begin{aligned} I(y) &= \int_{-\infty}^{y} (y-\theta)q(\theta)d\theta + \int_{y}^{\infty} (\theta-y)q(\theta)d\theta \\ &= yF_Q(y) - \int_{-\infty}^{y} \theta q(\theta)d\theta + \int_{y}^{\infty} \theta q(\theta)d\theta - y(1-F_Q(y)). \end{aligned}$$

Comme Q est à densité, $F_Q'(y) = q(y)$. La fonction $y \to I(y)$ est dérivable et d'après la relation ci-dessus

$$I'(y) = F_Q(y) + yq(y) - yq(y) - yq(y) + yq(y) - (1-F_Q(y)) = 2F_Q(y) - 1.$$

Donc $I'(y) = 0$ si $F_Q(y) = 1/2$, ce qui équivaut à $Q(\,(-\infty, y]\,) = 1/2$. Il suffit de prendre y égal à la médiane de la loi Q, c'est-à-dire la médiane a posteriori.

3 Bayes et fonction de perte de classification

On suppose dans ce paragraphe que l'espace des paramètres Θ peut naturellement s'écrire comme une partition

$$\Theta = \Theta_0 \cup \Theta_1,$$

avec $\Theta_0 \cap \Theta_1 = \emptyset$. On définit la *fonction de perte de classification* comme, pour $\theta, \theta' \in \Theta$,

$$L_c(\theta, \theta') = \mathbb{1}_{\theta\in\Theta_0,\ \theta'\in\Theta_1} + \mathbb{1}_{\theta\in\Theta_1,\ \theta'\in\Theta_0}.$$

Notons que $L_c(\theta, \theta') = 0$ si et seulement si θ, θ' sont dans la même région Θ_0 ou Θ_1.

Proposition 4. Soient $\theta_0 \in \Theta_0$ et $\theta_1 \in \Theta_1$ deux points fixés quelconques. Soit Π une loi a priori sur Θ. Un estimateur de Bayes pour la perte L_C est donné par

$$\hat{\theta}^C(X) = \theta_1 \mathbb{1}_{\Pi(\Theta_0\,|\,X)\le\Pi(\Theta_1\,|\,X)} + \theta_0 \mathbb{1}_{\Pi(\Theta_0\,|\,X)>\Pi(\Theta_1\,|\,X)}.$$

Preuve.

D'après le Théorème 1, il suffit de minimiser le risque a posteriori

$$\begin{aligned} \int L_C(\theta, T(X))d\Pi(\theta\,|\,X) &= \int \mathbb{1}_{\theta\in\Theta_0}\mathbb{1}_{T(X)\in\Theta_1} d\Pi(\theta\,|\,X) + \int \mathbb{1}_{\theta\in\Theta_1}\mathbb{1}_{T(X)\in\Theta_0} d\Pi(\theta\,|\,X) \\ &= \Pi[\Theta_0\,|\,X]\mathbb{1}_{T(X)\in\Theta_1} + \Pi[\Theta_1\,|\,X]\mathbb{1}_{T(X)\in\Theta_0}. \end{aligned}$$

Cette fonction est minimale si, pour chaque X fixé, on choisit $T(X) \in \Theta_1$ quelconque si $\Pi(\Theta_1\,|\,X) \ge \Pi[\Theta_0\,|\,X]$ et $T(X) \in \Theta_0$ quelconque si $\Pi(\Theta_1\,|\,X) < \Pi[\Theta_0\,|\,X]$.

Exemple. Dans le modèle $\mathcal{P} = \{P_0, P_1\} = \{P_\theta,\ \theta \in \Theta\}$ avec $\Theta = \{0, 1\}$, supposons $dP_0 = f_0 d\mu$, $dP_1 = f_1 d\mu$ et soit Π une loi de probabilité sur Θ définie par

$$\Pi[\{0\}] = \pi_0 > 0, \quad \Pi[\{1\}] = \pi_1 = 1 - \pi_0.$$

Un estimateur de Bayes pour la perte de classification est donné par

$$\hat{\theta}^C(X) = \mathbb{1}\{\Pi[\Theta_0 \mid X] \leq \Pi[\Theta_1 \mid X]\} = \mathbb{1}\left\{\frac{\pi_0 f_0(X)}{\pi_0 f_0(X) + \pi_1 f_1(X)} \leq \frac{\pi_1 f_1(X)}{\pi_0 f_0(X) + \pi_1 f_1(X)}\right\},$$

soit

$$\hat{\theta}^C(X) = \mathbb{1}\{\pi_0 f_0(X) \leq \pi_1 f_1(X)\}.$$

En particulier, dans le cas d'un a priori uniforme $\pi_0 = \pi_1 = 1/2$, on sélectionne $\theta = 1$ si et seulement si la densité sous P_0 au point observé (i.e. la vraisemblance !) est plus grande que la densité sous P_1 en ce point. Pour un a priori non uniforme, on pondère par π_0, π_1.

3.3 Relation entre critères de décision

A *Une inégalité très simple et très utile.*

Théorème 2. Pour toute loi a priori Π sur Θ,

$$R_B(\Pi) \leq R_M.$$

Le risque bayésien minore toujours le risque minimax.

Preuve.

Par définition $R_B(\Pi) = \inf_T \int R(\theta, T) d\Pi(\theta)$. Or comme $\Pi(\Theta) = 1$,

$$\int R(\theta, T) d\Pi(\theta) \leq \sup_{\theta \in \Theta} R(\theta, T) \int d\Pi(\theta) = \sup_{\theta \in \Theta} R(\theta, T).$$

En prenant l'infimum en T de part et d'autre, il vient $R_B(\Pi) \leq R_M$.

De nombreuses minorations de risques minimax reposent sur cette inégalité. Souvent, le risque minimax sur un modèle donné peut être obtenu en construisant une loi a priori 'la plus défavorable' pour laquelle $R_B(\Pi)$ est le plus grand possible. Nous verrons un exemple ci-dessous.

✍ Applications pratiques (note culturelle). Cette inégalité intervient dans de très nombreux cas où l'on veut évaluer ou minorer un risque minimax. Nous verrons des exemples en 3.4 ci-dessous et en TD.

B *Admissibilité, conditions suffisantes*

Definition 9. On dit que deux estimateurs T_1 et T_2 sont équivalents si leurs fonctions de risque sont égales, soit

$$R(\theta, T_1) = R(\theta, T_2), \qquad \forall\, \theta \in \Theta.$$

Théorème 3. [De Bayes et unique ⇒ admissible] Soit T_1 un estimateur de Bayes pour Π et supposons-le unique à équivalence près, c'est-à-dire que si T_2 est de Bayes pour Π alors $R(\theta, T_2) = R(\theta, T_1)$ pour tout $\theta \in \Theta$. Alors T_1 est admissible.

Preuve.

Supposons T_1 inadmissible. Alors il existerait T_2 et $\theta_1 \in \Theta$ avec

$$\begin{aligned} R(\theta, T_2) &\leq R(\theta, T_1) \qquad \forall\ \theta \in \Theta \\ R(\theta_1, T_2) &< R(\theta_1, T_1). \end{aligned}$$

On intègre la première identité par rapport à Π, ce qui donne $R_B(\Pi, T_2) \leq R_B(\Pi, T_1) = R_B(\Pi)$ car T_1 est de Bayes pour Π, et donc T_2 aussi. Par unicité à équivalence près, on en déduit $R(\theta, T_1) = R(\theta, T_2)$ pour tout $\theta \in \Theta$, ce qui contredit l'inégalité ci-dessus si $\theta = \theta_1$.

Cas de la perte quadratique. Supposons $\Theta \subset \mathbb{R}$, soit Π une loi a priori et ℓ la perte quadratique. Supposons que le risque de Bayes $R_B(\Pi)$ est fini et soient T_2 un estimateur de Bayes. On sait déjà que $T_1(X) = \overline{\theta}(X)$ est un estimateur de Bayes pour la perte quadratique et Π.

Une condition suffisante pour que T_1, T_2 soient équivalents pour la perte quadratique est que la loi marginale P_X de X domine toutes les lois P_θ pour $\theta \in \Theta$ (cela signifie que $P_X(A) = 0$ implique $P_\theta(A) = 0$ pour tout $\theta \in \Theta$).

En effet, d'après la preuve du théorème 1, si $m(x)$ désigne la densité marginale de X,

$$\begin{aligned} R_B(\Pi) = R_B(\Pi, T_2) &= \int E_\theta((T_2 - \theta)^2) d\Pi(\theta) \\ &= \int E[(T_2(x) - \theta)^2 \,|\, X = x] m(x) d\mu(x), \end{aligned}$$

où l'on avait noté $\rho(\Pi, T_2 \,|\, X) = E[(\theta - T_2(X))^2 \,|\, X = x]$. En écrivant $T_2(x) - \theta = T_2(x) - \overline{\theta}(x) + \overline{\theta}(x)$ et en développant le carré, on constate que l'espérance sachant $X = x$ du terme croisé est nulle car $E[\theta \,|\, X = x] = \overline{\theta}(x)$ par définition. On en déduit

$$R_B(\Pi, T_2) = R_B(\Pi, T_1) + \int (\overline{\theta}(x) - T_2(x))^2 m(x) d\mu(x).$$

Donc $\int(\overline{\theta}(x) - T_2(x))^2 dP_X(x) = 0$. Donc $\overline{\theta}(x) = T_2(x)$, P_X-presque sûrement. Comme P_X domine toutes les lois P_θ, on a $\overline{\theta}(x) = T_2(x)$, P_θ-presque sûrement. Cela implique

$R(\theta, \overline{\theta}) = R(\theta, T_2)$ puisque deux fonctions égales P_θ-p.s. ont la même intégrale contre P_θ. Ainsi T_2 et $\overline{\theta}$ sont équivalents.

Exemple du modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$. Si $\Pi = \Pi_{a,\sigma^2} = \mathcal{N}(a, \sigma^2)$ est prise comme loi a priori sur Π, l'estimateur

$$\overline{\theta}_{a,\sigma^2}(X) = \frac{\overline{X} + \frac{a}{n\sigma^2}}{1 + \frac{1}{n\sigma^2}}$$

est de Bayes pour Π_{a,σ^2}. D'après le critère ci-dessus, pour montrer qu'il est admissible, il suffit de vérifier que la loi de $(X_1, \ldots, X_n)$ domine toutes les lois $\mathcal{N}(\theta, 1)^{\otimes n}$. C'est le cas car toutes ces lois admettent des densités strictement positives par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, et donc n'importe laquelle de ces lois domine les autres. Ainsi $\overline{\theta}_{a,\sigma^2}(X)$ est admissible. Cela montre que tout estimateur de la forme $\alpha\overline{X} + \beta$ avec $\alpha \in (0, 1)$ et $\beta \in \mathbb{R}$ est admissible (dans ce modèle, pour la perte quadratique). On pourra vérifier en exercice que $0\overline{X} + \beta = \beta$ est aussi admissible (estimateur constant), alors que $1\overline{X} + \beta$ ne l'est pas (le comparer à $\overline{X}$).

Théorème 4. ['presque de Bayes' ⇒ admissible] Supposons que toutes les fonctions de risque $\theta \to R(\theta, T)$ finies sont continues. Soit T un estimateur de risque fini tel que pour tout $\varepsilon > 0$, pour tout ouvert U non vide de Θ, il existe une loi a priori $\Pi = \Pi_{\varepsilon,U}$ sur Θ telle que

$$R_B(\Pi, T) < R_B(\Pi) + \varepsilon\Pi(U).$$

Alors T est admissible.

Preuve.

On raisonne par l'absurde. Si T était inadmissible, il existerait T_2 et $\theta_1 \in \Theta$ avec

$$\begin{aligned} R(\theta, T_2) &\leq R(\theta, T) \qquad \forall\ \theta \in \Theta \\ R(\theta_1, T_2) &< R(\theta_1, T). \end{aligned}$$

Notons que T_2 est alors de risque fini car T l'est Par ailleurs, par continuité des fonctions $\theta \to R(\theta, T')$ pour $T' = T$ et $T' = T_2$, il existe V voisinage de θ_1 tel que pour tout $\theta \in V$,

$$R(\theta, T_2) \leq R(\theta, T) - \varepsilon.$$

Par ailleurs, pour toute loi a priori Π,

$$\begin{aligned} R_B(\Pi, T_2) &= \int_V R(\theta, T_2) d\Pi(\theta) + \int_{V^c} R(\theta, T_2) d\Pi(\theta) \\ &\leq \int_V R(\theta, T) d\Pi(\theta) - \varepsilon\Pi(V) + \int_{V^c} R(\theta, T_2) d\Pi(\theta) \\ &\leq R_B(\Pi, T) - \varepsilon\Pi(V). \end{aligned}$$

Par définition du risque bayésien, $R_B(\Pi) \leq R_B(\Pi, T_2)$, donc

$$R_B(\Pi) \leq R_B(\Pi, T) - \varepsilon\Pi(V)$$

soit $R_B(\Pi, T) \geq R_B(\Pi) + \varepsilon\Pi(V)$, ce qui contredit l'hypothèse du théorème.

Exemple du modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$. Montrons à l'aide du Théorème 3 que $\overline{X}$ est admissible pour la perte quadratique. Son risque vaut $1/n$ et est donc fini. La continuité des fonctions de risque finies pour ce modèle sera vue en TD. Soit $\varepsilon > 0$ fixé et U un ouvert non vide de $\mathbb{R}$. Soit $[c, d], c < d$ un intervalle inclus dans U. Soit $\Pi = \mathcal{N}(a, \sigma^2)$. Pour cette loi a priori, on peut calculer le risque bayésien (pour la perte quadratique) : il s'agit du risque quadratique de la moyenne a posteriori d'après la Proposition 2. On vérifie par le calcul (exercice) que

$$R_B(\Pi) = \frac{1}{n + \sigma^{-2}}.$$

Par ailleurs, puisque le risque de $\overline{X}$ est constant égal à $1/n$, on a $R_B(\Pi, \overline{X}) = 1/n$. Ainsi

$$R_B(\Pi, \overline{X}) - R_B(\Pi) = \frac{\sigma^{-2}}{n(n + \sigma^{-2})}.$$

Par ailleurs,

$$\Pi([c, d]) = \int_c^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}} dx \sim \frac{c - d}{\sqrt{2\pi}\sigma} \quad (\sigma \to \infty).$$

Comme $\sigma^{-2} = o(\sigma^{-1})$ quand $\sigma \to \infty$, on a donc

$$R_B(\Pi, \overline{X}) - R_B(\Pi) < \varepsilon\, \Pi([c, d])$$

pour σ assez grand. Donc $\overline{X}$ est admissible d'après le Théorème 3.

C *Minimaxité : conditions suffisantes*

Proposition 5. Un estimateur admissible, de risque constant, est minimax.

Application. Dans le modèle fondamental, $\overline{X}$ est minimax car admissible (voir plus haut) et de risque constant. Une preuve alternative de ce fait est également donnée plus bas.

Preuve.

Un tel estimateur T, étant de risque constant, vérifie $R_{max}(T) = R(\theta, T)$ pour tout $\theta \in \Theta$. S'il n'était pas minimax, on pourrait trouver un estimateur T_2 tel que

$$R(\theta, T_2) \leq R_{max}(T_2) < R_{max}(T) = R(\theta, T)$$

pour tout $\theta \in \Theta$ et donc T ne serait pas admissible. Contradiction.

Théorème 5. Soit T un estimateur de Bayes pour une loi a priori Π. Si T est de risque constant, alors T est minimax.

Application. Dans le modèle binomial $P_\theta = \text{Bin}(n, \theta)$, avec $\theta \in [0, 1]$, un estimateur minimax pour le risque quadratique peut s'obtenir comme suit. Soit $\Pi_{a,b}$ une loi a priori Beta(a, b) (voir TDs) sur θ. Pour tous a, b positifs, on peut calculer explicitement la moyenne a posteriori $\overline{\theta}_{a,b}(X)$ pour l'a priori $\Pi_{a,b}$. L'un de ces estimateurs a un risque quadratique constant (voir TDs pour les calculs), il est donc minimax.

Preuve.

Si T n'était pas minimax, on pourrait trouver un estimateur T_2 tel que

$$R_B(\Pi, T_2) \leq R_{max}(T_2) < R_{max}(T).$$

Mais comme T est de risque constant par hypothèse, $R_{max}(T) = R_B(T)$, ce qui implique avec l'identité précédente que T ne peut être de Bayes pour Π. Contradiction.

Théorème 6. Si l'estimateur T est tel qu'on puisse trouver une suite $(\Pi_k)_{k\geq 1}$ de lois a priori avec

$$R_{max}(T) \leq \overline{\lim_{k\to\infty}} R_B(\Pi_k),$$

alors T est minimax.

Application. $\overline{X}$ est minimax pour la perte quadratique (le retour !). Pour la loi a priori $\Pi_{a,\sigma^2} = \mathcal{N}(a, \sigma^2)$, le risque bayésien $R_B(\Pi_{a,\sigma^2})$ s'obtient en calculant le risque de Bayes de la moyenne a posteriori, puisqu'il s'agit d'un estimateur de Bayes pour la perte quadratique. On obtient $R_B(\Pi_{a,\sigma^2}) = (n + \sigma^{-2})^{-1}$. Or

$$\lim_{\sigma^2\to\infty} R_B(\Pi_{a,\sigma^2}) = \frac{1}{n} = R_{max}(\overline{X}),$$

ce qui montre à nouveau que $\overline{X}$ est minimax.

Preuve.

Tout risque bayésien est inférieur ou égal au risque minimax R_M, qui est lui-même inférieur ou égal à $R_{max}(T)$. Donc

$$R_{max}(T) \leq \overline{\lim_{k\to\infty}} R_B(\Pi_k) \leq R_M \leq R_{max}(T)$$

On en conclut $R_{max}(T) = R_M$ donc T est minimax.

Remarque. On peut noter que dans l'énoncé du théorème, si l'hypothèse est vérifiée, il y a nécessairement égalité entre le risque et la limsup. En effet, par définition du risque maximal,

$$\begin{aligned}
R_{max}(T) &\geq R(\theta, T) \\
R_{max}(T) &\geq R_B(\Pi_k, T) && \text{on intègre en } \Pi_k \\
&\geq R_B(\Pi_k) && \text{définition de } R_B(\Pi_k) \\
&\geq \overline{\lim_{k\to\infty}} R_B(\Pi_k) && \text{on prend la } \overline{\lim} \text{ en } k.
\end{aligned}$$

On en déduit grâce à l'hypothèse que $R_{max}(T) = \overline{\lim}_{k\to\infty} R_B(\Pi_k)$.

3.4 Minorations du risque minimax

De nombreux résultats de minoration du risque minimax reposent sur l'argument bayésien très simple suivant. On remarque que pour ℓ une fonction de perte donnée, *pour tous points* θ_1, θ_2 *quelconques de* Θ,

$$R(\theta_1, T) \leq R_{max}(T)$$
$$R(\theta_2, T) \leq R_{max}(T),$$

ce qui entraîne

$$\frac{1}{2}\left[R(\theta_1, T) + R(\theta_2, T)\right] \leq R_{max}(T).$$

Notons $\Pi_{1,2}$ la loi a priori sur Θ

$$\Pi_{1,2} = \frac{1}{2}\delta_{\theta_1} + \frac{1}{2}\delta_{\theta_2}.$$

On note alors que

$$\begin{aligned}\frac{1}{2}\left[R(\theta_1, T) + R(\theta_2, T)\right] &= \int \frac{1}{2}\ell(\theta_1, T(x))dP_{\theta_1}(x) + \int \frac{1}{2}\ell(\theta_2, T(x))dP_{\theta_2}(x) \\ &= \int\int \ell(\theta, T(x))dP_\theta(x)d\Pi_{1,2}(\theta) = R_B(\Pi_{1,2}, T).\end{aligned}$$

Ainsi $R_B(\Pi_{1,2}, T) \leq R_{max}(T)$ ce qui implique en prenant l'infimum en T que $R_B(\Pi_{1,2}, T) \leq R_M$ (on retrouve le Théorème 2), soit

$$R_M \geq \frac{1}{2}\left[R(\theta_1, T) + R(\theta_2, T)\right]. \tag{3}$$

Avant d'énoncer un résultat découlant de (3), notons que si $\ell(\cdot,\cdot)$ est une distance $d(\cdot,\cdot)$, alors par inégalité triangulaire,

$$d(\theta_1, T) + d(\theta_2, T) \geq d(\theta_1, \theta_2).$$

Si $\ell(\cdot,\cdot)$ est une distance au carré $d(\cdot,\cdot)^2$ (par exemple une norme au carré), alors l'inégalité $(a+b)^2 \leq 2(a^2+b^2)$ entraîne que

$$d(\theta_1, T)^2 + d(\theta_2, T)^2 \geq \frac{1}{2}d(\theta_1, \theta_2).$$

Condition sur la fonction de perte ℓ. Dans la suite, nous supposerons qu'il existe $\theta_1, \theta_2 \in \Theta$ tels que pour tout $t \in \Theta$,

$$\ell(\theta_1, T) + \ell(\theta_2, T) \geq \alpha(\theta_1, \theta_2), \tag{4}$$

où $\alpha(\theta_1, \theta_2)$ est une fonction quelconque de θ_1 et de θ_2 (mais qui ne dépend pas de T).

Nous avons vu ci-dessus que (4) est vérifiée pour ℓ une distance ou une distance au carré.

Definition 10. Soient P, Q deux mesures de probabilité avec $dP = pd\mu$ et $dQ = qd\mu$. On appelle affinité $\mathcal{A}(P, Q)$ entre P et Q la quantité, avec $p \wedge q = \min(p, q)$,

$$\mathcal{A}(P, Q) = \int (p \wedge q) d\mu = \int \min(p(x), q(x)) d\mu(x).$$

Théorème 7. Si (4) est vérifiée en des points θ_1, θ_2 de Θ, alors

$$R_M \geq \frac{\alpha(\theta_1, \theta_2)}{2} \mathcal{A}(P_{\theta_1}, P_{\theta_2}).$$

Preuve.

D'après (3), et en utilisant (4),

$$\begin{aligned} R_M &\geq \frac{1}{2} E_{\theta_1} \ell(\theta_1, T) + \frac{1}{2} E_{\theta_2} \ell(\theta_2, T) \\ &\geq \frac{1}{2} \int \ell(\theta_1, T(x)) p_{\theta_1}(x) d\mu(x) + \frac{1}{2} \int \ell(\theta_2, T(x)) p_{\theta_2}(x) d\mu(x) \\ &\geq \int \left[\frac{1}{2} \ell(\theta_1, T(x)) + \frac{1}{2} \ell(\theta_2, T(x)) \right] (p_{\theta_1} \wedge p_{\theta_2})(x) d\mu(x) \\ &\geq \int \frac{1}{2} \alpha(\theta_1, \theta_2) (p_{\theta_1} \wedge p_{\theta_2})(x) d\mu(x) = \frac{1}{2} \alpha(\theta_1, \theta_2) \mathcal{A}(P_{\theta_1}, P_{\theta_2}). \end{aligned}$$

Proposition 6. Soient P, Q deux mesures de probabilité avec $dP = pd\mu$ et $dQ = qd\mu$. Posons

$$\|P - Q\|_1 = \int |p - q| d\mu.$$

Alors on peut écrire

$$\mathcal{A}(P, Q) = 1 - \frac{1}{2} \|P - Q\|_1.$$

Preuve.

On note que $p \wedge q = \frac{1}{2}(p + q - |p - q|)$. Il suffit d'intégrer par rapport à μ en utilisant que p, q sont des densités par rapport à μ.

On déduit de ce qui précède que si (4) est vérifiée,

$$R_M \geq \frac{\alpha(\theta_1, \theta_2)}{2}\left(1 - \frac{1}{2}\|P_{\theta_1} - P_{\theta_2}\|_1\right). \tag{5}$$

L'objectif dans la suite de cette section va être, entre autres, de démontrer que la 'meilleure vitesse possible' au sens du risque minimax dans les modèles paramétriques réguliers est $\frac{C}{\sqrt{n}}$. De tels modèles sont de la forme $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$, avec $\Theta \subset \mathbb{R}^d$ et $d \geq 1$ fixé (par exemple, le modèle fondamental avec n observations).

Pour cela, d'après (5), il suffit de majorer $\|P_{\theta_1}^{\otimes n} - P_{\theta_2}^{\otimes n}\|_1$, pour des points θ_1, θ_2 bien choisis dans Θ.

Definition 11. Soient P, Q deux mesures de probabilité avec $dP = pd\mu$ et $dQ = qd\mu$. On définit l'affinité de Hellinger entre P et Q par

$$\rho(P, Q) = \int \sqrt{p}\sqrt{q}d\mu,$$

et on rappelle que la distance de Hellinger entre P et Q est

$$h(P, Q) = \left\{\int (\sqrt{p} - \sqrt{q})^2 d\mu\right\}^{1/2}.$$

On peut vérifier que les définitions ci-dessus sont indépendantes du choix de la mesure dominante μ. Les quantités ρ et h ont les propriétés suivantes

1. $h(P, Q)^2 = 2 - 2\rho(P, Q)$ par définition
2. $0 \leq h(P, Q) \leq \sqrt{2}$ on développe le carré et $\int p = \int q = 1$
3. $\|P - Q\|_1 \leq 2h(P, Q)$

   Preuve.

   En utilisant l'inégalité de Cauchy-Schwarz et $(a + b)^2 \leq 2a^2 + 2b^2$,

   $$\begin{aligned}\int |p - q|d\mu &= \int |\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}|d\mu \\ &\leq h(P, Q)\left(\int (2p + 2q)d\mu\right)^{\frac{1}{2}} = 2h(P, Q).\end{aligned}$$

4. Soient deux mesures produit P, Q données par

   $$P = \otimes_{i=1}^n P_i, \quad Q = \otimes_{i=1}^n Q_i,$$

où l'on suppose que $dP_i = p_i d\mu$ et $dQ_i = q_i d\mu$, pour tous indices i. Alors

$$\rho(P, Q) = \prod_{i=1}^{n} \rho(P_i, Q_i).$$

Preuve.

En effet, par définition de P,

$$dP(x_1, \dots, x_n) = p_1(x_1) \cdots p_n(x_n) d\mu(x_1) \cdots d\mu(x_n),$$

donc P a pour densité $p_1(x_1) \cdots p_n(x_n)$ par rapport à $\mu^{\otimes n}$. Donc via le théorème de Fubini,

$$\begin{aligned}
\rho(P, Q) &= \int \cdots \int \sqrt{\prod_{i=1}^{n} p_i(x_i)} \sqrt{\prod_{i=1}^{n} q_i(x_i)} d\mu(x_1) \cdots d\mu(x_n) \\
&= \prod_{i=1}^{n} \int \sqrt{p_i(x_i)} \sqrt{q_i(x_i)} d\mu(x_i) \\
&= \prod_{i=1}^{n} \rho(P_i, Q_i).
\end{aligned}$$

Pour P_{θ_1} et P_{θ_2} dans le modèle $\mathcal{P}$, on note $h(\theta_1, \theta_2) = h(P_{\theta_1}, P_{\theta_2})$ pour simplifier les notations.

Théorème 8. [Le Cam] Soit $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$ un modèle quelconque et ℓ une fonction de perte donnée par, pour tous s, t dans Θ,

$$\ell(s, t) = d(s, t)^p$$

pour $p \geq 1$ un entier et d une distance sur Θ. Alors pour tous $\theta_1, \theta_2 \in \Theta$,

$$\begin{aligned}
R_M &= \inf_T \sup_{\theta \in \Theta} E_\theta \left[\ell(\theta, T(X)) \right] \\
&\geq \frac{1}{2^p} (1 - \sqrt{n} h(\theta_1, \theta_2)) d(\theta_1, \theta_2)^p.
\end{aligned}$$

Preuve.

On commence par vérifier que d^p satisfait la propriété (4) avec

$$\alpha(\theta_1, \theta_2) = \frac{1}{2^{p-1}} d(\theta_1, \theta_2)^p.$$

Cela résulte de l'inégalité, par convexité de $x \to x^p$, pour $a, b \geq 0$,

$$\left(\frac{a+b}{2}\right)^p \leq \frac{a^p + b^p}{2}.$$

On applique maintenant (5) aux mesures $P_{\theta_1}^{\otimes n}$ et $P_{\theta_2}^{\otimes n}$,

$$R_M \geq \frac{1}{2}\frac{1}{2^{p-1}} d(\theta_1, \theta_2)^p \left\{1 - \frac{1}{2}\|P_{\theta_1}^{\otimes n} - P_{\theta_2}^{\otimes n}\|_1\right\}.$$

En utilisant les propriétés de ρ et h ci-dessus, on majore maintenant

$$\|P_{\theta_1}^{\otimes n} - P_{\theta_2}^{\otimes n}\|_1 \leq h(P_{\theta_1}^{\otimes n}, P_{\theta_2}^{\otimes n})$$

et on écrit, en notant pour simplifier $P_1 = P_{\theta_1}, P_2 = P_{\theta_2}$,

$$\begin{aligned} h(P_1^{\otimes n}, P_2^{\otimes n})^2 &= 2 - 2\rho(P_1^{\otimes n}, P_2^{\otimes n}) \\ &= 2 - 2\rho(P_1, P_2)^n \\ &= 2 - 2\left(1 - \frac{h(P_1, P_2)}{2}\right)^n. \end{aligned}$$

En utilisant l'inégalité $(1-x)^n \geq 1 - nx$ pour $x \geq 0$ (la fonction convexe $x \to (1-x)^n$ reste au-dessus de sa tangente en 0), on en déduit

$$\begin{aligned} h(P_1^{\otimes n}, P_2^{\otimes n})^2 &\leq 2 - 2(1 - n\frac{h^2(P_1, P_2)}{2}) \\ &\leq n h^2(P_1, P_2). \end{aligned}$$

On combine les inégalités ci-dessus pour obtenir le résultat annoncé.

3.5 Applications

Une première conséquence générale du Théorème 8 est que si $d = h$ distance de Hellinger, si l'on dispose de n observations i.i.d., et si que l'on peut trouver deux points quelconques dans le modèle à une distance de Hellinger égale à $\frac{c}{\sqrt{n}}$, avec $c < 1$, alors

$$R_M \geq \frac{1}{2}(1-c)\frac{c}{\sqrt{n}} = \frac{C}{\sqrt{n}}.$$

Exemple de calcul de h. Dans le modèle fondamental $P_\theta = \mathcal{N}(\theta, 1)$, on peut vérifier par le calcul (voir TD) que, si l'on note $h(\theta, \theta') := h(P_\theta, P_{\theta'})$,

$$h^2(\theta, \theta') = 2 - 2e^{-\frac{\theta - \theta'}{8}}.$$

On a $0 \leq h^2 \leq 2$ et cette distance vaut ici 0 si $\theta = \theta'$. Si θ et θ' sont proches on a $h^2(\theta, \theta') = (\theta - \theta')^2/4$. Si l'on pose $\theta_1 = 0$ et $\theta_2 = 1/\sqrt{n}$, d'après ce qui précède $R_M \geq \frac{C}{\sqrt{n}}$.

A *Modèles paramétriques réguliers.*

L'exemple précédent dans le modèle fondamental est une manifestation d'un phénomène plus général dans les modèles paramétriques 'réguliers' (nous ne définirons pas précisément cette notion dans le cadre de ce cours mais elle signifie grossièrement que l'on peut dériver suffisamment en θ la quantité $p_\theta(x)$). Pour un tel modèle, si disons $\Theta \subset \mathbb{R}$, si $\theta_0 \in \Theta$ est un point dans l'intérieur de Θ et que l'on suppose l'information de Fisher $I(\theta_0)$ en θ_0 strictement positive, soit $I(\theta_0) > 0$, on peut montrer (admis) que

$$\lim_{t\to 0} \frac{h^2(\theta_0 + t, \theta_0)}{t^2} = \frac{I(\theta_0)}{4}.$$

En particulier ceci implique que $h^2(\theta_0 + t, \theta_0) \asymp t^2$ quand $t \to 0$. Posons, avec θ_0 un point dans l'intérieur de Θ, et $a > 0$ assez petit,

$$\theta_1 = \theta_0, \qquad \theta_2 = \theta_0 + \frac{a}{\sqrt{n}}.$$

Par définition, $(\theta_2 - \theta_1)^2 = \frac{a^2}{n}$, donc d'après ce qui précède,

$$h^2(\theta_1, \theta_2) \sim \frac{I(\theta_0)}{4}\frac{a^2}{n}.$$

D'après le théorème 8, pour a assez petit et n assez grand,

$$\inf_T \sup_{\theta\in\Theta} E_\theta((T(X) - \theta)^2) \geq \frac{1}{4}(1 - C)\frac{a^2}{n} \geq Dn,$$

où l'on a posé $C = I(\theta_0)a^2/8$ et où l'on choisit a assez petit de sorte que $I(\theta_0)a^2 < 1$.

Le risque minimax pour la perte quadratique dans les modèles paramétriques réguliers est donc borné inférieurement par D/n, où D est une constante qui dépend de $I(\theta_0)^{-1}$. On peut en fait montrer que la 'constante optimale' autour du point θ_0 est $I(\theta_0)^{-1}$.

✍ Les points B et C ci-après sont des notes culturelles et ne sont pas exigibles à l'examen.

B *Un exemple de modèle non-paramétrique.*

On observe $X_1, \ldots, X_n \in [0, 1]$ i.i.d. de densité inconnue f sur l'intervalle $[0, 1]$. On suppose que f appartient à l'intersection $\mathcal{C}$ de l'ensemble de toutes les densités avec la classe de Hölder $\Sigma(1, L)$ définie par, pour $L > 0$,

$$\Sigma(1, L) = \{f : [0, 1] \to \mathbb{R},\ \|f\|_\infty + \|f'\|_\infty \leq L\}.$$

Le risque minimax pour l'estimation de f au point $x_0 \in (0, 1)$ pour la fonction de perte quadratique est minoré par, pour $X = (X_1, \ldots, X_n)$,

$$R_M = \inf_T \sup_{f\in\mathcal{C}} E_f(T(X) - f(x_0))^2 \geq C_0 n^{-2/3}$$

pour une constante $C_0 > 0$ dépendant seulement de L. Nous montrerons ce résultat en TD. Plus généralement, pour une classe de régularité Hölder $\Sigma(\beta, L)$ avec $\beta > 0$, on peut montrer

que le risque minimax R_M se minore par $n^{-\beta/(2\beta+1)}$. On peut également montrer que R_M est exactement de cet ordre. En effet, cette vitesse est atteinte par un estimateur 'à noyau' (cf un cours d'estimation non-paramétrique).

C *Un exemple de modèle de 'grande dimension'*

Considérons le modèle dit de *de suite gausienne parcimonieuse*

$$X_i = \theta_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

avec $\varepsilon_i \sim \mathcal{N}(0,1)$ indépendants, avec $\theta = (\theta_1, \ldots, \theta_n) \in \ell_0[s_n] =: \Theta$, où l'on pose

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n,\ \mathrm{Card}\{i,\ \theta_i \neq 0\} \leq s_n\}\,,$$

et on suppose que $s_n/n \to 0$ quand $n \to \infty$, c'est-à-dire que seul un petit nombre de coordonnées de θ sont non nulles (mais on ne sait pas lesquelles).

Donoho, Hoch, Johnstone et Stern (1994) ont montré que, lorsque $n \to \infty$,

$$\inf_T \sup_{\theta \in \ell_0[s_n]} E_\theta \|T(X) - \theta\|^2 = 2 s_n \log(n/s_n)(1 + o(1)),$$

où l'infimum porte sur tous les estimateurs possibles de θ dans le modèle ci-dessus.

CHAPITRE 4

Critères de choix de lois a priori

Nous examinons plusieurs critères possibles de choix de lois a priori. Certains sont dictés par des impératifs pratiques. Par exemple, certaines lois a priori induisent des lois a posteriori plus simples à calculer que d'autres. Nous voyons en particulier le cas des familles conjuguées. Des choix de lois a priori, comme l'a priori de Jeffreys, sont basés sur des notions d'invariance. D'autres critères encore cherchent à 'estimer' la loi a priori a partir des données, comme c'est le cas des méthodes bayésiennes empiriques. Il est également possible d'utiliser 'plusieurs niveaux' de lois a priori, ce qui mène à des méthodes dites hiérarchiques.

Dans l'approche bayésienne, un élément central est la loi a priori Π. Une fois Π choisie, toute l'inférence en découle : loi a posteriori $\Pi[\cdot \mid X]$, estimateurs 'aspects' de l'a posteriori $\overline{\theta}(X), \hat{\theta}^{med}(X), \hat{\theta}^{mode}(X)$, les régions de crédibilité ...

De plus, nous avons vu au Chapitre 3 que Π étant donnée, on pouvait fonder notre décision, si le risque bayésien est choisit, sur l'estimateur de Bayes, qui minimise ce risque.
Mais comment choisir Π en général ? Nous donnons ici quelques éléments de réponse, sans être exaustifs.

4.1 Information disponible avant l'expérience

Dans de nombreux cas, le statisticien dispose d'éléments (plus ou moins précis) sur le paramètre à estimer. Ces éléments peuvent être

* *qualitatifs* : on peut savoir à l'avance, par exemple, que le paramètre à estimer est positif. C'est le cas pour un certain nombre de grandeurs physiques (poids, taille). Il est alors naturel de prendre une loi a priori sur $\mathbb{R}^+$ plutôt que sur $\mathbb{R}$ tout entier. Parfois, des contraintes de formes sont connues à l'avance , comme la monotonicité ou la convexité de densités de lois apparaissant dans le modèle.

* *quantitatifs* : on peut parfois savoir qu'il est beaucoup plus probable (parce que, par exemple, on a observé de nombreuses expériences similaires) que le paramètre soit dans une certaine région de l'espace plutôt qu'une autre. L'exemple suivant sera vu en TD : on soupçonne un lancé de pièce (on est donc dans le cas d'une expérience de type $\{\text{Be}(\theta),\ \theta \in [0,1]\}$ d'être biaisé avec probabilité $2/3$ de donner 'pile'. Une possibilité dans ce cas est de prendre une loi a priori *mélange* sur θ, de type $a\delta_{2/3} + (1-a)\delta_{1/2}$, pour prendre en compte le fait que, grossièrement, soit le tirage est biaisé avec $\theta = 2/3$, soit il est non-biaisé. Un choix plus réaliste consiste à prendre une loi mélange $a\text{Beta}(2,4) + (1-a)\text{Beta}(3,3)$ comme a priori sur θ. Dans ce cas, les deux lois Beta sont d'espérance $2/3$ et $1/2$ mais mettent aussi un peu de masse a priori autour de ces deux quantités.

Pour certains des critères ci-dessus, on parle parfois d'information "subjective".
A ceux-ci s'ajoutent aussi souvent des critères *pratiques*, liés à la simulation de lois a posteriori et au temps de calcul correspondant. En effet, en dehors de cas simples comme celui de lois a priori conjuguées, la simulation d'échantillons distribués suivant la loi a posteriori (ou le calcul d'aspects comme la moyenne ou la médiane) peut être plus ou moins coûteuse suivant les lois a priori considérées. Nous en dirons un peu plus au Chapitre 7.

Quels critères "objectifs" de choix de Π peut-on proposer ?

4.2 Lois a priori conjuguées

Disposer d'une famille de lois conjuguée rend typiquement les calculs assez simples lorsque les paramètres a posteriori s'expriment explicitement à l'aide de ceux a priori et des données. De plus, si l'on sait simuler suivant les éléments de la famille considérée, la simulation suivant la loi a posteriori est un cas particulier, donc le temps ou complexité de calcul sont souvent réduits dans ce cas ce qui est souvent avantageux.

Modèle/vraisemblance	Famille de lois a priori conjuguée
$\mathcal{N}(\theta, 1)$	$\mathcal{N}(a, v),\ a \in \mathbb{R},\ v > 0$
$\mathcal{N}(\theta, \sigma^2),\ \sigma^2$ *connu*	$\mathcal{N}(a, v),\ a \in \mathbb{R},\ v > 0$
$\text{Be}(\theta)$ ou $\text{Bin}(n, \theta)$	$\text{Beta}(a, b),\ a, b > 0$
$\text{Mult}(p_0, \ldots, p_k)$	$\text{Dirichlet}(a_0, \ldots, a_k),\ a_0, \ldots, a_k > 0$
$\text{Poisson}(\lambda)$	$\text{Gamma}(a, b),\ a, b > 0$
$\text{Exp}(\lambda)$	$\text{Gamma}(a, b),\ a, b > 0$
$\text{Gamma}(k, \lambda),\ k$ *connu*	$\text{Gamma}(a, b),\ a, b > 0$

Exercice. Ecrire la loi a posteriori dans chaque cas et retrouver (s'aider au besoin des exercices de TD) la propriété de conjugaison.

La plupart des cas de lois conjuguées contenues dans le tableau précédent correspondent à un seul paramètre inconnu (à l'exception du modèle multinomial). Lorsque plusieurs paramètres sont inconnus, ce qui revient typiquement à dire que le paramètre est dans un sous-ensemble de $\mathbb{R}^d$, $d > 1$, trouver une loi conjuguée peut être plus délicat. Nous voyons deux exemples

classiques ci-dessous.

A *Modèle $\mathcal{N}(\mu, \sigma^2)$, moyenne et variance inconnus*

Lemme 1. [loi inverse-gamma] Soit $Y \sim \text{Gamma}(a, b)$, de densité $f_Y(y) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}$. Alors $Z = Y^{-1}$ a pour densité

$$f_Z(z) = \frac{b^a}{\Gamma(a)} z^{-a-1} e^{-\frac{b}{z}}.$$

La loi de Z s'appelle loi inverse-gamma $IG(a, b)$.

Preuve.

Le lemme s'obtient en calculant $E[f(Z)] = E[f(Y^{-1})]$ pour toute f mesurable bornée : en effectuant le changement de variable $z = y^{-1}$,

$$\begin{aligned} E[f(Z)] = E[f(Y^{-1})] &= \int_0^\infty f(y^{-1}) \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by} dy \\ &= \int_0^\infty f(z) \frac{b^a}{\Gamma(a)} z^{1-a} e^{-\frac{b}{z}} \frac{1}{z^2} dz \end{aligned}$$

et le résultat s'en déduit.

Idée. Cas où μ est connu et vaut 0. Montrons qu'alors la famille des lois inverse-gamma est conjuguée. Si $\sigma^2 \sim IG(a, b)$ et $X \,|\, \sigma^2 \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned} f_{\sigma^2 \,|\, X}(\sigma^2) &\propto \sigma^{-1} e^{-\frac{X^2}{2\sigma^2}} \sigma^{2-a-1} e^{-\frac{b}{\sigma^2}} \\ &\propto (\sigma^2)^{-a-\frac{3}{2}} e^{-\frac{1}{\sigma^2}(b + \frac{X^2}{2})} \end{aligned}$$

On obtient $\mathcal{L}(\sigma^2 \,|\, X) = IG(a + \frac{1}{2}, b + \frac{X^2}{2})$.

Exercice. Vérifier la propriété de conjugaison dans le cas de n observations.

Dans le cas où à la fois μ et σ^2 sont inconnues, on peut essayer de déjà utiliser une lois inverse-gamma sur σ^2. En revanche, l'idée qui consiste à proposer pour loi a priori sur le couple (μ, σ^2) une loi produit, donc de densité du type $g(\mu)h(\sigma^2)$ ne va pas fonctionner ; en effet, la vraisemblance s'écrit, déjà dans le cas d'une observation, $C\sigma^{-1} \exp\{\frac{1}{2\sigma^2}(X - \mu)^2\}$, qui est une expression qui *mélange* μ et σ^2.

Definition 1. [loi normale inverse-gamma]. On appelle loi $\text{NIG}(a, b, c, d)$, loi normale inverse-gamma la loi sur $\mathbb{R} \times \mathbb{R}^+$ définie par le schéma

$$\begin{aligned} \mu \,|\, \sigma^2 &\sim \mathcal{N}(a, \frac{\sigma^2}{b}) \\ \sigma^2 &\sim IG(c, d). \end{aligned}$$

La densité d'une loi NIG(a,b,c,d) est $\frac{d^c}{\sqrt{2\pi}\Gamma(c)}(\sigma^2)^{-c-\frac{3}{2}}e^{-\frac{d}{\sigma^2}}e^{-\frac{b(\mu-a)^2}{2\sigma^2}}$.

Théorème 1. Soit $X=(X_1,\dots,X_n)$ et considérons le cadre bayésien

$$\begin{aligned} X\,|\,\mu,\sigma^2 &\sim \mathcal{N}(\mu,\sigma^2)^{\otimes n} \\ (\mu,\sigma^2) &\sim NIG(a,b,c,d)=\Pi. \end{aligned}$$

La famille de toutes les lois NIG normales inverse-gamma est conjuguée et

$$\Pi[\cdot\,|\,X] \;=\; NIG\left(a_X,b_X,c_X,d_X\right),$$

avec, si l'on pose $S^2=\overline{X^2}-\overline{X}^2$,

$$\begin{aligned} a_X &= \frac{n\overline{X}+ab}{n+b}, & b_X &= b+n \\ c_X &= c+\frac{n}{2}, & d_X &= d+\frac{nS^2}{2}+\frac{nb}{2(n+b)}(\overline{X}-a)^2 \end{aligned}$$

Preuve.

La vraisemblance s'écrit

$$\begin{aligned} f_{\mu,\sigma^2}(X) &= (2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\{-\frac{n}{2\sigma^2}(\mu-\overline{X})^2-\frac{n}{2\sigma^2}S^2\}. \end{aligned}$$

La formule de Bayes donne

$$f_{\mu,\sigma^2\,|\,X}(\mu,\sigma^2)\propto(\sigma^2)^{-\frac{n}{2}-a-\frac{3}{2}}\exp\{-\frac{n}{2\sigma^2}(\mu-\overline{X})^2-\frac{n}{2\sigma^2}S^2-\frac{b(\mu-a)^2}{2\sigma^2}-\frac{d}{\sigma^2}\}.$$

Il suffit maintenant de regrouper les termes en μ en un seul trinôme,

$$\begin{aligned} n(\mu-\overline{X})^2+b(\mu-a)^2 &= (n+b)(\mu-\frac{n\overline{X}+ab}{n+b})^2+n\overline{X}^2+a^2b-\frac{(n\overline{X}+ab)^2}{n+b} \\ &= (n+b)(\mu-\frac{n\overline{X}+ab}{n+b})^2+\frac{nb}{n+b}(\overline{X}-a)^2. \end{aligned}$$

On en déduit la formule annoncée.

Dans la pratique, un a priori souvent utilisé est

$$d\Pi^*(\mu,\sigma^2)=\frac{1}{\sigma^2}d\mu d\sigma^2.$$

Il s'agit d'un a priori (doublement) impropre : $\int\int d\Pi^*(\mu,\sigma^2) = +\infty$ et chaque intégrale simple vaut déjà $+\infty$. Cet a priori rend les formules nettement plus simples. Il correspond en effet au cas limite $a = 0, b \to 0, c = 0, d = 0$ et la loi a posteriori est

$$\Pi^*[\cdot \,|\, X] \sim NIG\left(\overline{X}, n, \frac{n}{2}, \frac{nS^2}{2}\right).$$

Ce qu'il faut retenir :

- La loi a priori conjuguée standard pour σ^2 inconnu dans un modèle gaussien est une loi inverse gamma.
- A plusieurs dimensions, les lois a priori conjuguées ne sont pas toujours des produits de lois indépendantes.

B *Modèle $\mathcal{N}(\mu, \Sigma)$ gaussien multivarié, Σ connue*

Un autre cas important est celui de lois gaussiennes multidimensionnelles, où chaque observation est dans $\mathbb{R}^d$, $d \geq 1$. Nous traitons le cadre où la matrice de variance-covariance Σ est connue. Il est possible de l'étendre au cas où Σ est inconnue en suivant des idées similaires à celles vues en A.

Théorème 2. Soit $X = (X_1, \ldots, X_n)$ avec $X_i \in \mathbb{R}^d, d \geq 1$. Soit μ_1 un réel fixé et Σ, Σ_1 deux matrices symétrique définies positives fixées. Considérons le cadre bayésien

$$\begin{aligned} X \,|\, \mu &\sim \mathcal{N}(\mu, \Sigma)^{\otimes n} \\ \mu &\sim \mathcal{N}(\mu_1, \Sigma_1) = \Pi. \end{aligned}$$

La famille $\{\mathcal{N}(\mu_1, \Sigma_1),\ \mu_1 \in \mathbb{R},\ \Sigma_1 > 0\}$ est conjuguée et

$$\Pi[\cdot \,|\, X] = \mathcal{N}(\mu_X, \Sigma_X),$$

où l'on a posé

$$\begin{aligned} \Sigma_X &= (\Sigma_1^{-1} + n\Sigma^{-1})^{-1} \\ \mu_X &= \Sigma_X(\Sigma_1^{-1}\mu_1 + n\Sigma^{-1}\overline{X}). \end{aligned}$$

Preuve.

Pour simplifier les notations, nous traitons ici le cas d'une observation $n = 1$ et de $\mu_1 = 0$, soit $X \,|\, \mu \sim \mathcal{N}(\mu, \Sigma)$ et $\mu \sim \Pi = \mathcal{N}(0, \Sigma_1)$. La formule de Bayes donne

$$f_{\mu \,|\, X}(\mu) \propto \exp\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) - \frac{1}{2}\mu^T\Sigma_1^{-1}\mu\}.$$

Il suffit de regrouper les termes en μ pour former une nouvelle forme quadratique de la forme, dans l'exponentielle, de $-\frac{1}{2}(\mu - v)^T\Sigma_X^{-1}(\mu - v)$. On cherche donc v, Σ_X tels que, en

développant chaque forme quadratique

$$\mu^T(\Sigma^{-1} + \Sigma_1^{-1})\mu - 2X^T\Sigma^{-1}\mu = \mu^T\Sigma_X^{-1}\mu - 2v^T\Sigma_X^{-1}\mu.$$

En identifiant, il suffit donc de poser

$$\Sigma_X = (\Sigma^{-1} + \Sigma_1^{-1})^{-1}$$

et de choisir v tel que

$$2v^T\Sigma_X^{-1} = 2X^T\Sigma^{-1},$$

ce qui donne $v = \Sigma_X\Sigma^{-1}X$ comme annoncé (cas $n = 1, \mu_1 = 0$).

Exercice. Etendre la preuve au cas $n > 1$ et μ_1 quelconque.

Remarque. Le théorème 2 peut se voir comme un résultat de conditionnement sur les vecteurs gaussiens. Les lois de $X \,|\, \mu$ et de μ sont gaussiennes, donc la loi jointe de (X, μ) aussi, ainsi que la loi conditionnelle de $\mu \,|\, X$.

4.3 Lois invariantes : a priori de Jeffreys

Dans le but de trouver une loi a priori qui serait "universelle", Jeffreys (1946) propose de chercher Π qui soit invariant par changement de paramétrisation du problème $\eta = g(\theta)$, où η désigne le nouveau paramètre et θ le paramètre d'origine.

Exemple. Soit $\mathcal{P} = \{\text{Be}(\theta),\ \theta \in (0,1)\}$. Une loi a priori classique que nous avons rencontrée pour ce problème est $\Pi = \text{Unif}[0,1]$. Cette loi n'est pas invariante par reparamétrisation.

Définissons un nouveau paramètre η comme $\eta = \sqrt{\theta}$. On a $\theta = \eta^2$ soit $d\theta = 2\eta d\eta$. Ainsi $d\Pi(\theta) = d\theta = 2\eta d\eta$. La loi a priori en η a donc une densité 2η par rapport à la mesure de Lebesgue sur $[0,1]$. Ce n'est donc pas une loi uniforme : c'est une loi Beta$(2,1)$.

Notion d'information de Fisher. Soit $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ avec $dP_\theta = p_\theta d\mu$. On pose $\ell_\theta = \log p_\theta$.

- Cas de la dimension 1 : $\Theta \subset \mathbb{R}$. Si cette quantité existe, l'information de Fisher au point θ pour le modèle $\mathcal{P}$ est

$$I(\theta) = E_\theta[\ell_\theta'^{\,2}(X)], \qquad \text{avec } \ell_\theta'(X) = \frac{\partial}{\partial\theta}\ell_\theta(X) = \frac{p_\theta'}{p_\theta}(X).$$

- Cas de la dimension $d \geq 1 : \Theta \subset \mathbb{R}^d$. Si cette quantité existe, la matrice d'information de Fisher au point θ pour le modèle $\mathcal{P}$ est

$$I(\theta) = E_\theta[\nabla\ell_\theta(X)\nabla\ell_\theta(X)^T],$$

où $\nabla\ell_\theta$ est le vecteur gradient de $\theta \to \ell_\theta(X)$.

Intuition.

$I(\theta)$ correspond à la "quantité d'information" disponible pour le paramètre θ. Pour $\hat{\theta} = \hat{\theta}^{MV}$ estimateur du maximum de vraisemblance, pour un modèle *régulier*,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)^{-1})$$

sous P_θ. Plus $I(\theta)$ est grande, plus la variance asymptotique du maximum de vraisemblance est petit, et plus le modèle est "informatif" au point θ.

Exemple. Considérons le modèle $\mathcal{P} = \{\text{Be}(\theta),\ \theta \in (0,1)\}$.

La vraisemblance s'écrit $p_\theta(X) = \theta^X(1-\theta)^{1-X}$. On en déduit

$$\ell_\theta(X) = \frac{X}{\theta} - \frac{1-X}{1-\theta},$$

puis, en utilisant $X(1-X) = 0$, puisque X vaut 0 ou bien 1,

$$I(\theta) = E_\theta\left[\frac{X^{2\,2}}{\theta} + \frac{(1-X)^2}{(1-\theta)^2}\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

On remarque par ailleurs que l'estimateur du maximum de vraisemblance est ici $\hat{\theta}^{MV}(X) = \overline{X}$ et que sous P_θ, quand $n \to \infty$,

$$\sqrt{n}(\hat{\theta}^{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{Var}_\theta(X_1)) = \mathcal{N}(0, I(\theta)^{-1}).$$

Dans la définition qui suit, on suppose que la mesure dominante ν sur Θ est la (restriction à Θ de la) mesure de Lebesgue sur $\mathbb{R}$, respectivement $\mathbb{R}^d, d \geq 1$.

Definition 2. Pour $\Theta \subset \mathbb{R}$, l'a priori de Jeffreys est la mesure sur Θ de densité $\pi(\theta)$ par rapport à $\nu = \text{Leb}_{|\Theta}$ proportionnelle à $\sqrt{I(\theta)}$, avec

$$\pi(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}d\nu(\theta)} \qquad \text{si l'a priori est propre, i.e. } \int \sqrt{I(\theta)}d\nu(\theta) < +\infty.$$

$$\pi(\theta) = \sqrt{I(\theta)} \qquad \text{si l'a priori est impropre, i.e. } \int \sqrt{I(\theta)}d\nu(\theta) = +\infty.$$

Plus généralement, si $\Theta \subset \mathbb{R}^d$, $d \geq 1$, l'a priori de Jeffreys a une densité par rapport à $\nu = \text{Leb}_{|\Theta}$ proportionnelle à $\sqrt{\det(I(\theta))}$.

Exemples.

1. $\mathcal{P} = \{\text{Be}(\theta),\ \theta \in (0,1)\}$. La loi a priori de Jeffreys est celle dont la densité $\pi(\theta)$ est proportionnelle à, d'après le calcul de l'information de Fisher ci-dessus,

$$\pi(\theta) \propto \sqrt{I(\theta)} = \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.$$

On reconnait la densité d'une loi Beta$(\frac{1}{2}, \frac{1}{2})$.

2. $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$. Ici $I(\theta) = 1$ pour tout $\theta \in \mathbb{R}$. Comme $\int_{\mathbb{R}} 1 d\theta = +\infty$, l'a priori de Jeffreys est celui dont la densité $\pi(\theta)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est 1. Donc Π coïncide avec la mesure de Lebesgue sur $\mathbb{R}$. Il s'agit d'un a priori impropre.

Proposition 1. L'a priori de Jeffreys est invariant par re-paramétrisation lisse du modèle statistique. Plus précisément, si Π est l'a priori de Jeffreys dans le modèle paramétré par θ, si

$$\eta = g(\theta), \qquad g \text{ difféomorphisme},$$

et si $\Pi^* = \Pi \circ g^{-1}$ est la mesure image de Π par g, alors Π^* est l'a priori de Jeffreys dans le modèle paramétré par η.

Preuve.

Nous nous limiterons au cas $d = 1$. Commençons par écrire la log-vraisemblance dans le modèle de deux façons différentes suivant la paramétrisation

$$\ell_\theta = \log p_\theta = \log p_{g^{-1}(\eta)} \qquad \text{et} \qquad \lambda_\eta = \log \tilde{p}_\eta.$$

On a $\tilde{p}_\eta = p_{g^{-1}(\eta)}$ puisque la log-vraisemblance est calculée au même point. L'information de Fisher $I^*(\eta)$ dans le modèle paramétré par η s'exprime en fonction de celle $I(\theta)$ dans le modèle paramétré par θ. En effet, $p'_\theta = (\tilde{p}_{g(\theta)})' = g'(\theta)\tilde{p}'(g(\theta))$, donc

$$I(\theta) = \int \frac{{p'_\theta}^2}{p_\theta} d\mu = g'(\theta)^2 \int \frac{{\tilde{p}'_{g(\theta)}}^2}{\tilde{p}_{g(\theta)}} d\mu = g'(\theta)^2 I^*(g(\theta)).$$

Pour toute fonction mesurable bornée h, on a $\int h(g(\theta)) d\Pi(\theta) = \int h(\eta) d\Pi^*(\eta)$ (propriété de la mesure image). Il suffit de montrer que $\int h(\eta) d\Pi^*(\eta) = \int h(\eta) d\tilde{\Pi}(\eta)$, où $\tilde{\Pi}$ est l'a priori de Jeffreys dans le modèle paramétré par η. Pour cela,

$$\begin{aligned}
\int h(g(\theta))\pi(\theta) d\theta &= \int h(\eta)\pi(g^{-1}(\eta)) \frac{d\eta}{g'(g^{-1}(\eta))} \\
&= \int h(\eta) \frac{C\sqrt{I(g^{-1}(\eta))}}{g'(g^{-1}(\eta))} d\eta \\
&= \int h(\eta) C \sqrt{I^*(\eta)} d\eta = \int h(\eta) d\tilde{\Pi}(\eta),
\end{aligned}$$

où C est la constante de normalisation éventuelle dans le cas où Π est une loi a priori non impropre, ce qu'il fallait démontrer.

Exercice. Vérifier directement par le calcul que $\Pi = \text{Beta}(\frac{1}{2}, \frac{1}{2})$ est invariant dans le modèle $\mathcal{P} = \{\text{Be}(\theta),\ \theta \in (0, 1)\}$.

Remarques. 1. Parfois une paramétrisation donnée, cela dit, sera plus naturelle pour un problème donné, donc ce type d'invariance n'est pas toujours indispensable.

2. L' a priori de Jeffreys conduit parfois à des a priori impropres, comme c'est le cas pour le modèle fondamental ci-dessus.

4.4 L'approche bayésienne empirique ou *empirical Bayes*

Pour déterminer une loi a priori pour un problème donné, une approche très utilisée en pratique est la suivante

a) on se restreint à une classe de lois a priori $\{\Pi_\alpha\}$ plus ou moins grande, par exemple

   (a) toutes les lois $\{\mathcal{N}(a, \sigma^2),\ a \in \mathbb{R},\ \sigma^2 > 0\}$

   (b) toutes les lois $\{\mathcal{E}(\lambda),\ \lambda > 0\}$

   (c) toutes les lois $\{dP(x) = g(x)dx,\ g \text{ densité sur } [0,1]\}$.

b) on "estime" (voir ci-dessous) les paramètres de la loi a priori, disons α (dans les exemples $\alpha = (a, \sigma^2)$, $\alpha = \lambda$ et $\alpha = g$ respectivement)

c) on mène l'inférence bayésienne avec la loi a priori $\Pi_{\hat{\alpha}}$, ce qui résulte en une loi a posteriori $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$.

Comment faire l'étape b) d'estimation du paramètre ?

On suppose le cadre dominé avec les notations suivantes $d\Pi_\alpha(\theta) = \pi_\alpha d\nu(\theta)$ et $dP_\theta = p_\theta d\mu$.

Idée. Pour "estimer" α, l'idéal serait de pouvoir former une vraisemblance en α. Cela est possible en intégrant la vraisemblance usuelle par rapport à θ.

Méthode du maximum de vraisemblance marginal. La loi marginale de X dans le cadre bayésien

$$\begin{aligned} \theta &\sim \Pi_\alpha \\ X \,|\, \theta &\sim P_\theta \end{aligned}$$

a pour densité par rapport à μ

$$f_X(x) = \int p_{X \,|\, \theta}(x) d\Pi_\alpha(\theta) =: \varphi_\alpha(x).$$

On note $\varphi_\alpha(\cdot)$ cette dernière densité et on pose

$$\hat{\alpha}(X) = \operatorname*{argmax}_\alpha \varphi_\alpha(X).$$

Le principe est de marginaliser par rapport à la variable inconnue θ pour avoir une vraisemblance qui ne dépend que de α. On peut aussi écrire, en interprétant α d'un point de vue purement bayésien comme une variable pour laquelle on n'aurait pas encore spécifié de loi,

$$\begin{aligned} \theta \,|\, \alpha &\sim \Pi_\alpha \\ X \,|\, \theta = X \,|\, \theta, \alpha &\sim P_\theta \end{aligned}$$

d'où l'on déduit que $X \,|\, \alpha$ est la loi de densité $\int p_\theta(x)\pi_\alpha(\theta)d\nu(\theta)$.

Exemples.

1. Modèle fondamental

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

Déterminons un choix de loi a priori Π_μ par méthode bayésienne empirique.

La loi marginale de X_1 a pour densité

$$\begin{aligned} f_{X_1}(x) &\propto \int e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}(\theta-\mu)^2} d\theta \\ &\propto e^{-\frac{x^2}{2}} \int e^{-(\theta - \frac{x+\mu}{2})^2 + \frac{(x+\mu)^2}{4}} d\theta \\ &\propto e^{-\frac{x^2}{4} + \frac{\mu x}{2}} \propto e^{-\frac{1}{2}(x-\mu)^2}. \end{aligned}$$

Ainsi la loi marginale de X_1 (sachant μ) est une $\mathcal{N}(\mu, 1)$. Donc

$$\hat{\mu}(X_1) = \operatorname*{argmax}_{\mu \in \mathbb{R}} e^{-\frac{1}{2}(X_1-\mu)^2} = X_1.$$

On estime dont μ par l'observation X_1 et la loi a priori est $\Pi_{\hat{\mu}} = \mathcal{N}(X_1, 1)$. Plus généralement pour n observations nous verrons en TD que

$$\hat{\mu}(X_1, \dots, X_n) = \overline{X}.$$

La loi a priori par méthode bayésienne empirique (du maximum de vraisemblance marginal) est donc $\Pi_{\hat{\mu}} = \mathcal{N}(\overline{X}, 1)$. La loi a posteriori correspondante est $\Pi_{\hat{\mu}}[\cdot \,|\, X] = \mathcal{N}(\overline{X}, \frac{1}{n+1})$. On remarque que celle-ci est centrée exactement en $\overline{X}$.

2. Modèle de Poisson

$$\begin{aligned} X_1, \dots, X_n \,|\, \theta &\sim \mathcal{P}(\theta)^{\otimes n} \\ \theta &\sim \mathcal{E}(\lambda) = \Pi_\lambda. \end{aligned}$$

Dans ce cadre, on a $\hat{\lambda}^{EB}(X) = \overline{X}$, voir TDs, donc $\Pi_{\hat{\lambda}^{EB}(X)} = \mathcal{E}(\overline{X})$ et la loi a posteriori obtenue par la méthode bayésienne empirique ci-dessus est une loi Gamma$(n\overline{X} + 1, \frac{n\overline{X}+1}{\overline{X}})$. On note qu'à nouveau cette loi est centrée exactement en $\overline{X}$.

3. Exemple historique de Robbins (1955)

$$\begin{aligned} X_1, \dots, X_n \,|\, \theta_1 \dots, \theta_n &\sim \bigotimes_{i=1}^{n} \mathcal{P}(\theta_i) \\ (\theta_1, \dots, \theta_n) &\sim P_g^{\otimes n}, \end{aligned}$$

où P_g désigne une loi de densité (inconnue) g sur $\mathbb{R}^+$. Cet exemple est plus délicat : il s'agit d'un cadre 'non-paramétrique' car la quantité inconnue est une fonction, la densité g.

4.5 L'approche bayésienne hiérarchique ou *hierarchical Bayes*

Dans le même cadre qu'à la section précédente, si l'on dispose d'une famille de lois a priori $\{\Pi_\alpha,\ \alpha \in \mathcal{A}\}$, plutôt que d'estimer α par une méthode bayésienne empirique comme précédemment, on peut être plutôt complètement bayésien. Pour cela, il suffit de prendre α lui-même aléatoire et de choisir une loi a priori sur α !

Supposons que pour tout $\alpha \in \mathcal{A}$, on a $d\Pi_\alpha(\theta) = \pi_\alpha(\theta)d\nu(\theta)$, avec ν mesure σ-finie sur Θ et $dQ(\alpha) = q(\alpha)d\nu'(\alpha)$, avec ν' mesure σ-finie sur $\mathcal{A}$ (on suppose aussi une condition de mesurabilité comme au Chapitre 2). La mise en oeuvre de l'idée précédente, du point de vue de la loi a priori, s'écrit

$$\begin{aligned} \theta \,|\, \alpha &\sim \Pi_\alpha \\ \alpha &\sim Q. \end{aligned}$$

Ainsi, la loi marginale de θ s'interprète comme une loi *mélange*, dont la densité par rapport à ν est

$$f_\theta(\theta) = \int \pi_\alpha(\theta) dQ(\alpha).$$

Nous pouvons remarquer qu'il s'agit tout simplement de l'approche bayésienne habituelle, pour laquelle la densité de la loi a priori Π sur θ prend ici la forme du mélange ci-dessus, où Q est la loi mélangeante.

Exemples.

1. Considérons un exemple de tirage de pile ou face, où l'on soupçonne que soit les pièces sont équilibrées, soit elles sont biaisées avec probabilité 1/3 de tirer pile. On peut proposer le cadre suivant avec une loi a priori de type ci-dessus

$$\begin{aligned} X_1, \dots, X_n \,|\, \theta &\sim \mathrm{Be}(\theta)^{\otimes n} = P_\theta^{\otimes n} \\ \theta \,|\, \alpha &\sim \mathrm{Beta}(6-\alpha, 6+\alpha) = \Pi_\alpha \\ \alpha &\sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2 = Q. \end{aligned}$$

   On constate que la loi a priori Π induite sur θ n'est autre que la loi Π de densité $\pi(\theta)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par

$$\pi(\theta) = \frac{1}{2}q_0 + \frac{1}{2}q_1,$$

   où q_0 et q_1 sont les densités respectives des lois $\mathrm{Beta}(6,6)$ et $\mathrm{Beta}(4,8)$.

2. Certaines lois classiques s'expriment déjà naturellement sous forme de mélange, comme la loi de Laplace $\mathrm{Lap}(\lambda)$. Si $Y \sim \mathrm{Lap}(\lambda)$, on peut vérifier que $\mathcal{L}(Y) = \mathcal{L}(Z)$ avec la loi de Z donnée par le schéma

$$\begin{aligned} Y \,|\, \tau &\sim \mathcal{N}(0, \tau) \\ \tau &\sim \mathcal{E}(\frac{\lambda^2}{2}). \end{aligned}$$

CHAPITRE 5

Convergence de lois a posteriori

Nous voyons dans ce chapitre qu'il est possible d'étudier les lois a posteriori bayésiennes d'un point de vue fréquentiste. Nous définissons les notions de consistance et de convergence de ces lois dans un cadre asymptotique où le nombre d'observations tend vers l'infini. Ensuite, nous considérons la question de la forme limite des lois a posteriori et énonçons le théorème de Bernstein–von Mises. Nous en voyons des conséquences importantes, notamment pour la construction de régions de confiance.

Le tableau suivant présente certains modèles rencontrés précédemment avec lois a priori Π, et les expressions explicites de la loi a posteriori $\Pi[\cdot \,|\, X]$ et de la moyenne a posteriori $\overline{\theta}(X)$.

Modèle $\mathcal{P}$	A priori Π	A posteriori $\Pi[\cdot \,\vert\, X]$	$\overline{\theta}(X)$
$\{\mathcal{N}(\theta,1)^{\otimes n},\ \theta\in\mathbb{R}\}$	$\mathcal{N}(a,1)$	$\mathcal{N}(\frac{a+n\overline{X}}{n+1},\frac{1}{n+1})$	$\frac{a+n\overline{X}}{n+1}$
$\{\mathrm{Be}(\theta)^{\otimes n},\ \theta\in(0,1)\}$	$\mathrm{Beta}(a,b)$	$\mathrm{Beta}(a+n\overline{X},b+n-n\overline{X})$	$\frac{a+n\overline{X}}{a+b+n}$
$\{\mathrm{Poisson}(\theta)^{\otimes n},\ \theta>0\}$	$\mathcal{E}(a)$	$\mathrm{Gamma}(1+n\overline{X},n+a)$	$\frac{1+n\overline{X}}{n+a}$
$\{\mathcal{E}(\theta)^{\otimes n},\ \theta>0\}$	$\mathrm{Gamma}(a,b)$	$\mathrm{Gamma}(n+a,b+n\overline{X})$	$\frac{n+a}{b+n\overline{X}}$

La lecture des deux dernières colonnes du tableau suggère une proximité frappante de la loi a posteriori avec l'estimateur classique des moments $\overline{X}$ ($1/\overline{X}$ dans le modèle exponentiel, puisque $E[\mathcal{E}(\theta)] = 1/\theta$). En effet, dans chacun des exemples ci-dessus,

Que se passerait-il si l'on oubliait le cadre bayésien utilisé pour former la loi a posteriori et que l'on étudiait $\Pi[\cdot \,|\, X]$ *en probabilité sous* $P_{\theta_0}^{\otimes n}$*, c'est-à-dire que l'on supposait*

$$X_1,\dots,X_n \text{ i.i.d. } \sim P_{\theta_0},$$

soit l'hypothèse fréquentiste qu'il existe une vraie valeur θ_0 *du paramètre ?*

En particulier, il serait peut-être possible d'utiliser la loi a posteriori $\Pi[\cdot \,|\, X]$ ou un de ses aspects comme estimateur de θ_0. Par exemple, dans chacun des exemples ci-dessus, on déduit du fait que $\overline{X} \to \theta_0$ en probabilité sous P_{θ_0} que

$$\overline{\theta}(X) \to^{P_{\theta_0}} \theta_0.$$

De plus, on peut également vérifier dans chaque exemple que la variance a posteriori tend vers 0 (le faire en exercice). Cela devrait signifier que, sous P_{θ_0}, la masse a posteriori se concentre autour de θ_0. Nous allons préciser ceci ci-dessous. Enfin, peut-on dire quelque chose du niveau de confiance asymptotique des régions de crédibilité ?

Remarque technique sur le dénominateur de la formule de Bayes. Dans le cadre bayésien, si l'on pose $D(X) = \int \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta)$, on peut noter que $D(x)$ n'est autre que la densité marginale de X. Ainsi, si $A = \{D(X) = 0\}$, et P désigne la probabilité sous la loi jointe bayésienne de (X, θ),

$$P[A] = E[\mathbb{1}_{D(X)=0}] = \int \mathbb{1}_{D(x)=0} D(x) d\mu(x) = 0.$$

Ceci montre que le dénominateur de la formule de Bayes est non nul, P-presque sûrement. En revanche, rien n'interdit qu'il soit nul avec probabilité non nulle sous P_{θ_0}. Pour l'étude fréquentiste de l'a posteriori $\Pi[\cdot \,|\, X]$, nous supposerons que le dénominateur de la formule de Bayes est non-nul P_{θ_0}-presque sûrement, soit

$$P_{\theta_0}^{\otimes n}\left[\int \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta) > 0\right] = 1,$$

de sorte que la formule de Bayes est bien définie P_{θ_0}-presque sûrement. C'est de toute façon le cas dans une très grande majorité de modèles.

5.1 Consistance de lois a posteriori

Cadre. Dans toute la suite de ce chapitre, on considère le cadre d'un modèle $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$, $d \geq 1$. On munit Θ d'une loi a priori Π et, pour former la loi a posteriori $\Pi[\cdot \,|\, X]$, on suppose $X = (X_1, \ldots, X_n) \,|\, \theta$ de loi $P_\theta^{\otimes n}$. Une fois $\Pi[\cdot \,|\, X]$ formée, on l'étudie sous l'*hypothèse fréquentiste*

$$X = (X_1, \ldots, X_n) \ \sim P_{\theta_0}^{\otimes n}.$$

Comme nous nous limiterons ici au cas i.i.d., nous écrirons simplement pour simplifier dans la suite 'sous P_{θ_0}' au lieu de 'sous $P_{\theta_0}^{\otimes n}$'.

Definition 1. On dit que $\Pi[\cdot \,|\, X] = \Pi[\cdot \,|\, X_1, \ldots, X_n]$ est consistant au point $\theta_0 \in \Theta$ si, pour tout $\varepsilon > 0$,

$$\Pi\left[\{\theta :\ \|\theta - \theta_0\| \leq \varepsilon\} \,|\, X_1, \ldots, X_n\right] \to 1,$$

en probabilité sous P_{θ_0}.

✍ Pour Z_n une variable aléatoire telle que $0 \leq Z_n \leq 1$, on a

$$Z_n \xrightarrow{P} 0 \quad \Leftrightarrow \quad E[Z_n] \to 0 \quad (n \to \infty),$$

et de même $Z_n \to 1$ en probabilité si $E[Z_n] \to 1$ (exercice).

En particulier, pour montrer que l'a posteriori est consistant, il suffit de montrer que

$$E_{\theta_0} \Pi \left[\{\theta : \ \|\theta - \theta_0\| \leq \varepsilon\} \,|\, X \right] \to 0 \quad (n \to \infty),$$

ou le même résultat avec l'événement complémentaire et l'espérance qui tend vers 1.

Exemple d'a posteriori non consistant. Soit $\mathcal{P} = \{\mathcal{N}(\theta, 1), \ \theta \in \mathbb{R}\}$ et considérons l'a priori $\Pi = \text{Unif}[0, 1]$. La densité a posteriori est proportionnelle à $\exp\{-\sum_{i=1}^n (X_i - \theta)^2/2\} \mathbb{1}_{[0,1]}(\theta)$. En particulier la densité a posteriori est nulle à l'extérieur de $[0, 1]$. En particulier $\Pi[\cdot \,|\, X]$ est inconsistant en $\theta_0 = 2$ puisque

$$\Pi[[3/2, 5/2] \,|\, X] = 0.$$

Quelques exemples de consistance. Nous verrons par la suite un résultat général (le théorème BvM) qui implique la consistance, mais on peut retenir que typiquement il suffit que la loi a priori mette une masse strictement positive sur tout voisinage de θ_0 pour avoir consistance. Voyons maintenant deux exemples en détail.

Proposition 1. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \ \theta \in \mathbb{R}\}$ avec une loi a priori $\Pi = \mathcal{N}(a, 1)$ sur θ, la loi a posteriori $\Pi[\cdot \,|\, X]$ est consistante en tout point $\theta_0 \in \mathbb{R}$.

Preuve.

La loi a posteriori est une $\mathcal{N}(\overline{\theta}, \frac{1}{n+1})$, avec $\overline{\theta} = (a + n\overline{X})/(n + 1)$. En utilisant $\overline{X} \to \theta_0$ sous P_{θ_0} par la loi des grands nombres, on a $\overline{\theta} \to \theta_0$ sous P_{θ_0}. Pour tout θ_0 réel et $\varepsilon > 0$,

$$\begin{aligned}
\Pi[\,|\theta - \theta_0| > \varepsilon \,|\, X\,] &= \Pi[\,|\theta - \theta_0| > \varepsilon \,|\, X\,] \mathbb{1}_{|\overline{\theta} - \theta_0| < \frac{\varepsilon}{2}} + \Pi[\,|\theta - \theta_0| > \varepsilon \,|\, X\,] \mathbb{1}_{|\overline{\theta} - \theta_0| > \frac{\varepsilon}{2}} \\
&\leq \Pi\left[\,|\theta - \overline{\theta}| > \frac{\varepsilon}{2} \,|\, X\,\right] \mathbb{1}_{|\overline{\theta} - \theta_0| < \frac{\varepsilon}{2}} + \mathbb{1}_{|\overline{\theta} - \theta_0| > \frac{\varepsilon}{2}} \\
&\leq \Pi\left[\,|\theta - \overline{\theta}| > \frac{\varepsilon}{2} \,|\, X\,\right] + \mathbb{1}_{|\overline{\theta} - \theta_0| > \frac{\varepsilon}{2}} = (I) + (II),
\end{aligned}$$

où on a utilisé que $\Pi[A \,|\, X] \leq 1$ pour tout A mesurable. De plus,

$$E_{\theta_0}(II) = P_{\theta_0}[|\overline{\theta} - \theta_0| > \frac{\varepsilon}{2}] = o(1),$$

puisque $\overline{\theta}$ converge en proba vers θ_0. Enfin d'après l'expression explicite de la loi a posteriori,

$$\begin{aligned} E_{\theta_0}(I) &= E_{\theta_0} P\left[\left|\mathcal{N}(\overline{\theta}, \frac{1}{n+1}) - \overline{\theta}\right| > \frac{\varepsilon}{2} \,|\, X\right] \\ &= E_{\theta_0} P\left[|\mathcal{N}(0, \frac{1}{n+1})| > \frac{\varepsilon}{2}\right] \\ &= P\left[\frac{1}{\sqrt{n+1}}|\mathcal{N}(0,1)| > \frac{\varepsilon}{2}\right] = o(1), \end{aligned}$$

puisque le lemme de Slutsky donne $\frac{1}{\sqrt{n+1}}|\mathcal{N}(0,1)| \to 0$ en probabilité. Donc l'a posteriori est consistant en θ_0.

Consistance dans le cadre où Θ est discret et fini. Soit $\Theta = \{1, \dots, k\}$. Quitte à renommer les éléments de Θ, on peut toujours se ramener à ce cas. On considère le modèle

$$\mathcal{P} = \{P_1, \dots, P_k\} = \{P_\theta,\ \theta \in \Theta\}, \tag{6}$$

où P_j sont des mesures de probabilité sur un espace E. On suppose comme d'habitude que les P_j sont dominées : pour tout $j \in \{1, \dots, k\}$, on a $dP_j = p_j d\mu$ pour μ mesure dominante (ici on pourrait par exemple prendre $P_1 + \dots + P_k$). Soit Π une loi a priori sur Θ. Celle-ci est définie par la donnée de

$$\Pi[\theta = j] = \pi_j, \quad j = 1, \dots, k.$$

Proposition 2. Dans le cadre du modèle discret (6), supposons le modèle identifiable et soit Π une loi a priori sur Θ telle que $\pi_j > 0$ pour tout $j = 1, \dots, k$. Alors la loi a posteriori $\Pi[\cdot \,|\, X]$ est consistante en tout point $\theta_0 = l_0 \in \{1, \dots, k\}$.

Preuve.

Le vrai paramètre est $\theta_0 = l_0 \in \{1, \dots, k\}$, et il suffit de démontrer que $\Pi[\{l_0\} \,|\, X_1, \dots, X_n] \to 1$ en probabilité sous P_{l_0}.

1ère étape - formule de Bayes. La formule de Bayes s'écrit ici, Π étant une loi discrète, pour $B \subset \{1, \dots, k\}$,

$$\Pi[B \,|\, X] = \frac{\sum_{j \in B} \prod_{i=1}^n p_j(X_i)\pi_j}{\sum_{j=1}^k \prod_{i=1}^n p_j(X_i)\pi_j}.$$

Notons $\mathcal{V}(j, X) = \prod_{i=1}^n p_j(X_i)$, alors

$$\Pi[\{l_0\} \,|\, X] = \frac{\mathcal{V}(l_0, X)\pi_{l_0}}{\sum_{j=1}^k \mathcal{V}(j, X)\pi_j}.$$

On commence par minorer le dénominateur. Pour tout $j \neq l_0$, on a $\mathcal{V}(j,X) \leq \max_{j\neq l_0} \mathcal{V}(j,X)$. Comme $\sum_{j\neq l_0} \pi_j = 1 - \pi_{l_0}$, on en déduit

$$\begin{aligned}\Pi[\{l_0\} \,|\, X] &\geq \frac{\mathcal{V}(l_0,X)\pi_{l_0}}{\mathcal{V}(l_0,X)\pi_{l_0} + [\max_{j\neq l_0} \mathcal{V}(j,X)] \sum_{j\neq l_0} \pi_j} \\ &\geq \frac{\mathcal{V}(l_0,X)\pi_{l_0}}{\mathcal{V}(l_0,X)\pi_{l_0} + [\max_{j\neq l_0} \mathcal{V}(j,X)] \, (1-\pi_{l_0})}.\end{aligned}$$

2ème étape - comparaison des vraisemblances sous P_{l_0}.
L'inégalité de Markov avec la fonction $x \to \sqrt{x}$ pour une variable aléatoire Z donne $P[Z \geq t] \leq t^{-1/2} E[Z^{1/2}]$, pour tout $t > 0$. Ainsi

$$P_{l_0}\left[\mathcal{V}(j,X) \geq t\mathcal{V}(l_0,X)\right] \leq \frac{1}{\sqrt{t}} E_{l_0}\left[\sqrt{\frac{\mathcal{V}(j,X)}{\mathcal{V}(l_0,X)}}\right].$$

L'espérance dans cette dernière expression s'écrit

$$\begin{aligned}E_{l_0}\left[\sqrt{\frac{\mathcal{V}(j,X)}{\mathcal{V}(l_0,X)}}\right] &= \int \left[\frac{\prod_{i=1}^n f_j(x_i)}{\prod_{i=1}^n f_{l_0}(x_i)}\right]^{1/2} \prod_{i=1}^n f_{l_0}(x_i) d\mu(x_i) \\ &= \int \sqrt{\prod_{i=1}^n f_j(x_i) \prod_{i=1}^n f_{l_0}(x_i)} \prod_{i=1}^n d\mu(x_i) \\ &= \rho(P_j^{\otimes n}, P_{l_0}^{\otimes n}) = \rho(P_j, P_{l_0})^n \\ &\leq \left[\max_{j\neq l_0} \rho(P_j, P_{l_0})^n\right]^n := r^n,\end{aligned}$$

où l'on a utilisé la propriété de l'affinité de Hellinger ρ vue au chapitre 3 et où on définit r comme le maximum apparaissant dans la dernière expression. Le modèle étant identifiable, on a $\rho(P_j, P_{l_0}) < 1$ pour tout $j \neq l_0$ (sinon la distance de Hellinger entre les mesures serait nulle et elles seraient égales), donc $r < 1$.

3ème étape - minoration du dénominateur avec grande probabilité. Soit $\mathcal{A}$ l'événement

$$\mathcal{A} = \{\max_{j\neq l_0} \mathcal{V}(j,X) < r^n \mathcal{V}(l_0,X)\}.$$

La probabilité du complémentaire $\mathcal{A}^c$ de $\mathcal{A}$ est majorée par

$$\begin{aligned}P_{l_0}[\mathcal{A}^c] &= P_{l_0}\left[\exists k \in \{1,\dots,l\} \setminus \{l_0\},\ \mathcal{V}(j,X) \geq r^n \mathcal{V}(l_0,X)\right] \\ &\leq \sum_{j\neq l_0} P_{l_0}[\mathcal{V}(j,X) \geq r^n \mathcal{V}(l_0,X)] \\ &\leq \sum_{j\neq l_0} \frac{1}{\sqrt{r^n}} r^n = (k-1) r^{n/2}.\end{aligned}$$

Comme $r < 1$ et k est fixé, cette dernière quantité tend vers 0 quand $n \to \infty$. On en déduit

$$\begin{aligned}\Pi[\{l_0\} \,|\, X] \mathbb{1}_{\mathcal{A}} &\geq \frac{\pi_{l_0}\mathcal{V}(l_0,X)}{\pi_{l_0}\mathcal{V}(l_0,X) + r^n \mathcal{V}(l_0,X)(1-\pi_{l_0})} \mathbb{1}_{\mathcal{A}} \\ &\geq \frac{\pi_{l_0}}{\pi_{l_0} + r^n(1-\pi_{l_0})} \mathbb{1}_{\mathcal{A}}\end{aligned}$$

Par conséquent,

$$\begin{aligned} E_{l_0}\left[\Pi[\{l_0\}\,|\,X]\right] &= E_{l_0}\left[\Pi[\{l_0\}\,|\,X]\,\mathbb{1}_{\mathcal{A}} + \Pi[\{l_0\}\,|\,X]\mathbb{1}_{\mathcal{A}^c}\right] \\ &\geq \frac{\pi_{l_0}}{\pi_{l_0} + r^n(1-\pi_{l_0})} P_{l_0}[\mathcal{A}] = (1+o(1)), \end{aligned}$$

où on a utilisé que $r^n \to 1$ et $P[\mathcal{A}] = 1 - P[\mathcal{A}^c] = 1 + o(1)$, ce qui conclut la démonstration.

5.2 Vitesses de convergence

On peut étendre naturellement la notion de consistance en permettant à ε dans la définition de la consistance de varier, et typiquement de tendre vers 0, avec n.

Definition 2. On dit que l'a posteriori $\Pi[\cdot\,|\,X] = \Pi[\cdot\,|\,X_1,\dots,X_n]$ converge à vitesse ε_n au point $\theta_0 \in \Theta$ si,

$$\Pi\left[\{\theta :\ \|\theta - \theta_0\| \leq \varepsilon_n\}\,|\,X_1,\dots,X_n\right] \to 1,$$

en probabilité sous P_{θ_0}.

Proposition 3. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta,1)^{\otimes n},\ \theta \in \mathbb{R}\}$ avec une loi a priori $\Pi = \mathcal{N}(a,1)$ sur θ, la loi a posteriori $\Pi[\cdot\,|\,X]$ est converge en tout point $\theta_0 \in \mathbb{R}$, à vitesse

$$\varepsilon_n = \frac{M_n}{\sqrt{n}},$$

pour toute suite $M_n \to \infty$ quand $n \to \infty$.

Preuve.
Voir TD.

Dans les modèles paramétriques 'réguliers', la vitesse de convergence sera toujours celle de la proposition ci-dessus, donc essentiellement $1/\sqrt{n}$ à une suite arbitraire tendant vers l'infini près. Cela résulte du théorème vu à la section suivante.

5.3 Forme limite et théorème de Bernstein–von Mises

Nous allons énoncer un résultat de forme limite pour la loi a posteriori. Ce résultat peut être vu comme une sorte de théorème central limite, même si pour des objets beaucoup plus généraux qu'une moyenne empirique. Asymptotiquement, les loi a posteriori ressemblent typiquement à des lois gaussiennes, centrées en un estimateur 'optimal', et de variance une constante divisée par n. Pour montrer un tel résultat, nous est d'abord utile d'avoir une notion de 'proximité' pour deux lois.

Definition 3. Soient P, Q deux mesures de probabilité avec $dP = pd\mu$ et $dQ = qd\mu$. La distance L^1 entre P et Q, est $\|P - Q\|_1 = \int |p - q| d\mu$.

On peut vérifier que la définition ci-dessus ne dépend pas du choix de la mesure dominante, voir TD.

Exemple. Soit $P_n = \text{Unif}[0, 1 + \frac{1}{n}]$ et $P = \text{Unif}[0, 1]$. On calcule

$$\begin{aligned}\|P_n - P\|_1 &= \int_0^1 |\frac{1}{1+\frac{1}{n}} - 1| du + \int_1^{1+\frac{1}{n}} \frac{1}{1+\frac{1}{n}} du \\ &= \frac{2}{n+1} = o(1).\end{aligned}$$

Ainsi $\|P_n - P\|_1 \to 0$ quand $n \to \infty$.

Laplace, au début des années 1800, a remarqué et démontré que dans le modèle binomial $\{\text{Bin}(n, \theta),\ \theta \in (0, 1)\}$, avec une loi a priori uniforme sur θ (i.e. le modèle considéré par Bayes), la loi a posteriori est une loi $\text{Beta}(1 + X, 1 + n - X)$, et que cette loi ressemble étrangement à une loi $\mathcal{N}(\frac{X}{n}, \frac{1}{n})$. On notera que X/n se trouve être l'estimateur du maximum de vraisemblance dans ce modèle. Depuis, de nombreux statisticiens se sont intéressés à ce phénomène, parmi lesquels Bernstein, von Mises, Le Cam, ...

Théorème 1. [Bernstein-von Mises (BvM)] Soit $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta \subset \mathbb{R}^d\}$ un modèle régulier. Soit $\theta_0 \in \Theta$. On suppose que la loi a priori Π sur Θ vérifie

- Π a une densité π par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$.
- $\pi(\theta_0) > 0$ et $\pi(\cdot)$ continue au point θ_0.

On suppose l'information de Fisher $I(\theta_0)$ au point θ_0 inversible. Soit $\hat{\theta}^{MV}$ l'estimateur du maximum de vraisemblance dans ce modèle. Alors quand $n \to \infty$,

$$\left\| \Pi[\cdot \mid X] - \mathcal{N}\left(\hat{\theta}^{MV}, \frac{I(\theta_0)^{-1}}{n}\right)(\cdot) \right\|_1 \to 0,$$

en probabilité sous P_{θ_0}.

Ce résultat implique une 'dualité' remarquable entre lois limites fréquentistes et lois limites bayésiennes. En effet, le théorème BvM donne

$$\mathcal{L}(\theta - \hat{\theta}^{MV} \mid X) \approx \mathcal{N}\left(0, \frac{I(\theta_0)^{-1}}{n}\right).$$

Par ailleurs, un des résultats fondamentaux sur le maximum de vraisemblance dans les modèles

réguliers est que

$$\mathcal{L}(\hat{\theta}^{MV} - \theta_0) \approx \mathcal{N}\left(0, \frac{I(\theta_0)^{-1}}{n}\right).$$

On note qu'il s'agit de la même loi limite. Ceci a des conséquences spectaculaires en termes de régions de crédibilité, voir plus loin.

Preuve.

Nous faisons la preuve dans le modèle fondamental pour une loi a priori gaussienne. Pour une preuve générale sous de jolies conditions, voir le livre *Asymptotic Statistics* de van der Vaart, Chapitre 10 (plutôt niveau M2/thèse). On pose donc $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ et $\Pi = \mathcal{N}(a, 1)$ pour $a \in \mathbb{R}$ fixé. Il s'agit de montrer, au vu de l'expression explicite de la loi a posteriori et du fait que $I(\theta) = 1$ pour tout θ dans le modèle fondamental, que pour $\theta_0 \in \mathbb{R}$ vraie valeur du paramètre, quand $n \to \infty$,

$$E_{\theta_0} \left\| \mathcal{N}\left(\frac{a + n\overline{X}}{n+1}, \frac{1}{n+1}\right) - \mathcal{N}\left(\overline{X}, \frac{1}{n}\right) \right\|_1 \to 0.$$

Il y a plusieurs preuves possibles. Celle ci-dessous repose sur une comparaison de distances et un calcul explicite (on peut aussi passer par les densités et une version en probabilité du lemme de Scheffé). Deux lemmes sont utilisés, voir ci-dessous pour leur énoncé et preuve. Notons $\overline{\theta} = \frac{a+n\overline{X}}{n+1}$ la moyenne a posteriori.

$$\begin{aligned}
&\left\| \mathcal{N}\left(\overline{\theta}, \frac{1}{n+1}\right) - \mathcal{N}\left(\overline{X}, \frac{1}{n}\right) \right\|_1^2 \\
&= \left\| \mathcal{N}\left(\overline{\theta} - \overline{X}, \frac{1}{n+1}\right) - \mathcal{N}\left(0, \frac{1}{n}\right) \right\|_1^2 \\
&= \left\| \mathcal{N}\left(\sqrt{n}(\overline{\theta} - \overline{X}), \frac{n}{n+1}\right) - \mathcal{N}(0,1) \right\|_1^2 \\
&\leq 2\left\{ \left\| \mathcal{N}\left(\sqrt{n}(\overline{\theta} - \overline{X}), \frac{n}{n+1}\right) - \mathcal{N}\left(0, \frac{n}{n+1}\right) \right\|_1^2 + \left\| \mathcal{N}\left(0, \frac{n}{n+1}\right) - \mathcal{N}(0,1) \right\|_1^2 \right\},
\end{aligned}$$

où à la dernière ligne on a utilisé l'inégalité $(a+b)^2 \leq 2a^2 + 2b^2$. On combine maintenant les lemmes 1 et 2 ci-dessous : on majore chaque distance L^1 au carré par la divergence de Kullback-Leibler correspondante, que l'on écrit ensuite explicitement. On obtient

$$\begin{aligned}
&\left\| \mathcal{N}\left(\overline{\theta}, \frac{1}{n+1}\right) - \mathcal{N}\left(\overline{X}, \frac{1}{n}\right) \right\|_1^2 \\
&\leq 8\left\{ \frac{1}{2}\log\frac{n}{n+1} + \frac{1}{2}\left[\frac{1 - \frac{n}{n+1}}{\frac{n}{n+1}} \right] + \frac{n(\overline{\theta} - \overline{X})^2}{2\frac{n}{n+1}} \right\} \\
&\leq 4\log(1 - \frac{1}{n+1}) + \frac{4}{n} + \frac{n+1}{2}\left(\frac{a - \overline{X}}{n+1}\right)^2 \\
&\leq -\frac{4}{n+1} + \frac{4}{n} + \frac{4}{n(n+1)}(\sqrt{n}(\overline{X} - a))^2 \leq \frac{4}{n(n+1)}\left[1 + (\sqrt{n}(\overline{X} - a))^2\right].
\end{aligned}$$

Or, l'inégalité $(a+b)^2 \leq 2a^2 + 2b^2$ implique

$$n(\overline{X} - a)^2 \leq 2n(\overline{X} - \theta_0)^2 + 2n(\theta_0 - a)^2,$$

et comme $\sqrt{n}(\overline{X} - \theta_0)$ est égal en loi à une $\mathcal{N}(0,1)$ (car on est dans le modèle fondamental), on en conclut par le lemme de Slutsky que

$$\left\| \mathcal{N}\left(\overline{\theta}, \frac{1}{n+1}\right) - \mathcal{N}\left(\overline{X}, \frac{1}{n}\right) \right\|_1^2 \leq \frac{8}{n(n+1)}[n(\overline{X} - \theta_0)^2] + \frac{4}{n(n+1)}(1 + 2n(\theta_0 - a))$$

quand $n \to \infty$, ce qu'il fallait démontrer.

Lemme 1. Soient P, Q deux mesures de probabilité avec $dP = pd\mu$ et $dQ = qd\mu$. Alors

$$\|P - Q\|_1 \leq 2\sqrt{KL(P,Q)},$$

avec $KL(P,Q) = \int \log(p/q) p d\mu$ la divergence de Kullback-Leibler entre P et Q.

Preuve.

Nous avons vu au chapitre 3 l'inégalité $\|P - Q\|_1 \leq 2h(P,Q)$. On relie alors $KL(P,Q)$ à $h(P,Q)$ comme suit

$$\begin{aligned} KL(P,Q) &= \int p \log \frac{p}{q} d\mu = 2 \int p \log \sqrt{\frac{p}{q}} d\mu \\ &= -2 \int p \log \sqrt{\frac{q}{p}} d\mu = -2 \int p \log(1 + \sqrt{\frac{q}{p}} - 1) d\mu \\ &\geq -2 \int p (\sqrt{\frac{q}{p}} - 1) d\mu = 2 - 2 \int \sqrt{pq} d\mu = h(P,Q)^2. \end{aligned}$$

On en déduit le résultat en combinant les deux inégalités.

Lemme 2. Soit KL la divergence de Kullback-Leibler entre deux lois de probabilité définie au Lemme 1. Pour tout $\mu \in \mathbb{R}$ et $\sigma^2 > 0$,

$$\begin{aligned} KL(\mathcal{N}(0,\sigma^2), \mathcal{N}(\mu,\sigma^2)) &= \frac{\mu^2}{2\sigma^2} \\ KL(\mathcal{N}(0,1), \mathcal{N}(0,\sigma^2)) &= \log \sigma + \frac{1-\sigma^2}{2\sigma^2}. \end{aligned}$$

Preuve.

Cela résulte d'un calcul direct à partir de la définition, laissé en exercice.

5.4 Confiance asymptotique des régions de crédibilité

On se place en dimension 1, soit $\Theta \subset \mathbb{R}$. On suppose que l'on a construit une loi a posteriori $\Pi[\cdot \,|\, X]$ dans le modèle $\mathcal{P}$ à partir d'une loi a priori Π et d'observations X. On considère la région de crédibilité $[a_n(X), b_n(X)]$ de niveau $1-\alpha$ formée par les quantiles de la loi a posteriori

$$\Pi[(-\infty, a_n(X)] \,|\, X] = \frac{\alpha}{2} \tag{7}$$

$$\Pi[(b_n(X), +\infty] \,|\, X] = \frac{\alpha}{2}. \tag{8}$$

Dans la suite on note $o_P(1)$ une quantité arbitraire qui tend vers 0 en probabilité sous $P_{\theta_0}^{\otimes n}$.

Théorème 2. Soit $\alpha > 0$ et $z_\alpha = q_{1-\alpha/2}^{\mathcal{N}(0,1)}$ le quantile de niveau $1-\alpha$ d'une loi normale standard. Supposons le théorème BvM vérifié. Alors, pour $a_n(X), b_n(X)$ définis par (7)-(8), quand $n \to \infty$,

$$[a_n(X), b_n(X)] = \left[\hat{\theta}^{MV} - \frac{z_\alpha}{\sqrt{nI(\theta_0)}}(1 + o_P(1)) \,,\, \hat{\theta}^{MV} + \frac{z_\alpha}{\sqrt{nI(\theta_0)}}(1 + o_P(1)) \right].$$

Ce résultat donne un développement asymptotique à l'ordre 1 des bornes de l'intervalle de crédibilité $[a_n(X), b_n(X)]$ défini à partir des quantiles de la loi a posteriori. Notons que cet intervalle coïncide asymptotiquement avec l'intervalle *de confiance* "idéal" que l'on voudrait pouvoir construire à partir de $\hat{\theta}^{MV}$. En effet, si on suppose les conditions réunies (modèle régulier) pour que $\hat{\theta}^{MV}$ de la convergence en loi soit "efficace" au sens où

$$\sqrt{n}(\hat{\theta}^{MV} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta_0)^{-1}),$$

on a que l'intervalle idéal suivant, avec $z_\alpha = q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}$,

$$I^*(X) = \left[\hat{\theta}^{MV} \pm \frac{z_\alpha}{\sqrt{nI(\theta_0)}} \right]$$

a un niveau de confiance asymptotique $1-\alpha$, puisque

$$P_{\theta_0}\left[\sqrt{n} |\hat{\theta}^{MV} - \theta_0| \leq z_\alpha \right] \to P[|\mathcal{N}(0,1)| \leq z_\alpha] = 1 - \alpha.$$

En général cependant, l'information de Fisher $I(\theta_0)$ est inconnue puisqu'elle dépend de θ_0. Une solution standard consiste à remplacer $I(\theta_0)$ par un estimateur, par exemple $I(\hat{\theta}^{MV})$ (sous les conditions de régularité habituelles, $\theta \to I(\theta)$ est continue, donc la consistance de θ^{MV} implique celle de $I(\hat{\theta}^{MV})$ vers $I(\theta_0)$).

Un des intérêts de l'approche bayésienne est que l'obtention de la région de crédibilité est *automatique* (pas besoin d'estimer $I(\theta_0)$, cela est fait 'automatiquement' par la procédure). De plus, nous allons voir ci-dessous que l'on peut l'utiliser comme région de confiance.

Preuve.

Soient A_1, A_2 les ensembles mesurables définis par

$$A_1 = (-\infty, A_n(X)], \qquad A_2 = (B_n(X), +\infty).$$

Par définition de A_n et B_n, on a

$$\Pi[A_1 \,|\, X] = \Pi[A_2 \,|\, X] = \frac{\alpha}{2}.$$

Le théorème BvM est vérifié par hypothèse et d'après le Lemme 3, on a donc

$$\sup_{A \in \mathcal{A}} \left| \Pi[A \,|\, X] - \mathcal{N}\left(\hat{\theta}^{MV}, \frac{I(\theta_0)^{-1}}{n}\right)(A) \right| = o_P(1).$$

En particulier, en appliquant ceci $A = A_1$, on en déduit que

$$\mathcal{N}\left(\hat{\theta}^{MV}, \frac{I(\theta_0)^{-1}}{n}\right)((-\infty, A_n(X)]) = \frac{\alpha}{2} + o_P(1).$$

En notant $P = P_X$ la loi de probabilité induite par la loi normale ci-dessus (à X fixé), alors

$$P\left[\mathcal{N}(0,1) \leq \sqrt{nI(\theta_0)}(A_n(X) - \hat{\theta}^{MV})\right] = \frac{\alpha}{2} + o_P(1).$$

On en déduit, en notant $\Phi(x) = P[\mathcal{N}(0,1) \leq x]$ la fonction de répartition de la loi normale standard et Φ^{-1} sa réciproque,

$$\sqrt{nI(\theta_0)}(A_n(X) - \hat{\theta}^{MV}) = \Phi^{-1}(\frac{\alpha}{2} + o_P(1)).$$

Or Φ^{-1} est continue, donc par théorème de l'image continue on en déduit que l'expression précédente converge en probabilité vers $\Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2) = -z_\alpha$. Par conséquent on obtient

$$A_n(X) = \hat{\theta}^{MV} - \frac{z_\alpha(1 + o_P(1))}{\sqrt{nI(\theta_0)}},$$

et le résultat pour $B_n(X)$ s'obtient de même.

Lemme 3. Soient P et Q deux mesures de probabilité sur $\mathcal{X}$ muni d'une tribu $\mathcal{A}$. Alors

$$\|P - Q\|_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|,$$

et le supremum dans l'expression s'appelle *distance en variation totale* entre P et Q.

Preuve.

Soit μ une mesure dominante, telle que $dP = pd\mu$ et $dQ = qd\mu$. Notons déjà que

$$\sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \sup_{A\in\mathcal{A}} P(A) - Q(A),$$

puisque pour toute valeur r de $P(A) - Q(A)$, la valeur $-r$ est atteinte si l'on remplace A par son complémentaire A^c. Puis on note que

$$\sup_{A\in\mathcal{A}} P(A) - Q(A) = \sup_{A\in\mathcal{A}} \int_A (p-q)d\mu = \int \mathbb{1}_{p>q}(p-q)d\mu.$$

En effet, comme $(p-q) \leq 0$ si $p \leq q$, on a

$$\int_A (p-q)d\mu \leq \int_{A\cap\{p>q\}} (p-q)d\mu \leq \int_{\{p>q\}} (p-q)d\mu.$$

De plus, en échangeant les rôles de p et de q par symétrie de l'argument en p et q, on a aussi

$$\sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \int \mathbb{1}_{p<q}(q-p)d\mu = \int \mathbb{1}_{p>q}(p-q)d\mu.$$

(on peut aussi le démontrer en notant que la différence vaut $\int(\mathbb{1}_{p<q} + \mathbb{1}_{p\geq q})qd\mu + \int(\mathbb{1}_{p<q} + \mathbb{1}_{p\geq q})pd\mu = 1 - 1 = 0$). On en conclut

$$2\sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \int \mathbb{1}_{p<q}(q-p)d\mu + \int \mathbb{1}_{p>q}(p-q)d\mu = \int |p-q|d\mu.$$

Théorème 3. [Confiance asymptotique des régions de crédibilité] Supposons le théorème BvM vérifié. Alors l'intervalle de crédibilité $I(X) = [a_n(X), b_n(X)]$ définit par (7)-(8) est un intervalle de confiance asymptotique au niveau $1-\alpha$, c'est-à-dire

$$P_{\theta_0}\left[\,\theta_0 \in [a_n(X), b_n(X)]\,\right] \to 1-\alpha \quad (n\to\infty).$$

Preuve.

Il suffit de montrer que $P_{\theta_0}\left[\,\theta_0 \in (-\infty, a_n(X)]\,\right] \to \alpha/2$ et que $P_{\theta_0}\left[\,\theta_0 \in (b_n(X), +\infty)\,\right] \to \alpha/2$. Pour cela, on utilise les développements asymptotiques obtenus au Theorème 2. Quand $n\to\infty$,

$$\begin{aligned}
P_{\theta_0}\left[\,\theta_0 \in (-\infty, a_n(X)]\,\right] &= P_{\theta_0}\left[\,\theta_0 < \hat{\theta}^{MV} - \frac{z_\alpha}{\sqrt{nI(\theta_0)}}(1+o_P(1))\,\right] \\
&= P_{\theta_0}\left[\,\sqrt{nI(\theta_0)}(\hat{\theta}^{MV} - \theta_0) > z_\alpha(1+o_P(1))\,\right] \\
&= P_{\theta_0}\left[\,\sqrt{nI(\theta_0)}(\hat{\theta}^{MV} - \theta_0) - z_\alpha o_P(1) > z_\alpha\,\right].
\end{aligned}$$

Comme la quantité à gauche du signe $>$ à la dernière ligne de l'expression ci-dessus converge en loi vers une variable $\mathcal{N}(0,1)$, on en déduit que l'expression converge vers $\alpha/2$. On fait de même pour l'intervalle $(b_n(X), +\infty)$, ce qui conclut la démonstration.

CHAPITRE 6

Tests bayésiens

Nous voyons dans ce chapitre quelques propriétés des tests d'un point de vue bayésien. Nous rappelons les définitions de base sur les tests, puis voyons comment construire un test bayésien à partir de la loi a posteriori. Nous donnons enfin brièvement quelques éléments de consistance asymptotique des tests bayésiens.

Dans le cadre d'un modèle $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, avec $dP_\theta = p_\theta d\mu$ et $\Theta \subset \mathbb{R}^d, d \geq 1$, l'objectif va maintenant être de *tester une propriété* du paramètre θ, c'est-à-dire que l'on voudrait savoir, à partir des données, si θ appartient à une certaine région de l'espace Θ_0 ou non. Plus précisément, dans la suite nous considèrerons deux *hypothèses*, celle que θ appartient à $\Theta_0 \subset \Theta$ et celle que θ appartient à $\Theta_1 \subset \Theta$. Dans la suite, on supposera

$$\Theta_0 \cap \Theta_1 = \emptyset.$$

Parfois, on pourra avoir $\Theta_1 = \Theta_0^c$, mais pas nécessairement.

6.1 Définitions

Types d'hypothèses.

L'hypothèse que θ appartient à Θ_0 s'appelle *hypothèse nulle*, l'hypothèse que θ appartient à Θ_1 s'appelle *hypothèse alternative*.

Une hypothèse réduite à un singleton, par exemple $\Theta = 0 = \{\theta_0\}$ ou $\Theta_1 = \theta_1\}$, est dite hypothèse *simple*. Sinon, on parle d'hypothèse *composite*.

Definition 1. Un test est une fonction mesurable $\phi(X_1, \ldots, X_n)$ des observations, à valeurs dans $\{0, 1\}$. Parfois, on autorise le test à prendre des valeurs dans l'intervalle $[0, 1]$, auquel cas on parle de test généralisé.

6.2 L'approche fréquentiste

Soit φ un test de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \theta_1$. Il y a deux types d'erreurs possibles

1. Rejeter H_0 alors que $\theta \in \Theta_0$, dans ce cas φ vaut 1 alors que les données $X_1, \dots, X_n$ ont été générées i.i.d. de loi P_θ avec $\theta \in \Theta_0$.

2. Accepter H_0 alors que $\theta \in \Theta_1$, dans ce cas φ vaut 0 alors que les données $X_1, \dots, X_n$ ont été générées i.i.d. de loi P_θ avec $\theta \in \Theta_1$.

On parle aussi parfois d'*erreur de première espèce* pour le premier type d'erreur ci-dessus et *erreur de deuxième espèce* pour le second.
Remarquons que du point de vue pratique, les deux types d'hypothèses H_0 et H_1 ne sont en général pas symétriques. Souvent, H_0 correspond à l'hypothèse de 'base' (par exemple, qu'un individu n'est pas malade), tandis que H_1 est l'"autre" hypothèse (par exemple, qu'un individu est malade).

Definition 2. On appelle niveau d'un test φ la quantité

$$\sup_{\theta \in \Theta_0} E_\theta \varphi = \sup_{\theta \in \Theta_0} P_\theta[\varphi = 1].$$

La fonction de $\Theta_1 \to [0,1]$ définie par

$$\theta \to E_\theta[\varphi]$$

s'appelle fonction puissance.

L'approche fréquentiste des tests consiste à chercher à construire un test φ dont le niveau est au plus α et ensuite, parmi ces tests (de niveau α), à en chercher un dont la puissance est la plus grande possible.

Exemple. Soit $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$, et posons $\Theta_0 = \{0\}$ et $\Theta_1 = \{1\}$. Le test

$$T(X) = \mathbb{1}_{\{\sqrt{n}\overline{X} \geq q_\alpha\}}, \qquad \text{où} \quad t_\alpha = q_{1-\alpha}^{\mathcal{N}(0,1)},$$

est un test de niveau α. On peut montrer, en utilisant le théorème de Neyman-Pearson, qui s'applique puisqu'il s'agit de deux hypothèses simples, qu'il s'agit d'un test Uniformément le Plus Puissant (UPP).

6.3 L'approche bayésienne

Soit Π une loi a priori sur $\Theta_0 \cup \Theta_1$. Notons qu'avec cette supposition, Π n'est pas forcément défini sur Θ tout entier. L'approche bayésienne des tests s'introduit naturellement à partir du cadre de théorie de la décision étudié au Chapitre 3.

Fonction de perte. On étend légèrement la définition vue au chapitre 3 pour tenir compte du fait que nous travaillons avec des tests. Ici une fonction de perte L sera une application

$$\begin{array}{rccc} L: & \Theta \times \{0,1\} & \to & \mathbb{R}^+ \\ & (\theta, \varphi) & \to & L(\theta, \varphi). \end{array}$$

Fonction de perte équilibrée. Souvent, on travaillera avec la perte du 0-1 suivante

$$L(\theta, \varphi) = \begin{cases} 1 & \text{si } \theta \in \Theta_0,\ \varphi = 1, \quad \text{ou } \theta \in \Theta_1,\ \varphi = 0, \\ 0 & \text{sinon.} \end{cases}$$

Proposition 1. L'estimateur de Bayes pour la fonction de perte du $0-1$ est

$$\varphi(X) = \mathbb{1}_{\Pi(\Theta_0 \,|\, X) \leq \Pi(\Theta_1 \,|\, X)} = \mathbb{1}_{\Pi(\Theta_0 \,|\, X) \leq \frac{1}{2}}.$$

Il s'agit d'un test, appelé test de Bayes pour la fonction de perte du $0-1$.

Preuve.

L'estimateur de Bayes minimise $\varphi \to \int L(\theta, \varphi) d\Pi(\theta \,|\, X)$. Cette fonction s'écrit

$$\begin{aligned} \int L(\theta, \varphi) d\Pi(\theta \,|\, X) &= \int (\mathbb{1}_{\theta \in \Theta_0,\ \varphi = 1} + \mathbb{1}_{\theta \in \Theta_1,\ \varphi = 0} \\ &= \Pi(\Theta_0 \,|\, X) \mathbb{1}_{\varphi = 1} + \Pi(\Theta_1 \,|\, X) \mathbb{1}_{\varphi = 0}. \end{aligned}$$

Cette fonction est minimale pour $\varphi = \mathbb{1}_{\Pi(\Theta_0 \,|\, X) \leq \Pi(\Theta_1 \,|\, X)}$.

Fonction de perte pondérée. On définit

$$L(\theta, \varphi) = \begin{cases} a_0 & \text{si } \theta \in \Theta_0,\ \varphi = 1 \\ a_1 & \text{si } \theta \in \Theta_1,\ \varphi = 0, \\ 0 & \text{sinon.} \end{cases}$$

Le test de Bayes pour la fonction de perte pondérée est, par la même preuve que ci-dessus,

$$\varphi(X) = \mathbb{1}_{a_0 \Pi(\Theta_0 \,|\, X) \leq a_1 \Pi(\Theta_1 \,|\, X)}.$$

Exemples. Modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$.

A Cas d'hypothèses simples.

On veut tester

$$H_0 \ : \ \{\theta = 0\} \qquad \text{contre} \qquad H_1 \ : \ \{\theta = 1\}$$

Tout d'abord, il s'agit de construire une loi a priori Π sur l'ensemble $\Theta_0 \cup \Theta_1$ soit ici $\{0, 1\}$. L'a priori Π est donc de la forme, avec $\pi_0 + \pi_1 = 1$,

$$\Pi = \pi_0 \delta_0 + \pi_1 \delta_1.$$

On calcule alors

$$\Pi[\{0\}\,|\,X] = \frac{\pi_0 \prod_{i=1}^n e^{-\frac{1}{2}(X_i-0)^2}}{\pi_0 \prod_{i=1}^n e^{-\frac{1}{2}(X_i-0)^2} + \pi_1 \prod_{i=1}^n e^{-\frac{1}{2}(X_i-1)^2}} = \frac{\pi_0}{\pi_0 + \pi_1 e^{n\overline{X}-n/2}},$$

et $\Pi[\{1\}\,|\,X] = 1 - \Pi[\{0\}\,|\,X]$. On en déduit que le test bayésien pour la fonction de perte pondéré s'écrit

$$\varphi(X) = \mathbb{1}_{a_0\pi_0 \le a_1\pi_1 e^{n\overline{X}-\frac{n}{2}}} = \mathbb{1}\left\{\overline{X} \ge \frac{1}{2} + \frac{1}{n}\log\left(\frac{a_0\pi_0}{a_1\pi_1}\right)\right\}.$$

On remarque que le test se met sous la forme $\{\overline{X} \ge t_n\}$, comme pour le test de Neyman-Pearson ci-dessus, et que si la fonction de perte est celle du $0-1$ et que l'a priori est symétrique, soit $\pi_0 = \pi_1 = 1/2$, le test est $\varphi(X) = \mathbb{1}_{\overline{X}\ge 1/2}$. Dans ce dernier cas, les hypothèses H_0 et H_1 jouent des rôles symétriques.

B Cas d'hypothèses composites.

Supposons que l'on veuille tester

$$H_0 \ : \ \{\theta < 0\} \qquad \text{contre} \qquad H_1 \ : \ \{\theta \ge 0\}$$

Il s'agit ici de construire une loi a priori Π sur $\mathbb{R}_-^* \cup \mathbb{R}_+ = \mathbb{R}$. Choisissons par exemple $\Pi = \mathcal{N}(\mu, 1)$. La loi a posteriori est dans ce cas $\Pi[\cdot\,|\,X] = \mathcal{N}(\overline{\theta}_X, \frac{1}{n+1})$, avec $\overline{\theta}_X = \frac{\mu+n\overline{X}}{n+1}$. Le test

$$\varphi = \mathbb{1}_{a_0\Pi[\Theta_0\,|\,X] \le a_1\Pi[\Theta_1\,|\,X]}$$

est bayésien pour la fonction de perte pondérée, et on a

$$\Pi[\Theta_0\,|\,X] = P\left[\overline{\theta}_X + \frac{1}{\sqrt{n+1}}\mathcal{N}(0,1) < 0\right] = \Phi(-\sqrt{n+1}\,\overline{\theta}_X),$$

où Φ est la fonction de répartition d'une loi normale standard.

Remarques. Le test bayésien pour les fonctions de perte du $0-1$ ou pondérée se met toujours sous la forme $\mathbb{1}\{\Pi[\Theta_0\,|\,X] \le t\}$ pour un réel t qui dépend de a_0 a_1 (et $t = 1/2$ pour la fonction de perte du $0-1$).
La fonction risque associée aux fonctions de perte ci-dessus est

$$\begin{aligned} E_\theta[L_{a_0,a_1}(\theta,\varphi)] &= E_\theta\left[a_0\mathbb{1}_{\varphi=1}\mathbb{1}_{\theta\in\Theta_0} + a_1\mathbb{1}_{\varphi=0}\mathbb{1}_{\theta\in\Theta_1}.\right] \\ &= a_0 P_\theta[\varphi=1]\mathbb{1}_{\theta\in\Theta_0} + a_1 P_\theta[\varphi=0]\mathbb{1}_{\theta\in\Theta_1}. \end{aligned}$$

Le risque bayésien associé s'écrit

$$R_B(\Pi,\varphi) = \int E_\theta[L_{a_0,a_1}(\theta,\varphi)]d\Pi(\theta) = a_0\int_{\Theta_0} P_\theta[\varphi=1]d\Pi(\theta) + a_1\int_{\Theta_1} P_\theta[\varphi=0]d\Pi(\theta).$$

Le test bayésien minimise par définition ce risque. Les erreurs de premières espèces sont moyennées par rapport à la loi a priori. Les constantes a_0, a_1 introduisent une pondération

éventuelle supplémentaire.

Notion de facteur de Bayes

Definition 3. Le facteur de Bayes se définit comme

$$B_{01}^{\Pi}(X) = \frac{\frac{\Pi[\Theta_0 \mid X]}{\Pi[\Theta_0]}}{\frac{\Pi[\Theta_1 \mid X]}{\Pi[\Theta_1]}} = \frac{\Pi[\Theta_0 \mid X]}{\Pi[\Theta_1 \mid X]} \frac{\Pi[\Theta_1]}{\Pi[\Theta_0]}.$$

On définit de même $B_{10}^{\Pi}(X) = B_{01}^{\Pi}(X)^{-1}$.

Le facteur de Bayes s'interprète naturellement comme un "rapport de vraisemblance bayésien".

a. *Cas où* $\Theta_0 = \{\theta_0\}$ *et* $\Theta_1 = \{\theta_1\}$, *et* $\Pi = \pi_0\delta_{\theta_0} + \pi_1\delta_{\theta_1}$. Avec l'expression de l'a posteriori obtenue ci-dessus, on obtient

$$B_{01}^{\Pi}(X) = \frac{\pi_0 f_{\theta_0}(X)}{\pi_0 f_{\theta_0}(X) + \pi_1 f_{\theta_1}(X)} \frac{\Pi[\Theta_1]}{\Pi[\Theta_0]} = \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}.$$

Il s'agit donc du rapport de vraisemblance classique.

b. *Cas général.* Soit $m(X) = \int f_\theta(X) d\Pi(\theta)$. On a

$$\begin{aligned} B_{01}^{\Pi}(X) &= \frac{\int_{\Theta_0} f_\theta(X) d\Pi(\theta)}{m(X)\Pi(\Theta_0)} \times \left(\frac{\int_{\Theta_1} f_\theta(X) d\Pi(\theta)}{m(X)\Pi[\Theta_1]} \right)^{-1} \\ &= \frac{\int_{\Theta_0} f_\theta(X) d\overline{\Pi}_0(\theta)}{\int_{\Theta_1} f_\theta(X) d\overline{\Pi}_1(\theta)} = \frac{m^{\overline{\Pi}_0}(X)}{m^{\overline{\Pi}_1}(X)}, \end{aligned}$$

où $\overline{\Pi}_0, \overline{\Pi}_1$ sont les restrictions de Π_0 et Π_1 à Θ_0, Θ_1,

$$\overline{\Pi}_0 = \frac{\Pi[\cdot \cap \Theta_0]}{\Pi[\Theta_0]}, \qquad \overline{\Pi}_1 = \frac{\Pi[\cdot \cap \Theta_1]}{\Pi[\Theta_1]}.$$

Le facteur de Bayes $B_{01}^{\Pi}(X)$ s'interprète donc comme le quotient des *vraisemblances marginales* de X si l'a priori est respectivement $\overline{\Pi}_0 = \Pi_{|\Theta_0}$ et $\overline{\Pi}_1 = \Pi_{|\Theta_1}$.

Comment lit-on sur le facteur de Bayes si l'on rejette ou non H_0 ? Rappelons que le test bayésien pour la fonction de perte pondérée rejette H_0 si $\Pi[\Theta_0 \mid X]/\Pi[\Theta_1 \mid X] \leq a_1/a_0$ donc on rejette H_0 si

$$B_{01}^{\Pi}(X) \leq \frac{a_1 \Pi[\Theta_1]}{a_0 \Pi[\Theta_0]}.$$

En pratique, on compare souvent $B_{01}^{\Pi}(X)$ à 1, ce qui mathématiquement correspond au cas "équilibré" où $a_0 = a_1 = \Pi[\Theta_0] = \Pi[\Theta_1]$, pour avoir une idée de l'évidence en faveur de H_0 ou de H_1.

Si $B_{01}^{\Pi}(X) \ll 1$, on rejette clairement H_0

Si $B_{01}^{\Pi}(X) \gg 1$, on ne rejette clairement pas H_0

Cas d'une hypothèse nulle ponctuelle.

Considérons le problème de test des hypothèses, pour $\theta_0 \notin \Theta_1$,

$$H_0 : \{\theta = \theta_0\} \qquad \text{contre} \qquad H_1 : \{\theta \in \Theta_1\}$$

Dans ce cadre, il faut définir une loi a priori Π sur $\{\theta_0\} \cup \Theta_1$. Dans le cas où $\theta \in \mathbb{R}$ et $\Theta_1 = \{\theta_0\}^c = \mathbb{R} \setminus \{\theta_0\}$, un choix qui pourrait sembler à première vue naturel serait celui d'une loi Π à densité par rapport à la mesure de Lebesgue. Cependant, dans ce cas on aurait $\Pi[\Theta_0] = \Pi[\{0\}] = 0$ et donc on rejèterais toujours H_0.

D'un point de vue bayésien, si l'on suppose que l'hypothèse nulle est un singleton $\{\theta_0\}$, c'est que l'on se donne pour garanti que θ peut valoir *exactement* θ_0, donc il est naturel d'intégrer cette information à la loi a priori. On pose

$$\Pi = \pi_0 \delta_{\theta_0} + (1 - \pi_0) G,$$

où G est une mesure de probabilité telle que $G(\Theta_1) = G(\{\theta \neq \theta_0\}) = 1$.

On distingue deux cas suivant la forme de Θ_1.

a) Si $\Theta_1 = \{\theta_1\}$, on pose

$$G = \delta_{\theta_1}.$$

Dans ce cas, la formule de Bayes s'écrit (il s'agit du cas discret pour Θ, avec pour mesure dominante $\nu = \delta_{\theta_0} + \delta_{\theta_1}$ sur Θ)

$$\Pi[\{\theta_0\} \,|\, X] = \frac{\pi_0 p_{\theta_0}(X)}{\pi_0 p_{\theta_0}(X) + (1 - \pi_0) p_{\theta_1}(X)}.$$

b) Si $\Theta_1 \subset \mathbb{R}^d$, d'intérieur non vide, on pose

$$dG(\theta) = g(\theta) d\theta,$$

avec g positive mesurable telle que $\int_{\Theta_1} g(\theta) d\theta = 1$. La formule de Bayes s'écrit, pour tout A mesurable,

$$\begin{aligned}
\Pi[A \,|\, X] &= \frac{\int_A p_\theta(X) d\Pi(\theta)}{\int p_\theta(X) d\Pi(\theta)} \\
&= \frac{\pi_0 p_{\theta_0}(X) \mathbb{1}_{\theta_0 \in A} + (1 - \pi_0) \int_{A \cap \{\theta :\ \theta \neq \theta_0\}} p_\theta(X) g(\theta) d\theta}{\pi_0 p_{\theta_0}(X) + (1 - \pi_0) \int p_\theta(X) g(\theta) d\theta} \\
&= \frac{\pi_0 p_{\theta_0}(X) \mathbb{1}_{\theta_0 \in A} + (1 - \pi_0) \int_A p_\theta(X) g(\theta) d\theta}{\pi_0 p_{\theta_0}(X) + (1 - \pi_0) \int p_\theta(X) g(\theta) d\theta},
\end{aligned}$$

puisque $\int_{A \cap \{\theta :\ \theta \neq \theta_0\}} p_\theta(X) g(\theta) d\theta = \int_A p_\theta(X) g(\theta) d\theta$ car ajouter un point à l'ensemble d'intégration ne change pas l'intégrale par rapport à la mesure de Lebesgue.

Exemple. Dans le modèle $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$, on veut tester

$$H_0 : \{\theta = 0\} \qquad \text{contre} \qquad H_1 : \{\theta \neq 0\}.$$

En suivant le principe précédent, une loi a priori raisonnable est

$$\Pi = \pi_0 \delta_0 + (1 - \pi_0)\mathcal{N}(0, 1).$$

La formule de Bayes ci-dessus donne

$$\Pi[\theta = 0 \,|\, X] = \frac{\pi_0 p_0(X)}{\pi_0 p_0(X) + (1 - \pi_0) \int p_\theta(X) g(\theta) d\theta}.$$

Le facteur de Bayes $B_{01}(X)$ vaut, par définition,

$$\begin{aligned} B_{01}(X) &= \frac{\Pi[\Theta_0 \,|\, X]}{\Pi[\Theta_1 \,|\, X]} \frac{\Pi[\Theta_1]}{\Pi[\Theta_0]} \\ &= \frac{p_0(X)\pi_0}{(1 - \pi_0) \int p_\theta(X) g(\theta) d\theta} \frac{1 - \pi_0}{\pi_0} \\ &= \frac{p_0(X)}{\int p_\theta(X) g(\theta) d\theta}. \end{aligned}$$

On remarque comme noté ci-dessus que le facteur de Bayes est le quotient de la densité marginale de X sous $H_0 = \{\theta = 0\}$ et $\Pi_{|\Theta_0}$ et de la densité marginale de X sous $H_1 = \{\theta \neq 0\}$ et $\Pi_{|\Theta_1}$. Le calcul de cette dernière s'écrit

$$\begin{aligned} \int p_\theta(X) g(\theta) d\theta &= \int \exp\{-\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2 - \frac{\theta^2}{2}\} \frac{d\theta}{(\sqrt{2\pi})^{n+1}} \\ &= \frac{\exp\{-\frac{n}{2}\overline{X^2}}{\sqrt{2\pi}^n} \frac{1}{\sqrt{2\pi}} \int \exp\{-\frac{n+1}{2}(\theta - \frac{n\overline{X}}{n+1})^2 + \frac{n^2}{2(n+1)}\overline{X}^2\} d\theta \\ &= f_0(X) \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2\pi}{n+1}} \exp\{\frac{n^2}{2(n+1)}\overline{X}^2\} \\ &= f_0(X) \frac{1}{\sqrt{n+1}} \exp\{\frac{n^2}{2(n+1)}\overline{X}^2\}. \end{aligned}$$

On en conclut que

$$B_{01}^{\Pi}(X) = \frac{\sqrt{n+1}}{\exp\{\frac{n^2}{2(n+1)}\overline{X}^2}.$$

On rejette H_0 si

$$B_{01}^{\Pi}(X) < \frac{a_1}{a_0} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)} = \frac{a_1(1 - \pi_0)}{a_0 \pi_0},$$

ce qui se simplifie en $B_{01}^{\Pi}(X) < 1$ dans le cas 'équilibré'.

6.4 Analyse asymptotique des tests bayésiens

Definition 4. Un test bayésien

$$\varphi(X) = \mathbb{1}_{a_0\Pi(\Theta_0\,|\,X)\leq a_1\Pi(\Theta_1\,|\,X)} = \mathbb{1}_{B_{01}^{\Pi}(X)\leq \frac{a_1\Pi(\Theta_1)}{a_0\Pi(\Theta_0)}}$$

est dit consistant au sens fréquentiste si

$$\begin{aligned} &\forall\,\theta\in\Theta_0, \qquad E_\theta\varphi(X)\to 0 \qquad (n\to\infty)\\ &\forall\,\theta\in\Theta_1, \qquad E_\theta[1-\varphi(X)]\to 0 \qquad (n\to\infty). \end{aligned}$$

Si Θ_0 et Θ_1 sont "bien séparés", alors un test bayésien est typiquement consistant.

Cas où $\Theta\subset\mathbb{R}$. Supposons

$$\inf\{|\theta_1-\theta_2|,\ \theta_1\in\Theta_1,\theta_2\in\Theta_2\} =: \rho > 0.$$

Dans ce cas, pour que le test bayésien soit consistant, il suffit que la loi a posteriori soit consistante.

Preuve.

En effet, si c'est le cas, alors si $\theta\in\Theta_0$, d'après l'hypothèse ci-dessus $\Theta_1\subset\{\theta',\ |\theta'-\theta|>\rho/2\}$. Comme la loi a posteriori est consistante par hypothèse, on a donc $\Pi[\Theta_1\,|\,X] = o_{P_\theta}(1)$ quand $n\to\infty$. On en déduit $\Pi[\Theta_0\,|\,X]\to 1$ en probabilité sous P_θ. Cela entraine, comme alors $\Pi[\Theta_1\,|\,X]/\Pi[\Theta_0\,|\,X] = o_{P_\theta}(1)$, que $\varphi(X) = o_{P_\theta}(1)$ soit $E_\theta\varphi(X) = o(1)$. Un raisonnement similaire montre que si $\theta\in\Theta_1$, alors $E_\theta[1-\varphi(X)] = o(1)$.

Si l'hypothèse de séparation ci-dessus n'est pas vérifée, la théorie est un peu plus délicate. Dans le cadre de l'exemple du test de $H_0 : \{\theta=\theta_0\}$ contre $H_1 : \{\theta\neq\theta_0\}$ ci-dessus, avec $\Pi = (1-\pi_0)\mathcal{N}(0,1)+\pi_0\delta_0$, on constate d'après le calcul précédent que

$$\begin{aligned} &B_{01}^{\Pi}(X)\to+\infty \quad \text{sous } P_\theta,\ \theta=\theta_0\\ &B_{01}^{\Pi}(X)\to 0 \quad \text{sous } P_\theta,\ \theta\neq\theta_0. \end{aligned}$$

Dans ce cas, on dit que le facteur de Bayes est consistant. Cela implique que le test est consistant.

CHAPITRE 7

Simulations de lois a posteriori

*Dans ce dernier chapitre, nous faisons un bref tour d'horizon de méthodes permettant de calculer des intégrales, ainsi que de simuler suivant une loi de densité donnée. Nous introduisons ainsi les méthodes de Monte-Carlo, l'échantillonage d'importance (*importance sampling*)*, et les méthodes MCMC.

Pour avoir accès à la loi a posteriori, ou à certains de ses aspects (moyenne, médiane, quantile etc.), il faut être capable d'évaluer des intégrales, au minimum

$$\int p_\theta(X) d\Pi(\theta) = \int p_\theta(X)\pi(\theta)d\theta,$$

disons si Π admet une densité par rapport à la mesure de Lebesgue. Ceci, si l'on souhaite calculer exactement $\Pi[A \,|\, X]$ ou $\int \theta d\Pi(\theta \,|\, X)$ par exemple. On peut aussi vouloir simuler suivant la loi $\imath[\cdot \,|\, X]$ ou une loi qui s'en approche.

7.1 Méthodes de calcul d'intégrales

Soit à calculer

$$I = \int_C f(x)dx,$$

pour $f : \mathbb{R}^d \to \mathbb{R}$ une fonction connue pour laquelle il est facile d'avoir accès à toutes les valeurs $f(y)$ pour $y \in \mathbb{R}^d$, et C un sous-ensemble de $\mathbb{R}^d$.

A Méthodes déterministes

Une façon simple d'approcher l'intégrale I est de découper l'ensemble d'intégration C en sous-ensembles plus petits C_i et d'approcher sur chaque C_i la fonction f par une fonction plus simple, par exemple une constante, ou une fonction affine. Si on prend le cas où on approche

la fonction par une constante, on retrouve le cas des sommes de Riemann, pour lesquelles on approche I par

$$\frac{1}{N}\sum_{i=1}^{N} f(\xi_i),$$

avec par exemple $\xi_i = i/N$. Une difficulté avec ce type de méthode est que si l'on se place en dimension d, si l'on a besoin de N points pour atteindre une précision ε donnée en dimension 1, alors le nombre de points nécessaires pour obtenir la même précision en dimension d est typiquement de l'ordre de N^d. Lorsque la dimension d vaut au moins 3, on se retrouve rapidement confrontés à un très grand nombre de calculs. De façon surprenante au premier abord, l'utilisation d'une méthode introduisant de l'aléatoire va permettre de s'affranchir de la dépendance en la dimension.

B Méthodes de Monte-Carlo

Au lieu de prendre des points fixés ξ_i à la base de notre approximation, on peut les tirer au hasard. Par exemple, si $d = 1$ et que l'on cherche à estimer

$$I = \int_0^1 f(u)du,$$

la loi des grands nombres, pour f intégrable, donne, si X_i sont i.i.d. de loi uniforme sur $[0, 1]$,

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i) \to I,$$

où la converge est presque sûre. Plus généralement, supposons, ce qui est le cas en statistique bayésienne avec le dénominateur de la formule de Bayes, que l'on veuille calculer

$$J = \int f(x)g(x)dx,$$

où f est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ et g une densité sur $\mathbb{R}^d$, avec $\int |f|g < \infty$. La loi des grands nombres donne

$$\hat{J}_N := \frac{1}{N}\sum_{i=1}^{N} f(X_i) \to \int f(x)g(x)dx = J, \tag{9}$$

presque sûrement. De plus, si $\int f^2 g < \infty$, on a aussi par le théorème central limite, quand $N \to \infty$,

$$\sqrt{N}(\hat{J}_N - J) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}_g(f)),$$

où $\mathrm{Var}_g(f) = \int (f - J)^2 g$. Un grand avantage de cette approche par rapport aux méthodes déterministes citées plus haut est que la vitesse de convergence dans le résultat limite ci-dessus est $N^{-1/2}$, indépendemment de la dimension d du problème.

Difficulté possible de l'approche (9). Pour pouvoir appliquer (9) efficacement, il faut savoir simuler de manière rapide des variables aléatoires de loi de densité g, avec dans le cas de l'application aux statistiques bayésiennes g égale à la densité a priori. Pour certaines lois a

priori (par exemple des mélanges à plusieurs niveaux de hiérarchie), cette simulation peut être délicate et/ou coûteuse en temps de calcul.

C Méthodes de Monte-Carlo par *Importance Sampling* (ou Échantillonnage d'importance)

Soit h une densité sur $\mathbb{R}^d$, pour laquelle on sait simuler efficacement des variables de loi de densité h. Soit $Y_1, \dots, Y_N$ un tirage i.i.d. suivant la loi de densité h. On pose

$$\tilde{J}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{f(Y_i)g(Y_i)}{h(Y_i)}.$$

Sous la condition d'intégrabilité $\int |f|g < \infty$, la loi des grands nombres donne

$$\tilde{J}_N \to J.$$

On note que l'on ne doit plus simuler suivant g mais sous h, que l'on choisit. Cependant, il faut, si l'on veut avoir un théorème central limite, pouvoir vérifier la condition de moment d'ordre 2, c'est-à-dire

$$\int \left(\frac{fg}{h}\right)^2 h = \int \frac{f^2 g^2}{h} < \infty.$$

Pour cela il suffit typiquement que les queues de distribution de h soient plus lourdes que celles de g.

Application de l'Importance Sampling. Supposons que l'on veuille estimer, avec ϕ la densité d'une loi $\mathcal{N}(0,1)$,

$$J = P[\mathcal{N}(0,1) > 3].$$

On remarque que J se met sous la forme $J = \int \mathbb{1}_{u>3}\phi(u)du = \int f(u)g(u)du$. On peut envisager deux méthodes de Monte-Carlo

1. Méthode de Monte-Carlo simple. On pose tout simplement

$$\hat{J}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i>3},$$

où les variables X_i sont i.i.d. $\mathcal{N}(0,1)$. Un inconvénient de cette approche est qu'il faut des N assez grands pour observer des indicatrices non nulles. Intuitivement, le nombre d'observations 'utiles' est assez faible.

2. Méthode de Monte-Carlo par Importance Sampling. On pose

$$\tilde{J}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}_{Y_i>3}\phi(Y_i)}{h(Y_i)},$$

où h est la densité d'une loi $\mathcal{N}(4,1)$ et $Y_1, \dots, Y_N$ sont i.i.d. de loi $\mathcal{N}(4,1)$. On vérifie

que la condition de moment est bien vérifiée, puisque

$$\begin{aligned}\int \mathbb{1}_{y>3}\frac{\phi(y)^2}{h(y)} &= \int_3^\infty (e^{-4y+8})^2\phi(y-4)dy \\ &= \int_3^\infty e^{-8(y-4)-16}\phi(y-4)dy = \int_{-1}^\infty e^{-8t-16}\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt. \\ &= \int_{-1}^\infty e^{-\frac{(t+8)^2}{2}+16}\frac{dt}{\sqrt{2\pi}} = e^{16}P[\mathcal{N}(-8,1)\geq -1] = e^{16}P[\mathcal{N}(0,1)\geq 7] < \infty.\end{aligned}$$

On peut vérifier que la variance correspondante est inférieure à celle de la méthode simple ci-dessus.

Une question naturelle est de savoir quelle est le choix optimal de la fonction h. La proposition suivante a un intérêt surtout théorique car ce choix optimal dépend d'une intégrale similaire à la quantité qu'on cherche à obtenir. En revanche elle est utile pour suggérer des formes de densités. Dans l'exemple précédent par exemple, elle donne une densité optimale qui est la loi normale conditionnée à être supérieure à 3.

Proposition 1. Le choix optimal théorique de la densité h pour la méthode d'importance sampling est donné par

$$h^* = \frac{|f|g}{\int |f|g}.$$

Preuve.

La variance s'écrit

$$\mathrm{Var}_h\left[\frac{fg}{h}\right] = \int (fg/h)^2 h - (\int (fg/h)h)^2 = \int \frac{f^2g^2}{h} - (\int fg)^2.$$

Le terme $(\int fg)^2$ ne dépend pas de h donc il suffit de minimiser le premier terme. Or

$$\int \frac{f^2g^2}{h} = E_h[(\frac{fg}{h})^2] \geq (E_h|\frac{fg}{h}|)^2 \geq (\int \frac{|f|g}{h}h)^2 = (\int |f|g)^2.$$

Posons $h^* = |f|g/(\int |f|g)$. C'est une densité et elle atteint la borne ci dessus car

$$\int \frac{f^2g^2}{h^*} = (\int |f|g)^2.$$

7.2 Les méthodes MCMC

L'abbréviation MCMC signifie Monte Carlo Markov Chain. Il s'agit typiquement d'approcher une loi ou une intégrale à l'aide d'une chaîne de Markov.

7.2.1 Un bref aperçu sur les chaînes de Markov

Definition 1. La suite de variables aléatoires $X_1, \ldots, X_n, \ldots$ est une chaîne de Markov (CM) d'espace d'états $\mathcal{X}$ si l'espace d'arrivée des variables X_i est $\mathcal{X}$ et, pour tout $n \geq 1$,

$$\mathcal{L}(X_n \,|\, X_0, \ldots, X_{n-1}) = \mathcal{L}(X_n \,|\, X_{n-1}).$$

Pour fixer les idées, on prendra dans la suite $\mathcal{X} = \mathbb{R}$.
Une CM est dite homogène si pour tout $n \geq 1$,

$$\mathcal{L}(X_n \,|\, X_{n-1} = y) = \mathcal{L}(X_1 \,|\, X_0 = y) \quad \forall y \in \mathcal{X}.$$

Dans la suite, on travaillera toujours avec des CM homogènes.
Pour une CM homogène, la loi des données est caractérisée par la collection de densités

$$\begin{aligned} p(x, y) &:= f_{X_{n+1} \,|\, X_n = x}(y) = \text{‘densité en } y \text{ de } X_{n+1} \text{ sachant } X_n = x' \\ &= f_{X_1 \,|\, X_0 = x}(y). \end{aligned}$$

Les densités $(p(x, \cdot))_{x \in \mathcal{X}}$ s'appellent densités de transition. On a en particulier $\int p(x, y)dy = 1$.

Exemple. $X_{n+1} = X_n + \xi_{n+1}$ avec (ξ_i) i.i.d. $\mathcal{N}(0, 1)$. Dans ce cas $p(x, y) = \phi(y - x)$ et $p(x, \cdot)$ est la densité d'une loi $\mathcal{N}(x, 1)$. Il s'agit d'une marche aléatoire avec une longueur de pas variable.

Definition 2. On dit que f est une densité stationnaire pour la CM $(X_n)_{n \geq 0}$ si pour tout $x \in \mathcal{X}$,

$$f(x) = \int f(u)p(u, x)du.$$

On dit alors que la loi de densité f est stationnaire.

Interprétation. Supposons que pour un certain entier k, l'observation X_k admet une loi de densité f stationnaire. Alors X_{k+1} suit encore une loi de densité f.

Preuve.

> En effet, $f(u)p(u, x)$ n'est autre dans ce cas que la densité du couple (X_k, X_{k+1}) en tant que produit de la densité marginale de X_k au point u qui est $f(u)$ par hypothèse et de la densité conditionnelle au point u de X_{k+1} sachant $X_k = x$ qui par définition est $p(u, x)$.

Remarque. Une densité stationnaire n'existe pas toujours. Ainsi, on peut montrer qu'il n'y en a pas dans l'exemple très simple de marche aléatoire ci-dessus, car celle-ci “diffuse” : la loi à l'instant $n + 1$ est plus dispersée que la loi à l'instant n et donc on ne peut avoir la même loi. Nous verrons plus loin des exemples de chaînes avec densité stationnaire. Dans l'exemple de la marche aléatoire, il suffit d'autoriser la marche à s'arrêter à certains instants avec une

certaine probabilité au lieu de marcher sans relâche ...

Definition 3. La CM $(X_n)_{n\geq 0}$ vérifie la condition d'équilibre ponctuel si pour tous x, y dans $\mathcal{X}$,

$$f(x)p(x,y) = f(y)p(y,x).$$

Proposition 2. Sous la condition d'équilibre ponctuel, f est une densité stationnaire.

Preuve.

On intègre la condition d'équilibre ponctuel par rapport à y. D'une part

$$\int f(x)p(x,y)dy = f(x)\int p(x,y)dy = f(x).$$

D'autre part avec l'équilibre ponctuel ceci coïncide avec $\int f(y)p(y,x)dy$. Donc f est une densité stationnaire.

Les deux faits suivants, dont nous ne détaillons pas les hypothèses précises, ce qui serait l'objet d'un cours sur les chaînes de Markov, peuvent être vus comme des analogues de la loi des grands nombres pour des observations i.i.d. On note

$$(Pg)(y) = \int p(x,y)g(x)dx.$$

L'objet P opère sur les fonctions, il s'agit de ce qu'on appelle un opérateur à noyau. Notons P^j l'itérée $j$ème de P (i.e. on compose P avec lui-même un nombre j de fois).

Fait 1. Sous certaines conditions, en particulier l'existence d'une unique densité stationnaire f, on a que pour toute densité g bornée,

$$\|P^N g - f\|_1 \to 0 \qquad (N \to \infty).$$

En particulier, partant de $\mathcal{L}(X_0) = g$, la loi de $\mathcal{L}(X_N)$ après N itérations de la chaîne est très proche de la loi stationnaire f.

Fait 2. Sous certaines conditions, si (X_n) est une CM de densité stationnaire f, quand $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \to \int h(u)f(u)du = E_f[h(X)],$$

où la convergence est presque sûre.

Une conséquence importante du Fait 2 est que si l'on a à disposition une CM de densité stationnaire f, alors on sait approcher l'intégrale $\int h(u)f(u)du$! Il se trouve de plus qu'il existe des algorithmes très simples permettant de construire une CM de densité stationnaire f.

7.2.2 Algorithmes MCMC

Le cadre est le suivant. Supposons que l'on veuille soit simuler (disons approximativement) suivant une loi de densité f, ou bien que l'on veuille évaluer une intégrale du type $\int h(u)f(u)du$, comme c'est le cas en statistiques bayésiennes pour f la densité a priori. Il suffit de construire une chaîne Markov (X_n) de densité stationnaire f, car alors, d'après les deux faits ci-dessus, la loi de X_N avec N grand sera proche d'une loi de densité f, tandis que la moyenne $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$ approchera l'intégrale cherchée.

A *L'algorithme de Metropolis-Hastings*

Soit $q(\cdot \,|\, x)$ une collection de densités conditionnelles. Par exemple, on peut prendre $q(\cdot \,|\, x)$ la densité d'une loi $\mathcal{N}(x, v)$, avec v une constante positive à choisir. Pour ce choix précis, l'algorithme s'appelle *Random Walk Metropolis Hastings*. Notons qu'il sera utile pour l'algorithme de savoir simuler rapidement suivant les lois de cette collection, ce qui est bien sûr le cas pour les lois normales.

Soit f la densité suivant laquelle on veut simuler, ou pour laquelle on veut calculer $\int h(u)f(u)du$.

Algorithme de Metropolis Hastings.

Soit $X_0 \in \mathcal{X}$ quelconque.

Supposons que $X_1, \ldots, X_i$ ont été générées. On génère X_{i+1} comme suit

1. Générer $Y \sim q(\cdot \,|\, X_i)$
2. Soit $r = r(X_i, Y)$, où
$$r(x, y) = \min\left(\frac{f(y)q(x \,|\, y)}{f(x)q(y \,|\, x)}, 1\right)$$
3. Poser
$$X_{i+1} = \begin{cases} Y & \text{avec probabilité } r \\ X_i & \text{avec probabilité } 1 - r. \end{cases}$$

Par construction (X_n) est une chaîne de Markov, homogène, puisque $\mathcal{L}(X_{i+1} \,|\, X_i = x)$ dépend seulement de x et pas de i.

Notons $p(x, \cdot)$ les densités de transition de cette chaîne.

Théorème 1. La chaîne de Markov générée par l'algorithme de Metropolis-Hastings ci-dessus a pour densité stationnaire f.

Preuve.

Il suffit de vérifier la condition d'équilibre ponctuel, d'après la proposition ci-dessus. Soient x, y dans $\mathcal{X}$ avec $x \neq y$. Par symétrie on peut toujours supposer $f(y)q(x \mid y) \leq f(x)q(y \mid x)$, quitte à échanger les rôles de x et y (la condition d'équilibre ponctuel ne change pas si on permute x et y). Dans ce cas notons que

$$r(x, y) = \frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}, \qquad r(y, x) = 1.$$

Rappelons que $p(x, y)$ est la densité de transition, i.e. la probabilité instantanée d'aller en y en partant de x. Comme $x \neq y$, pour passer de x à y avec la chaîne définie par l'algorithme, il faut deux choses : a) générer y avec $q(y \mid x)$ à l'étape 1. et b) accepter le mouvement de x à y à l'étape 3. Ainsi

$$p(x, y) = q(y \mid x)r(x, y) = \frac{f(y)q(x \mid y)}{f(x)}.$$

On en déduit que $f(x)p(x, y) = f(y)q(x \mid y)$.

Par le même argument, on calcule $p(y, x)$, qui vaut $q(x \mid y)r(y, x)$. Cette fois $r(y, x) = 1$, donc $q(x \mid y) = p(y, x)$.

Les deux identités précédentes mises ensemble donnent la condition d'équilibre ponctuel pour $x \neq y$, et celle-ci est immédiate si $x = y$.

Application aux statistiques bayésiennes. Dans un cadre bayésien, où nous noterons Z les données pour ne pas confondre avec X ci-dessus, on cherche typiquement à simuler suivant la loi a posteriori, ou à calculer des intégrales de type $\int h(\theta)d\pi(\theta \mid Z)$. Donc la loi cible est $f(\theta) = f_{\theta \mid Z}(\theta)$.

Pour utiliser l'algorithme de Metropolis-Hastings, il faut savoir simuler suivant $q(\cdot \mid x)$ pour tout x. Comme on a le choix de $q(\cdot \mid x)$, par exemple une loi gaussienne $\mathcal{N}(x, 1)$, ce n'est pas un problème. Il faut aussi pouvoir calculer le quotient dans la probabilité d'acceptation $r(x, y)$. C'est en principe un problème, car $f_{\theta \mid Z}(\theta)$ est typiquement difficile à évaluer car son expression contient le dénominateur $\int p_\theta(Z)\pi(\theta)d\theta$. Le point remarquable ici est que cette

quantité se simplifie ! En effet ici

$$\begin{aligned} r(x,y) &= \frac{f_{\theta\,|\,Z}(y)}{f_{\theta\,|\,Z}(x)} \frac{q(x\,|\,y)}{q(y\,|\,x)} \\ &= \frac{\frac{p_y(Z)\pi(y)}{\int p_y(Z)\pi(y)dy}}{\frac{p_x(Z)\pi(x)}{\int p_x(Z)\pi(x)dx}} \frac{q(x\,|\,y)}{q(y\,|\,x)} \\ &= \frac{p_y(Z)\pi(y)}{p_x(Z)\pi(x)} \frac{q(x\,|\,y)}{q(y\,|\,x)}. \end{aligned}$$

Cette expression se calcule directement, du moins si l'expression de la fonction π n'est pas trop complexe.

Un cas un peu plus délicat est celui où la loi a priori est un mélange car l'expression de π peut alors être un peu compliquée. Dans ce cas on peut utiliser l'algorithme de Gibbs ci-dessous.

B *L'algorithme de Gibbs*

Supposons que l'on veuille simuler suivant la loi d'un couple (X, Y), dans un cadre où il est facile de simuler suivant les lois conditionnelles $\mathcal{L}(X\,|\,Y)$ et $\mathcal{L}(Y\,|\,X)$.

Algorithme de Gibbs.

Soient $X_0, Y_0 \in \mathcal{X}$ quelconques.

Supposons que $(X_1, Y_1) \ldots, (X_i, Y_i)$ ont été générées. On génère (X_{i+1}, Y_{i+1}) comme suit

1. Générer $X_{i+1} \sim \mathcal{L}(X\,|\,Y = Y_i)$
2. Générer $Y_{i+1} \sim \mathcal{L}(Y\,|\,X = X_{i+1})$.

La suite $(X_i, Y_i)_{i\geq 1}$ est une chaîne de Markov homogène, dont $\mathcal{L}((X, Y))$ est une loi stationnaire, cf. TD. Ainsi pour N assez grand, la loi de (X_N, Y_N) sera une bonne approximation de la loi de (X, Y).

Application aux statistiques bayésiennes. L'algorithme de Gibbs est particulièrement utile dans les modèles hiérarchiques, pour lesquels il est typiquement facile de simuler suivant une variable sachant toutes les autres. Ainsi, dans le modèle hiérarchique avec les notations du Chapitre 4, si l'on sait simuler suivant les lois $\mathcal{L}(\theta\,|\,\alpha, Z)$ et $\mathcal{L}(\alpha\,|\,\theta, Z)$, l'algorithme de Gibbs permet de simuler suivant une approximation de $\mathcal{L}((\theta, \alpha)\,|\,Z)$.