Master 1 Mathématiques et Applications
Sorbonne Université
2023-2024

Statistiques bayésiennes

Anna Ben-Hamou

Table des matières

Chapitre 0

Introduction

Dans ce chapitre, nous introduisons les notions de base de probabilités et de statistique utiles pour la suite, parmi lesquelles les notions d'espace probabilisé, de variable aléatoire, de convergences de variables aléatoires, d'expérience statistique, de modèle, d'estimateur et de régions de confiance. Enfin nous définissons la notion de loi conditionnelle qui joue un rôle central dans la suite.

1. Outils de probabilité

1.1. Espace probabilisé, variable aléatoire.

Définition 0.1. Soit Ω un ensemble. Une tribu $\mathcal{F}$ sur Ω est un ensemble de parties de Ω tel que

— $\mathcal{F}$ est non-vide ;
— $\mathcal{F}$ est stable par complémentaire ;
— $\mathcal{F}$ est stable par union dénombrable.

Le couple $(\Omega, \mathcal{F})$ est appelé espace mesurable.

Définition 0.2. Soit $(\Omega, \mathcal{F})$ un espace mesurable. Une mesure sur Ω est une application $\mu : \mathcal{F} \to [0, +\infty]$ telle que

— $\mu(\emptyset) = 0$;
— si $(E_n)_{n\geq 1}$ est une suite de parties disjointes de Ω appartenant à $\mathcal{F}$, alors

$$\mu\left(\bigcup_{n\geq 1} E_n\right) = \sum_{n\geq 1} \mu(E_n)\,.$$

Cette propriété s'appelle la σ-additivité.

Si de plus $\mu(\Omega) = 1$, on dit que μ est une mesure de probabilité.

Si μ est une mesure sur $(\Omega, \mathcal{F})$, le triplet $(\Omega, \mathcal{F}, \mu)$ est alors appelé espace mesuré. Si $\mathbb{P}$ est une mesure de probabilité, le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé espace probabilisé ou encore espace de probabilité.

Définition 0.3. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $(E, \mathcal{E})$ un espace mesurable. Une variable aléatoire X est une fonction mesurable de Ω dans E, i.e.

$$\forall A \in \mathcal{E},\ X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega,\, X(\omega) \in A\} \in \mathcal{F}.$$

Définition 0.4. Si X est une variable aléatoire de $(\Omega, \mathcal{F}, \mathbb{P})$ dans $(E, \mathcal{E})$, on dit que X est de loi Q, et l'on note $X \sim Q$, si pour tout $A \in \mathcal{E}$,

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = Q(A).$$

Autrement dit, Q est la mesure image de $\mathbb{P}$ par X. De manière équivalente, cela signifie que pour toute fonction φ intégrable par rapport à Q, soit $\varphi \in L^1(Q)$,

$$\int_\Omega \varphi\left(X(\omega)\right) \mathrm{d}\mathbb{P}(\omega) = \int_E \varphi(x) \mathrm{d}Q(x) = \mathbb{E}[\varphi(X)].$$

1.2. Lois à densité.

Définition 0.5. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré. La mesure μ est dite σ-finie s'il existe une suite $(E_n)_{n\geq 1}$ d'éléments de $\mathcal{E}$ de mesure finie (i.e. pour tout $n \geq 1$, $\mu(E_n) < \infty$) telle que

$$E = \bigcup_{n\geq 1} E_n.$$

Exercice 0.1. Montrer que la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et la mesure de comptage sur $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ sont toutes les deux des mesures σ-finies.

Définition 0.6. Soient P et μ deux mesures σ-finies sur un espace mesurable $(E, \mathcal{E})$. On dit que P est absolument continue par rapport à μ, et l'on note $P \ll \mu$, si

$$\forall A \in \mathcal{E},\ \mu(A) = 0 \quad \Rightarrow \quad P(A) = 0.$$

Proposition 0.1 (Théorème de Radon-Nikodym)**.** *Soient P et μ deux mesures σ-finies sur un espace mesurable $(E, \mathcal{E})$. Si $P \ll \mu$, alors P a une densité par rapport à μ, c'est-à-dire qu'il existe une fonction mesurable positive p telle que pour tout $A \in \mathcal{E}$,*

$$P(A) = \int_A p(x) \mathrm{d}\mu(x).$$

La fonction p est appelée dérivée de Radon-Nikodym de P par rapport à μ, et est notée $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$. Cette notation se comprend bien :

$$P(A) = \int_A \mathrm{d}P(x) = \int_A \frac{\mathrm{d}P(x)}{\mathrm{d}\mu(x)} \mathrm{d}\mu(x) = \int_A p(x) \mathrm{d}\mu(x).$$

Exemple 0.1. 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 $E = \{0, 1\}$, la loi de Bernoulli $P_\theta = \mathcal{B}(\theta)$ admet une densité par rapport à la mesure $\mu = \delta_0 + \delta_1$. En effet, pour tout $A \subset \{0, 1\}$, on peut écrire,

$$\begin{aligned} P_\theta(A) &= (1-\theta)\delta_0(A) + \theta\delta_1(A) \\ &= (1-\theta) \int_A \delta_0(\mathrm{d}x) + \theta \int_A \delta_1(\mathrm{d}x) \\ &= \int_A \left\{(1-\theta)\mathbb{1}_{x=0} + \theta\mathbb{1}_{x=1}\right\} [\delta_0 + \delta_1](\mathrm{d}x)\,. \end{aligned}$$

- Sur $E = \{0, 1, \dots, n\}$, la loi binomiale $P_\theta = \mathcal{B}(n, \theta)$ admet une densité par rapport à la mesure $\mu = \sum_{i=0}^{n} \delta_i$ donnée par
$$k \mapsto \theta^k (1-\theta)^{n-k}\,.$$
- Sur $E = \mathbb{N}^*$, la loi géométrique $P_\theta = \mathcal{G}(\theta)$ admet une densité par rapport à la mesure de comptage sur $\mathbb{N}^*$, $\sum_{i\geq 1} \delta_i$, donnée par
$$k \mapsto (1-p)^{k-1} p\,.$$
- La loi normale $\mathcal{N}(\mu, \sigma^2)$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, donnée par
$$x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$
- Sur $E = \mathbb{R}$, la loi exponentielle $P_\theta = \mathcal{E}(\theta)$, $\theta > 0$, admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, donnée par
$$x \mapsto \theta\, \mathrm{e}^{-\theta x} \mathbb{1}_{x\geq 0}\,.$$

1.3. 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 $B \in \mathcal{F}$. Si P a une densité p par rapport à une mesure μ sur E et Q une densité q par rapport à une mesure ν sur F, alors $P \otimes Q$ a pour densité $p \times q$ par rapport à $\mu \otimes \nu$
$$\mathrm{d}(P \otimes Q) = pq\mathrm{d}(\mu \otimes \nu) = pq\mathrm{d}\mu\mathrm{d}\nu.$$

Deux variables aléatoires X et Y sont indépendantes si et seulement si la loi du couple (X, Y) est le produit de la loi de X et de la loi de Y.

Exemple 0.2. La loi sur $\mathbb{R}^2$ dont la densité par rapport à la mesure produit $\mathrm{Leb}(\mathbb{R}) \otimes \mathrm{Leb}(\mathbb{R})$ est
$$\frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$
est une loi produit. En effet, on reconnaît le produit des densités de deux lois normales standards $\mathcal{N}(0, 1)$. 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 $dP = p\mathrm{d}\mu$, alors $P^{\otimes n}$ a une densité sur E^n par rapport à $\mu^{\otimes n}$, égale à $q(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i)$.

1.4. Vecteurs gaussiens.

Définition 0.7. Soit $d \geq 1$ un entier. Un vecteur aléatoire X de $\mathbb{R}^d$ est dit gaussien si toute combinaison linéaire de ses coordonnées est une variable gaussienne réelle. Un vecteur gaussien est caractérisé par son vecteur d'espérances $\mu \in \mathbb{R}^d$ et sa matrice de covariance $\Sigma \in \mathcal{M}_d(\mathbb{R})$,

symétrique et semi-définie positive. On note alors $X \sim \mathcal{N}(\mu, \Sigma)$. Si Σ est définie positive, alors X possède une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ donnée par

$$x \mapsto \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}\, {}^t(x-\mu)\Sigma^{-1}(x-\mu)\right\}.$$

Notons en particulier que si Σ est une matrice diagonale (et donc également Σ^{-1}), la densité de la loi $\mathcal{N}(\mu, \Sigma)$ s'exprime comme un produit de densités coordonnée par coordonnée. D'après ce qui précède, cela signifie que les coordonnées X_i de X sont indépendantes. Si en revanche Σ n'est pas diagonale, Σ^{-1} non plus et la densité ne s'écrit pas comme un produit : les coordonnées X_i sont corrélées. Si $X = (X_1, \ldots, X_d) \sim \mathcal{N}(\mu, \Sigma)$, on a $\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$.

1.5. Lois Beta, Gamma, Dirichlet.

Définition 0.8. Pour $p > 0$ et $\lambda > 0$, la loi Gamma $\Gamma(p, \lambda)$ est la loi dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est donnée par

$$x \mapsto \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \mathbb{1}_{[0,+\infty[}(x),$$

où

$$\Gamma(p) = \int_0^{+\infty} z^{p-1} e^{-z} \mathrm{d}z\,.$$

Notons que la loi $\Gamma(1, \lambda)$ correspond à la loi exponentielle $\mathrm{Exp}(\lambda)$.

Exercice 0.2. Soit $Z \sim \Gamma(p, \lambda)$. Montrer que

$$\mathbb{E}Z = \frac{p}{\lambda} \qquad \text{et} \qquad \mathrm{Var}(Z) = \frac{p}{\lambda^2}\,.$$

Définition 0.9. Pour $a > 0$ et $b > 0$, la loi $\mathrm{Beta}(a, b)$ est la loi dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est donnée par

$$x \mapsto \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{[0,1]}(x),$$

où

$$B(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1} \mathrm{d}z = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,.$$

Notons que la loi $\mathrm{Beta}(1,1)$ correspond à la loi uniforme sur $[0,1]$.

Exercice 0.3. Soit $X \sim \mathrm{Beta}(a, b)$. Montrer que

$$\mathbb{E}X = \frac{a}{a+b} \qquad \text{et} \qquad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Proposition 0.2 (Propriétés des lois Gamma et Beta)**.**

— *Si* $Y \sim \Gamma(p, \lambda)$ *et* $Z \sim \Gamma(q, \lambda)$ *sont indépendantes, alors*

$$Y + Z \sim \Gamma(p + q, \lambda).$$

En particulier, si $E_1, \dots, E_n$ *sont des variables i.i.d. de loi* $\mathrm{Exp}(\lambda)$*, alors*

$$\sum_{i=1}^{n} E_i \sim \Gamma(n, \lambda).$$

— *Si* $Y \sim \Gamma(p, \lambda)$ *alors, pour* $t > 0$*,*

$$tY \sim \Gamma\left(p, \frac{\lambda}{t}\right).$$

— *Si* $X \sim \Gamma(a, \lambda)$ *et* $Y \sim \Gamma(b, \lambda)$ *sont indépendantes, alors*

$$\frac{X}{X + Y} \sim \mathrm{Beta}(a, b).$$

Exercice 0.4. Montrer que si $E_1 \sim \mathrm{Exp}(\lambda)$ et $E_2 \sim \mathrm{Exp}(\lambda)$, alors la variable $\frac{E_1}{E_1+E_2}$ est uniformément distribuée sur $[0, 1]$.

Définition 0.10. Soit $K \geq 2$ un entier, et $\mathcal{S}_{K-1}$ le simplexe de dimension $K - 1$, i.e.

$$\mathcal{S}_{K-1} = \left\{ z = (z_1, \dots, z_{K-1}) \in \mathbb{R}^{K-1},\ z_1, \dots, z_{K-1} > 0,\ \sum_{i=1}^{K-1} z_i \leq 1 \right\}.$$

Soient $\alpha_1, \dots \alpha_K > 0$. La loi de Dirichlet de paramètre $(\alpha_1, \dots, \alpha_K)$, notée $\mathrm{Dir}(\alpha_1, \dots, \alpha_K)$, est la loi dont la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^{K-1}$ est donnée par

$$x = (x_1, \dots, x_{K-1}) \mapsto \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \mathbb{1}_{\{x \in \mathcal{S}_{K-1}\}},$$

où $x_K = 1 - x_1 - \cdots - x_{K-1}$.

Remarque 0.3. En fait, on dira souvent que c'est le vecteur $X = (X_1, \dots, X_K)$, où $X_K = 1 - X_1 - \cdots - X_{K-1}$, qui suit la loi $\mathrm{Dir}(\alpha_1, \dots, \alpha_K)$, mais il faut bien comprendre qu'il n'y a que $K - 1$ degrés de liberté (pour la loi à K paramètres).

La loi de Dirichlet peut être vue comme une généralisation de la loi Beta au cas multi-dimensionnel. On observe en particulier que pour $K = 2$, $\mathrm{Dir}(a, b) = \mathrm{Beta}(a, b)$. La loi de Dirichlet a pour support l'ensemble des vecteurs de taille K qui définissent une loi de probabilité sur un ensemble à K éléments.

Proposition 0.3 (Propriétés de la loi de Dirichlet)**.**

— *Les lois marginales d'une loi de Dirichlet sont des lois Beta. Plus précisément, si* $X = (X_1, \dots, X_K) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$*, alors pour* $i \in \{1, \dots, K\}$*,*

$$X_i \sim \mathrm{Beta}\left(\alpha_i, \sum_{k=1}^{K} \alpha_k - \alpha_i\right).$$

En particulier,

$$\mathbb{E}X_i = \frac{\alpha_i}{\sum_{k=1}^K \alpha_k}$$

— *Si* $Z_1 \sim \Gamma(\alpha_1, \lambda)$, ..., $Z_K \sim \Gamma(\alpha_K, \lambda)$ *sont indépendantes, alors, en notant* $Z = Z_1 + \cdots + Z_K$, *on a*

$$\left(\frac{Z_1}{Z}, \ldots, \frac{Z_K}{Z}\right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K).$$

1.6. Inégalités classiques. On sera souvent amené à contrôler la probabilité qu'une variable aléatoire soit plus grande qu'un certain seuil, ou bien qu'elle s'écarte de son espérance de plus d'un certain seuil. Pour cela, deux inégalités seront utiles.

Proposition 0.4 (Inégalité de Markov)**.** *Soit* X *une variable aléatoire réelle positive et* $a \in \mathbb{R}_+^*$. *On a*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

En particulier, pour X une variable aléatoire réelle et $p \in \mathbb{N}^*$, comme la fonction $x \mapsto x^p$ est croissante sur $\mathbb{R}_+$, on obtient

$$\mathbb{P}(|X| \geq a) = \mathbb{P}(|X|^p \geq a^p) \leq a^{-p}\mathbb{E}[|X|^p].$$

Un corollaire immédiat de l'inégalité de Markov est l'inégalité de Bienaymé-Tchebychev.

Proposition 0.5 (Inégalité de Bienaymé-Tchebychev)**.** *Soit* X *une variable aléatoire réelle et* $a \in \mathbb{R}_+^*$. *On a*

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq a\right) \leq \frac{\mathrm{Var}(X)}{a^2},$$

où $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$.

Exemple 0.4. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de loi de Bernoulli $\mathcal{B}(p)$ et soit $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Comme $\mathbb{E}\overline{X}_n = p$, on a, par l'inégalité de Bienaymé-Tchebychev, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{p(1-p)}{n\varepsilon^2}.$$

Pour $\varepsilon > 0$ et $p \in [0,1]$ fixés, on obtient donc une probabilité qui décroît en $1/n$. On peut obtenir une décroissance bien meilleure via l'inégalité ci-dessous.

Proposition 0.6 (Inégalité de Hoeffding)**.** *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes et bornées au sens où pour tout* $i = 1, \ldots, n$, *il existe des réels* $a_i \leq b_i$ *tels que* $a_i \leq X_i \leq b_i$ *p.s. Alors, pour tout* $\varepsilon \geq 0$,

$$\mathbb{P}\left(\overline{X}_n - \mathbb{E}\overline{X}_n \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Exemple 0.5. En reprenant l'exemple précédent, on peut prendre $a_i = 0$ et $b_i = 1$, et l'on obtient par l'inégalité de Hoeffding

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 n\right).$$

Pour $\varepsilon > 0$ et $p \in [0,1]$ fixés, on obtient donc une probabilité qui décroît exponentiellement vite en n, ce qui est bien plus rapide que la décroissance en $1/n$ obtenue via Bienaymé-Tchebychev.

1.7. Convergences. Pour $x \in \mathbb{R}^d,\ d \geq 1$, on note $\|x\| = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$ la norme euclidienne.

Définition 0.11. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. La suite (X_n) converge en probabilité vers X, ce que l'on note $X_n \xrightarrow{\mathbb{P}} X$, si

$$\forall\, \varepsilon > 0, \qquad \mathbb{P}\left(\|X_n - X\| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Définition 0.12. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. La suite (X_n) converge dans $\mathbb{L}^2$ vers X, ce que l'on note $X_n \xrightarrow{\mathbb{L}^2} X$, si

$$\mathbb{E}\left[\|X_n - X\|^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Définition 0.13. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. La suite (X_n) converge presque sûrement vers X, ce que l'on note $X_n \xrightarrow{\text{p.s.}} X$, si

$$\mathbb{P}\left(\left\{\omega \in \Omega,\ X_n(\omega) \underset{n\to\infty}{\longrightarrow} X(\omega)\right\}\right) = 1.$$

Proposition 0.7. *On a*

$$X_n \xrightarrow{\text{p.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{\mathbb{P}} X,$$

et

$$X_n \xrightarrow{\mathbb{L}^2} X \quad \Rightarrow \quad X_n \xrightarrow{\mathbb{P}} X,$$

Exercice 0.5. Démontrer la Proposition 0.7 (pour la deuxième implication, on pourra utiliser l'inégalité de Bienaymé-Tchebychev).

Proposition 0.8 (Loi forte des grands nombres)**.** *Soit $(X_n)_{n\geq 1}$ une suite de variables i.i.d. à valeurs dans $\mathbb{R}^d$, $d \geq 1$, avec $\mathbb{E}[\|X_1\|] < \infty$. Alors*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\text{p.s.}} \mathbb{E}X_1.$$

Définition 0.14. 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

$f : \mathbb{R}^d \to \mathbb{R}$ continue bornée,

$$\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)].$$

De même, on dira que (X_n) converge en loi vers une loi P si $\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)]$ pour $X \sim P$, pour toute fonction f continue bornée.

On rappelle que pour $A \subset \mathbb{R}^d$, la frontière de A est $\partial A = \overline{A} \setminus \overset{\circ}{A}$.

Proposition 0.9. *$X_n \xrightarrow{\mathcal{L}} X$ dans $\mathbb{R}^d$ si et seulement si pour tout borélien A de $\mathbb{R}^d$ pour lequel $\mathbb{P}(X \in \partial A) = 0$, on a*

$$\mathbb{P}(X_n \in A) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \in A).$$

Remarque 0.6. Si la loi de X est à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, alors la condition $\mathbb{P}(X \in \partial A) = 0$ est vérifiée pour tous les boréliens A de $\mathbb{R}^d$. Par exemple, si $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, alors pour tout intervalle I de $\mathbb{R}$,

$$\mathbb{P}(Z_n \in I) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(\mathcal{N}(0,1) \in I).$$

Notons aussi que si les variable X_n et X sont à valeurs dans $\mathbb{R}$, de fonctions de répartition respectives F_n et F, alors la convergence en loi est équivalente à la convergence simple des fonctions de répartitions en tout point de continuité de F : $X_n \xrightarrow{\mathcal{L}} X$ si et seulement si pour tout $x \in \mathbb{R}$ tel que F est continue en x, on a

$$F_n(x) \underset{n\to\infty}{\longrightarrow} F(x).$$

Proposition 0.10 (TCL dans $\mathbb{R}^d$)**.** *Soit (X_n) une suite de variables aléatoires i.i.d. dans $\mathbb{R}^d$, avec $\mathbb{E}[\|X_1\|^2] < \infty$. Soit $\mu = \mathbb{E}X_1$ et $\Sigma = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])^T\right]$. Alors*

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

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

Proposition 0.11 (Théorème de l'image continue)**.** *Soient X_n, X des variables aléatoires à valeurs dans $\mathbb{R}^d$. Soit $g : \mathbb{R}^d \to \mathbb{R}^k$ une fonction continue. Alors $X_n \xrightarrow{\mathcal{L}} X$ implique $g(X_n) \xrightarrow{\mathcal{L}} g(X)$. De même, $X_n \xrightarrow{\mathbb{P}} X$ implique $g(X_n) \xrightarrow{\mathbb{P}} g(X)$ et $X_n \xrightarrow{\text{p.s.}} X$ implique $g(X_n) \xrightarrow{\text{p.s.}} g(X)$.*

Proposition 0.12 (Lemme de Slutsky)**.** *Soient X_n, Y_n des suites de variables aléatoires réelles, X une variable aléatoire réelle, et $a \in \mathbb{R}$. Si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathbb{P}} a$, alors $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, a)$.*

Remarque 0.7. Pour a constante, $Z_n \xrightarrow{\mathcal{L}} a$ si et seulement si $Z_n \xrightarrow{\mathbb{P}} a$.

2. Outils de 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 $X_1, \dots, X_n$, dont la loi est inconnue.

2.1. Modèles statistiques.

Définition 0.15. Une expérience statistique est la donnée de
- une variable aléatoire $\mathbf{X}$ définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans un espace mesurable $(E, \mathcal{E})$.
- 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.

Dans l'approche fréquentiste, on suppose que la loi de $\mathbf{X}$ appartient au modèle, c'est-à-dire qu'il existe $\theta \in \Theta$ tel que $\mathbf{X}$ est de loi P_θ. L'inférence statistique consiste à chercher à estimer θ à partir d'une réalisation de la variable aléatoire $\mathbf{X}$. Souvent, $\mathbf{X}$ consiste en un n-uplet $\mathbf{X} = (X_1, \dots, X_n)$. Dans ce cas, l'espace $(E, \mathcal{E})$ et le modèle $\mathcal{P}$ dépendent de n (attention, cette dépendance en n ne sera pas toujours explicitée dans les notations).

Modèle du n-échantillon. Lorsque $\mathbf{X} = (X_1, \dots, X_n)$, on prendra souvent un modèle de la forme
$$\mathcal{P}_n = \left\{P_\theta^{\otimes n},\, \theta \in \Theta\right\},$$
où $P_\theta^{\otimes n} = P_\theta \otimes \cdots \otimes P_\theta$. Autrement dit, les variables $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées (en abrégé i.i.d.) selon la loi P_θ.

Si $\mathbf{X} \sim P_\theta^{\otimes n}$, on notera parfois (de manière abusive) $\mathbb{P}_\theta(\mathbf{X} \in A)$ au lieu de $\mathbb{P}(\mathbf{X} \in A)$, pour bien mettre en valeur le fait qu'il s'agit de la probabilité de l'événement $\{\mathbf{X} \in A\}$ quand $\mathbf{X} \sim P_\theta^{\otimes n}$. De même, pour $\varphi : E \to \mathbb{R}$ mesurable, on notera $\mathbb{E}_\theta\varphi(\mathbf{X})$ au lieu de $\mathbb{E}\varphi(\mathbf{X})$.

Définition 0.16. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est identifiable si pour tous $\theta, \theta' \in \Theta$,
$$P_\theta = P_{\theta'} \quad \Rightarrow \quad \theta = \theta'.$$
Autrement dit, la fonction $\theta \mapsto P_\theta$ est injective.

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

Définition 0.17. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est dominé s'il existe une mesure σ-finie μ sur E telle que, pour tout $\theta \in \Theta$, $P_\theta \ll \mu$. Toutes les lois P_θ admettent alors une densité p_θ par rapport à μ, soit
$$\mathrm{d}P_\theta(x) = p_\theta(x)\mathrm{d}\mu(x).$$

Dans la suite, nous travaillerons toujours avec des modèles dominés, et paramétriques au sens où $\Theta \subset \mathbb{R}^d$.

Exemples de modèles

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

- Le modèle des lois de Bernoulli (tirage à pile ou face) :
$$\mathcal{P} = \{\mathcal{B}(\theta),\ \theta \in [0,1]\},$$
où $\mathcal{B}(\theta)$ est la loi de Bernoulli de paramètre θ. C'est la loi discrète définie par :
$$\mathbb{P}(X=1) = \theta, \qquad \mathbb{P}(X=0) = 1-\theta,$$
ce que l'on note aussi $\mathcal{B}(\theta) = (1-\theta)\delta_0 + \theta\delta_1$, où δ_a est la mesure de Dirac en $a \in \mathbb{R}$. C'est un modèle dominé par $\mu = \delta_0 + \delta_1$, de densité $p_\theta(x) = (1-\theta)\mathbb{1}_{x=0} + \theta\mathbb{1}_{x=1}$. Le modèle est identifiable. Une façon de le voir est de remarquer que si $P_\theta = P_{\theta'}$, alors $\mathbb{E}_\theta X = \mathbb{E}_{\theta'} X$ (si deux lois sont égales, tous leurs moments sont égaux). Or $\mathbb{E}_\theta = \theta$. Donc $\theta = \theta'$.
- Le modèle gaussien :
$$\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}.$$
C'est un modèle dominé par μ la mesure de Lebesgue sur $\mathbb{R}$: $\mathrm{d}P_\theta(x) = p_\theta(x)\mathrm{d}x$ avec
$$p_\theta(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\theta)^2}{2}}.$$
Il s'agit aussi d'un modèle identifiable. En effet, par le même argument que pour le modèle des lois de Bernoulli, on peut remarquer que si $P_\theta = P_{\theta'}$, alors $\mathbb{E}_\theta X = \theta = \mathbb{E}_{\theta'} X = \theta'$. On peut aussi utiliser le fait que si deux lois à densité par rapport à μ sont égales, alors leurs densités 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.
- Le modèle gaussien avec moyenne *et* variance inconnues :
$$\mathcal{P} = \left\{\mathcal{N}(\mu, \sigma^2),\ (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\right\}.$$
Le paramètre du modèle est $\theta = (\mu, \sigma^2)$ et l'espace des paramètres est $\Theta = \mathbb{R} \times \mathbb{R}_+^*$. (Montrer qu'il s'agit d'un modèle identifiable.)
- Le modèle gaussien en dimension $d \geq 1$. Il s'agit de l'ensemble des lois $\mathcal{N}(\mu, \Sigma)$ avec $\mu \in \mathbb{R}^d$ réels et Σ une matrice $d \times d$ symétrique semi-définie positive.
- Les 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 une variable aléatoire réelle de densité f fixée connue. (Montrer que la densité d'une telle variable X est $\sigma^{-1} f(\frac{\cdot - \mu}{\sigma})$.)
- Le modèle des lois gamma $\Gamma(p, \lambda)$ avec $p > 0$ fixé :
$$\mathcal{P} = \left\{\Gamma(p, \lambda),\ \lambda \in \mathbb{R}_+^*\right\}.$$
- Le modèle des lois de Poisson :
$$\mathcal{P} = \{\mathcal{P}(\theta),\ \theta > 0\},$$

où $\mathcal{P}(\theta)$ est la loi de Poisson de paramètre θ, définie par

$$\forall k \in \mathbb{N},\ \mathbb{P}(X = k) = \frac{e^{-\theta}\theta^k}{k!}\cdot$$

- Le modèle « non-lisse » des lois uniformes est

$$\mathcal{P} = \left\{\text{Unif}[0,\theta],\ \theta \in \mathbb{R}_+^*\right\}$$

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

2.2. Estimateur, consistance, normalité asymptotique.

Définition 0.18. Dans une expérience statistique $(\mathbf{X}, \mathcal{P})$, où $\mathbf{X}$ est à valeurs dans $(E, \mathcal{E})$, et $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est une famille de lois sur E, un estimateur $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ est une fonction mesurable de $\mathbf{X}$, à valeurs dans l'espace des paramètres Θ (plus précisément, la fonction $\widehat{\theta}$ est mesurable de $(E, \mathcal{E})$ dans $(\Theta, \mathcal{B}(\Theta))$ où $\mathcal{B}(\Theta)$ est la tribu des boréliens).

En pratique, nous disposerons d'une suite d'expériences statistiques $\left(\mathbf{X}^{(n)}, \mathcal{P}_n\right)$, indicée par n la taille de l'échantillon. Cela nous conduira donc à construire des suites d'estimateurs $\widehat{\theta}_n$. Lorsqu'il n'y aura pas d'ambiguïté, cette dépendance en n sera cependant souvent omise des notations.

Exemple 0.8. Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$, si l'on dispose d'observations $\mathbf{X} = (X_1, \ldots, X_n)$, alors $\widehat{\theta}_n(\mathbf{X}) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $\widetilde{\theta}_n(\mathbf{X}) = 1$ sont tous les deux des estimateurs de θ.

Définition 0.19. Dans une suite d'expériences statistiques $(\mathbf{X}^{(n)}, \mathcal{P}_n)$ avec $\mathcal{P}_n = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$, la suite d'estimateurs $\widehat{\theta}_n$ est dite consistante si, pour tout $\theta \in \Theta$, quand $\mathbf{X}^{(n)} \sim P_\theta^{\otimes n}$,

$$\widehat{\theta}_n(\mathbf{X}^{(n)}) \xrightarrow{\mathbb{P}} \theta.$$

Définition 0.20. Dans une suite d'expériences statistiques $(\mathbf{X}^{(n)}, \mathcal{P}_n)$ avec $\mathcal{P}_n = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$, la suite d'estimateurs $\widehat{\theta}_n$ est dite asymptotiquement normale si pour tout $\theta \in \Theta$, il existe une matrice Σ_θ symétrique définie positive, telle que, quand $\mathbf{X}^{(n)} \sim P_\theta^{\otimes n}$,

$$\sqrt{n}\left(\widehat{\theta}_n(\mathbf{X}^{(n)}) - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta).$$

Par abus de langage, on dira souvent que l'estimateur $\widehat{\theta}_n$ est consistant ou asymptotiquement normal pour dire que la suite d'estimateurs est consistante ou asymptotiquement normale.

Exercice 0.6. Montrer que si $\widehat{\theta}_n$ est asymptotiquement normal, alors $\widehat{\theta}_n$ est consistant.

2.3. Le risque quadratique.

Nous introduisons maintenant une notion de risque qui sera précisée dans la suite du cours.

Définition 0.21. Le risque quadratique d'un estimateur $\widehat{\theta}$ est la fonction $\theta \mapsto \mathbf{R}(\theta, \widehat{\theta})$ définie sur Θ par

$$\mathbf{R}(\theta, \widehat{\theta}) = \mathbb{E}_\theta \left[\|\widehat{\theta}(\mathbf{X}) - \theta\|^2 \right] = \int_E \|\widehat{\theta}(x) - \theta\|^2 \mathrm{d}P_\theta(x).$$

Lorsque $\Theta \subset \mathbb{R}$, le risque quadratique s'écrit

$$\mathbf{R}(\theta, \widehat{\theta}) = \mathbb{E}_\theta \left[\left(\widehat{\theta}(\mathbf{X}) - \theta \right)^2 \right] = \int_E (\widehat{\theta}(x) - \theta)^2 \mathrm{d}P_\theta(x).$$

Typiquement, un « bon » estimateur est un estimateur qui a un petit risque quadratique. Il ne faut cependant pas oublier que le risque quadratique est une fonction définie sur l'ensemble Θ : le risque quadratique peut être petit pour certaines valeurs de θ, grand pour d'autres.

Exemple 0.9. Considérons l'expérience $(\mathbf{X}, \mathcal{P})$ avec $\mathcal{P} = \{\mathcal{B}(\theta)^{\otimes n},\ \theta \in [0,1]\}$, i.e. $\mathbf{X} = (X_1, \ldots, X_n)$ est une suite i.i.d. de variables de Bernoulli de paramètre θ. Un estimateur naturel est la moyenne empirique

$$\widehat{\theta}_n(\mathbf{X}) = \overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Notons d'abord que cet estimateur est consistant (loi des grands nombres), et asymptotiquement normal (TCL). Cherchons à calculer son risque quadratique. Soit $\theta \in [0,1]$ et $\mathbf{X} \sim P_\theta^{\otimes n}$. En remarquant que $\mathbb{E}\widehat{\theta}_n(\mathbf{X}) = \theta$, on a

$$\mathbf{R}(\theta, \widehat{\theta}_n) = \mathbb{E}\left[(\widehat{\theta}_n(\mathbf{X}) - \theta)^2 \right] = \mathrm{Var}\left(\widehat{\theta}_n(\mathbf{X}) \right) = \frac{\theta(1-\theta)}{n}\,.$$

On remarque ainsi que le risque est minimal pour $\theta = 0$ et $\theta = 1$. Cela est naturel puisque dans ces deux cas, il n'y a pas d'aléatoire : presque sûrement, on n'observe soit que des piles, soit que des faces et l'estimateur $\widehat{\theta}_n(\mathbf{X})$ ne peut pas se tromper. Le risque quadratique est maximal en $\theta = 1/2$. On s'intéresse bien sûr aussi à la manière dont le risque dépend de n, la taille de l'échantillon. Ici, on voit que, pour tout $\theta \in \Theta$, le risque $\mathbf{R}(\theta, \widehat{\theta}_n)$ décroît en $1/n$.

Contrôler le risque quadratique permet notamment de contrôler la probabilité que l'estimateur $\widehat{\theta}(\mathbf{X})$ soit « loin » de θ. En effet, par l'inégalité de Markov, on a, pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta \left(|\widehat{\theta}(\mathbf{X}) - \theta| \geq \varepsilon \right) \leq \frac{\mathbb{E}_\theta[(\widehat{\theta}(\mathbf{X}) - \theta)^2]}{\varepsilon^2} = \frac{\mathbf{R}(\theta, \widehat{\theta})}{\varepsilon^2}.$$

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

Proposition 0.13 (Décomposition biais-variance)**.** *Soit $(\mathbf{X}, \mathcal{P})$ une expérience statistique avec $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ et soit $\widehat{\theta}(\mathbf{X})$ un estimateur. Alors, pour tout $\theta \in \Theta$, si $\mathbf{X} \sim P_\theta$,*

$$\mathbf{R}(\theta, \widehat{\theta}) = \mathbb{E}\left[\left\| \widehat{\theta}(\mathbf{X}) - \mathbb{E}\left[\widehat{\theta}(\mathbf{X}) \right] \right\|^2 \right] + \left\| \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta \right\|^2.$$

La fonction $\theta \mapsto \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta$ s'appelle le biais de $\widehat{\theta}$.

A noter que lorsque $\Theta \subset \mathbb{R}$, le risque quadratique s'écrit comme la somme de la variance et du biais au carré, i.e.

$$\mathbf{R}(\theta, \widehat{\theta}) = \mathrm{Var}_\theta\left(\widehat{\theta}(\mathbf{X})\right) + \left(\mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right)^2.$$

DÉMONSTRATION. Soit $\theta \in \Theta$ et $\mathbf{X} \sim P_\theta$. On peut toujours décomposer

$$\widehat{\theta}(\mathbf{X}) - \theta = \widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X}) + \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta.$$

En prenant la norme au carré puis l'espérance, et en utilisant la linéarité de l'espérance, on obtient

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\theta}(\mathbf{X}) - \theta\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X})\right\|^2\right] + \left\|\mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right\|^2 + 2\mathbb{E}\left[\left\langle \widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X}), \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right\rangle\right] \\ &= \mathbb{E}\left[\left\|\widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X})\right\|^2\right] + \left\|\mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right\|^2.\end{aligned}$$

En effet, l'espérance du produit scalaire est nulle car le terme de biais $\mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta$ est déterministe et par linéarité

$$\mathbb{E}\left[\left\langle \widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X}), \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right\rangle\right] = \left\langle \mathbb{E}\left[\widehat{\theta}(\mathbf{X}) - \mathbb{E}\widehat{\theta}(\mathbf{X})\right], \mathbb{E}\widehat{\theta}(\mathbf{X}) - \theta\right\rangle = 0\,.$$

■

Exemple 0.10. Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\theta, 1)$, pour $\theta \in \mathbb{R}$.

- Considérons l'estimateur constant $\widetilde{\theta}_n = \theta_0$ pour un certain $\theta_0 \in \mathbb{R}$ fixé. On a

  $$\mathbf{R}(\theta, \widetilde{\theta}_n) = \mathbb{E}(\theta - \theta_0)^2 = (\theta - \theta_0)^2.$$

  Le risque est imbattable si $\theta = \theta_0$ puisqu'il est nul, mais dès que $\theta \neq \theta_0$ il est strictement positif et ne décroît pas avec la taille de l'échantillon (pire, il explose pour $\theta \to \pm\infty$).

- Considérons maintenant l'estimateur $\widehat{\theta}_n(\mathbf{X}) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On remarque d'abord que cet estimateur est sans biais : pour tout $\theta \in \mathbb{R}$, si $\mathbf{X} \sim \mathcal{N}(\theta, 1)^{\otimes n}$, alors $\mathbb{E}\overline{X}_n = \theta$. Ainsi

  $$\mathbf{R}(\theta, \widehat{\theta}_n) = \mathrm{Var}(\widehat{\theta}_n(\mathbf{X})) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{1}{n}\,\mathrm{Var}(X_1) = \frac{1}{n},$$

  en utilisant que pour des variables indépendantes, la variance de la somme est la somme des variances. Le risque de $\widehat{\theta}_n$ ne dépend par de θ et tend vers 0 à vitesse $1/n$ quand $n \to \infty$.

Exercice 0.7. Soit $X \sim \mathcal{B}(n, \theta)$, la loi binomiale de paramètres $n \in \mathbb{N}^*$, et $\theta \in [0, 1]$ (on suppose n connu, seul θ est inconnu). On rappelle que X a la même loi que $\sum_{i=1}^n \varepsilon_i$ où les ε_i sont i.i.d. de loi de Bernoulli $\mathcal{B}(\theta)$. Pour $\widehat{\theta} = X/n$,

(1) écrire la décomposition biais-variance.

(2) montrer que pour tout $\theta \in [0, 1]$, $\mathbf{R}(\theta, \widehat{\theta}) \leq 1/(4n)$.

2.4. Intervalles et régions de confiance.

Définition 0.22. Soit $\alpha \in]0,1[$.

— Cas $\Theta \subset \mathbb{R}$. Un intervalle de confiance de niveau (au moins) $1-\alpha$ est un intervalle aléatoire $I(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$ où $a(\mathbf{X}), b(\mathbf{X})$ sont des statistiques à valeurs dans $\mathbb{R}$ vérifiant

$$\forall\, \theta \in \Theta,\ \mathbb{P}_\theta\left(\theta \in I(\mathbf{X})\right) \geq 1-\alpha.$$

— Cas $\Theta \subset \mathbb{R}^d$. Une région de confiance de niveau (au moins) $1-\alpha$ est un sous-ensemble aléatoire $\mathcal{R}(\mathbf{X}) \subset \Theta$ vérifiant

$$\forall\, \theta \in \Theta,\ \mathbb{P}_\theta\left(\theta \in \mathcal{R}(\mathbf{X})\right) \geq 1-\alpha.$$

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 telle que le niveau de confiance reste au moins de $1-\alpha$.

Pour construire un intervalle de confiance à partir d'un estimateur $\widehat{\theta}(\mathbf{X})$, on peut chercher à contrôler la probabilité de déviation de $\widehat{\theta}(\mathbf{X})$ par rapport à θ. Comme vu précédemment, par l'inégalité de Markov, on a, pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta\left(|\widehat{\theta}(\mathbf{X}) - \theta| > \varepsilon\right) \leq \frac{\mathbf{R}(\widehat{\theta}, \theta)}{\varepsilon^2}.$$

Exemple 0.11. Modèle de Bernoulli : on observe $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. de loi $\mathcal{B}(\theta)$, $\theta \in [0,1]$. On pose $\widehat{\theta}(\mathbf{X}) = \overline{X}_n$. D'après l'exercice ci-dessus, $\mathbf{R}(\widehat{\theta}, \theta) \leq \frac{1}{4n}$, donc pour tout $\varepsilon > 0$,

$$\mathbb{P}_\theta\left(|\widehat{\theta}(\mathbf{X}) - \theta| > \varepsilon\right) \leq \frac{1}{4n\varepsilon^2}$$

soit aussi, en prenant l'événement complémentaire,

$$\mathbb{P}_\theta\left(\theta \in [\widehat{\theta}(\mathbf{X}) - \varepsilon, \widehat{\theta}(\mathbf{X}) + \varepsilon]\right) \geq 1 - \frac{1}{4n\varepsilon^2}.$$

Pour obtenir un intervalle de confiance de niveau $1-\alpha$, il suffit de choisir ε de sorte que $\alpha = 1/(4n\varepsilon^2)$, soit $\varepsilon = 1/\sqrt{4n\alpha}$. On a donc obtenu que

$$I(\mathbf{X}) = \left[\widehat{\theta}(\mathbf{X}) - \frac{1}{\sqrt{4n\alpha}}, \widehat{\theta}(\mathbf{X}) + \frac{1}{\sqrt{4n\alpha}}\right] = \left[\widehat{\theta}(\mathbf{X}) \pm \frac{1}{\sqrt{4n\alpha}}\right]$$

est un intervalle de confiance de niveau au moins $1-\alpha$.

L'intervalle $I(\mathbf{X})$ ne peut bien sûr pas dépendre de θ que l'on ne connaît pas. Il ne doit dépendre que de quantités connues. Par exemple de α, la probabilité d'erreur recherchée, ou de n, la taille de l'échantillon, ou bien sûr de $\mathbf{X}$, les données observées. Or en général, le risque $\mathbf{R}(\widehat{\theta}, \theta)$ dépend de θ et ne permet donc pas toujours de construire directement un intervalle de confiance. Par exemple dans le modèle de Bernoulli ci-dessus, il vaut $\theta(1-\theta)/n$. C'est pourquoi on a dû le majorer pour obtenir une quantité indépendante de θ.

On peut aussi utiliser d'autres inégalités plus fines que celle de Markov, comme par exemple l'inégalité de Hoeffding (voir TDs). On peut parfois aussi utiliser directement la loi de $\widehat{\theta}$ si elle

est connue (modulo bien sûr la connaissance de θ). Ceci n'est pas très fréquent, mais un cas particulier est le modèle gaussien.

Exemple 0.12. Modèle gaussien : on observe $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. de loi $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. On pose $\widehat{\theta}(\mathbf{X}) = \overline{X}_n$. On sait que $\overline{X}_n \sim \mathcal{N}(\theta, \frac{1}{n})$, soit $\sqrt{n}(\overline{X}_n - \theta) \sim \mathcal{N}(0, 1)$. Ainsi, en notant Φ la fonction de répartition d'une variable $\mathcal{N}(0, 1)$ et en utilisant la symétrie de la loi normale, on a

$$\mathbb{P}_\theta \left(\sqrt{n} |\widehat{\theta}(\mathbf{X}) - \theta| > \Phi^{-1}(1 - \alpha/2) \right) = \alpha \, .$$

On a donc obtenu que

$$I(\mathbf{X}) = \left[\widehat{\theta}(\mathbf{X}) \pm \frac{\Phi^{-1}(1 - \alpha/2)}{\sqrt{n}} \right]$$

est un intervalle de confiance de niveau (exactement) $1 - \alpha$.

Parfois, on ne connaît pas la loi de $\widehat{\theta}_n$ pour n fixé mais on connaît sa loi limite quand $n \to \infty$. Cela permet de construire des intervalles de confiance dits *asymptotiques*.

Définition 0.23. Soit $\alpha > 0$.

— Cas $\Theta \subset \mathbb{R}$. Un intervalle de confiance asymptotique de niveau (au moins) $1 - \alpha$ est un intervalle aléatoire $I(\mathbf{X}^{(n)})$ vérifiant

$$\forall \, \theta \in \Theta, \ \liminf_{n \to \infty} \mathbb{P}_\theta \left(\theta \in I(\mathbf{X}^{(n)}) \right) \geq 1 - \alpha.$$

— Cas $\Theta \subset \mathbb{R}^d$. Une région de confiance asymptotique de niveau (au moins) $1 - \alpha$ est un sous-ensemble aléatoire $\mathcal{R}(\mathbf{X}^{(n)}) \subset \Theta$ vérifiant

$$\forall \, \theta \in \Theta, \ \liminf_{n \to \infty} \mathbb{P}_\theta \left(\theta \in \mathcal{R}(\mathbf{X}^{(n)}) \right) \geq 1 - \alpha.$$

Exemple 0.13. Revenons à l'exemple du modèle de Bernoulli : $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. de loi $\mathcal{B}(\theta)$, $\theta \in [0, 1]$ et $\widehat{\theta}_n(\mathbf{X}) = \overline{X}_n$. D'après le TCL, on sait que

$$\sqrt{n}(\overline{X}_n - \theta) \ \xrightarrow{\mathcal{L}} \ \mathcal{N}(0, \theta(1 - \theta)),$$

soit

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

D'autre part, on sait par la loi des grands nombres que l'estimateur $\widehat{\theta}_n$ est consistant : $\widehat{\theta}_n \ \xrightarrow{\mathbb{P}} \ \theta$, et par le théorème de l'image continue,

$$\sqrt{\frac{\theta(1 - \theta)}{\widehat{\theta}_n(1 - \widehat{\theta}_n)}} \ \xrightarrow{\mathbb{P}} \ 1.$$

Ainsi, par le lemme de Slutsky,

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

On a donc, en notant $q_\alpha = \Phi^{-1}(1-\alpha/2)$,

$$\mathbb{P}_\theta\left(\frac{\sqrt{n}\left|\widehat{\theta}_n - \theta\right|}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} \leq q_\alpha\right) \underset{n\to\infty}{\longrightarrow} 1-\alpha.$$

L'intervalle

$$I(\mathbf{X}) = \left[\widehat{\theta}_n \pm \frac{q_\alpha\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}}{\sqrt{n}}\right]$$

est donc un intervalle de confiance asymptotique de niveau $1-\alpha$.

De façon plus générale, on a le résultat suivant.

Proposition 0.14. *Supposons $\Theta \subset \mathbb{R}$ et soit $\widehat{\theta}_n = \widehat{\theta}_n(\mathbf{X})$ un estimateur asymptotiquement normal, i.e.*

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

Supposons de plus que la fonction $\theta \mapsto \sigma^2(\theta)$ est continue et notons $q_\alpha = \Phi^{-1}(1-\alpha/2)$ pour $\alpha \in]0,1[$. Alors

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

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

$$\liminf_{n\to\infty} \mathbb{P}_\theta\left(\theta \in I(\mathbf{X})\right) = 1-\alpha.$$

Démonstration. On constate que

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

Comme $\widehat{\theta}_n$ est asymptotiquement normal, il est consistant (voir exercice en Section 1), donc $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$. Par image continue (Proposition 0.11), on en déduit que $\sigma(\widehat{\theta}_n) \xrightarrow{\mathbb{P}} \sigma(\theta)$. Par ailleurs, on sait que

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

Grâce au lemme de Slutsky (Proposition 0.12), on en déduit

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

La proposition 0.9 permet d'en déduire

$$\mathbb{P}_\theta\left(-q_\alpha \leq \frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sigma(\widehat{\theta}_n)} \leq q_\alpha\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(|\mathcal{N}(0,1)| \leq q_\alpha\right) = 1-\alpha,$$

soit $\mathbb{P}_\theta\left(\theta \in I(\mathbf{X})\right) \underset{n\to\infty}{\longrightarrow} 1-\alpha$, ce qu'il fallait démontrer. ■

2.5. Vraisemblance. Supposons le modèle dominé par rapport à une mesure dominante μ, i.e. $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$, et soit $\mathbf{X} = (X_1, \dots, X_n) \sim P_\theta^{\otimes n}$. La densité du n-uplet $\mathbf{X}$ par rapport à $\mu^{\otimes n}$ est donc $p_\theta(x_1) \dots p_\theta(x_n)$. Cette densité, vue comme une fonction de θ, et évaluée aux points d'observation $X_1, \dots, X_n$, s'appelle *vraisemblance*.

Définition 0.24. Soit $\mathbf{X} = (X_1, \dots, X_n) \sim P_\theta^{\otimes n}$, avec $\theta \in \Theta$. La vraisemblance en $\mathbf{X}$ est la fonction de Θ dans $[0, 1]$ définie par

$$\theta \mapsto L_\theta(\mathbf{X}) = \prod_{i=1}^{n} p_\theta(X_i).$$

On manipule généralement plus facilement la log-vraisemblance, définie par

$$\theta \mapsto \ell_\theta(\mathbf{X}) = \sum_{i=1}^{n} \log(p_\theta(X_i)).$$

Définition 0.25. Dans un modèle dominé, un estimateur du maximum de vraisemblance (EMV) est, sous réserve d'existence, un élément $\widehat{\theta}(\mathbf{X})$ de Θ qui vérifie

$$\widehat{\theta}(\mathbf{X}) \in \arg\max_{\theta \in \Theta} L_\theta(\mathbf{X}),$$

ou de manière équivalente

$$\widehat{\theta}(\mathbf{X}) \in \arg\max_{\theta \in \Theta} \ell_\theta(\mathbf{X}).$$

Exercice 0.8. Montrer que, dans le modèle de Bernoulli $\mathcal{P} = \{\mathcal{B}(\theta)^{\otimes n},\ \theta \in [0, 1]\}$, l'EMV est unique et est donné par $\widehat{\theta}(\mathbf{X}) = \overline{X}_n$.

3. Lois conditionnelles

On commence par rappeler que, pour A, B des événements avec $\mathbb{P}(B) > 0$, la probabilité de l'événement A sachant que l'événement B est réalisé est définie par

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

3.1. Le cas discret. Soient E et F deux ensembles dénombrables (on peut penser à $\mathbb{N}$ pour fixer les idées), et soient X et Y deux variables aléatoires à valeurs dans E et F respectivement.

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 $\mathbb{P}(X = e)$ et $\mathbb{P}(Y = f)$ pour tous les éléments possibles $e \in E$ et $f \in F$. Si, pour $x \in E$, Q_x est la loi $\mathcal{L}(Y \mid X = x)$ que l'on cherche à définir, il suffit donc aussi de se donner $Q_x(\{y\})$ pour tout $y \in F$. On définit tout simplement ces quantités à l'aide de la formule ci-dessus pour la probabilité conditionnelle d'un événement sachant un autre événement.

Définition 0.26. Soit $x \in E$ tel que $\mathbb{P}(X = x) > 0$. La loi conditionnelle de Y sachant $\{X = x\}$, i.e. $\mathbb{P}(Y \in \cdot \mid X = x)$, parfois notée $\mathcal{L}(Y \mid X = x)$, est définie, pour tout $y \in F$, par

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}.$$

Exemple 0.14. Soient Y, Z deux variables aléatoires indépendantes de lois $Y \sim \mathcal{B}(1/2)$ et $Z \sim \mathcal{B}(1/2)$. 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,

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

Par ailleurs, comme Y ne prend que les valeurs 0 ou 1, on en déduit que $\mathbb{P}(Y = 0 \mid X = 1) = 1 - \mathbb{P}(Y = 1 \mid X = 1) = \frac{1}{2}$. Ainsi

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

Exercice 0.9. En procédant de la même manière, montrer que

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

Par extension, on définit la loi conditionnelle de Y sachant X, i.e. $\mathbb{P}(Y \in \cdot \mid X)$, parfois 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,

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

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

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

3.2. Le cas à densité. 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 $(E, \mathcal{E})$ et une mesure β positive σ-finie sur $(F, \mathcal{F})$;
— une variable aléatoire X sur E et une variable aléatoire Y sur F.
On suppose que le couple (X, Y) admet une densité notée $h(x, y)$ par rapport à $\alpha \otimes \beta$, ce que l'on écrit aussi, si P désigne la loi du couple,

$$\mathrm{d}P(x, y) = h(x, y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

Proposition 0.15. *Dans le cadre ci-dessus, la loi de X seule, appelée loi marginale de X, est la loi de densité f donnée par*

$$f(x) = \int_F h(x, y)\mathrm{d}\beta(y).$$

Démonstration. Pour toute fonction φ mesurable bornée, en utilisant le théorème de Fubini,

$$\begin{aligned}\mathbb{E}[\varphi(X)] &= \int_{E\times F} \varphi(x)\mathrm{d}P(x,y)\\ &= \int_{E\times F} \varphi(x)h(x,y)\mathrm{d}(\alpha\otimes\beta)(x,y)\\ &= \int_E \varphi(x)\left\{\int_F h(x,y)\mathrm{d}\beta(y)\right\}\mathrm{d}\alpha(x) = \int_E \varphi(x)f(x)\mathrm{d}\alpha(x).\end{aligned}$$

■

De même, la loi marginale de Y est la loi dont la densité sur F par rapport à β est donnée par $g(y) = \int_E h(x,y)\mathrm{d}\alpha(x)$. À partir de la loi du couple (X,Y), on peut donc facilement déduire les lois individuelles de X et Y. En revanche, la donnée des deux lois marginales ne permet absolument pas de déterminer la loi du couple.

Dans le cas général de variables à densité, l'événement $\{X = x\}$, pour $x \in E$, peut être de probabilité nulle. Par exemple, si X admet une densité f par rapport à la mesure de Lebesgue sur $E = \mathbb{R}$, alors $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$. On ne peut donc pas conditionner par rapport à cette événement. Cependant, si $f(x) > 0$, on peut définir ce qu'on appelle la densité conditionnelle de Y sachant $X = x$.

Définition 0.27. Soit $x \in E$ tel que $f(x) > 0$. La loi conditionnelle de Y sachant $X = x$, notée $\mathcal{L}(Y \mid X = x)$, est la loi dont la densité sur F par rapport à β est donnée par

$$g_x(y) = \frac{h(x,y)}{f(x)} = \frac{h(x,y)}{\int_F h(x,y)\mathrm{d}\beta(y)}.$$

On notera parfois $g(y \mid x)$ au lieu de $g_x(y)$. Notons que par définition, la fonction $y \mapsto g(y \mid x)$ est une densité par rapport à β, soit $\int_F g(y \mid x)\mathrm{d}\beta(y) = 1$.

Pour avoir une quantité définie pour tous les x de E, on peut étendre la définition de $g_x(y)$ au cas où $f(x) = 0$ en posant le quotient ci-dessus égal à une valeur quelconque (par exemple 0) lorsque $f(x) = 0$. Ces points x n'auront typiquement pas d'incidence dans les calculs car l'ensemble des x tels que $f(x) = 0$ est un ensemble de mesure nulle sous $\mathcal{L}(X)$ ($\int_E \mathbb{1}_{f(x)=0} f(x)\mathrm{d}\alpha(x) = 0$).

Exercice 0.10. Vérifier que le cas discret est un cas particulier de la formule ci-dessus, pour lequel E et F sont dénombrables, et α, β sont les mesures de comptage sur E et F respectivement, $\alpha = \sum_{e\in E} \delta_e$ et $\beta = \sum_{f\in F} \delta_f$.

Comme dans le cas discret, on définit par extension la loi de Y sachant X comme la loi égale à $\mathcal{L}(Y \mid X = x)$ quand $X = x$. La densité conditionnelle de Y sachant X est définie comme la densité sur F par rapport à β donnée par

$$g_X(y) = g(y \mid X) = \begin{cases} \frac{h(X,y)}{f(X)} & \text{si } f(X) > 0,\\ 0 & \text{si } f(X) = 0.\end{cases}$$

En fait, on écrira simplement $g_X(y) = \frac{h(X,y)}{f(X)}$ puisque $f(X) > 0$ presque sûrement. À partir de la densité conditionnelle de $Y \mid X$ et de la densité marginale de X, on retrouve la densité jointe du couple (X, Y), puisque par définition $h(x, y) = g_x(y)f(x)$.

Exemple 0.15. Soit un couple (X, Y) de variables aléatoires sur $\mathbb{R}_+ \times \mathbb{R}_+$ de densité

$$h(x, y) = x\,\mathrm{e}^{-x(y+1)}$$

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 $h(x, y)$ par la densité marginale

$$f(x) = \int_0^\infty x\,\mathrm{e}^{-x(y+1)}\mathrm{d}y = \mathrm{e}^{-x}.$$

Ainsi

$$g_x(y) = \frac{x\,\mathrm{e}^{-x(y+1)}}{\mathrm{e}^{-x}} = x\,\mathrm{e}^{-xy}$$

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

Exercice 0.11. Déterminer la densité de la loi marginale de Y et montrer que la loi conditionnelle de $X \mid Y$ est une loi Gamma $\Gamma(2, Y + 1)$.

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

$$g_x(y) = \frac{h(x, y)}{f(x)} \propto h(x, y),$$

et

$$h(x, y) = x\,\mathrm{e}^{-x}\,\mathrm{e}^{-xy} \propto \mathrm{e}^{-xy}.$$

La densité de Y sachant $X = x$ est donc proportionnelle à e^{-xy}. Or la loi dont la densité en y est proportionnelle à e^{-xy} est bien la loi $\mathcal{E}(x)$, de densité $y \mapsto xe^{-xy}$. Cette méthode évite de devoir calculer la densité marginale $f(x)$. Dans cet exemple, ce calcul était facile mais ce n'est pas toujours le cas, nous verrons d'autres exemples au prochain chapitre. Il faut cependant faire attention quand on utilise ce symbole. En effet, nous aurons parfois affaire à des fonctions dépendant de nombreuses variables, et il faut bien savoir quelles variables on considère comme étant des constantes par rapport à la variable d'intérêt. C'est pourquoi nous préciserons parfois la notation en écrivant $f(x, y) \propto_y g(x, y)$ pour bien signifier que l'on considère les fonctions $y \mapsto f(x, y)$ et $y \mapsto g(x, y)$, et que c'est tout ce qui ne dépend pas de y qui est considéré comme constant. Cela signifie que pour tout x, il existe $c(x)$ tel que pour tout y,

$$f(x, y) = c(x)g(x, y).$$

Notons que si la fonction $y \mapsto f(x, y)$ est une densité sur F par rapport à une mesure β (comme ci-dessus pour $f(x, y) = g_x(y)$), alors la constante $c(x)$ est simple à retrouver après

coup : comme

$$\int_F f(x,y)\mathrm{d}\beta(y) = 1,$$

on a nécessairement

$$c(x) = \frac{1}{\int_F g(x,y)\mathrm{d}\beta(y)}.$$

3.3. Espérance conditionnelle. On rappelle l'abréviation $g(y \mid x)$ pour la densité de Y sachant $X = x$.

Définition 0.28. Soit $\varphi : F \to \mathbb{R}$ une fonction mesurable telle que $\mathbb{E}\varphi(Y) < \infty$. On définit

$$\mathbb{E}[\varphi(Y) \mid X] = \int_F \varphi(y) g(y \mid X)\mathrm{d}\beta(y).$$

Proposition 0.16. *Pour toute fonction $\psi : E \times F \to \mathbb{R}$ mesurable, telle que $\psi(X,Y)$ est intégrable, on a*

$$\begin{aligned}\mathbb{E}[\psi(X,Y)] &= \mathbb{E}\left[\mathbb{E}[\psi(X,Y) \mid X]\right] \\ &= \int_E \int_F \psi(x,y) g(y \mid x)\mathrm{d}\beta(y) f(x)\mathrm{d}\alpha(x)\,.\end{aligned}$$

En particulier, si $\psi(X,Y) = \psi_1(X)\psi_2(Y)$, avec ψ_1, ψ_2 mesurables et telles que $\psi_1(X)$, $\psi_2(Y)$ et $\psi_1(X)\psi_2(Y)$ sont intégrables, on a

$$\mathbb{E}\left[\psi_1(X)\psi_2(Y)\right] = \mathbb{E}\left[\mathbb{E}[\psi_2(Y) \mid X]\psi_1(X)\right].$$

Démonstration. Par le théorème de Fubini, on a

$$\begin{aligned}\mathbb{E}[\psi(X,Y)] &= \int_{E\times F} \psi(x,y) h(x,y)\mathrm{d}(\alpha\otimes\beta)(x,y) \\ &= \int_E \int_F \psi(x,y)\frac{h(x,y)}{f(x)}\mathrm{d}\beta(y) f(x)\mathrm{d}\alpha(x) \\ &= \int_E \left\{\int_F \psi(x,y) g(y \mid x)\mathrm{d}\beta(y)\right\} f(x)\mathrm{d}\alpha(x) \\ &= \mathbb{E}\left[\mathbb{E}\left[\psi(X,Y) \mid X\right]\right]\end{aligned}$$

■

Proposition 0.17. *Dans le cadre précédent, soit (X,Y) un couple de variables aléatoires à valeurs dans $E \times F$ avec $F = \mathbb{R}$, de densité $h(x,y)$ par rapport à $\alpha\otimes\beta$. Supposons Y de carré intégrable : $\mathbb{E}[Y^2] < \infty$. Alors*

$$\inf_{\substack{\varphi: E\to\mathbb{R} \\ \mathbb{E}[\varphi(X)^2]<\infty}} \mathbb{E}\left[(Y-\varphi(X))^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right].$$

Démonstration. On note que pour toute fonction $\varphi : E \to \mathbb{R}$ telle que $\mathbb{E}[\varphi(X)^2] < \infty$,

$$\mathbb{E}\left[(Y-\varphi(X))^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X] - \varphi(X))^2\right].$$

En effet, le double produit est nul puisque

$$\mathbb{E}\left[(Y-\mathbb{E}[Y\mid X])(\mathbb{E}[Y\mid X]-\varphi(X))\right]=\mathbb{E}\left[\mathbb{E}\left[Y-\mathbb{E}[Y\mid X]\mid X\right](\mathbb{E}[Y\mid X]-\varphi(X))\right]=0.$$

Ainsi, pour toute fonction φ telle que $\mathbb{E}[\varphi(X)^2]<\infty$, on a $\mathbb{E}\left[(Y-\varphi(X))^2\right]\geq\mathbb{E}\left[(Y-\mathbb{E}[Y\mid X])^2\right]$. Pour conclure il suffit de montrer que $\mathbb{E}\left[\mathbb{E}[Y\mid X]^2\right]<\infty$. Or, par l'inégalité de Jensen conditionnelle,

$$\mathbb{E}\left[\mathbb{E}[Y\mid X]^2\right]\leq\mathbb{E}\left[\mathbb{E}[Y^2\mid X]\right]=\mathbb{E}[Y^2]<\infty.$$

■

4. Approches statistiques

Nous introduisons les 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 $\mathbf{X}$.

4.1. Approche fréquentiste. Dans l'approche fréquentiste, on suppose

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

Typiquement, θ_0 est inconnu et l'on cherche à l'estimer à l'aide des données $\mathbf{X}$. Par exemple, dans le modèle gaussien, $\mathbf{X}=(X_1,\dots,X_n)$ 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 $\mathcal{N}(\theta_0,1)$. On peut alors estimer θ_0 par la moyenne empirique $\overline{X}_n$. Ce choix se justifie par exemple par la loi des grands nombres qui assure que $\overline{X}_n\xrightarrow{\mathbb{P}}\theta_0$.

Les grandes questions dans le cadre fréquentiste sont celles abordées dans la Section 2 :

(1) **Estimation.** Il s'agit de construire un estimateur $\widehat{\theta}(\mathbf{X})$ qui soit proche de la vraie valeur θ_0 du paramètre θ. Typiquement, on souhaite souvent qu'un estimateur soit consistant, asymptotiquement normal, et que son risque quadratique tende vers 0 assez vite.

(2) **Intervalles/régions de confiance.** On cherche à construire un sous-ensemble aléatoire $\mathcal{R}(\mathbf{X})$ de Θ tel que $\theta_0\in\mathcal{R}(\mathbf{X})$ avec grande probabilité (sous P_{θ_0}).

(3) **Tests.** On veut répondre par « vrai » ou « faux » à une propriété donnée de θ_0 en construisant un *test* $\varphi(\mathbf{X})$ à valeurs dans $\{0,1\}$.

Comme nous le verrons dans ce cours, ces questions peuvent aussi être posées dans le cadre bayésien.

4.2. Approche bayésienne. 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.

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. Avant d'observer l'échantillon, nous avons une certaine connaissance a priori, ou une certaine

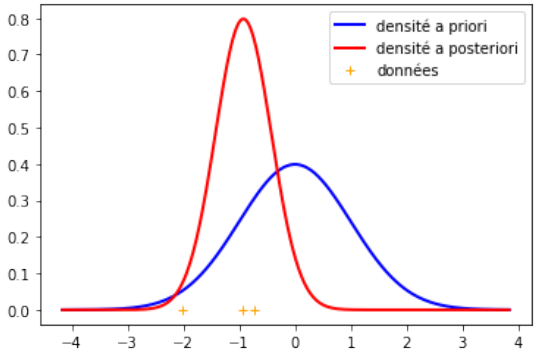
croyance, sur le paramètre (par exemple on sait qu'il est positif, ou bien on se doute qu'il est proche de 0, ou bien on sait qu'il est contenu dans un certain intervalle). Cette connaissance a priori est modélisée par une loi, la *loi a priori*, appelée aussi le *prior*. L'observation de l'échantillon nous permet ensuite de *mettre à jour* cette loi en formant ce qu'on appelle la *loi a posteriori*. L'objet central de l'approche bayésienne est ainsi une loi qui évolue avec la taille de l'échantillon : on part d'une loi a priori, et la prise en compte de chaque nouvelle observation fait évoluer cette loi : même si l'on part d'une certaine croyance a priori, l'observation du réel nous amène à modifier nos croyances.

Par exemple, imaginons que l'on cherche à savoir quelle est la probabilité $\theta \in [0,1]$ qu'une certaine pièce de monnaie tombe sur pile. L'approche fréquentiste va essentiellement faire appel à la loi des grands nombres et au théorème central limite : si on lance la pièce un grand nombre n de fois, et que l'on observe une certaine fréquence $\overline{X}_n$ de lancers donnant pile, alors on peut raisonnablement penser que la valeur $\overline{X}_n$ devient de plus en plus proche (quand n grandit) de la vraie valeur de θ (loi des grands nombres), et qu'avec grande probabilité, la vraie valeur de θ se trouve dans un certain intervalle centré en $\overline{X}_n$ dont la taille est donnée par des fluctuations gaussiennes (théorème central limite). L'approche bayésienne serait plutôt la suivante : a priori, si l'on n'a effectué aucun lancer, on ne sait pas grand chose du paramètre, mais l'on sait cependant qu'il appartient à l'intervalle $[0,1]$. Si c'est là notre seule information a priori sur θ, on peut commencer par dire que θ est distribué selon une loi uniforme sur $[0,1]$. C'est notre loi a priori. Choisir une loi uniforme revient à ne privilégier aucune valeur de $[0,1]$ par rapport aux autres (si l'on pense que la pièce n'est sûrement pas trop biaisée, on aurait pu plutôt choisir une loi qui met plus de poids autour de $1/2$). Ensuite on commence à lancer la pièce. Les résultats des lancers vont permettre de mettre à jour la loi initiale. Par exemple, si l'on observe bien plus de piles que de faces, on ne maintiendra pas notre a priori uniforme, mais on mettra à jour notre connaissance en formant une loi a posteriori qui mettra plus de poids au-dessus de $1/2$ qu'en dessous.

Dans l'approche bayésienne, on suppose donc que le paramètre inconnu θ du modèle est lui-même aléatoire, de loi donnée par la loi a priori. Cette loi reflète notre connaissance préalable (éventuelle) du paramètre. Ensuite, une fois des données $X_1, \dots, X_n$ 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 1. On obtient alors une nouvelle loi, la loi a posteriori. Notons déjà que si l'on choisit comme loi a priori une mesure de Dirac en un point, alors l'observation des données ne changera rien. La loi a posteriori restera toujours cette même mesure de Dirac. Dans ce cas, on a une connaissance certaine de la vraie valeur du paramètre et aucune donnée ne la modifiera. Ce cas extrême n'a donc pas beaucoup d'intérêt d'un point de vue statistique : si l'on est sûr de quelque chose, les statistiques ne servent à rien.

Illustrons les idées ci-dessus dans le cadre du modèle gaussien $\{\mathcal{N}(\theta,1)^{\otimes n},\ \theta \in \mathbb{R}\}$, avec pour loi a priori sur θ la loi $\mathcal{N}(0,1)$. Nous verrons au Chapitre 1 qu'après avoir observé n données $X_1, \dots, X_n$, la loi a posteriori est $\Pi_n = \mathcal{N}\left(\frac{n}{n+1}\overline{X}_n, \frac{1}{n+1}\right)$. La Figure 1 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.

FIGURE 1. Densités a priori et a posteriori



Notons l'effet de la mise à jour sur l'espérance de la loi a posteriori : au départ, on avait une loi a priori dont l'espérance était nulle. Puis l'observation des données a fait évoluer cette espérance : au temps n, l'espérance de la loi a posteriori est $\frac{1}{n+1} \times 0 + \frac{n}{n+1}\overline{X}_n$. Plus n grandit, plus l'espérance se rapproche de $\overline{X}_n$. Mais il reste toujours un effet marginal de l'a priori : l'espérance a posteriori s'écrit comme une moyenne pondérée entre l'espérance de la loi a priori 0 (avec une pondération $\frac{1}{n+1}$) et la moyenne empirique de l'échantillon $\overline{X}_n$ (avec une pondération $\frac{n}{n+1}$). La connaissance a priori « s'efface » donc de plus en plus, au profit de ce qui est observé. La mise à jour a aussi un effet sur la variance de la loi a posteriori qui ici décroît en $\frac{1}{n+1}$: la loi a posteriori devient de plus en plus concentrée.

Chapitre 1

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. Puis nous traitons du problème du choix de la loi a priori. Enfin, nous définissons certains aspects importants de la loi a posteriori (moyenne, médiane, variance a posteriori), ainsi que la notion de régions de crédibilité.

1. Le cadre bayésien

Le point de départ est toujours une expérience statistique : on se donne $\mathbf{X}$ un 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.

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. Ainsi le paramètre est une variable aléatoire $\boldsymbol{\theta}$, de loi Π.

Remarque 1.1. Il ne faut pas confondre la variable aléatoire $\boldsymbol{\theta}$ et les éléments $\theta \in \Theta$. Au tableau, ce sera difficile de différencier ces deux notations mais dans ce polycopié, nous nous efforcerons de bien utiliser la notation $\boldsymbol{\theta}$ (en gras) lorsqu'il s'agit de la variable aléatoire.

On suppose toujours dans la suite que

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

$$\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$$

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

$$\mathrm{d}\Pi = \pi \mathrm{d}\nu$$

L'étape suivante consiste à dire comment intervient $\mathbf{X}$. Plus précisément, nous allons spécifier la loi du couple $(\mathbf{X}, \boldsymbol{\theta})$. Pour que les quantités qui suivent soient bien définies, nous supposerons que l'application

$$\begin{array}{ccc} E \times \Theta & \to & \mathbb{R}_+ \\ (x, \theta) & \mapsto & p_\theta(x) \end{array} \tag{1.1}$$

est mesurable, où $E \times \Theta$ est muni de la tribu produit $\mathcal{E} \times \mathcal{B}(\Theta)$.

Proposition 1.1. *Supposons l'application* (1.1) *mesurable. Alors la fonction*

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

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

DÉMONSTRATION. Grâce à (1.1), l'application $(\theta, x) \mapsto \pi(\theta)p_\theta(x)$ est mesurable comme produit de fonctions mesurables, et positive par définition. Le théorème de Fubini donne alors que

$$\int_{\Theta\times E} \pi(\theta)p_\theta(x)\mathrm{d}(\nu\otimes\mu)(\theta,x) = \int_\Theta \left[\int_E p_\theta(x)\mathrm{d}\mu(x)\right]\pi(\theta)\mathrm{d}\nu(\theta) = \int_\Theta \pi(\theta)\mathrm{d}\nu(\theta) = 1.$$

■

Définition 1.1. Dans le cadre bayésien, on suppose l'application (1.1) mesurable et l'on définit la loi $\mathcal{L}(\boldsymbol{\theta}, \mathbf{X})$ du couple $(\boldsymbol{\theta}, \mathbf{X})$ comme la loi de densité $(\theta, x) \mapsto \pi(\theta)p_\theta(x)$ par rapport à $\nu \otimes \mu$. Autrement dit, la loi de $\boldsymbol{\theta}$ et la loi conditionnelle $\mathcal{L}(\mathbf{X} \mid \boldsymbol{\theta})$ sont données par

$$\begin{aligned} \boldsymbol{\theta} &\sim \Pi \\ \mathbf{X} \mid \boldsymbol{\theta} &\sim P_{\boldsymbol{\theta}}. \end{aligned} \tag{1.2}$$

Vérifions que les lois de $\boldsymbol{\theta}$ et de $\mathbf{X} \mid \boldsymbol{\theta}$ sont bien celles données dans la définition. La densité de $\boldsymbol{\theta}$ s'obtient en intégrant la densité jointe

$$\forall \theta \in \Theta\,, \quad \int_E \pi(\theta)p_\theta(x)\mathrm{d}\mu(x) = \pi(\theta),$$

donc $\boldsymbol{\theta} \sim \Pi$. La densité de $\mathbf{X} \mid \boldsymbol{\theta}$ s'obtient par la formule de la densité conditionnelle

$$\forall x \in E\,, \quad \frac{\pi(\boldsymbol{\theta})p_{\boldsymbol{\theta}}(x)}{\int_E \pi(\boldsymbol{\theta})p_{\boldsymbol{\theta}}(x)\mathrm{d}\mu(x)} = p_{\boldsymbol{\theta}}(x),$$

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

La loi marginale de $\mathbf{X}$ s'obtient également par intégration de la densité jointe. C'est la loi de densité f par rapport à μ donnée par

$$f : x \mapsto \int_\Theta p_\theta(x)\pi(\theta)\mathrm{d}\nu(\theta)\,.$$

Remarque 1.2. Attention ! Dans le cadre bayésien, la loi de $\mathbf{X}$ n'est donc pas P_θ, qui est la loi de $\mathbf{X}$ sachant $\boldsymbol{\theta} = \theta$.

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'observation, c'est-à-dire $\mathbf{X}$. On obtient ainsi la définition suivante.

Définition 1.2. La loi a posteriori est la loi conditionnelle $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{X})$. C'est une loi sur Θ, qui est notée $\Pi(\cdot \mid \mathbf{X})$.

Notons que sous l'hypothèse (1.1) que nous supposerons vérifiée dans la suite, il est équivalent de se donner la loi jointe de $(\boldsymbol{\theta}, \mathbf{X})$ ou les deux lois de $\boldsymbol{\theta}$ et de $\mathbf{X} \mid \boldsymbol{\theta}$ suivant (1.2). Nous ferons donc simplement référence à (1.2) quand nous parlerons de formalisme ou de cadre bayésien.

Theorème 1.2 (Formule de Bayes)**.** *La loi a posteriori* $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{X})$ *a une densité par rapport à* ν *donnée par*

$$\forall \theta \in \Theta\,,\ \pi(\theta \mid \mathbf{X}) = \frac{p_\theta(\mathbf{X})\pi(\theta)}{f(\mathbf{X})}\,, \quad \text{où} \quad f(\mathbf{X}) = \int_\Theta \pi(\theta) p_\theta(\mathbf{X}) \mathrm{d}\nu(\theta)\,.$$

Démonstration. Il suffit de combiner la définition 1.1 et la formule de la densité conditionnelle de la définition 0.27. ■

Cas du modèle d'échantillonnage. Soit une expérience statistique d'échantillonnage où $\mathbf{X} = (X_1, \dots, X_n)$ et $\mathcal{P} = \{P_\theta^{\otimes n}\,,\ \theta \in \Theta\}$. Le formalisme bayésien s'écrit

$$\boldsymbol{\theta} \sim \Pi$$
$$X_1, \dots, X_n \mid \boldsymbol{\theta} \sim P_{\boldsymbol{\theta}}^{\otimes n}.$$

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

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

La loi marginale de $\mathbf{X} = (X_1, \dots, X_n)$ a elle pour densité

$$f : (x_1, \dots, x_n) \mapsto \int_\Theta \pi(\theta) \prod_{i=1}^n p_\theta(x_i) \mathrm{d}\nu(\theta).$$

La formule de Bayes donne donc pour densité conditionnelle de $\boldsymbol{\theta}$ sachant $\mathbf{X}$

$$\forall \theta \in \Theta\,,\ \pi(\theta \mid \mathbf{X}) = \pi(\theta \mid X_1, \dots, X_n) = \frac{\prod_{i=1}^n p_\theta(X_i)\pi(\theta)}{f(X_1, \dots, X_n)}\,,$$

où $f(X_1, \dots, X_n) = \int_\Theta \pi(\theta) \prod_{i=1}^n p_\theta(X_i) \mathrm{d}\nu(\theta)$.
Interprétation. La densité a posteriori en tant que fonction de θ est proportionnelle à

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

Cette quantité est le produit de la vraisemblance (cf. Chapitre 0) 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.

Exemple 1.3 (L'exemple historique de Bayes)**.** Thomas Bayes (dans son célèbre *Essay Towards Solving a Problem in the Doctrine of Chances* publié de manière posthume en 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 avant la première boule. Bayes se demande : connaissant X, quelle inférence peut-on mener sur p ?

Exercice 1.1. Dans cette expérience, quel est l'ensemble Θ ? La loi a priori ? La famille de lois $(P_\theta)_{\theta \in \Theta}$? Répondre à la question de Bayes en calculant la densité a posteriori.

Exemple 1.4 (Le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$)**.**

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

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

Choisissons comme loi a priori $\Pi = \mathcal{N}(0, 1)$. Les mesures μ et ν sont toutes les deux la mesure de Lebesgue sur $\mathbb{R}$. On a

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

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

$$\pi(\theta \mid X) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X-\theta)^2}{2}}}{f(X)}, \quad \text{où} \quad f(X) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X-\theta)^2}{2}} \mathrm{d}\theta\,.$$

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

Méthode 1 – 'on écrit tout'

$$\pi(\theta \mid X) = \frac{e^{-\theta^2 + \theta X - \frac{X^2}{2}}}{\displaystyle\int_{\mathbb{R}} e^{-\theta^2 + \theta X - \frac{X^2}{2}} \mathrm{d}\theta} = \frac{e^{-\left(\theta - \frac{X}{2}\right)^2 - \frac{X^2}{4}}}{\displaystyle\int_{\mathbb{R}} e^{-\left(\theta - \frac{X}{2}\right)^2 - \frac{X^2}{4}} \mathrm{d}\theta} = \frac{e^{-\left(\theta - \frac{X}{2}\right)^2}}{\displaystyle\int_{\mathbb{R}} e^{-\left(\theta - \frac{X}{2}\right)^2} \mathrm{d}\theta}.$$

L'intégrale au dénominateur est égale à $\int_{\mathbb{R}} e^{-u^2} du$, qui vaut $\sqrt{\pi}$. Ainsi

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

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

Méthode 2 – 'proportionnel à'. On constate qu'il n'est pas utile de garder l'intégrale $f(X)$ au dénominateur dans les calculs, puisque c'est une expression qui dépend de X seulement et pas de θ, et intervient donc seulement en termes de constante de normalisation. Le symbole $\propto$ ci-dessous signifie 'à constante de proportionnalité près', où cette constante peut dépendre de tout sauf de θ.

$$\pi(\theta \mid X) \propto \pi(\theta) p_\theta(X) \propto e^{-\theta^2 + \theta X} \propto e^{-\left(\theta - \frac{X}{2}\right)^2}.$$

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

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

$$\begin{aligned} \mathbf{X} = (X_1, \ldots, X_n) \mid \boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\theta}, 1)^{\otimes n} \\ \boldsymbol{\theta} &\sim \Pi = \mathcal{N}(0, 1) \end{aligned}$$

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

$$\pi(\theta \mid X_1, \dots, X_n) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \left\{ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} \right\}}{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \left\{ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} \right\} \mathrm{d}\theta}.$$

Déterminons la densité à constante multiplicative près :

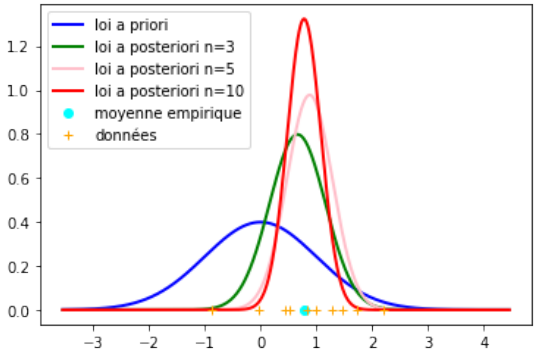
$$\begin{aligned}
\pi(\theta \mid X_1, \dots, X_n) &\propto \exp\left(-\sum_{i=1}^{n} \frac{1}{2}(X_i - \theta)^2 - \frac{\theta^2}{2} \right) \\
&\propto \exp\left(-\frac{n+1}{2}\theta^2 + n\overline{X}_n \theta \right) \\
&\propto \exp\left(-\frac{n+1}{2} \left(\theta - \frac{n\overline{X}_n}{n+1}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

On en conclut

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

La figure 1 trace la loi a priori, les données, et les lois a posteriori correspondant à $n = 3, 5, 10$ observations dans le cadre du modèle gaussien. On constate que la loi a posteriori se concentre près de $\overline{X}_n$ et que l' incertitude, que l'on peut décrire comme l'écart-type de la loi a posteriori, décroît comme $1/\sqrt{n}$ quand n augmente.

FIGURE 1. Densités a priori et a posteriori



Exercice 1.2. Dans le modèle gaussien avec n observations, si la loi a priori Π sur $\boldsymbol{\theta}$ est une $\mathcal{N}(a, v)$, montrer que

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

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, en précisant les poids alloués à chacune des deux moyennes.

2. Aspects de la loi a posteriori

Dans l'exemple du modèle gaussien ci-dessus, nous constatons que la moyenne de la loi a posteriori (sachant $\mathbf{X}$) vaut

$$\int_\Theta \theta \mathrm{d}\Pi(\theta \mid \mathbf{X}) = \mathbb{E}\left[\mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right) \mid \mathbf{X}\right] = \frac{n\overline{X}_n}{n+1}.$$

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

Définition 1.3. 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, notée $m_{\mathbf{X}}$:

$$m_{\mathbf{X}} = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \int_\Theta \theta \mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

— le mode a posteriori : c'est un point où le maximum de la densité a posteriori $\theta \mapsto \pi(\theta \mid \mathbf{X})$ est atteint. On le note

$$\mathsf{mode}(\boldsymbol{\theta} \mid \mathbf{X}) \in \arg\max_{\theta \in \Theta} \pi(\theta \mid \mathbf{X}).$$

— la variance a posteriori (pour $\Theta \subset \mathbb{R}$), notée $v_{\mathbf{X}}$: c'est la variance de la loi a posteriori, soit

$$v_{\mathbf{X}} = \mathrm{Var}(\boldsymbol{\theta} \mid \mathbf{X}) = \mathbb{E}\left[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}])^2 \mid \mathbf{X}\right] = \int_\Theta (\theta - m_{\mathbf{X}})^2 \mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

Si $\Theta \subset \mathbb{R}^d,\ d \geq 2$, on peut définir la matrice de variance-covariance a posteriori :

$$\Sigma_{\mathbf{X}} = \int_\Theta (\theta - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}])(\theta - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}])^T \mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

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.

Définition 1.4. Dans le cadre précédent, si $\Theta \subset \mathbb{R}$, soit $F_{\mathbf{X}}$ la fonction de répartition de la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$. On note $F_{\mathbf{X}}^{-1}$ l'inverse généralisée de $F_{\mathbf{X}}$, définie pour tout $u \in [0,1]$ par

$$F_{\mathbf{X}}^{-1}(u) = \inf\left\{\theta \in \Theta,\ F_{\mathbf{X}}(\theta) \geq u\right\}.$$

Pour $p \in [0,1]$, on définit alors le quantile a posteriori d'ordre p comme $F_{\mathbf{X}}^{-1}(p)$. Le quantile a posteriori d'ordre $1/2$ s'appelle la médiane a posteriori.

Si la fonction $F_{\mathbf{X}}$ 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 $F_{\mathbf{X}}^{-1}$ est simplement la réciproque de $F_{\mathbf{X}}$.

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

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \mathsf{mode}(\boldsymbol{\theta} \mid \mathbf{X}) = F_{\mathbf{X}}^{-1}(1/2) = \frac{n\overline{X}_n}{n+1} \quad \text{et} \quad \mathrm{Var}(\boldsymbol{\theta} \mid \mathbf{X}) = \frac{1}{n+1}.$$

Notons que les statistiques $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]$, $\mathsf{mode}(\boldsymbol{\theta} \mid \mathbf{X})$ et $F_{\mathbf{X}}^{-1}(1/2)$ sont des estimateurs ponctuels au sens usuel du terme. Dans l'exemple du modèle gaussien, ils sont même très proches de $\overline{X}_n$. Nous en dirons plus sur ce sujet aux Chapitres 3 et 5.

3. Le choix de la loi a priori

Il existe 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 verrons en particulier le cas des familles conjuguées. Certains 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 à 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 de nombreux cas, le statisticien dispose d'abord 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. Ou bien 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 lancer de pièce d'être biaisé avec probabilité 2/3 de donner 'pile'. On est dans un modèle de Bernoulli $\{\mathcal{B}(\theta),\ \theta \in [0,1]\}$. 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 $a\,\mathrm{Beta}(4,2) + (1-a)\,\mathrm{Beta}(3,3)$ 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. À 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. Cette question cruciale fera l'objet du Chapitre 2.

Commençons déjà par souligner que si l'on n'a que très peu d'informations a priori, il peut être tentant de choisir une loi a priori Π qui ne favorise aucune région par rapport à une autre, i.e. de densité constante. On parle d'a priori non-informatif. Mais cela ne donne pas forcément une mesure de probabilité au sens où on l'on peut alors avoir $\Pi(\Theta) = +\infty$ (par exemple si $\Theta = \mathbb{R}$ et Π la mesure de Lebesgue). C'est ce qu'on appelle un a priori impropre et nous allons voir qu'il est quand même possible de définir une loi a posteriori (qui elle est bien une mesure de probabilité), à condition que l'a priori ne soit pas « trop » impropre.

3.1. A priori impropres.

Définition 1.5. 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. On supposera néanmoins que Π est une mesure σ-finie, absolument continue par rapport à ν, et l'on notera π sa dérivée de Radon-Nikodym par rapport ν.

Définition 1.6. Dans le cadre d'une expérience $(\mathbf{X}, \mathcal{P})$ avec $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, et $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$, si l'on met un a priori impropre Π sur θ avec $\mathrm{d}\Pi = \pi \mathrm{d}\nu$, et si $\int_\Theta p_\theta(\mathbf{X}) \mathrm{d}\Pi(\theta)$ est finie p.s., alors la loi a posteriori correspondante $\Pi[\cdot \mid \mathbf{X}]$ est la loi sur Θ de densité par rapport à ν égale à

$$\theta \mapsto \pi(\theta \mid \mathbf{X}) = \frac{p_\theta(\mathbf{X})\pi(\theta)}{\displaystyle\int_\Theta p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\nu(\theta)}.$$

Exemple 1.5.

▶ Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, prenons comme loi a priori la mesure de Lebesgue $\Pi = \mathrm{Leb}_\mathbb{R}$ sur $\mathbb{R}$. Notons que pour tout $x \in \mathbb{R}$,

$$\int_\mathbb{R} p_\theta(x)\mathrm{d}\Pi(\theta) = \int_\mathbb{R} \frac{e^{-\frac{(x-\theta)^2}{2}}}{\sqrt{2\pi}}\mathrm{d}\theta = 1 < \infty.$$

Pour n observations, la densité a posteriori est

$$\theta \mapsto \pi(\theta \mid X_1, \ldots, X_n) = \frac{\displaystyle\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}}}{\displaystyle\int_\mathbb{R} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} \mathrm{d}\theta} \ \propto\ e^{-\frac{n}{2}(\theta - \overline{X}_n)^2}.$$

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

▶ Le problème du tramway.

Le problème suivant est mentionné par Harold Jeffreys dans son livre de 1961. Un voyageur arrive dans une ville qu'il ne connaît pas. La première chose qu'il voit est un tramway numéroté 100. Cela peut-il l'aider à avoir une idée du nombre total de tramways dans cette ville ?

Essayons de modéliser ce problème dans un cadre bayésien. On prend $\Theta = \mathbb{N}^*$ l'ensemble des valeurs possibles pour le nombre total de tramways dans la ville. Sachant qu'il y a n tramways dans la ville, le numéro du premier tramway observé est supposé uniformément

distribué entre 1 et n. Plus formellement

$$\begin{aligned} X \mid n &\sim \mathrm{Unif}(\{1,\dots,n\}\\ n &\sim \Pi, \end{aligned}$$

où Π est une loi a priori sur $\mathbb{N}^*$ à choisir. Si l'on ne veut vraiment rien supposer a priori, on est tenté de prendre l'a priori impropre $\Pi = \sum_{n\geq 1} \delta_n$, i.e. attribuer à chaque nombre de tramways possible le même poids 1. Le problème avec ce choix est que, pour tout $X \in \mathbb{N}^*$, on a

$$\int_{\mathbb{N}^*} p_n(X)\mathrm{d}\Pi(n) = \sum_{n\geq 1} \frac{1}{n}\mathbb{1}_{X\in\{1,\dots,n\}} = \sum_{n\geq X} \frac{1}{n} = +\infty\,.$$

On peut quand même penser qu'il ne peut pas y avoir trop de tramways dans la ville, et prendre une loi a priori correspondant à des poids décroissants. Par exemple

$$\Pi = \sum_{n\geq 1} \frac{1}{n^\gamma}\delta_n\,,$$

avec $\gamma > 0$. Si $\gamma \leq 1$, il s'agit d'un a priori impropre car on a alors $\Pi(\mathbb{N}^*) = \sum_{n\geq 1} 1/n^\gamma = +\infty$. En revanche, contrairement au cas précédent, on a

$$\int_{\mathbb{N}^*} p_n(X)\mathrm{d}\Pi(n) = \sum_{n\geq 1} \frac{\mathbb{1}_{X\in\{1,\dots,n\}}}{n^{1+\gamma}} = \sum_{n\geq X} \frac{1}{n^{1+\gamma}} < \infty.$$

Si le premier tramway observé porte le numéro $X \in \mathbb{N}^*$, quelle va alors être notre loi a posteriori ? Par la définition 1.6, on obtient

$$\pi(n \mid X) = \frac{\frac{1}{n^{1+\gamma}}\mathbb{1}_{X\leq n}}{\sum_{k\geq X}\frac{1}{k^{1+\gamma}}}\,.$$

3.2. Conjugaison.

Définition 1.7. 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 $\Pi \in \mathcal{F}$, si Π est prise comme loi a priori dans le cadre bayésien de ce modèle, alors la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ associée appartient aussi à $\mathcal{F}$.

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 gaussien $\mathcal{P} = \{\mathcal{N}(\theta, v),\ \theta \in \mathbb{R}\}$ (pour tout $v > 0$ fixé).
- la famille des lois Beta $\mathcal{F} = \{\mathrm{Beta}(a,b),\ a > 0,\ b > 0\}$ est conjuguée pour le modèle des lois de Bernoulli. Pour $a > 0$ et $b > 0$ fixés, on considère

$$\begin{aligned} \mathbf{X} = (X_1,\dots,X_n) \mid \boldsymbol{\theta} &\sim \mathcal{B}(\boldsymbol{\theta})^{\otimes n}\\ \boldsymbol{\theta} &\sim \mathrm{Beta}(a,b) \end{aligned}$$

On obtient comme densité a posteriori (détails en TDs)

$$\pi(\theta \mid \mathbf{X}) \propto \theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{[0,1]}(\theta).\prod_{i=1}^{n}\left(\theta^{X_i}(1-\theta)^{n-X_i}\right)$$
$$\propto \theta^{a+n\overline{X}_n-1}(1-\theta)^{b+n-n\overline{X}_n-1}\mathbb{1}_{[0,1]}(\theta).$$

La loi dont la densité est proportionnelle à cette expression est la loi $\text{Beta}(a+n\overline{X}_n, b+n-n\overline{X}_n)$. Ainsi

$$\Pi[\cdot \mid \mathbf{X}] = \text{Beta}(a+n\overline{X}_n, b+n-n\overline{X}_n) \in \mathcal{F}.$$

- la famille des lois de Dirichlet est conjuguée pour le modèle multinomial, voir TDs.
- la famille des lois Gamma $\mathcal{F} = \{\Gamma(p,\lambda),\, p>0,\, \lambda>0\}$ est conjuguée pour le modèle des lois de Poisson. Pour $p>0$ et $\lambda>0$ fixés, on considère

$$\mathbf{X} = (X_1,\dots,X_n) \mid \boldsymbol{\theta} \sim \mathcal{P}(\boldsymbol{\theta})^{\otimes n}$$
$$\boldsymbol{\theta} \sim \Gamma(p,\lambda)$$

On obtient (détails en TDs)

$$\pi(\theta \mid \mathbf{X}) \propto \theta^{p-1}e^{-\lambda\theta}\prod_{i=1}^{n}\left(e^{-\theta}\theta^{X_i}\right)\mathbb{1}_{\theta\geq 0} \propto \theta^{p+n\overline{X}_n-1}e^{-(\lambda+n)\theta}\mathbb{1}_{\theta\geq 0}.$$

La loi dont la densité est proportionnelle à cette expression est la loi $\Gamma(p+n\overline{X}_n, \lambda+n)$. Ainsi

$$\Pi[\cdot \mid \mathbf{X}] = \Gamma(p+n\overline{X}_n, \lambda+n) \in \mathcal{F}.$$

- la famille des lois Gamma $\mathcal{F} = \{\Gamma(p,\lambda),\, p>0,\, \lambda>0\}$ est conjuguée pour le modèle des lois $\text{Gamma}(k,\theta)$ (pour tout $k>0$ fixé).
- la famille des lois de Pareto $\mathcal{F} = \{\mathcal{P}(\alpha,r),\, \alpha>0,\, r>0\}$ est conjuguée pour le modèle des lois uniformes. Pour $\alpha>0$ et $r>0$ fixés, on considère

$$\mathbf{X} = (X_1,\dots,X_n) \mid \boldsymbol{\theta} \sim \text{Unif}[0,\boldsymbol{\theta}]^{\otimes n}$$
$$\boldsymbol{\theta} \sim \mathcal{P}(\alpha,r)$$

On rappelle que la densité de la loi $\mathcal{P}(\alpha,r)$ par rapport à la mesure de Lebesgue est donnée par $z \mapsto \alpha r^{\alpha} z^{-(\alpha+1)}\mathbb{1}_{[r,+\infty[}(z)$. On obtient

$$\pi(\theta \mid \mathbf{X}) \propto \theta^{-(\alpha+1)}\mathbb{1}_{\theta\geq r}\frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{1}_{0\leq X_i\leq\theta}$$
$$\propto \theta^{-(\alpha+n+1)}\mathbb{1}_{\theta\geq\max\{r,X_1,\dots,X_n\}}.$$

La loi dont la densité est proportionnelle à cette expression est la loi $\mathcal{P}(\alpha+n, r_{\mathbf{X}})$, où $r_{\mathbf{X}} = \max\{r,X_1,\dots,X_n\}$. Ainsi

$$\Pi[\cdot \mid \mathbf{X}] = \mathcal{P}(\alpha+n, r_{\mathbf{X}}) \in \mathcal{F}.$$

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 lois de la famille considérée, la simulation suivant la loi a posteriori est un cas particulier, donc le temps ou la complexité de calcul sont réduits dans

ce cas ce qui est souvent avantageux (voir chapitre Simulation).

La plupart des cas de lois conjuguées citées ci-dessus 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.

Le modèle $\mathcal{N}(\mu, \sigma^2)$, moyenne et variance inconnues

Lemme 1.3. *Soit $Y \sim \text{Gamma}(a, b)$, de densité $f_Y(y) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by} \mathbb{1}_{\mathbb{R}_+^*}(y)$. Alors $Z = Y^{-1}$ a pour densité*

$$f_Z(z) = \frac{b^a}{\Gamma(a)} z^{-a-1} e^{-\frac{b}{z}} \mathbb{1}_{\mathbb{R}_+^*}(z).$$

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

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

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

et le résultat s'en déduit. ■

Dans le cas où μ est connu et vaut 0, la famille des lois inverse-gamma est conjuguée pour le modèle $\{\mathcal{N}(0, \sigma^2),\ \sigma^2 > 0\}$ (le paramètre étant $\theta = \sigma^2$). En effet, si $\boldsymbol{\sigma}^2 \sim \text{IG}(a, b) = \Pi$ et $X \mid \boldsymbol{\sigma}^2 \sim \mathcal{N}(0, \boldsymbol{\sigma}^2)$, la densité a posteriori est donnée par

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

On obtient $\mathcal{L}(\boldsymbol{\sigma}^2 \mid X) = \text{IG}\left(a + \frac{1}{2}, b + \frac{X^2}{2}\right)$.

Exercice 1.3. Vérifier la propriété de conjugaison dans le cas de n observations.

Dans le cas où à la fois μ et σ^2 sont inconnus, on peut déjà essayer d'utiliser une loi inverse-gamma sur σ^2. En revanche, l'idée qui consiste à proposer une loi produit comme loi a priori sur le couple (μ, σ^2), 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.

Définition 1.8. On appelle loi $\mathrm{NIG}(a,b,c,d)$, loi normale inverse-gamma la loi sur $\mathbb{R}\times\mathbb{R}_+^*$ définie par le schéma

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

La densité d'une loi $\mathrm{NIG}(a,b,c,d)$ est

$$(\mu,\sigma^2)\mapsto \frac{d^c}{\Gamma(c)}\sqrt{\frac{b}{2\pi}}(\sigma^2)^{-c-\frac{3}{2}}e^{-\frac{d}{\sigma^2}}e^{-\frac{b(\mu-a)^2}{2\sigma^2}}.$$

Theorème 1.4. *Soit* $\mathbf{X}=(X_1,\dots,X_n)$ *et considérons le cadre bayésien*

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

La famille de toutes les lois NIG *normales inverse-gamma est conjuguée pour ce modèle et*

$$\Pi[\cdot\ \big|\ \mathbf{X}]\ =\ \mathrm{NIG}\left(a_{\mathbf{X}},b_{\mathbf{X}},c_{\mathbf{X}},d_{\mathbf{X}}\right),$$

avec, si l'on pose $s_{\mathbf{X}}=\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i^2-\overline{X}_n^2=\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)^2$,

$$\begin{aligned}a_{\mathbf{X}}&=\frac{n\overline{X}_n+ab}{n+b}, & b_{\mathbf{X}}&=b+n\\ c_{\mathbf{X}}&=c+\frac{n}{2}, & d_{\mathbf{X}}&=d+\frac{ns_{\mathbf{X}}}{2}+\frac{nb}{2(n+b)}(\overline{X}_n-a)^2\end{aligned}$$

Démonstration. La vraisemblance s'écrit

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

La formule de Bayes donne donc pour la densité a posteriori

$$\pi\left((\mu,\sigma^2)\ \big|\ \mathbf{X}\right)\propto(\sigma^2)^{-\frac{n}{2}-c-\frac{3}{2}}\exp\left\{-\frac{n}{2\sigma^2}(\mu-\overline{X}_n)^2-\frac{n}{2\sigma^2}s_{\mathbf{X}}-\frac{b(\mu-a)^2}{2\sigma^2}-\frac{d}{\sigma^2}\right\}.$$

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

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

On en déduit la formule annoncée. ∎

Dans la pratique, un a priori souvent utilisé est

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

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 et l'on peut vérifier (voir TDs) que la loi a posteriori est

$$\Pi^*[\cdot \mid \mathbf{X}] \sim \text{NIG}\left(\overline{X}_n, n, \frac{n-1}{2}, \frac{ns_{\mathbf{X}}}{2}\right).$$

Le modèle $\mathcal{N}(\mu,\Sigma)$ en dimension $d \geq 1$, 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 à la section précédente pour le cas uni-dimensionnel.

Theorème 1.5. *Soit $\mathbf{X} = (X_1,\dots,X_n)$ avec $X_i \in \mathbb{R}^d, d \geq 1$. Soit $\mu_0 \in \mathbb{R}^d$ fixé et Σ, Σ_0 deux matrices symétriques définies positives fixées. Considérons le cadre bayésien*

$$\begin{aligned} \mathbf{X} \mid \boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu},\Sigma)^{\otimes n} \\ \boldsymbol{\mu} &\sim \mathcal{N}(\mu_0,\Sigma_0) = \Pi. \end{aligned}$$

La famille $\{\mathcal{N}(\mu_0,\Sigma_0),\ \mu_0 \in \mathbb{R}^d,\ \Sigma_0$ symétrique définie positive$\}$ est conjuguée et

$$\Pi[\cdot \mid \mathbf{X}] = \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}),$$

avec

$$\begin{aligned} \Sigma_{\mathbf{X}} &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} \\ \mu_{\mathbf{X}} &= \Sigma_{\mathbf{X}}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\overline{X}_n). \end{aligned}$$

Démonstration. Voir TDs. ■

Remarque 1.6. Le Théorème 1.5 peut se voir comme un résultat de conditionnement sur les vecteurs gaussiens. Les lois de $X \mid \boldsymbol{\mu}$ et de $\boldsymbol{\mu}$ sont gaussiennes, donc la loi jointe de $(X,\boldsymbol{\mu})$ aussi, ainsi que la loi conditionnelle de $\boldsymbol{\mu} \mid X$.

3.3. Lois invariantes : a priori de Jeffreys. Dans le but de trouver une loi a priori qui serait « universelle », Jeffreys (1946) propose de chercher une loi Π qui soit invariante par changement de paramétrisation du problème $\eta = \varphi(\theta)$, où η désigne le nouveau paramètre et θ le paramètre d'origine.

Nous allons nous restreindre au cas où Θ est un intervalle ouvert de $\mathbb{R}$, et nous placer dans le cadre des modèles dits réguliers. On rappelle que la log-vraisemblance est la fonction $\theta \mapsto \ell_\theta(\mathbf{X}) = \log p_\theta(\mathbf{X})$. La dérivée par rapport à θ de la log-vraisemblance (si elle existe) est notée (de façon trompeuse) $\ell'_\theta(\mathbf{X})$ et s'appelle le score. On a

$$\ell'_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ell_\theta(\mathbf{X}) = \frac{p'_\theta(\mathbf{X})}{p_\theta(\mathbf{X})},$$

où $p'_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}p_\theta(\mathbf{X})$.

Définition 1.9. Soit Θ un intervalle ouvert de $\mathbb{R}$. Le modèle $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ avec $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$ sera dit régulier si

(1) pour μ-presque tout $x \in E$, la fonction $\theta \mapsto p_\theta(x)$ est absolument continue sur Θ ;

(2) pour tout $\theta_0 \in \Theta$, il existe $E_0 \subset E$ avec $\mu(E \setminus E_0) = 0$ tel que pour tout $x \in E_0$, la fonction $\theta \mapsto p'_\theta(x)$ est continue en θ_0 ;

(3) pour tout $\theta \in \Theta$, le score $\ell'_\theta(\mathbf{X})$ possède un moment d'ordre 2, et la fonction $\theta \mapsto \mathbf{I}(\theta) = \mathbb{E}_\theta\left[\ell'_\theta(\mathbf{X})^2\right]$ est continue sur Θ.

La quantité $\mathbf{I}(\theta)$ est alors appelée information de Fisher du modèle au point $\theta \in \Theta$.

Remarque 1.7. Il est facile de voir que si le modèle $\mathcal{P}^{(1)} = \{P_\theta,\, \theta \in \Theta\}$ est régulier, alors pour tout $n \geq 1$, le modèle $\mathcal{P}^{(n)} = \{P_\theta^{\otimes n},\, \theta \in \Theta\}$ l'est aussi, et que, si $\mathbf{I}_n(\theta)$ est l'information de Fisher du modèle $\mathcal{P}^{(n)}$,

$$\mathbf{I}_n(\theta) = n\mathbf{I}(\theta)\,,$$

où $\mathbf{I}(\theta) = \mathbf{I}_1(\theta)$.

Intuition. Informellement, $\mathbf{I}(\theta)$ correspond à la quantité d'information disponible pour le paramètre θ. On peut montrer que, dans un modèle régulier, avec des hypothèses supplémentaires, s'il existe une suite consistante $(\widehat{\theta}_n)$ d'estimateurs du maximum de vraisemblance, alors pour tout $\theta \in \Theta$ avec $\mathbf{I}(\theta) > 0$,

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

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

Exemple 1.8. Considérons le modèle $\mathcal{P} = \{\mathcal{B}(\theta),\ \theta \in (0,1)\}$. Pour le modèle à une observation, 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 1,

$$\mathbf{I}(\theta) = \mathbb{E}_\theta\left[\frac{X^2}{\theta^2} + \frac{(1-X)^2}{(1-\theta)^2}\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

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

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta)) = \mathcal{N}(0, \mathbf{I}(\theta)^{-1}).$$

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 $\mathbb{R}$.

Définition 1.10. Dans un modèle régulier avec $\Theta \subset \mathbb{R}$ ouvert, l'a priori de Jeffreys est la mesure sur Θ de densité π par rapport à $\nu = \mathrm{Leb}_{|\Theta}$ proportionnelle à $\sqrt{\mathbf{I}(\cdot)}$. Plus précisément,

pour tout $\theta \in \Theta$, $\pi(\theta) = \frac{1}{\Lambda}\sqrt{\mathbf{I}(\theta)}$ avec

$$\Lambda = \begin{cases} \int_\Theta \sqrt{\mathbf{I}(\theta)}\mathrm{d}\theta & \text{si } \int_\Theta \sqrt{\mathbf{I}(\theta)}\mathrm{d}\theta < +\infty, \\ 1 & \text{si } \int_\Theta \sqrt{\mathbf{I}(\theta)}\mathrm{d}\theta = +\infty. \end{cases}$$

Exemple 1.9. (1) $\mathcal{P} = \{\mathcal{B}(\theta),\ \theta \in (0,1)\}$. D'après le calcul de l'information de Fisher ci-dessus, la loi a priori de Jeffreys est celle dont la densité $\pi(\theta)$ est proportionnelle à

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

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

(2) $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$. Ici $\mathbf{I}(\theta) = 1$ pour tout $\theta \in \mathbb{R}$. L'a priori de Jeffreys est donc la mesure de Lebesgue sur $\mathbb{R}$. Il s'agit d'un a priori impropre.

Proposition 1.6. *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* $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, *si* φ *est un* $\mathcal{C}^1$*-difféomorphisme de* Θ *dans* $\varphi(\Theta)$, *alors* $\Pi \circ \varphi^{-1}$, *la mesure image de* Π *par* φ, *est l'a priori de Jeffreys dans le modèle* $\mathcal{Q} = \{P_{\varphi^{-1}(\eta)},\ \eta \in \varphi(\Theta)\}$.

Démonstration. Notons $\psi = \varphi^{-1}$ et $q_\eta = p_{\psi(\eta)}$, soit $p_\theta = q_{\varphi(\theta)}$. L'information de Fisher $\mathbf{J}(\eta)$ dans le modèle paramétré par η s'exprime en fonction de celle $\mathbf{I}(\theta)$ dans le modèle paramétré par θ. En effet, $q'_\eta = \psi'(\eta)p'_{\psi(\eta)}$, donc

$$\mathbf{J}(\eta) = \int_E \frac{q'_\eta(x)^2}{q_\eta(x)}\mathrm{d}\mu(x) = \psi'(\eta)^2 \int_E \frac{p'_{\psi(\eta)}(x)^2}{p_{\psi(\eta)}(x)}\mathrm{d}\mu(x) = \psi'(\eta)^2\mathbf{I}(\psi(\eta))\,.$$

Prenons Π l'a priori de Jeffreys dans le modèle paramétré par θ, de densité $\pi(\theta) = \frac{1}{\Lambda}\sqrt{\mathbf{I}(\theta)}$ avec Λ donné dans la définition 1.10. Cherchons la densité de $\Pi \circ \varphi^{-1}$, la mesure image de Π par φ. Soit f une fonction mesurable bornée. Par le théorème de transfert et la définition de Π, on a

$$\int_{\varphi(\Theta)} f(\eta)\mathrm{d}\Pi \circ \varphi^{-1}(\eta) = \int_\Theta f\left(\varphi(\theta)\right)\mathrm{d}\Pi(\theta) = \frac{1}{\Lambda}\int_\Theta f\left(\varphi(\theta)\right)\sqrt{\mathbf{I}(\theta)}\mathrm{d}\theta\,.$$

En effectuant le changement de variable $\eta = \varphi(\theta)$, on obtient

$$\begin{aligned} \frac{1}{\Lambda}\int_\Theta f\left(\varphi(\theta)\right)\sqrt{\mathbf{I}(\theta)}\mathrm{d}\theta &= \frac{1}{\Lambda}\int_{\varphi(\Theta)} f(\eta)\sqrt{\mathbf{I}\left(\varphi^{-1}(\eta)\right)}\left|(\varphi^{-1})'(\eta)\right|\mathrm{d}\eta \\ &= \frac{1}{\Lambda}\int_{\varphi(\Theta)} f(\eta)\sqrt{\mathbf{J}(\eta)}\mathrm{d}\eta\,, \end{aligned}$$

où l'on a utilisé l'expression de $\mathbf{J}(\eta)$ donnée plus haut. Ainsi la mesure image $\Pi \circ \varphi^{-1}$ a pour densité $\eta \mapsto \frac{1}{\Lambda}\sqrt{\mathbf{J}(\eta)}$. C'est donc bien l'a priori de Jeffreys dans le modèle paramétré par η. ■

Exercice 1.4. Vérifier directement par le calcul que $\Pi = \text{Beta}(\frac{1}{2}, \frac{1}{2})$ est invariant dans le modèle $\mathcal{P} = \{\mathcal{B}(\theta),\ \theta \in (0,1)\}$.

3.4. L'approche bayésienne hiérarchique ou *hierarchical Bayes*. Dans l'approche bayésienne hiérarchique, on se donne une famille paramétrée de lois a priori $\{\Pi_\alpha,\ \alpha \in \mathcal{A}\}$, par exemple

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

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

(3) toutes les lois $\{\text{Beta}(a, b),\ a > 0,\ b > 0\}$.

Une solution naturelle consiste alors à mettre une loi sur le paramètre α de la loi a priori. On ajoute ainsi un troisième niveau décrivant la loi du paramètre $\boldsymbol{\alpha}$ qui devient lui aussi une variable aléatoire (comme pour θ, on notera $\boldsymbol{\alpha}$ lorsqu'il s'agit de la variable aléatoire, et α pour un élément fixé de $\mathcal{A}$). Supposons que pour tout $\alpha \in \mathcal{A}$, on a $\mathrm{d}\Pi_\alpha(\theta) = \pi_\alpha(\theta)\mathrm{d}\nu(\theta)$, avec ν mesure σ-finie sur Θ et $\mathrm{d}Q(\alpha) = q(\alpha)\mathrm{d}\nu'(\alpha)$, avec ν' mesure σ-finie sur $\mathcal{A}$ (on suppose aussi que l'application $(\alpha, \theta) \mapsto \pi_\alpha(\theta)$ est mesurable). La mise en œuvre de l'idée précédente, du point de vue de la loi a priori, s'écrit

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\alpha} &\sim \Pi_{\boldsymbol{\alpha}} \\ \boldsymbol{\alpha} &\sim Q. \end{aligned}$$

Ainsi, la loi marginale de $\boldsymbol{\theta}$ s'interprète comme une loi *mélange*, dont la densité par rapport à ν est

$$\pi(\theta) = \int_{\mathcal{A}} \pi_\alpha(\theta)\mathrm{d}Q(\alpha).$$

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.

Exemple 1.10. 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

$$\begin{aligned} X_1, \ldots, X_n \mid \boldsymbol{\theta}, \boldsymbol{\alpha} &\sim \mathcal{B}(\boldsymbol{\theta})^{\otimes n} = P_{\boldsymbol{\theta}}^{\otimes n} \\ \boldsymbol{\theta} \mid \boldsymbol{\alpha} &\sim \text{Beta}(6 - \boldsymbol{\alpha}, 6 + \boldsymbol{\alpha}) = \Pi_{\boldsymbol{\alpha}} \\ \boldsymbol{\alpha} &\sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2 = Q. \end{aligned}$$

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

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

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

3.5. 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) comme dans la section précédente, on se restreint à une classe de lois a priori $(\Pi_\alpha)_{\alpha \in \mathcal{A}}$;

b) on « estime » (voir ci-dessous) le paramètre α de la loi a priori ;

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

On suppose que toutes les lois Π_α sont dominées par une même mesure ν avec $\mathrm{d}\Pi_\alpha = \pi_\alpha \mathrm{d}\nu$ et, comme d'habitude, que toutes les lois P_θ sont dominées par une même mesure μ avec $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$.

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 : c'est le principe de la méthode du maximum de vraisemblance marginal.

Pour $\alpha \in \mathcal{A}$, la loi marginale de $\mathbf{X}$ dans le cadre bayésien

$$\begin{aligned} \boldsymbol{\theta} &\sim \Pi_\alpha \\ \mathbf{X} \mid \boldsymbol{\theta} &\sim P_\theta \end{aligned}$$

a pour densité par rapport à μ

$$f_\alpha(x) = \int_\Theta p_\theta(x) \mathrm{d}\Pi_\alpha(\theta).$$

Sous réserve d'existence, un estimateur du maximum de vraisemblance marginal est alors donné par

$$\widehat{\alpha}(\mathbf{X}) \in \arg\max_{\alpha \in \mathcal{A}} f_\alpha(\mathbf{X}).$$

Le principe est de marginaliser par rapport à la variable inconnue θ pour avoir une vraisemblance qui ne dépend que de α, puis de déterminer $\widehat{\alpha}$ qui maximise cette vraisemblance. Notons que cette étape d'estimation sort du cadre purement bayésien. Si la loi a posteriori pour la loi a priori Π_α est donnée par $\Pi_\alpha[\cdot \mid \mathbf{X}]$, alors la « loi a posteriori » donnée par la méthode bayésienne empirique est obtenue par *plug-in* : on remplace α par $\widehat{\alpha}$ et l'on considère la loi $\Pi_{\widehat{\alpha}}[\cdot \mid \mathbf{X}]$. Cette loi est généralement bien définie mais il y a abus de langage à l'appeler loi a posteriori : on a fait comme si cela ne changeait pas le modèle de prendre une loi a priori qui dépend des données. On l'appellera plutôt pseudo-loi a posteriori.

Notons que la méthode bayésienne empirique peut se voir comme une approximation de la méthode bayésienne hiérarchique, où la loi a posteriori de $\boldsymbol{\alpha}$ sachant $\mathbf{X}$ est approchée par une loi de Dirac en l'EMV.

Exemple 1.11. (1) Modèle gaussien

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

Déterminons un choix de loi a priori Π_μ par méthode bayésienne empirique, déjà dans le cas d'une observation $n = 1$ (le cas général sera vu en TD). La loi marginale de X_1 a pour densité

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

Ainsi la loi marginale de X_1 (lorsque l'a priori est Π_μ) est une $\mathcal{N}(\mu, 2)$. Donc

$$\widehat{\mu}(X_1) = \arg\max_{\mu \in \mathbb{R}} e^{-\frac{1}{4}(X_1 - \mu)^2} = X_1.$$

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

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

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

(2) Modèle exponentiel

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

Dans ce cadre, la méthode du maximum de vraisemblance marginal donne $\widehat{\lambda}(\mathbf{X}) = \overline{X}_n$ (voir TDs), donc $\Pi_{\widehat{\lambda}} = \mathcal{E}(\overline{X}_n)$ et la pseudo-loi a posteriori obtenue par la méthode bayésienne empirique ci-dessus est une loi $\text{Gamma}\big(n+1, (n+1)\overline{X}_n\big)$. On note que cette loi est centrée exactement en $1/\overline{X}_n$.

4. Régions de crédibilité

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

Par rapport à l'approche fréquentiste où l'on considère typiquement un estimateur $\widehat{\theta}(\mathbf{X})$ à valeurs dans Θ, on obtient ici une loi de probabilité aléatoire, $\Pi\left(\cdot \mid \mathbf{X}\right)$, à valeurs dans l'ensemble des mesures de probabilité sur Θ.

Nous avons vu à la Définition 1.3 que l'on pouvait à partir de la loi a posteriori construire des estimateurs ponctuels comme la moyenne, la médiane, 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. Ne pourrait-on pas utiliser $\Pi\left(\cdot \mid \mathbf{X}\right)$ pour obtenir des intervalles ou des régions de confiance ? Cette question motive la définition suivante.

Définition 1.11. Une région de crédibilité $A \subset \Theta$ de niveau (au moins) $1-\alpha$ pour $\Pi\left(\cdot \mid \mathbf{X}\right)$ est un ensemble p.s. mesurable [1] $A = A(\mathbf{X})$ tel que

$$\Pi\left(A \mid \mathbf{X}\right) \geq 1 - \alpha.$$

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 comme dans le cadre fréquentiste. Nous verrons cependant au Chapitre 5 qu'il est possible de faire une analyse

1. Au sens où $\mathbb{P}\left(\{\omega \in \Omega,\ A(\mathbf{X}(\omega)) \text{ est un borélien}\}\right) = 1$.

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. Ci-dessous nous voyons en détails deux constructions classiques.

4.1. Construction via des quantiles a posteriori. On suppose ici que $\Theta \subset \mathbb{R}$ et que la fonction de répartition a posteriori

$$t \mapsto F_{\mathbf{X}}(t) = \Pi\left(]-\infty, t] \mid \mathbf{X}\right)$$

est continue sur $\mathbb{R}$. Son inverse généralisée $F_{\mathbf{X}}^{-1}$ vérifie donc

$$\forall u \in]0,1]\,,\ F_{\mathbf{X}} \circ F_{\mathbf{X}}^{-1}(u) = u\,.$$

Dans ce cadre, en posant

$$a(\mathbf{X}) = F_{\mathbf{X}}^{-1}(\alpha/2) \qquad \text{et} \qquad b(\mathbf{X}) = F_{\mathbf{X}}^{-1}(1-\alpha/2)\,,$$

on a

$$\begin{aligned}\Pi\Big(]a_n(\mathbf{X}), b_n(\mathbf{X})] \mid \mathbf{X}\Big) &= \Pi\Big(]-\infty, b_n(\mathbf{X})] \mid \mathbf{X}\Big) - \Pi\Big(]-\infty, a_n(\mathbf{X})] \mid \mathbf{X}\Big) \\ &= F_{\mathbf{X}} \circ F_{\mathbf{X}}^{-1}(1-\alpha/2) - F_{\mathbf{X}} \circ F_{\mathbf{X}}^{-1}(\alpha/2) \\ &= 1 - \alpha/2 - \alpha/2 = 1-\alpha.\end{aligned}$$

Sous les hypothèses précédentes, nous avons donc construit un intervalle de crédibilité de niveau (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 $J(\mathbf{X}) =]-\infty, c_n(\mathbf{X})]$ avec $c_n(\mathbf{X}) = F_{\mathbf{X}}^{-1}(1-\alpha)$.

4.2. Régions de plus haute densité. Soit Q une loi de probabilité sur Θ de densité q par rapport à une mesure ν. On commence par définir les ensembles de niveau pour Q. Pour tout $y \geq 0$, on définit

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

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

Définition 1.12. Soit $\alpha \in]0,1[$. La région de plus haute densité (PHD) au niveau $1-\alpha$ pour une loi Q de densité q est la région $\mathcal{H} \subset \Theta$ donnée par

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

avec

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

Notons que comme $\alpha < 1$, on a $y_\alpha < +\infty$.

Lemme 1.7. *Soit $\mathcal{H}$ une région PHD au niveau $1-\alpha$ pour une loi Q sur Θ de densité q. Alors*

$$Q\left(\mathcal{H}\right) \geq 1-\alpha.$$

Démonstration. Notons

$$\mathcal{E}_\alpha = \{y \in \mathbb{R}_+, \quad Q(\mathcal{L}(y)) \geq 1-\alpha\} .$$

Ainsi $y_\alpha = \sup(\mathcal{E}_\alpha)$ et $\mathcal{H} = \mathcal{L}(y_\alpha)$. Soit (y_n) est une suite croissante d'éléments de $\mathcal{E}_\alpha$ qui converge vers y_α (on peut en trouver une par définition de la borne supérieure d'un ensemble). Par définition de $\mathcal{E}_\alpha$, on a, pour tout $n \geq 1$,

$$Q(\mathcal{L}(y_n)) \geq 1-\alpha. \tag{1.3}$$

Par croissance de (y_n) et par définition des ensembles de niveau, les ensembles $\mathcal{L}(y_n)$ sont emboîtés, i.e.

$$\mathcal{L}(y_1) \supset \mathcal{L}(y_2) \supset \dots$$

Le théorème de la limite monotone donne alors

$$Q(\mathcal{L}(y_n)) \underset{n\to\infty}{\longrightarrow} Q\left(\bigcap_{n\geq 1} \mathcal{L}(y_n)\right).$$

Montrons que $\bigcap_{n\geq 1} \mathcal{L}(y_n) = \mathcal{L}(y_\alpha)$. Soit $x \in \mathcal{L}(y_\alpha)$. Alors $q(x) \geq y_\alpha$, et comme $y_\alpha = \sup_{n\geq 1} y_n$, on a, pour tout $n \geq 1$, $q(x) \geq y_n$, i.e. $x \in \bigcap_{n\geq 1} \mathcal{L}(y_n)$. Inversement, si $x \in \bigcap_{n\geq 1} \mathcal{L}(y_n)$, alors $q(x) \geq y_n$ pour tout $n \geq 1$, et en passant à la limite dans l'inégalité, on obtient $q(x) \geq y_\alpha$, i.e. $x \in \mathcal{L}(y_\alpha)$. Le passage à la limite dans (1.3) donne donc $Q(\mathcal{L}(y_\alpha)) \geq 1-\alpha$. ■

La région de plus haute densité 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 illustre la définition précédente.

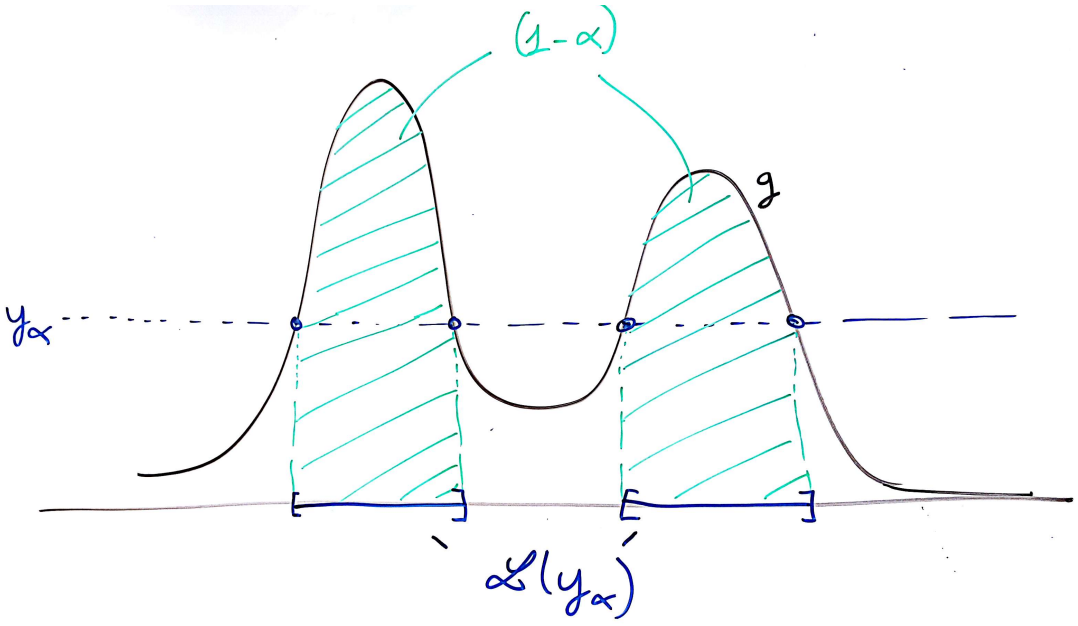


Figure 2. La réunion des deux intervalles en bleu sur l'axe des abscisses est la région PHD au niveau $1-\alpha$ pour la densité g dessinée. La région hachurée en vert a une aire égale à $(1-\alpha)$.

Dans l'énoncé ci-dessous, le volume d'un ensemble mesurable A est un synonyme pour $\nu(A) = \int_A \mathrm{d}\nu(\theta)$ (si ν est la mesure de Lebesgue, alors $\nu(A)$ est le volume usuel dans $\mathbb{R}^d$).

Theorème 1.8. *Dans le cadre de la Définition 1.12, la région PHD au niveau $1-\alpha$ est de volume minimal parmi les régions de même probabilité sous Q.*

Démonstration. Soit $\mathcal{H}$ une région PHD de niveau $1-\alpha$. Par définition, $\mathcal{H}$ est de la forme

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

Il suffit de montrer que si une région $A \subset \Theta$ a une probabilité au moins aussi grande que $\mathcal{H}$, soit $Q(A) \geq Q(\mathcal{H})$, alors $\nu(A) \geq \nu(\mathcal{H})$. Notons que

$$Q(A) = Q(A \cap \mathcal{H}) + Q(A \cap \mathcal{H}^c)$$
$$Q(\mathcal{H}) = Q(A \cap \mathcal{H}) + Q(\mathcal{H} \cap A^c).$$

Si $Q(A) \geq Q(\mathcal{H})$, on a donc $Q(A \cap \mathcal{H}^c) \geq Q(\mathcal{H} \cap A^c)$. D'autre part, on a

$$\nu(\mathcal{H} \cap A^c) = \int_{\mathcal{H}\cap A^c} \mathrm{d}\nu(\theta) \leq \int_{\mathcal{H}\cap A^c} \frac{q(\theta)}{y_\alpha} \mathrm{d}\nu(\theta)\,,$$

puisque si $\theta \in \mathcal{H}$, alors $q(\theta) \geq y_\alpha$. On obtient donc

$$\nu(\mathcal{H} \cap A^c) \leq \frac{Q(\mathcal{H} \cap A^c)}{y_\alpha} \leq \frac{Q(\mathcal{H}^c \cap A)}{y_\alpha} = \int_{\mathcal{H}^c\cap A} \frac{q(\theta)}{y_\alpha} \mathrm{d}\nu(\theta)\,.$$

Et comme, si $\theta \in \mathcal{H}^c$, alors $q(\theta) \leq y_\alpha$, on a

$$\int_{\mathcal{H}^c\cap A} \frac{q(\theta)}{y_\alpha} \mathrm{d}\nu(\theta) \leq \int_{\mathcal{H}^c\cap A} \mathrm{d}\nu(\theta) = \nu(\mathcal{H}^c \cap A)\,.$$

Ainsi $\nu(\mathcal{H} \cap A^c) \leq \nu(\mathcal{H}^c \cap A)$ et

$$\nu(\mathcal{H}) = \nu(\mathcal{H} \cap A) + \nu(\mathcal{H} \cap A^c) \leq \nu(cH \cap A) + \nu(A \cap \mathcal{H}^c) = \nu(A)\,,$$

ce qu'il fallait démontrer. ■

Remarque 1.12. Attention ! La région PHD au niveau $1-\alpha$ n'est pas nécessairement de volume minimal parmi les régions de masse $1-\alpha$. Par exemple, si Q est la loi uniforme sur $[0,1]$, alors la région PHD au niveau $1-\alpha$ correspond à tout l'intervalle $[0,1]$.

Définition 1.13. Dans une expérience statistique $(\mathbf{X}, \mathcal{P})$ avec une loi a priori Π sur $\Theta \subset \mathbb{R}^d$, soit $\Pi\left(\cdot \,\middle|\, \mathbf{X}\right)$ la loi a posteriori. La région PHD a posteriori au niveau $1-\alpha$ est la région PHD au niveau $1-\alpha$ pour la loi $\Pi\left(\cdot \,\middle|\, \mathbf{X}\right)$.

En général, les deux constructions 4.1 et 4.2 (par les quantiles et par les régions PHD) donnent des régions différentes. Un exemple est donné par la figure 2, où la région HPD est une union de deux intervalles disjoints, donc est nécessairement différente d'une région obtenue par quantiles qui correspond à 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 par les quantiles est souvent plus facile a mettre en œuvre, car elle nécessite seulement de connaître deux des quantiles a posteriori, tandis que la construction de régions PHD nécessite de travailler avec les ensembles de niveau de la densité a posteriori.

Chapitre 2

Simulation de la loi a posteriori

En pratique, la loi a posteriori est très souvent un objet extrêmement compliqué qui nécessite de savoir calculer des intégrales difficiles. Si l'on n'a pas d'accès direct à cette loi dont la densité n'a pas de forme explicite, on peut néanmoins chercher à la simuler, c'est-à-dire à échantillonner des valeurs selon cette loi.

Sauf dans certains cas particuliers (famille de lois conjuguée...), il peut être difficile de déterminer explicitement la loi a posteriori. Il peut alors s'avérer compliqué d'évaluer des quantités comme la moyenne, la médiane ou les quantiles a posteriori. Par exemple, la moyenne a posteriori s'écrit comme une intégrale contre la loi a posteriori :

$$\int_\Theta \theta \mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

Comment évaluer ce genre d'intégrales si l'on ne connaît pas précisément la loi a posteriori ?

Dans un premier temps, nous verrons comment, à partir d'un générateur de la loi uniforme, on peut simuler de nombreuses lois, pourvu qu'elles ne soient pas trop méchantes (nous verrons au Chapitre 6 comment affronter des lois plus méchantes). Puis on verra que si l'on sait simuler selon une loi P, alors on peut approcher des intégrales de la forme $\int \phi(x)\mathrm{d}P(x)$. Pour ceux qui souhaitent en savoir plus sur la simulation de variables aléatoires et les méthodes de Monte-Carlo, une référence classique est l'ouvrage de Luc Devroye, *Non-uniform random variate generation*, disponible là : http://luc.devroye.org/rnbookindex.html.

1. Simulation de lois gentilles

Dans toute méthode de simulation de variables aléatoires, on suppose toujours que l'on dispose d'un ingrédient de base : un générateur de la loi uniforme sur $[0, 1]$ auquel on peut faire appel pour obtenir des réalisations indépendantes. Tout ordinateur est équipé d'un tel générateur. Certes en réalité, il ne s'agit jamais de réalisations vraiment aléatoires, on parle de nombres *pseudo-aléatoires*. Dans le cadre de ce cours cependant, on supposera que l'on peut vraiment simuler des variables aléatoires uniformes. Pour plus de détails sur les générateurs de nombres aléatoires, la page wikipedia https://fr.wikipedia.org/wiki/Générateur_de_nombres_pseudo-aléatoires présente bien les grandes méthodes.

1.1. Méthode de la transformée inverse. La première grande méthode de simulation est la méthode de la transformée inverse. On souhaite simuler une variable aléatoire réelle X, de loi de fonction de répartition F. Soit F^{-1} l'inverse généralisée de F, i.e.

$$\forall u \in [0, 1]\,,\ F^{-1}(u) = \inf\{x \in \mathbb{R},\, F(x) \geq u\}\ ,$$

avec les conventions $\inf \mathbb{R} = -\infty$ et $\inf \emptyset = +\infty$. Remarquons que, même si l'on n'a pas en général l'équivalence $F^{-1}(u) = x \Leftrightarrow F(x) = u$, on a néanmoins toujours $F^{-1}(u) \leq x \Leftrightarrow u \leq$

$F(x)$. En effet, si $F(x) \geq u$, alors par définition de F^{-1}, $F^{-1}(u) \leq x$. Et si $F(x) < u$, comme F est continue à droite, il existe $\varepsilon > 0$ tel que $F(x+\varepsilon) < u$. Mais alors $F^{-1}(u) \geq x + \varepsilon > x$.

Soit maintenant $U \sim \text{Unif}[0,1]$. Pour tout $x \in \mathbb{R}$, on a

$$\mathbb{P}\left(F^{-1}(U) \leq x\right) = \mathbb{P}\left(U \leq F(x)\right) = F(x)\,.$$

Autrement dit, $F^{-1}(U) \sim X$. Ainsi, si l'on sait calculer F^{-1}, on sait simuler une variable aléatoire de loi F.

Cette méthode simple n'est pas toujours réalisable en pratique. En effet, elle nécessite de savoir inverser la fonction de répartition, ce qui n'est pas toujours possible. Il y a aussi des cas où la fonction de répartition elle-même n'est pas accessible autrement que sous forme d'une intégrale de la densité, l'exemple typique étant la loi normale. La méthode de rejet permet, dans certaines situations, de simuler des variables aléatoires dont on ne connaît explicitement que la densité, voire la densité à une constante multiplicative près.

1.2. Méthode de rejet. On souhaite simuler une variable aléatoire à valeurs dans un espace mesuré $(E, \mathcal{E}, \mu)$, de densité f par rapport à μ. On suppose que l'on sait simuler selon une autre densité g (par rapport à μ aussi) et que cette densité est telle qu'il existe une constante $m \geq 1$ telle que pour tout $y \in E$, $f(y) \leq mg(y)$, et que l'on sait calculer, pour tout $y \in E$ tel que $g(y) > 0$, le rapport $r(y) = \frac{f(y)}{mg(y)}$ (si $g(y) = 0$ alors $f(y) = 0$ aussi et l'on pose $r(y) = 0$).

Soient $(U_i)_{i\geq 1}$ une suite i.i.d. de loi Unif$([0,1])$ et $(Y_i)_{i\geq 1}$ une suite i.i.d. de loi de densité g, indépendante de $(U_i)_{i\geq 1}$.

Proposition 2.1 (Algorithme de rejet)**.** *Soit $\tau = \inf\{i \in \mathbb{N}^*,\, r(Y_i) \geq U_i\}$. La variable $X = Y_\tau$ est de densité f. Par ailleurs, τ est de loi géométrique de paramètre $1/m$ et est indépendante de X.*

Démonstration. Soit A un élément de la tribu $\mathcal{E}$ et $n \in \mathbb{N}^*$. On a, par indépendance des tirages,

$$\begin{aligned}\mathbb{P}\left(Y_\tau \in A, \tau = n\right) &= \mathbb{P}\left(r(Y_1) < U_1, \ldots, r(Y_{n-1}) < U_{n-1}, r(Y_n) \geq U_n, Y_n \in A\right)\\ &= \mathbb{P}(r(Y) < U)^{n-1}\mathbb{P}(r(Y) \geq U, Y \in A)\,,\end{aligned}$$

où Y est de densité g et U uniforme sur $[0,1]$, indépendante de Y. Par le théorème de Fubini et le fait que g et f sont des densités, on a

$$\mathbb{P}(r(Y) < U) = \int_E \int_0^1 \mathbb{1}_{\{r(y)<u\}} g(y)\mathrm{d}u\mathrm{d}\mu(y) = \int_E \left(1 - r(y)\right) g(y)\mathrm{d}\mu(y) = 1 - \int_E \frac{f(y)}{m}dy = 1 - \frac{1}{m}\,.$$

De même

$$\mathbb{P}(r(Y) \geq U, Y \in A) = \int_E \int_0^1 \mathbb{1}_{\{r(y)\geq u\}} \mathbb{1}_{\{y\in A\}} g(y)\mathrm{d}u\mathrm{d}\mu(y) = \int_A r(y)g(y)\mathrm{d}\mu(y) = \frac{1}{m}\int_A f(y)\mathrm{d}\mu(y)\,.$$

Ainsi

$$\mathbb{P}\left(Y_\tau \in A, \tau = n\right) = \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{m} \int_A f(y)\mathrm{d}\mu(y)\,.$$

Pour $A = E$, on a donc

$$\mathbb{P}\left(\tau = n\right) = \mathbb{P}\left(Y_\tau \in E, \tau = n\right) = \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{m}\,.$$

La variable τ est donc géométrique de paramètre $1/m$. En particulier, $\tau < +\infty$ p.s. et en sommant sur $n \in \mathbb{N}^*$, on obtient

$$\mathbb{P}\left(Y_\tau \in A\right) = \int_A f(y)\mathrm{d}\mu(y)\,.$$

Autrement dit, la variable Y_τ est bien de densité f, et comme $\mathbb{P}\left(Y_\tau \in A, \tau = n\right) = \mathbb{P}(Y_\tau \in A)\mathbb{P}(\tau = n)$, les variables τ et Y_τ sont bien indépendantes. ■

Remarque 2.1. L'espérance de τ étant égale à m, il faut en moyenne attendre m essais avant d'obtenir une simulation de X. Pour limiter le nombre moyen de rejets, il est donc important de faire en sorte que m soit aussi petit possible, i.e. de choisir g aussi proche que possible de f.

Exemple 2.2 (Simulation d'une variable uniforme sur un sous-ensemble du cube)**.** Soit A un ensemble borélien du cube $[0,1]^d$. S'il est très facile de simuler une variable uniforme sur $[0,1]^d$ (c'est un vecteur de d variables uniformes indépendantes sur $[0,1]$), il peut être bien plus complexe de simuler directement une variable uniforme sur A. Supposons cependant que l'on sache dire, pour tout élément du cube, s'il est dans A ou non. L'algorithme de rejet consiste alors à tirer des variables $Y_1, Y_2, \ldots$ indépendantes uniformes sur $[0,1]^d$, jusqu'au premier temps τ où $Y_\tau \in A$. La variable Y_τ est alors uniformément distribuée sur A. En effet, en notant $\lambda(A)$ le volume de A (sa mesure de Lebesgue), on a

$$\forall y \in \mathbb{R}^d,\ f(y) = \frac{1}{\lambda(A)}\mathbb{1}_{\{y\in A\}} \leq \frac{1}{\lambda(A)}\mathbb{1}_{\{y\in[0,1]^d\}}\,.$$

Ainsi, on peut prendre $m = \frac{1}{\lambda(A)}$ et le rapport r s'écrit alors simplement $r(y) = \mathbb{1}_{\{y\in A\}}$. Il n'y a donc pas besoin de tirer les variables (U_i) puisque l'on sait que le premier succès est précisément le premier temps où l'on tombe dans A. Notons qu'il n'y a pas non plus besoin de connaître $\lambda(A)$ pour implémenter l'algorithme. Remarquons cependant que si $\lambda(A)$ est très petit, cette méthode n'est pas satisfaisante puisque très coûteuse en temps de calcul (il faut attendre très longtemps avant de tomber dans A).

Exemple 2.3 (Loi gamma)**.** La loi $\Gamma(p,\lambda)$, avec $p,\lambda > 0$, est typiquement un exemple de loi dont la densité est relativement simple mais dont la fonction de répartition n'a pas d'expression explicite, donc la méthode de la transformée inverse ne convient pas. Quand $p > 1$, on peut utiliser l'algorithme de rejet avec comme densité auxiliaire la densité d'une variable exponentielle de paramètre μ, soit $g(y) = \mu e^{-\mu y}\mathbb{1}_{y\geq 0}$ (qui se simule bien par la méthode de la transformée inverse). Notons

$$m_\mu = \sup_{y\in\mathbb{R}_+} \frac{\lambda^p y^{p-1} e^{-\lambda y}}{\Gamma(p)\mu e^{-\mu y}}\,.$$

Un peu de calcul montre que, si $\mu < \lambda$, alors le supremum est atteint en $y = \frac{p-1}{\lambda-\mu}$ et vaut

$$m_\mu = \frac{\lambda^p (p-1)^{p-1} e^{1-p}}{\Gamma(p)\mu(\lambda-\mu)^{p-1}}\,,$$

et encore un peu de calcul montre que la constante m_μ est minimale pour $\mu^* = \frac{\lambda}{p}$, et que

$$m_{\mu^*} = \frac{p^p e^{1-p}}{\Gamma(p)}\,.$$

Notons que par la formule de Stirling, $m_{\mu^*} \sim_{p\to+\infty} \frac{e\sqrt{p}}{\sqrt{2\pi}}$. Donc plus p est grand, moins cette méthode est performante.

2. Méthodes de Monte-Carlo pour le calcul d'intégrales

Soit P une loi sur $(E,\mathcal{E})$ et soit $\phi : E \to \mathbb{R}$ une fonction mesurable P-intégrable connue (i.e. il est facile d'avoir accès aux valeurs $\phi(x)$ pour tout $x \in E$). On souhaite calculer

$$\mathbf{I} = \int_E \phi(x)\mathrm{d}P(x)\,.$$

Supposons par exemple que $E = [0,1]^d$ et que P admet une densité f par rapport à la mesure de Lebesgue sur $[0,1]^d$. Alors si f est connue, une façon simple d'approcher l'intégrale $\mathbf{I}$ est de découper le cube $[0,1]^d$ en N^d sous-cubes plus petits et d'approcher sur chaque sous-cube la fonction $\phi \times f$ par une fonction plus simple, par exemple une constante, ou une fonction affine. Si l'on prend le cas où l'on approche la fonction par une constante, on retrouve les sommes de Riemann, pour lesquelles on approche $\mathbf{I}$ par

$$\frac{1}{N^d} \sum_{i\in[\![1,N]\!]^d} (\phi \times f)(x_i)\ ,$$

avec par exemple $x_i = \frac{1}{N}(i_1,\dots,i_d)$. 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. C'est le principe des méthodes de Monte-Carlo.

2.1. Monte-Carlo standard. Au lieu de prendre des points fixés x_i à la base de notre approximation, on peut les tirer au hasard. Soit $X_1,\dots,X_N$ des variables aléatoires i.i.d. de loi P. Alors par la loi des grands nombres,

$$\mathbf{I}_N = \frac{1}{N}\sum_{i=1}^N \phi(X_i) \xrightarrow{\text{p.s.}} \int \phi(x)\mathrm{d}P(x) = \mathbf{I}\,. \tag{2.1}$$

De plus, si $\int \phi^2 \mathrm{d}P < \infty$, on a aussi par le théorème central limite, quand $N \to \infty$,

$$\sqrt{N}(\mathbf{I}_N - \mathbf{I}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(\phi(X))),$$

où $\mathrm{Var}(\phi(X))) = \int(\phi(x) - \mathbf{I})^2 \mathrm{d}P(x)$. 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 $N^{-1/2}$, indépendamment de la dimension d du problème. De plus, cette méthode ne suppose pas de connaître la densité f explicitement, mais simplement de savoir simuler des variables de loi P.

Dans certains cas, cette approche peut cependant s'avérer problématique. D'une part, on ne sait pas forcément simuler des variables selon P. D'autre part, le nombre de tirages nécessaires avant d'avoir une bonne précision peut être extrêmement grand. Par exemple, supposons que P est une loi $\mathcal{N}(0,1)$ et que $\phi(x) = \mathbb{1}_{x>3}$. Ainsi, $\mathbf{I} = \mathbb{P}(X > 3)$ où $X \sim \mathcal{N}(0,1)$. Si l'on tire $X_1, \ldots, X_N$ i.i.d. de loi $\mathcal{N}(0,1)$, il faut prendre N extrêmement grand avant d'obtenir une observation qui soit plus grande que 3. Dans le cas où l'intégrale à approcher correspond à un événement rare (comme l'événement qu'une variable gaussienne $\mathcal{N}(0,1)$ soit supérieure à 3), on a plutôt intérêt à simuler des variables selon une autre loi pour laquelle l'événement en question est « moins rare », et à évaluer, pour chaque observation, une fonction modifiée qui prenne en compte ce changement de mesure. Cette approche s'appelle l'échantillonnage d'importance (*importance sampling* en anglais), ou échantillonnage préférentiel.

2.2. Monte-Carlo par *Importance Sampling*. On cherche toujours à apprecher l'intégrale $\mathbf{I} = \int \phi \mathrm{d}P$. On suppose que P possède une densité p par rapport à une mesure σ-finie μ sur E. Soit Q une autre loi sur E, de densité q par rapport à μ, selon laquelle on sait simuler efficacement, et qui vérifie :

$$\forall x \in E\,,\ q(x) = 0 \ \Rightarrow\ \phi(x)p(x) = 0\,. \tag{2.2}$$

Notons que si $Y \sim Q$, alors, sous la condition de P-intégrabilité de ϕ, on a

$$\mathbb{E}\left[\phi(Y)\frac{p(Y)}{q(Y)}\right] = \int_E \phi(y)\frac{p(y)}{q(y)}q(y)\mathrm{d}\mu(y) = \int_E \phi(x)p(x)\mathrm{d}\mu(x) = \mathbf{I}\,.$$

Soit $Y_1, \ldots, Y_N$ un tirage i.i.d. suivant la loi Q. On pose

$$\mathbf{J}_N = \frac{1}{N}\sum_{i=1}^{N} \phi(Y_i)\frac{p(Y_i)}{q(Y_i)}\,.$$

La loi des grands nombres donne alors

$$\mathbf{J}_N \xrightarrow{\text{p.s.}} \mathbf{I}.$$

On note que l'on ne doit plus simuler suivant la loi P mais suivant la loi Q, que l'on choisit.

Si l'on veut avoir un théorème central limite, il faut pouvoir vérifier la condition de moment d'ordre 2, c'est-à-dire

$$\mathbb{E}\left[\left(\phi(Y)\frac{p(Y)}{q(Y)}\right)^2\right] = \int_E \frac{\phi(y)^2 p(y)^2}{q(y)}\mathrm{d}\mu(y) < \infty.$$

Reprenons l'exemple de l'approximation de $\mathbf{I} = \mathbb{P}(X > 3)$ où $X \sim \mathcal{N}(0,1)$. Si l'on utilise la méthode de Monte-Carlo simple, on pose tout simplement

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

où les variables X_i sont i.i.d. $P = \mathcal{N}(0,1)$. Dans ce cas, on a

$$\sqrt{N}(\mathbf{I}_N - \mathbf{I}) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \sigma^2_{\mathrm{MC}}\big)\,,$$

avec $\sigma^2_{\mathrm{MC}} = \mathbf{I}(1-\mathbf{I}) \approx \mathbb{P}(\mathcal{N}(0,1) > 3)$. En utilisant la méthode de Monte-Carlo par échantillonnage d'importance avec $Q = \mathcal{N}(3,1)$, on pose

$$\mathbf{J}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{p(Y_i)}{q(Y_i)} \mathbb{1}_{Y_i>3},$$

où $Y_1, \ldots, Y_N$ sont i.i.d. de loi Q. On a alors

$$\sqrt{N}(\mathbf{J}_N - \mathbf{I}) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \sigma^2_{\mathrm{IS}}\big)\,,$$

avec

$$\sigma^2_{\mathrm{IS}} \leq \int_3^\infty \frac{p(y)^2}{q(y)}\mathrm{d}y = \int_3^\infty \frac{\mathrm{e}^{-y^2+\frac{(y-3)^2}{2}}}{\sqrt{2\pi}}\mathrm{d}y = \int_3^\infty \frac{\mathrm{e}^{-\frac{1}{2}(y+3)^2+9}}{\sqrt{2\pi}}\mathrm{d}y = \mathrm{e}^9 \int_6^\infty \frac{\mathrm{e}^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\mathrm{d}y\,.$$

On peut constater numériquement que la variance obtenue par échantillonnage d'importance est bien plus petite. On a en effet

$$\sigma^2_{\mathrm{MC}} = \mathbf{I}(1-\mathbf{I}) \approx 10^{-3} \gg \sigma^2_{\mathrm{IS}} \approx \mathrm{e}^9 \mathbb{P}(\mathcal{N}(0,1) > 6) \approx 10^{-6}.$$

Une question naturelle est celle du choix optimal de la loi Q. 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 2.2. *Le choix optimal théorique de la densité q pour la méthode d'échantillonnage d'importance est donné par*

$$\forall y \in E\,,\ q^\star(y) = \frac{|\phi(y)|p(y)}{\int_E |\phi(x)|p(x)\mathrm{d}\mu(x)}.$$

Démonstration. Pour un choix de loi Q vérifiant l'hypothèse (2.2), la variance s'écrit

$$\mathbb{E}\left[\left(\phi(Y)\frac{p(Y)}{q(Y)}\right)^2\right] - \mathbf{I}^2\,,$$

où $Y \sim Q$. Le terme $\mathbf{I}^2$ ne dépendant pas de Q, il suffit de minimiser le premier terme. Or, par l'inégalité de Jensen,

$$\mathbb{E}\left[\left(\phi(Y)\frac{p(Y)}{q(Y)}\right)^2\right] \geq \mathbb{E}\left[\left|\phi(Y)\frac{p(Y)}{q(Y)}\right|\right]^2 = \mathbb{E}\left[|\phi(X)|\right]^2\,,$$

où $X \sim P$. Définissons la loi $Q^\star$ de densité $q^\star(x) = \frac{|\phi(x)|}{\mathbb{E}[|\phi(X)|]}p(x)$. C'est bien une loi de probabilité, elle satisfait la condition (2.2), et elle atteint la borne ci-dessus car, pour $Y \sim Q^\star$, on a

$$\mathbb{E}\left[\left(\phi(Y)\frac{p(Y)}{q^\star(Y)}\right)^2\right] = \int_E \phi(y)^2\frac{p(y)^2}{q^\star(y)}\mathrm{d}\mu(y) = \mathbb{E}\left[|\phi(X)|\right]\int_E |\phi(y)|p(y)\mathrm{d}\mu(y) = \mathbb{E}\left[|\phi(X)|\right]^2\,.$$

■

Remarque 2.4. Si $\phi \geq 0$, alors $\mathrm{d}Q^\star(x) = \frac{\phi(x)}{\mathbb{E}[\phi(X)]}\mathrm{d}P(x)$ et la variance est nulle. Dans ce cas un seul tirage suffit : $\mathbf{J}_1 = \mathbf{I}$. Mais c'est complètement irréaliste de supposer que l'on puisse simuler selon Q^* puisque c'est précisément $\mathbb{E}[\phi(X)]$ que l'on souhaite estimer.

Exemple 2.5. Dans de nombreuses situations, Q est la mesure uniforme sur un ensemble discret E « très grand ». Pour pouvoir échantillonner selon Q, il faudrait pouvoir énumérer les éléments de E, et cela serait bien trop coûteux. En fait, l'échantillonnage sur E et l'estimation de la taille de E sont deux problèmes étroitement liés. Considérons Λ_n l'ensemble des chemins auto-évitants (qui ne repassent pas deux fois sur le même sommet) sur la grille $[\![0, n]\!]^2$, qui partent de $(0, 0)$ et arrivent en (n, n). L'ensemble Λ_n est complexe et il n'existe pas de façon simple de compter ses éléments. En 1976, Donald Knuth a proposé un estimateur pour $|\Lambda_n|$ dont le principe repose sur l'échantillonnage d'importance. L'idée est la suivante : on ne sait pas tirer uniformément dans Λ_n mais l'on sait facilement tirer un chemin auto-évitant selon une autre mesure : on peut construire séquentiellement un chemin auto-évitant aléatoire en partant de $(0, 0)$ et en choisissant, à chaque temps, un sommet voisin « acceptable » (i.e. tel que le chemin obtenu reste auto-évitant et puisse être étendu en un chemin auto-évitant jusqu'en (n, n)) uniformément au hasard (à chaque temps, il y a toujours 1, 2, ou 3 voisins acceptables). Si l'on touche un des bords haut ou droit, on étend de façon canonique le chemin obtenu en un chemin auto-évitant qui arrive jusqu'en (n, n) (on longe le bord jusqu'en (n, n)). Soit W un élément aléatoire de Λ_n obtenu ainsi, et, pour $w \in \Lambda_n$, posons $p(w) = \mathbb{P}(W = w)$, et notons que $p(w)$ se calcule facilement. Voir Figure 2.5. Observons que

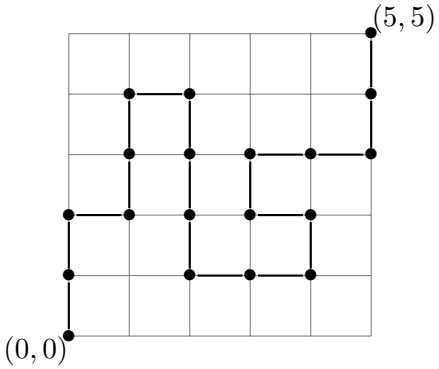
$$\mathbb{E}\left[\frac{1}{p(W)}\right] = \sum_{w\in\Lambda_n} p(w)\frac{1}{p(w)} = |\Lambda_n|\,.$$

Soient alors $W_1, \ldots, W_n$ des chemins aléatoires de même loi que W. Par la loi des grands nombres, l'estimateur

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p(W_i)}$$

est un estimateur consistant de $|\Lambda_n|$.

FIGURE 1. Un chemin auto-évitant w de $[\![0, 5]\!]^2$, avec $p(w) = \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{3}\right)^4 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{3}\right)^4 \times 1 \times \left(\frac{1}{2}\right)^2$.



Ainsi, passer par une autre mesure de probabilité, et considérer une fonctionnelle calculable d'échantillons issus de cette nouvelle loi, nous a permis d'estimer la taille de l'ensemble considéré.

2.3. Application : estimation de la moyenne a posteriori. Sauf dans quelques cas particuliers, la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$, avec $\mathbf{X} = (X_1, \dots, X_n)$, est un objet extrêmement complexe. Rappelons que l'on a

$$\forall \theta \in \Theta\,,\ \pi(\theta \mid \mathbf{X}) = \frac{\pi(\theta)\prod_{i=1}^n p_\theta(X_i)}{\int_\Theta \pi(\theta)\prod_{i=1}^n p_\theta(X_i) d\nu(\theta)}\,.$$

Simuler selon la loi a posteriori peut être très compliqué. Mais si l'on cherche non pas à simuler selon $\Pi[\cdot \mid \mathbf{X}]$ mais à approcher une intégrale contre cette loi, alors le problème se simplifie, et l'on peut remarquer qu'il suffit en fait de savoir simuler selon la loi a priori Π. En effet, soit $\phi : \Theta \to \mathbb{R}$ et supposons que l'on veuille évaluer l'intégrale $\int_\Theta \phi(\theta)\mathrm{d}\Pi(\theta \mid \mathbf{X})$ (par exemple la moyenne a posteriori pour $\phi = \mathsf{id}$). En notant pour alléger les notations $p_\theta(\mathbf{X}) = \prod_{i=1}^n p_\theta(X_i)$, on a

$$\int_\Theta \phi(\theta)\mathrm{d}\Pi(\theta \mid \mathbf{X}) = \frac{\int_\Theta \phi(\theta)p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\nu(\theta)}{\int_\Theta p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\nu(\theta)}\,.$$

Maintenant si l'on sait générer des variables i.i.d. $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m$ de loi Π, on a, par la loi des grands nombres, quand $m \to +\infty$,

$$\frac{1}{m}\sum_{j=1}^m \phi(\boldsymbol{\theta}_j)p_{\boldsymbol{\theta}_j}(\mathbf{X}) \xrightarrow[m\to+\infty]{\text{p.s.}} \int_\Theta \phi(\theta)p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\nu(\theta)\,,$$

et

$$\frac{1}{m}\sum_{j=1}^m p_{\boldsymbol{\theta}_j}(\mathbf{X}) \xrightarrow[m\to+\infty]{\text{p.s.}} \int_\Theta p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\nu(\theta)\,.$$

Ainsi, par continuité l'estimateur

$$\widehat{\phi}_n^{(m)} = \frac{\sum_{j=1}^m \phi(\boldsymbol{\theta}_j)p_{\boldsymbol{\theta}_j}(\mathbf{X})}{\sum_{j=1}^m p_{\boldsymbol{\theta}_j}(\mathbf{X})}$$

est fortement consistant : $\widehat{\phi}_n^{(m)} \xrightarrow[m\to+\infty]{\text{p.s.}} \int_\Theta \phi(\theta)\mathrm{d}\Pi(\theta \mid \mathbf{X})$. Attention, dans cette convergence, n est fixé et c'est m qui tend vers $+\infty$. En utilisant une méthode Delta en dimension 2, on peut aussi montrer que $\widehat{\phi}_n^{(m)}$ est asymptotiquement normal.

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 deux critères classiques : le risque de Bayes et le risque minimax, ainsi que certaines relations 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 ?

1. Risque ponctuel, risque bayésien, risque maximal

On se place dans le cadre d'une expérience $(\mathbf{X}, \mathcal{P})$ avec $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, le modèle étant dominé par une mesure μ sur E ($\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$), et l'on suppose que l'application $(\theta, x) \mapsto p_\theta(x)$ est mesurable.

Définition 3.1. Une fonction de perte ℓ est une fonction $\ell : \Theta \times \Theta \to \mathbb{R}_+$ mesurable avec

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

Exemple 3.1.

- Si $\Theta \subset \mathbb{R}$, la fonction $\ell(\theta, \theta') = (\theta - \theta')^2$ s'appelle la perte quadratique.
- Plus généralement, la perte quadratique dans $\Theta \subset \mathbb{R}^d$ est donnée par

$$\ell(\theta, \theta') = \|\theta - \theta'\|^2 = \sum_{i=1}^{d} (\theta_i - \theta'_i)^2.$$

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

Notons que les deux derniers exemples donnent bien des fonctions de perte à condition que le modèle soit identifiable.

On parlera parfois de fonction de perte pour une fonction $\ell : \Theta \times \Theta \to \mathbb{R}_+$ mesurable qui vérifie seulement

$$\forall\, \theta, \theta' \in \Theta, \quad \theta = \theta' \Rightarrow \ell(\theta, \theta') = 0.$$

Cela permet notamment d'inclure la perte de classification.

Exemple 3.2. Supposons que

$$\Theta = \Theta_0 \cup \Theta_1, \qquad \Theta_0 \cap \Theta_1 = \emptyset.$$

On définit la fonction de perte de classification par

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

Notons que $L_C(\theta, \theta') = 0$ si et seulement si θ et θ' sont dans la même région Θ_0 ou Θ_1. Il s'agit de la perte naturellement utilisée lorsque l'on veut construire un test pour répondre à une question binaire sur θ. Voir Chapitre 4.

1.1. Fonction de risque. On rappelle qu'un estimateur T est une fonction mesurable $T : E \to \Theta$.

Définition 3.2. La fonction de risque (ou simplement le risque) d'un estimateur T pour la fonction de perte ℓ est l'application

$$\begin{array}{rccl} \mathbf{R}(\cdot, T) & : & \Theta & \to & \mathbb{R}_+ \\ & & \theta & \mapsto & \mathbf{R}(\theta, T) = \mathbb{E}_\theta\left[\,\ell(\theta, T(\mathbf{X}))\,\right] = \int_E \ell(\theta, T(x))\mathrm{d}P_\theta(x). \end{array}$$

Le risque au point θ de l'estimateur T est donc la perte moyenne de T en θ (on parle de risque ponctuel).

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* en reprenant l'exemple du modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ pour la perte quadratique : l'estimateur constant égal à $\theta_0 \in \mathbb{R}$ a un risque nul en θ_0 donc est meilleur que n'importe quel autre estimateur en ce point, mais pour tous les θ tels que $(\theta - \theta_0)^2 > 1/n$, on préfère l'estimateur $\overline{X}_n$ qui a un risque constant sur $\mathbb{R}$ égal à $1/n$.

Dans la suite, on fixe une fonction de perte. Les définitions et résultats qui suivent s'entendent donc à fonction de perte fixée, même si, pour alléger les notations, on ne rappellera pas tout le temps cette dépendance.

Intuition. La notion de risque bayésien définie ci-dessous va nous donner une réponse possible à la question de trouver un estimateur de risque optimal. Cependant, cette notion dépendra de l'a priori choisi, ce qui n'en fait pas une réponse « universelle ». Le risque minimax défini ensuite est lui plus universel au sens où il ne dépend pas d'un prior particulier, mais correspond à une vision un peu pessimiste (on cherche un estimateur T qui minimise le pire risque possible, soit $\sup_{\theta\in\Theta} \mathbf{R}(\theta, T)$).

1.2. Risque bayésien et estimateurs de Bayes. Soit Π une loi a priori donnée sur Θ, de densité π par rapport à ν. Rappelons que nous travaillons également à fonction de perte ℓ donnée. Ainsi les définitions ci-dessous dépendent implicitement de ℓ.

Définition 3.3. On appelle risque de Bayes ou parfois risque bayésien pour l'estimateur T et la loi a priori Π la quantité

$$\begin{aligned}\mathbf{R}_B(\Pi, T) &= \mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X}))\right] \\ &= \int_\Theta \int_E \ell(\theta, T(x)) \mathrm{d}P_\theta(x) \mathrm{d}\Pi(\theta) \\ &= \int_\Theta \mathbf{R}(\theta, T) \mathrm{d}\Pi(\theta) \\ &= \mathbb{E}\left[\mathbf{R}(\boldsymbol{\theta}, T)\right] ,\end{aligned}$$

où la deuxième égalité vient du théorème de Fubini. En effet, en se rappelant que le couple $(\mathbf{X}, \boldsymbol{\theta})$ a pour densité $(x, \theta) \mapsto p_\theta(x)\pi(\theta)$ par rapport à $\mu \otimes \nu$ et en utilisant le théorème de Fubini, on a

$$\begin{aligned}\mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X}))\right] &= \int_{E\times\Theta} \ell(\theta, T(x)) p_\theta(x)\pi(\theta) \mathrm{d}(\mu \otimes \nu)(x, \theta) \\ &= \int_\Theta \left(\int_E \ell(\theta, T(x)) p_\theta(x) \mathrm{d}\mu(x) \right) \pi(\theta) \mathrm{d}\nu(\theta) \\ &= \int_\Theta \left(\int_E \ell(\theta, T(x)) \mathrm{d}P_\theta(x) \right) \mathrm{d}\Pi(\theta) .\end{aligned}$$

Une autre façon de le voir est par conditionnement : en remarquant que $\mathbf{R}(\boldsymbol{\theta}, T) = \mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X}) \,\middle|\, \boldsymbol{\theta}\right]$, on a

$$\mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X}))\right] = \mathbb{E}\left[\mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X})) \,\middle|\, \boldsymbol{\theta}\right]\right] = \mathbb{E}\left[\mathbf{R}(\boldsymbol{\theta}, T)\right] .$$

Définition 3.4. Un estimateur $T^\star$ est dit de Bayes pour la loi a priori Π si

$$\mathbf{R}_B(\Pi, T^\star) = \inf_T \mathbf{R}_B(\Pi, T),$$

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

$$\mathbf{R}_B(\Pi) = \inf_T \mathbf{R}_B(\Pi, T)$$

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 contre la loi a priori Π sur Θ. Un tel estimateur minimise donc un risque « en moyenne selon Π ».

Exemple 3.3. Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$, considérons la loi a priori $\Pi = \mathcal{N}(0, 1)$ et la fonction de perte quadratique $\ell(\theta, \theta') = (\theta - \theta')^2$. Calculons le risque de Bayes pour Π des estimateurs suivants

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

Pour l'estimateur constant T_1, on a

$$\begin{aligned}\mathbf{R}_B(\Pi, T_1) &= \int_\Theta \mathbf{R}(\theta, T_1)\mathrm{d}\Pi(\theta)\\ &= \int_\Theta \int_E (\theta - T_1(x))^2 \mathrm{d}P_\theta(x)\mathrm{d}\Pi(\theta)\\ &= \int_\Theta \int_E \theta^2 \mathrm{d}P_\theta(x)\mathrm{d}\Pi(\theta) = \int_\Theta \theta^2 \mathrm{d}\Pi(\theta) = 1.\end{aligned}$$

Pour T_2, rappelons d'abord que sous la loi P_θ, la variable $\overline{X}_n - \theta$ suit la loi $\mathcal{N}(0, 1/n)$. Ainsi, comme nous l'avons déjà vu, pour tout $\theta \in \Theta$, $\mathbf{R}(\theta, T_2) = \mathbb{E}_\theta[(\overline{X}_n - \theta)^2] = 1/n$, et

$$\mathbf{R}_B(\Pi, T_2) = \int_\Theta \frac{1}{n}\mathrm{d}\Pi(\theta) = \frac{1}{n}.$$

Pour T_3, nous calculons d'abord, pour $\theta \in \Theta$,

$$\begin{aligned}\mathbf{R}(\theta, T_3) &= \mathbb{E}_\theta\left[\left(\frac{n}{n+1}\overline{X}_n - \theta\right)^2\right]\\ &= \mathbb{E}_\theta\left[\left(\frac{n}{n+1}(\overline{X}_n - \theta) - \frac{\theta}{n+1}\right)^2\right]\\ &= \left(\frac{n}{n+1}\right)^2 \mathbb{E}_\theta\left[(\overline{X}_n - \theta)^2\right] + \left(\frac{\theta}{n+1}\right)^2 \qquad \text{car } \mathbb{E}_\theta[\overline{X}_n - \theta] = 0\\ &= \frac{n}{(n+1)^2} + \frac{\theta^2}{(n+1)^2},\end{aligned}$$

où la dernière égalité vient du fait que $\mathbb{E}_\theta\left[(\overline{X}_n - \theta)^2\right] = 1/n$. On obtient

$$\begin{aligned}\mathbf{R}_B(\Pi, T_3) &= \frac{n}{(n+1)^2} + \frac{1}{(n+1)^2}\int_\Theta \theta^2 \mathrm{d}\Pi(\theta)\\ &= \frac{n}{(n+1)^2} + \frac{1}{(n+1)^2} = \frac{1}{n+1}.\end{aligned}$$

On constate que pour tout $n \geq 2$, $\mathbf{R}_B(\Pi, T_3) < \mathbf{R}_B(\Pi, T_2) < \mathbf{R}_B(\Pi, T_1)$. Nous verrons par la suite que T_3 est en fait un estimateur de Bayes pour Π et la fonction de perte quadratique.

1.3. Risque maximal et estimateurs minimax.

Définition 3.5. Le risque maximal d'un estimateur T est

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

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.

Définition 3.6. Le risque minimax $\mathbf{R}_M$ est défini comme

$$\mathbf{R}_M = \inf_T \mathbf{R}_{\max}(T) = \inf_T \sup_{\theta\in\Theta} \mathbf{R}(\theta, T),$$

où l'infimum porte sur tous les estimateurs possibles T. Un estimateur $T^\star$ est minimax si

$$\mathbf{R}_{\max}(T^\star) = \mathbf{R}_M.$$

Puisque $\mathbf{R}_{\max}(T)$ 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 optimal dans le pire des cas alors qu'un estimateur de Bayes est optimal en moyenne. En ce sens, le critère minimax est plus pessimiste que le critère de Bayes, mais il a l'avantage d'être plus universel en ce qu'il ne dépend pas de la loi a priori Π.

Exemple 3.4. Reprenons l'exemple précédent du modèle gaussien avec les estimateurs T_1, T_2 et T_3 et calculons le risque maximal de chacun.

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

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 notions de risque de Bayes et de risque minimax peuvent être reliées entre elles sous certaines hypothèses, comme nous le verrons dans la suite.

2. Construction d'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 $\mathbf{R}_B(\Pi, T) = \int_\Theta \mathbf{R}(\theta, T)\mathrm{d}\Pi(\theta)$.

Définition 3.7. Soient ℓ une fonction de perte, Π une loi a priori et T un estimateur. Le risque a posteriori $\boldsymbol{\rho}(\Pi, T \mid \mathbf{X})$ est défini par

$$\boldsymbol{\rho}(\Pi, T \mid \mathbf{X}) = \mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X})) \mid \mathbf{X}\right] = \int_\Theta \ell(\theta, T(\mathbf{X}))\mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

Au lieu de prendre la moyenne de la fonction de perte par rapport à la loi de $(\boldsymbol{\theta}, \mathbf{X})$ comme pour le risque bayésien de la Définition 3.4, le risque a posteriori s'obtient conditionnellement à $\mathbf{X}$ en prenant la moyenne de la fonction de perte par rapport à la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$. Le risque a posteriori $\boldsymbol{\rho}(\Pi, T \mid \mathbf{X})$ est donc une variable aléatoire qui dépend de $\mathbf{X}$.

Exercice 3.1. Dans le modèle gaussien avec a priori $\Pi = \mathcal{N}(0, 1)$, calculer les risques a posteriori pour les estimateurs T_1, T_2, T_3 de l'exemple 3.3 et la perte quadratique.

Theorème 3.1. *Une fonction de perte ℓ et une loi a priori Π étant données, un élément*

$$T^\star(\mathbf{X}) \in \arg\min_T \boldsymbol{\rho}(\Pi, T \mid \mathbf{X}),$$

s'il existe, est un estimateur de Bayes pour Π.

On peut légitimement se demander en quoi le résultat du Théorème 3.1 est une simplification par rapport à la définition d'un estimateur de Bayes, qui introduit aussi un minimum. À supposer que l'on ait pu déterminer la loi a posteriori, le problème de minimisation du Théorème 3.1 est généralement plus simple à résoudre explicitement, en ce qu'il ne fait intervenir qu'une seule intégrale et non deux :

$$\boldsymbol{\rho}(\Pi, T \mid \mathbf{X}) = \int_\Theta \ell(\theta, T(\mathbf{X}))\mathrm{d}\Pi(\theta \mid \mathbf{X}),$$

alors que

$$\mathbf{R}_B(\Pi, T) = \int_\Theta \int_E \ell(\theta, T(x))\mathrm{d}P_\theta(x)\mathrm{d}\Pi(\theta).$$

Démonstration du Théorème 3.1. On peut supposer qu'il existe un estimateur T tel que $\mathbf{R}_B(\Pi, T)$ soit fini. Si ce n'est pas le cas, alors tout estimateur a un risque de Bayes pour Π infini, donc tout estimateur est de Bayes. Pour tout T tel que $\mathbf{R}_B(\Pi, T)$ est fini, on a

$$\begin{aligned}\mathbf{R}_B(\Pi, T) &= \mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X}))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\ell(\boldsymbol{\theta}, T(\mathbf{X})) \mid \mathbf{X}\right]\right] \\ &= \mathbb{E}\left[\boldsymbol{\rho}(\Pi, T \mid \mathbf{X})\right].\end{aligned}$$

Par définition, $\boldsymbol{\rho}(\Pi, T \mid \mathbf{X}) \geq \boldsymbol{\rho}(\Pi, T^\star \mid \mathbf{X})$. On en déduit que

$$\mathbf{R}_B(\Pi, T) \geq \mathbb{E}\left[\boldsymbol{\rho}(\Pi, T^\star \mid \mathbf{X})\right] = \mathbf{R}_B(\Pi, T^\star).$$

Ainsi, $T^\star$ est de Bayes, ce qu'il fallait démontrer. ■

Examinons maintenant les conséquences du Théorème 3.1 dans le cas de plusieurs fonctions de perte classiques.

2.1. 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 3.2. *Soit ℓ la perte quadratique et soit Π une loi a priori sur $\Theta \subset \mathbb{R}$. On suppose $\mathbb{E}[\boldsymbol{\theta}^2 \mid \mathbf{X}] < \infty$ p.s. Un estimateur de Bayes pour ℓ et la loi Π est donné par*

$$T^\star(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \int_\Theta \theta \mathrm{d}\Pi(\theta \mid \mathbf{X}),$$

la moyenne a posteriori pour la loi a priori Π.

Remarque 3.5. On suppose dans la proposition que $\mathbb{E}[\boldsymbol{\theta}^2 \mid \mathbf{X}] < \infty$ p.s. On peut montrer que si $\mathbb{E}[\boldsymbol{\theta}^2 \mid \mathbf{X}] = +\infty$ avec probabilité strictement positive, alors le risque a posteriori de tout estimateur $T(\mathbf{X})$ est infini avec probabilité strictement positive, et donc le risque de Bayes de tout estimateur est infini. Tout estimateur est donc de Bayes.

Démonstration de la Proposition 3.2. D'après le Théorème 3.1, il suffit de chercher un estimateur de Bayes sous la forme

$$T^\star(\mathbf{X}) = \arg\min_T \int_\Theta (T(\mathbf{X}) - \theta)^2 \mathrm{d}\Pi(\theta \mid \mathbf{X}).$$

Pour une variable aléatoire Z de carré intégrable, la fonction $\psi : a \mapsto \mathbb{E}[(Z-a)^2]$ est minimale pour $a = \mathbb{E}[Z]$ car

$$\psi(a) = \mathbb{E}[(Z - \mathbb{E}Z)^2] + (\mathbb{E}[Z] - a)^2 \geq \psi(\mathbb{E}[Z]).$$

Il suffit d'appliquer cette remarque à Z de loi $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{X})$ pour conclure, en notant que $\mathbb{E}[\boldsymbol{\theta}^2 \mid \mathbf{X}] < \infty$ par hypothèse, et que $\mathbb{E}[Z]$ est alors $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]$, la moyenne a posteriori. ■

Remarque 3.6. Pour calculer le risque de Bayes $\mathbf{R}_B(\Pi)$ (pour la perte quadratique), il y a deux manières de procéder. Généralement, le plus simple est de calculer la fonction de risque de l'estimateur de Bayes $T^\star$:

$$\theta \mapsto \mathbb{E}_\theta \left[(T^\star(\mathbf{X}) - \theta)^2 \right],$$

puis de l'intégrer contre Π. Mais dans certains cas, il peut être plus judicieux de remarquer que, pour la perte quadratique, le risque de Bayes est l'espérance de la variance a posteriori

$$\mathbf{R}_B(\Pi) = \mathbb{E}[v_{\mathbf{X}}],$$

où $v_{\mathbf{X}} = \mathbb{E}\left[\left(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]\right)^2 \mid \mathbf{X} \right]$. En effet,

$$\mathbf{R}_B(\Pi) = \mathbf{R}_B(\Pi, \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]) = \mathbb{E}\left[\left(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]\right)^2 \right] = \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]\right)^2 \mid \mathbf{X} \right] \right],$$

Calculer $\mathbb{E}[v_{\mathbf{X}}]$ est souvent difficile car il faut déterminer la loi marginale de $\mathbf{X}$. Mais dans certains cas, c'est très simple, notamment lorsque $v_{\mathbf{X}}$ ne dépend pas de $\mathbf{X}$, comme c'est le cas dans le modèle gaussien ci-dessous.

Exemple 3.7.

- Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \theta \in \mathbb{R}\}$ avec $\Pi = \mathcal{N}(0,1)$, nous avons vu au Chapitre 1 que

  $$\Pi[\cdot \mid \mathbf{X}] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1} \right).$$

  On en déduit avec la Proposition 3.2 qu'un estimateur de Bayes pour Π et la perte quadratique est donné par $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \frac{n\overline{X}_n}{n+1}$, comme annoncé plus haut. Et par la remarque ci-dessus, on a alors $\mathbf{R}_B(\Pi) = \mathbb{E}[v_{\mathbf{X}}] = \frac{1}{n+1}$.
- Dans le modèle de Bernoulli $\mathcal{P} = \{\mathcal{B}(\theta)^{\otimes n}, \theta \in [0,1]\}$ avec $\Pi = \mathrm{Unif}([0,1]) = \mathrm{Beta}(1,1)$, nous avons vu que

  $$\Pi[\cdot \mid \mathbf{X}] = \mathrm{Beta}(n\overline{X}_n + 1, n - n\overline{X}_n + 1).$$

  La moyenne a posteriori est

  $$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \frac{n\overline{X}_n + 1}{n\overline{X}_n + 1 + n - n\overline{X}_n + 1} = \frac{n\overline{X}_n + 1}{n+2}.$$

Par la Proposition 3.2, c'est un estimateur de Bayes pour Π et la perte quadratique. Calculons le risque quadratique de cet estimateur. Pour tout $\theta \in [0,1]$,

$$\begin{aligned}\mathbf{R}\left(\theta, \frac{n\overline{X}_n + 1}{n+2}\right) &= \mathbb{E}_\theta\left[\left(\frac{n\overline{X}_n + 1}{n+2} - \theta\right)^2\right] \\ &= \frac{1}{(n+2)^2}\mathbb{E}_\theta\left[\left(n(\overline{X}_n - \theta) + 1 - 2\theta\right)^2\right] \\ &= \frac{1}{(n+2)^2}\left(\mathrm{Var}(n\overline{X}_n) + (1-2\theta)^2\right) \\ &= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}.\end{aligned}$$

Le risque de Bayes pour $\Pi = \mathrm{Unif}[0,1]$ est donc

$$\mathbf{R}_B(\Pi) = \mathbf{R}_B\left(\Pi, \frac{n\overline{X}_n + 1}{n+2}\right) = \int_0^1 \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}\mathrm{d}\theta.$$

Après quelques calculs, on trouve

$$\mathbf{R}_B(\Pi) = \frac{1}{6(n+2)}.$$

2.2. 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 3.3. *Soit ℓ la perte en valeur absolue et soit Π une loi a priori sur $\Theta \subset \mathbb{R}$. On suppose $\mathbb{E}[|\boldsymbol{\theta}| \mid \mathbf{X}] < \infty$. Un estimateur de Bayes pour ℓ et la loi Π est donné par*

$$T^\star(\mathbf{X}) = F^{-1}_{\mathbf{X}}(1/2),$$

la médiane a posteriori pour la loi a priori Π.

Démonstration de la Proposition 3.3. D'après le Théorème 3.1, il suffit de chercher un minimiseur de la fonction

$$T(\mathbf{X}) \mapsto \mathbb{E}\left[|\boldsymbol{\theta} - T(\mathbf{X})| \mid \mathbf{X}\right].$$

Montrons que pour une variable aléatoire réelle Z intégrable de fonction de répartition F et de médiane $m = F^{-1}(1/2)$, on a

$$m \in \arg\min_{a \in \mathbb{R}} \mathbb{E}\left[|Z - a|\right]. \tag{3.1}$$

Pour tout $a \in \mathbb{R}$, on a

$$\begin{aligned}\mathbb{E}\left[|Z-a|\right] &= \int_0^{+\infty} \mathbb{P}(|Z-a| > t)\mathrm{d}t \\ &= \int_0^{+\infty} \mathbb{P}(Z > a+t)\mathrm{d}t + \int_0^{+\infty} \mathbb{P}(Z < a-t)\mathrm{d}t \\ &= \int_a^{+\infty} \mathbb{P}(Z > t)\mathrm{d}t + \int_{-\infty}^{a} \mathbb{P}(Z < t)\mathrm{d}t\,.\end{aligned}$$

Ainsi, si $a < m$, on a

$$\mathbb{E}\left[|Z-a|\right] - \mathbb{E}\left[|Z-m|\right] = \int_a^m \{\mathbb{P}(Z>t) - \mathbb{P}(Z<t)\}\mathrm{d}t\,.$$

Or pour $t < m$, on a $\mathbb{P}(Z > t) = 1 - \mathbb{P}(Z \leq t) > 1/2$, donc l'intégrale ci-dessus est positive. Et si $a > m$, on a

$$\mathbb{E}\left[|Z-a|\right] - \mathbb{E}\left[|Z-m|\right] = \int_m^a \{\mathbb{P}(Z<t) - \mathbb{P}(Z>t)\}\mathrm{d}t\,.$$

Or pour $t > m$, on a $\mathbb{P}(Z < t) \geq \mathbb{P}(Z \leq m) \geq 1/2$, donc dans ce cas aussi, l'intégrale est positive, ce qui établit (3.1). En appliquant ce résultat à la loi a posteriori, on obtient bien

$$F_{\mathbf{X}}^{-1}(1/2) \in \arg\min_T \mathbb{E}\left[|\boldsymbol{\theta} - T(\mathbf{X})| \;\middle|\; \mathbf{X}\right]\,.$$

■

3. Relation entre critères de décision

3.1. Une inégalité très simple et très utile.

Theorème 3.4. *Pour toute loi a priori* Π *sur* Θ *et toute fonction de perte, le risque bayésien minore toujours le risque minimax :*

$$\mathbf{R}_B(\Pi) \leq \mathbf{R}_M\,.$$

Démonstration. Par définition $\mathbf{R}_B(\Pi) = \inf_T \int \mathbf{R}(\theta, T)\mathrm{d}\Pi(\theta)$. Or comme $\Pi(\Theta) = 1$,

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

En prenant l'infimum en T de part et d'autre, il vient $\mathbf{R}_B(\Pi) \leq \mathbf{R}_M$. ■

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 », i.e. pour laquelle $\mathbf{R}_B(\Pi)$ est le plus grand possible. Nous verrons un exemple ci-dessous.

3.2. Minimaxité : conditions suffisantes.

Theorème 3.5. *Soit* T *un estimateur de Bayes pour une loi a priori* Π*. Si* T *est de risque constant, alors* T *est minimax.*

Démonstration. Soit T' un estimateur. Comme T est de Bayes pour Π, on a

$$\mathbf{R}_{\max}(T') \geq \mathbf{R}_B(\Pi, T') \geq \mathbf{R}_B(\Pi, T)\,.$$

Mais comme T est de risque constant, $\mathbf{R}_B(\Pi, T) = \mathbf{R}_{\max}(T)$. Ainsi pour tout estimateur T', on $\mathbf{R}_{\max}(T') \geq \mathbf{R}_{\max}(T)$, donc T est minimax. ■

Application. Dans le modèle binomial $P_\theta = \mathcal{B}(n, \theta)$, 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 $\mathrm{Beta}(a, b)$ (voir TDs) sur θ. Pour tous $a, b > 0$, on peut calculer explicitement la moyenne a posteriori $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]$

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.

Theorème 3.6. *Si un estimateur T est tel qu'on puisse trouver une suite $(\Pi_k)_{k\geq 1}$ de lois a priori avec*

$$\mathbf{R}_{\max}(T) \;=\; \overline{\lim_{k\to\infty}}\; \mathbf{R}_B(\Pi_k),$$

alors T est minimax.

Démonstration. Tout risque bayésien est inférieur ou égal au risque minimax $\mathbf{R}_M$, qui est lui-même inférieur ou égal à $\mathbf{R}_{\max}(T)$. Donc on a

$$\mathbf{R}_{\max}(T) = \overline{\lim_{k\to\infty}}\; \mathbf{R}_B(\Pi_k) \leq \mathbf{R}_M \leq \mathbf{R}_{\max}(T)$$

On en conclut $\mathbf{R}_{\max}(T) = \mathbf{R}_M$ donc T est minimax. ■

Application. Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta,1)^{\otimes n},\, \theta\in\mathbb{R}\}$, l'estimateur $\overline{X}_n$ est minimax pour la perte quadratique. Pour la loi a priori $\Pi_{\sigma^2} = \mathcal{N}(0,\sigma^2)$, le risque bayésien $\mathbf{R}_B(\Pi_{\sigma^2})$ 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 $\mathbf{R}_B(\Pi_{\sigma^2}) = \frac{1}{n+\sigma^{-2}}$. Or

$$\lim_{\sigma^2\to\infty} \frac{1}{n+\sigma^{-2}} = \frac{1}{n} = \mathbf{R}_{\max}(\overline{X}_n),$$

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

4. Minorations du risque minimax

Sauf dans certains cas particuliers, montrer qu'un estimateur est minimax s'avère être une tâche difficile. Généralement, on dispose d'un estimateur dont on sait calculer, ou au moins majorer, le risque. Mais rien ne nous dit que cet estimateur est minimax, c'est-à-dire que l'on ne pourrait pas trouver un autre estimateur qui « ferait mieux dans le pire des cas ». Il est donc important de savoir minorer le risque minimax, pour pouvoir affirmer que le risque maximal de tout estimateur doit nécessairement être plus grand qu'une certaine valeur.

Dans cette section, nous présentons la méthode de Le Cam, qui permet en fait d'obtenir une minoration non seulement du risque minimax, mais même d'une quantité plus précise que l'on appelera le risque minimax local.

Définition 3.8. Soit $\theta_0\Theta$. On appelle risque minimax local en θ_0 la quantité

$$\mathbf{R}_M^{\theta_0} = \sup_{\theta_1\in\Theta} \inf_T \max\left\{\mathbf{R}(\theta_0,T), \mathbf{R}(\theta_1,T)\right\}\,,$$

où l'infimum porte sur tous les estimateurs possibles.

Notons que si l'on sait minorer le risque minimax local pour un certain $\theta_0\in\Theta$ quelconques, alors on obtient immédiatement une borne inférieure sur le risque minimax puisque

$$\mathbf{R}_M \geq \sup_{\theta_0\in\Theta} \mathbf{R}_M^{\theta_0}\,.$$

La quantité $\mathbf{R}_M^{\theta_0}$ fournit néanmoins une information plus précise que $\mathbf{R}_M$: elle nous donne une idée du risque minimax localement autour de $\{\theta_0\}$.

4.1. Le Théorème de Le Cam. De nombreux résultats de minoration du risque minimax, comme la borne inférieure de Le Cam présentée dans cette section, reposent sur l'argument bayésien suivant. On remarque que pour ℓ une fonction de perte donnée, pour tous points θ_0, θ_1 de Θ, et pour tout estimateur T,

$$\max\{\mathbf{R}(\theta_0, T), \mathbf{R}(\theta_1, T)\} \geq \frac{1}{2}\left(\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)\right) .$$

Cela revient à utiliser l'inégalité $\mathbf{R}(\Pi, T) \leq \mathbf{R}_{\max}(T)$ pour $\Theta = \{\theta_0, \theta_1\}$ et l'a priori

$$\Pi = \frac{1}{2}\delta_{\theta_0} + \frac{1}{2}\delta_{\theta_1}.$$

En prenant l'infimum en T, on obtient

$$\mathbf{R}_M^{\theta_0,\theta_1} \geq \inf_T \frac{1}{2}\left(\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)\right) . \tag{3.2}$$

L'inégalité (3.2) constitue la première étape de minoration . Ensuite, on aimerait minorer $\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)$ par un terme qui ne dépende plus de T. Pour cela nous allons avoir besoin de l'hypothèse suivante (vérifiée pour toutes les fonctions de perte considérées dans ce cours).

Hypothèse. On suppose que la fonction de perte ℓ s'écrit

$$\ell(\cdot, \cdot) = d(\cdot, \cdot)^p, \tag{3.3}$$

où $d(\cdot, \cdot)$ est une distance et $p \geq 1$ un réel.

Lemme 3.7. *Si l'hypothèse (3.3) est vérifiée, alors, pour tous t, θ_0,$\theta_1 \in \Theta$,*

$$\frac{1}{2}\left(\ell(\theta_0, t) + \ell(\theta_1, t)\right) \geq \frac{1}{2^p}\ell(\theta_0, \theta_1).$$

Démonstration. Supposons que $\ell(\cdot, \cdot) = d(\cdot, \cdot)^p$, avec $p \geq 1$ et d une distance. Par l'inégalité triangulaire on a

$$\ell(\theta_0, \theta_1) \leq \left(d(\theta_0, t) + d(\theta_1, t)\right)^p .$$

Comme $p \geq 1$, la fonction $x \mapsto x^p$ est convexe sur $\mathbb{R}_+$, et, par l'inégalité de Jensen,

$$\left(d(\theta_0, t) + d(\theta_1, t)\right)^p = 2^p \left(\frac{d(\theta_0, t) + d(\theta_1, t)}{2}\right)^p \leq 2^{p-1}\left(d(\theta_0, t)^p + d(\theta_1, t)^p\right) .$$

Ainsi,

$$\ell(\theta_0, t) + \ell(\theta_1, t) \geq \frac{\ell(\theta_0, \theta_1)}{2^{p-1}}.$$

■

Le Lemme 3.7 ne permet pas immédiatement de minorer la quantité $\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)$ par une quantité ne dépendant plus de T. En effet, l'intégration de la fonction de perte dans $\mathbf{R}(\theta_0, T)$ et dans $\mathbf{R}(\theta_1, T)$ se fait contre des mesures différentes, P_{θ_0} et P_{θ_1}. La notion d'affinité présentée ci-dessous va nous permettre de résoudre ce problème.

Définition 3.9. Soient P, Q deux mesures de probabilité sur $(E, \mathcal{E})$ avec $\mathrm{d}P = p\mathrm{d}\mu$ et $\mathrm{d}Q = q\mathrm{d}\mu$. La distance en variation totale entre P et Q est définie par

$$d_{\mathrm{VT}}(P, Q) = \sup_{A \in \mathcal{E}} \{P(A) - Q(A)\} .$$

Notons $p \wedge q$ la fonction $x \mapsto (p \wedge q)(x) = \min\{p(x), q(x)\}$. On appelle affinité (en variation totale) entre P et Q la quantité

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

Proposition 3.8. *Soient P, Q deux mesures de probabilité avec $\mathrm{d}P = p\mathrm{d}\mu$ et $\mathrm{d}Q = q\mathrm{d}\mu$. On a*

$$d_{\mathrm{VT}}(P, Q) = \frac{1}{2} \int_E |p(x) - q(x)| \mathrm{d}\mu(x) = \int_E (p(x) - q(x))_+ \, \mathrm{d}\mu(x),$$

et

$$\mathcal{A}(P, Q) = 1 - d_{\mathrm{VT}}(P, Q).$$

Démonstration. Commençons par montrer que le supremum dans la définition de d_{VT} est atteint par

$$\Lambda = \{x \in E,\, p(x) > q(x)\}.$$

En effet, soit A un sous-ensemble quelconque de E. On a

$$P(A) - Q(A) \leq P(A \cap \Lambda) - Q(A \cap \Lambda) \leq P(\Lambda) - Q(\Lambda).$$

Ainsi

$$d_{\mathrm{VT}}(P, Q) = P(\Lambda) - Q(\Lambda) = \int_E (p(x) - q(x))_+ \, \mathrm{d}\mu(x),$$

et comme $P(\Lambda) - Q(\Lambda) = Q(\Lambda^c) - P(\Lambda^c)$, on a aussi

$$d_{\mathrm{VT}}(P, Q) = \frac{1}{2} \left(P(\Lambda) - Q(\Lambda) + Q(\Lambda^c) - P(\Lambda^c)\right) = \frac{1}{2} \int_E |p(x) - q(x)| \mathrm{d}\mu(x).$$

Pour la dernière égalité de l'énoncé, on note que pour tout $x \in E$,

$$(p \wedge q)(x) = \frac{1}{2} \left(p(x) + q(x) - |p(x) - q(x)|\right).$$

Il suffit ensuite d'intégrer par rapport à μ en utilisant que p et q sont des densités par rapport à μ. ∎

Theorème 3.9. *Si (3.3) est vérifiée avec $\ell(\cdot, \cdot) = d(\cdot, \cdot)^p$, alors, pour tous $\theta_0, \theta_1 \in \Theta$,*

$$\inf_T \frac{1}{2} \left(\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)\right) \geq \frac{d(\theta_0, \theta_1)^p}{2^p} \mathcal{A}(P_{\theta_0}, P_{\theta_1}).$$

Démonstration. Soit T un estimateur. En utilisant le Lemme 3.7 avec $t = T(x)$,

$$\begin{aligned}
\frac{1}{2} \left(\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)\right) &= \frac{1}{2} \left\{ \int_E \ell(\theta_0, T(x)) p_{\theta_0}(x) \mathrm{d}\mu(x) + \int_E \ell(\theta_1, T(x)) p_{\theta_1}(x) \mathrm{d}\mu(x) \right\} \\
&\geq \frac{1}{2} \int_E \Big\{ \ell(\theta_0, T(x)) + \ell(\theta_1, T(x)) \Big\} (p_{\theta_0} \wedge p_{\theta_1})(x) \mathrm{d}\mu(x) \\
&\geq \frac{\ell(\theta_0, \theta_1)}{2^p} \int_E (p_{\theta_0} \wedge p_{\theta_1})(x) \mathrm{d}\mu(x) = \frac{\ell(\theta_0, \theta_1)}{2^p} \mathcal{A}(P_{\theta_0}, P_{\theta_1}).
\end{aligned}$$

Cette inégalité étant valable pour tout estimateur T, on a

$$\inf_T \frac{1}{2} \left(\mathbf{R}(\theta_0, T) + \mathbf{R}(\theta_1, T)\right) \geq \frac{\ell(\theta_0, \theta_1)}{2^p} \mathcal{A}(P_{\theta_0}, P_{\theta_1})$$

■

Autrement dit, si (3.3) est vérifiée,

$$\inf_T \frac{1}{2}\left(\mathbf{R}(\theta_0,T)+\mathbf{R}(\theta_1,T)\right) \geq \frac{d(\theta_0,\theta_1)^p}{2^p}\left(1-d_{\mathrm{VT}}(P_{\theta_0},P_{\theta_1})\right). \tag{3.4}$$

L'objectif dans la suite de cette section va être, entre autres, de démontrer que la « meilleure vitesse possible » au sens du risque minimax local pour la perte quadratique dans les modèles paramétriques réguliers est de l'ordre de $1/n$. 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 gaussien avec n observations). Pour cela, d'après (3.4), il suffit de majorer $d_{\mathrm{VT}}(P_{\theta_0}^{\otimes n},P_{\theta_1}^{\otimes n})$, pour des points θ_0,θ_1 bien choisis dans Θ.

Définition 3.10. Soient P,Q deux mesures de probabilité sur $(E,\mathcal{E})$ avec $\mathrm{d}P=p\mathrm{d}\mu$ et $\mathrm{d}Q=q\mathrm{d}\mu$. La distance de Hellinger entre P et Q est définie par

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

On définit l'affinité de Hellinger entre P et Q par

$$\rho(P,Q)=\int_E\sqrt{p(x)q(x)}\mathrm{d}\mu(x).$$

On peut vérifier que les définitions ci-dessus sont indépendantes du choix de la mesure dominante μ.

Proposition 3.10. *Les quantités ρ et h ont les propriétés suivantes :*

(1) $h(P,Q)^2=2-2\rho(P,Q)$.

(2) $0\leq h(P,Q)\leq\sqrt{2}$.

(3) $d_{\mathrm{VT}}(P,Q)\leq h(P,Q)$.

(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 $\mathrm{d}P_i=p_i\mathrm{d}\mu$ et $\mathrm{d}Q_i=q_i\mathrm{d}\mu$, pour tout $i\in[\![1,n]\!]$. Alors

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

Démonstration. (1) Par définition, et comme p et q sont des densités,

$$h(P,Q)^2=\int_E(\sqrt{p}-\sqrt{q})^2\mathrm{d}\mu=\int_E\left(p(x)+q(x)-2\sqrt{p(x)q(x)}\right)\mathrm{d}\mu(x)=2-2\rho(P,Q).$$

(2) Clairement $h(P,Q)\geq 0$ et $h(P,Q)^2=2-2\rho(P,Q)\leq 2$.

(3) En utilisant l'inégalité de Cauchy-Schwarz,

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

En utilisant l'inégalité $(a+b)^2 \leq 2(a^2+b^2)$ et le fait que p et q sont des densités, on obtient bien l'inégalité voulue.

(4) En effet, par définition de P,

$$\mathrm{d}P(x_1,\dots,x_n) = p_1(x_1)\cdots p_n(x_n)\mathrm{d}\mu(x_1)\cdots\mathrm{d}\mu(x_n),$$

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

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

■

Proposition 3.11. *Soient P,Q deux mesures de probabilité sur $(E,\mathcal{E})$ avec $\mathrm{d}P = p\mathrm{d}\mu$ et $\mathrm{d}Q = q\mathrm{d}\mu$. On a*

$$d_{\mathrm{VT}}(P^{\otimes n},Q^{\otimes n}) \leq \sqrt{n}h(P,Q)\,.$$

Démonstration. Par la propriété (3) de la Proposition 3.10, on a

$$d_{\mathrm{VT}}(P^{\otimes n},Q^{\otimes n}) \leq h(P^{\otimes n},Q^{\otimes n}) = \sqrt{2}\sqrt{1-\rho(P^{\otimes n},Q^{\otimes n})}\,.$$

Or, par la propriété (4),

$$\rho(P^{\otimes n},Q^{\otimes n}) = \rho(P,Q)^n = \left(1-\frac{h^2(P,Q)}{2}\right)^n.$$

En utilisant l'inégalité $(1-x)^n \geq 1-nx$ pour $x\in[0,1]$, on en déduit

$$\rho(P^{\otimes n},Q^{\otimes n}) \geq 1-\frac{nh^2(P,Q)}{2}\,,$$

puis que

$$d_{\mathrm{VT}}(P^{\otimes n},Q^{\otimes n}) \leq \sqrt{2}\sqrt{\frac{nh^2(P,Q)}{2}} = \sqrt{n}h(P,Q)\,.$$

■

En combinant l'inégalité 3.4 et la Proposition 3.11, on en déduit le résultat suivant, appélé Théorème de Le Cam.

Theorème 3.12 (Le Cam). *Soit $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$ un modèle quelconque et ℓ une fonction de perte donnée par*

$$\ell(\cdot,\cdot) = d(\cdot,\cdot)^p$$

pour $p \geq 1$ et d une distance sur Θ. Alors pour tous $\theta_0, \theta_1 \in \Theta$,

$$\mathbf{R}_M \geq \mathbf{R}_M^{\theta_0} \geq \frac{d(\theta_0,\theta_1)^p}{2^p}\left(1 - \sqrt{n}h(\theta_0,\theta_1)\right).$$

4.2. Applications. Si l'on s'intéresse au risque quadratique $\ell(\theta,\theta') = \|\theta - \theta'\|^2$ (i.e. d est la distance euclidienne et $p = 2$), le théorème de Le Cam donne, pour tous $\theta_0, \theta_1 \in \Theta$,

$$\mathbf{R}_M^{\theta_0} \geq \sup_{\theta_1 \in \Theta} \frac{\|\theta_0 - \theta_1\|^2}{4}\left(1 - \sqrt{n}h(\theta_0,\theta_1)\right).$$

On voit que si l'on peut trouver θ_0, θ_1 à distance de l'ordre de $1/\sqrt{n}$ à la fois pour la distance euclidienne et pour la distance de Hellinger, alors on obtient une borne inférieure en $1/n$ pour le risque minimax.

Exemple 3.8. Dans le modèle gaussien $P_\theta = \mathcal{N}(\theta, 1)$, on peut vérifier par le calcul (voir TD) que

$$h^2(\theta,\theta') = 2\left(1 - e^{-\frac{(\theta-\theta')^2}{8}}\right).$$

Pour tout $\theta_0 \in \mathbb{R}$, si l'on prend $\theta_1 = \theta_0 + 1/\sqrt{n}$, en utilisant $1 - e^{-x} \leq x$, on a

$$h^2(\theta_0,\theta_1) = 2\left(1 - e^{-\frac{1}{8n}}\right) \leq \frac{1}{4n},$$

et ainsi, par le Théorème 3.12, le risque minimax local peut être minoré par

$$\mathbf{R}_M^{\theta_0} \geq \frac{1}{4n}\left(1 - \sqrt{n}\cdot\frac{1}{2\sqrt{n}}\right) = \frac{1}{8n}.$$

En particulier, $\mathbf{R}_M \geq \frac{1}{8n}$.

Modèles paramétriques réguliers. L'exemple précédent dans le modèle gaussien est une manifestation d'un phénomène plus général dans les modèles paramétriques réguliers. Dans un modèle régulier, si $\Theta \subset \mathbb{R}$, si $\theta_0 \in \Theta$ est un point dans l'intérieur de Θ et si l'on suppose que l'information de Fisher en θ_0, notée $\mathbf{I}(\theta_0)$, est strictement positive, on peut montrer (admis) que

$$\lim_{\varepsilon \to 0} \frac{h(\theta_0 + \varepsilon, \theta_0)}{\varepsilon} = \frac{\sqrt{\mathbf{I}(\theta_0)}}{2}.$$

En prenant $\varepsilon = \frac{c}{\sqrt{\mathbf{I}(\theta_0)n}}$ pour $c > 0$, on a donc

$$h(\theta_0, \theta_0 + \varepsilon) \underset{n\to+\infty}{\sim} \frac{c}{2\sqrt{n}}.$$

D'après le Théorème 3.12, pour $\theta_1 = \theta_0 + \varepsilon$, on a

$$\mathbf{R}_M^{\theta_0} \geq \frac{c^2}{4\mathbf{I}(\theta_0)n}\left(1 - \frac{c}{2} + o(1)\right).$$

En optimisant sur $c > 0$, on obtient, pour $c = \frac{4}{3}$,

$$\mathbf{R}_M^{\theta_0} \geq \frac{4}{27\mathbf{I}(\theta_0)n}(1 + o(1)).$$

Dans les modèles réguliers, pour tout $\theta_0 \in \Theta$, lorsque l'on choisit θ_1 à distance de l'ordre de $\frac{1}{\sqrt{\mathbf{I}(\theta_0)n}}$ de θ_0, le théorème de Le Cam donne une borne inférieure sur le risque minimax local (pour la perte quadratique) de l'ordre de $\frac{1}{n}$, avec une constante donnée par l'inverse de l'information de Fisher en θ_0.

Chapitre 4

Les tests bayésiens

Dans le cadre d'un modèle $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, avec $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$ 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 région $\Theta_0 \subset \Theta$ ou à une autre région $\Theta_1 \subset \Theta$, avec $\Theta_0 \cap \Theta_1 = \emptyset$. Contrairement à l'approche fréquentiste, on ne supposera pas toujours que $\Theta_0 \cup \Theta_1 = \Theta$. En revanche, $\Theta_0 \cup \Theta_1$ correspondra toujours au support de la loi a priori Π.

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 $\Theta_0 = \{\theta_0\}$, est dite hypothèse simple. Sinon, on parle d'hypothèse composite.

Définition 4.1. Un test est une fonction mesurable $\varphi(X_1, \ldots, X_n)$ des observations, à valeurs dans $\{0, 1\}$.

Commençons par quelques rappels sur l'approche fréquentiste des tests.

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 $\varphi(\mathbf{X}) = 1$ alors que les données $\mathbf{X} = (X_1, \ldots, X_n)$ ont été générées de façon i.i.d. selon une loi P_θ avec $\theta \in \Theta_0$. On appelle erreur de première espèce la fonction

$$\begin{aligned} \Theta_0 &\to [0, 1] \\ \theta &\mapsto \mathbb{P}_\theta(\varphi(\mathbf{X}) = 1). \end{aligned}$$

(2) Accepter H_0 alors que $\theta \in \Theta_1$: dans ce cas $\varphi(\mathbf{X}) = 0$ alors que les données $\mathbf{X} = (X_1, \ldots, X_n)$ ont été générées de façon i.i.d. selon une loi P_θ avec $\theta \in \Theta_1$. On appelle erreur de deuxième espèce la fonction

$$\begin{aligned} \Theta_1 &\to [0, 1] \\ \theta &\mapsto \mathbb{P}_\theta(\varphi(\mathbf{X}) = 0). \end{aligned}$$

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, celle que l'on maintient à moins d'avoir assez d'éléments pour la rejeter (penser à la présomption d'innocence dans un procès), tandis que H_1 est celle qui ne sera retenue que si les données fournissent assez d'éléments dans son sens (dans l'analogie juridique, la culpabilité).

Définition 4.2. On appelle taille d'un test φ la quantité

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \varphi(\mathbf{X}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\varphi(\mathbf{X}) = 1).$$

On dit qu'un test φ est de niveau α si sa taille est majorée par α. La fonction de $\Theta \to [0,1]$ définie par

$$\pi : \theta \mapsto \mathbb{E}_\theta[\varphi(\mathbf{X})]$$

s'appelle fonction puissance.

L'approche fréquentiste des tests consiste, pour un α donné, à chercher 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 proche de 1 sur Θ_1.

Exemple 4.1. Soit $\mathcal{P} = \{\mathcal{N}(\theta,1)^{\otimes n},\ \theta \in \mathbb{R}\}$, et posons $\Theta_0 = \mathbb{R}_-$ et $\Theta_1 = \mathbb{R}_+^*$. Le test

$$\varphi(\mathbf{X}) = \mathbb{1}_{\{\sqrt{n}\overline{X}_n > q_{1-\alpha}\}},$$

avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$, est un test de niveau α. On peut montrer que ce test est *uniformément plus puissant* parmi les tests de niveau α, c'est-à-dire que pour tout test $\widetilde{\varphi}$ de niveau α, on a

$$\forall \theta \in \Theta_1,\ \pi(\theta) \geq \widetilde{\pi}(\theta),$$

où π est la fonction puissance de φ et $\widetilde{\pi}$ celle de $\widetilde{\varphi}$.

L'approche bayésienne des tests consiste à choisir une loi a priori Π à support $\Theta_0 \cup \Theta_1$, et donc en particulier $\Pi(\Theta_0 \cup \Theta_1) = 1$. Notons qu'avec cette supposition, Π n'est pas forcément défini sur Θ tout entier.

1. Tests de Bayes

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

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

Définition 4.3. La fonction de perte du 0-1 est définie par

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

On parle aussi de fonction de perte équilibrée.

Choisir une telle fonction de perte revient à ne ne privilégier aucune hypothèse par rapport à l'autre : on pénalise par 1 le fait de se tromper (dans un sens ou dans l'autre).

Proposition 4.1. *L'estimateur de Bayes pour la fonction de perte du* 0-1 *est*

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

Il s'agit du test de Bayes pour la fonction de perte du 0–1.

DÉMONSTRATION. Par le Théorème 3.1, un estimateur de Bayes minimise le risque a posteriori $\int L(\theta,\varphi)\mathrm{d}\Pi(\theta \mid \mathbf{X})$. Ici on a

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

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

Cas de deux hypothèses simples de même poids. Supposons que $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, et notons $P_0 = P_{\theta_0}$ et $P_1 = P_{\theta_1}$, de densités respectives p_0 et p_1. Prenons comme loi a priori $\Pi = \frac{1}{2}\delta_{\theta_0} + \frac{1}{2}\delta_{\theta_1}$. Le test de Bayes s'écrit alors

$$\varphi^\star(\mathbf{X}) = \mathbb{1}_{p_0(\mathbf{X})\leq p_1(\mathbf{X})}\,.$$

De plus le risque de Bayes, c'est-à-dire le risque de Bayes de φ est égal à

$$\mathbf{R}_B(\Pi) = \frac{1}{2}\big(1 - d_{\mathrm{VT}}(P_0,P_1)\big)\,.$$

En effet,

$$\begin{aligned}\mathbf{R}_B(\Pi) &= \mathbb{E}\left[L(\boldsymbol{\theta},\varphi^\star(\mathbf{X}))\right] \\ &= \frac{1}{2}\mathbb{P}_0\left(p_0(\mathbf{X}) \leq p_1(\mathbf{X})\right) + \frac{1}{2}\mathbb{P}_1\left(p_0(\mathbf{X}) > p_1(\mathbf{X})\right) \\ &= \frac{1}{2}\Big(1 - \big(\mathbb{P}_0\left(p_0(\mathbf{X}) > p_1(\mathbf{X})\right) - \mathbb{P}_1\left(p_0(\mathbf{X}) > p_1(\mathbf{X})\right)\big)\Big) \\ &= \frac{1}{2}\Big(1 - \int_E \mathbb{1}_{\{p_0(x)>p_1(x)\}}\left(p_0(x) - p_1(x)\right)\mathrm{d}\mu(x)\Big) \\ &= \frac{1}{2}\big(1 - d_{\mathrm{VT}}(P_0,P_1)\big)\,.\end{aligned}$$

Si l'on veut pénaliser différemment les deux manières de se tromper, on peut plutôt choisir une fonction de perte pondérée.

Définition 4.4. Une fonction de perte est dite pondérée si elle s'écrit

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

avec $a_0, a_1 \in \mathbb{R}_+$.

Remarque 4.2. La fonction de risque pour la perte pondérée associée à un test φ est

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

et le risque bayésien de φ est

$$\begin{aligned}\mathbf{R}_B(\Pi,\varphi) &= \int_\Theta \mathbb{E}_\theta[L_{a_0,a_1}(\theta,\varphi(\mathbf{X}))]\mathrm{d}\Pi(\theta)\\ &= a_0\int_{\Theta_0}\mathbb{P}_\theta(\varphi(\mathbf{X})=1)\mathrm{d}\Pi(\theta)+a_1\int_{\Theta_1}\mathbb{P}_\theta(\varphi(\mathbf{X})=0)\mathrm{d}\Pi(\theta)\,.\end{aligned}$$

Par définition, le test de Bayes minimise ce risque. Les erreurs de première et de deuxième espèces sont ainsi moyennées par rapport à la loi a priori, et les constantes a_0, a_1 introduisent une pondération éventuelle supplémentaire.

Proposition 4.2. *Le test de Bayes pour la fonction de perte pondérée est*

$$\varphi^\star(\mathbf{X}) = \mathbb{1}_{a_0\Pi(\Theta_0|\mathbf{X})\le a_1\Pi(\Theta_1|\mathbf{X})} = \mathbb{1}_{\left\{\Pi(\Theta_0|\mathbf{X})\le\frac{a_1}{a_0+a_1}\right\}}.$$

Démonstration. Laissée en exercice. ■

Exemple 4.3. Considérons le modèle gaussien $\mathcal{P}=\{\mathcal{N}(\theta,1)^{\otimes n},\ \theta\in\mathbb{R}\}$.

(1) **Cas de deux hypothèses simples :** on veut tester

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

Tout d'abord, il s'agit de se donner 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

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

avec $\pi_0+\pi_1=1$. On calcule alors

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

et $\Pi(\{1\}\,\big|\,\mathbf{X})=1-\Pi(\{0\}\,\big|\,\mathbf{X})$. On en déduit que le test bayésien pour la fonction de perte pondérée s'écrit

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

On remarque que le test se met sous la forme $\{\overline{X}_n\ge t_n\}$ 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 $\varphi^\star(\mathbf{X})=\mathbb{1}_{\overline{X}_n\ge1/2}$. Dans ce dernier cas, les hypothèses H_0 et H_1 jouent des rôles symétriques.

(2) **Cas de deux hypothèses composites :** on veut tester

$$H_0\ :\ \{\theta\le0\}\qquad\text{contre}\qquad H_1\ :\ \{\theta>0\}$$

On doit d'abord choisir une loi a priori sur $\mathbb{R}$. Choisissons par exemple $\Pi=\mathcal{N}(\mu,1)$. La loi a posteriori est dans ce cas

$$\Pi[\cdot\,\big|\,\mathbf{X}] = \mathcal{N}\left(m_{\mathbf{X}},\frac{1}{n+1}\right),\quad\text{avec}\quad m_{\mathbf{X}}=\frac{\mu+n\overline{X}_n}{n+1}.$$

Le test

$$\varphi^\star(\mathbf{X}) = \mathbb{1}_{\{a_0\Pi(\Theta_0|\mathbf{X})\le a_1\Pi(\Theta_1|\mathbf{X})\}}$$

est un test de Bayes pour la fonction de perte pondérée, et on a

$$\Pi(\Theta_0 \mid \mathbf{X}) = \mathbb{P}\left(m_{\mathbf{X}} + \frac{1}{\sqrt{n+1}}\mathcal{N}(0,1) \leq 0 \mid \mathbf{X}\right) = \Phi\left(-\sqrt{n+1}\, m_{\mathbf{X}}\right),$$

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

(3) **Cas d'une hypothèse simple et une hypothèse composite :** on veut tester

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

Comme $\{0\} \cup \mathbb{R}^* = \mathbb{R}$, un choix qui pourrait sembler à première vue naturel serait celui d'une loi Π à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Cependant, dans ce cas on aurait $\Pi(\{0\}) = 0$ et donc on rejetterait toujours H_0. D'un point de vue bayésien, si l'hypothèse nulle correspond à un singleton $\{\theta_0\}$, c'est que l'on suppose que $\boldsymbol{\theta}$ peut valoir exactement θ_0, donc il est naturel d'intégrer cette information à la loi a priori. Par exemple, une loi a priori raisonnable est

$$\Pi = \pi_0\delta_0 + \pi_1\mathcal{N}(0,1),$$

avec $\pi_0 + \pi_1 = 1$. La formule de Bayes donne, pour q la densité d'une $\mathcal{N}(0,1)$,

$$\Pi(\{0\} \mid \mathbf{X}) = \frac{\pi_0 p_0(\mathbf{X})}{\pi_0 p_0(\mathbf{X}) + \pi_1 \int p_\theta(\mathbf{X}) q(\theta)\mathrm{d}\theta}.$$

On a

$$p_0(\mathbf{X}) = \frac{1}{(\sqrt{2\pi})^n}\, \mathrm{e}^{-\sum_{i=1}^n \frac{X_i^2}{2}},$$

et

$$\begin{aligned}
\int p_\theta(\mathbf{X}) q(\theta)\mathrm{d}\theta &= \frac{1}{(\sqrt{2\pi})^{n+1}} \int \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2 - \frac{\theta^2}{2}\right)\mathrm{d}\theta \\
&= \frac{\exp\left(-\frac{1}{2}\sum X_i^2\right)}{(\sqrt{2\pi})^{n+1}} \int \exp\left(-\frac{n+1}{2}\theta^2 + n\overline{X}_n\theta\right)\mathrm{d}\theta \\
&= \frac{p_0(\mathbf{X}) \exp\left(\frac{(n\overline{X}_n)^2}{2(n+1)}\right)}{\sqrt{2\pi}} \int \exp\left(-\frac{n+1}{2}\left(\theta - \frac{n\overline{X}_n}{n+1}\right)^2\right)\mathrm{d}\theta \\
&= \frac{p_0(\mathbf{X}) \exp\left(\frac{(n\overline{X}_n)^2}{2(n+1)}\right)}{\sqrt{n+1}}.
\end{aligned}$$

Ainsi le test de Bayes pour la fonction de perte pondérée consiste à rejeter H_0 si

$$a_0\pi_0 \leq a_1\pi_1 \frac{\exp\left(\frac{(n\overline{X}_n)^2}{2(n+1)}\right)}{\sqrt{n+1}},$$

c'est-à-dire si

$$\frac{|n\overline{X}_n|}{\sqrt{n+1}} \geq \sqrt{\ln(n+1) + 2\ln\left(\frac{a_0\pi_0}{a_1\pi_1}\right)}.$$

2. Tests bayésiens et apprentissage statistique (∗)

Sortons quelques instants du cadre bayésien et considérons le problème de classification suivant : soit $(\mathbf{X}, Y)$ une variable aléatoire à valeurs dans $E \times \{0,1\}$. La variable $\mathbf{X}$ est souvent appelée variable explicative, et Y le label. Le problème de classification consiste à prédire le label Y à partir de $\mathbf{X}$. Par exemple, si $\mathbf{X}$ est un vecteur contenant le nombre de fois où apparaissent certains mots dans un mail, on peut vouloir chercher à prédire si ce mail est un spam ou non.

Un classifieur est une fonction mesurable $f : E \to \{0,1\}$. On espère que la prédiction $f(\mathbf{X})$ sera proche du label Y. On définit le risque de classification d'un classifieur f par

$$\mathbf{R}(f) = \mathbb{P}(Y \neq f(\mathbf{X}))\,.$$

Ce risque n'est en fait rien d'autre que le risque bayésien de f dans le problème de test bayésien suivant : la loi du couple $(\mathbf{X}, Y)$ peut être décrite par

$$\begin{aligned} Y &\sim \Pi \\ \mathbf{X} \;\big|\; Y &\sim P_Y\,. \end{aligned}$$

où Π est une loi sur $\{0,1\}$ (en termes bayésiens, on interprète la loi marginale de Y comme la loi a priori). En considérant la fonction de perte du $0-1$, donnée par $L(y, f(x)) = \mathbb{1}_{y \neq f(x)}$, le risque bayésien d'un test f (pour la loi a priori Π) s'écrit

$$\mathbb{E}\left[L(Y, f(\mathbf{X}))\right] = \mathbb{P}(Y \neq f(\mathbf{X})) = \mathbf{R}(f)\,.$$

On appelle alors classifieur de Bayes le test de Bayes pour ce problème de test, i.e. le classifieur $f^\star$ qui minimise le risque a posteriori $\mathbb{P}(Y \neq f(\mathbf{X}) \;|\; \mathbf{X})$. Si l'on pose

$$\eta(\mathbf{X}) = \mathbb{P}(Y = 1 \;|\; \mathbf{X})\,,$$

alors $\mathbb{P}(Y \neq f(\mathbf{X}) \;|\; \mathbf{X}) = \mathbb{1}_{f(\mathbf{X})=0}\eta(\mathbf{X}) + \mathbb{1}_{f(\mathbf{X})=1}(1-\eta(\mathbf{X}))$ et $f^\star$ est donné par

$$f^\star(\mathbf{X}) = \mathbb{1}_{\{\eta(\mathbf{X}) \geq 1-\eta(\mathbf{X})\}} = \mathbb{1}_{\left\{\eta(\mathbf{X}) \geq \frac{1}{2}\right\}}\,.$$

Proposition 4.3. *Soit $f^\star$ le classifieur de Bayes donné par $f^\star(\mathbf{X}) = \mathbb{1}_{\left\{\eta(\mathbf{X}) \geq \frac{1}{2}\right\}}$. Alors*

$$\mathbf{R}(f^\star) = \mathbb{E}\left[\min\left\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\right\}\right] \leq \frac{1}{2}\,.$$

De plus, pour tout classifieur f, on a

$$\mathbf{R}(f) - \mathbf{R}(f^\star) = \mathbb{E}\left[|2\eta(\mathbf{X}) - 1|\, \mathbb{1}_{\{f(\mathbf{X}) \neq f^\star(\mathbf{X})\}}\right]\,.$$

Démonstration. Pour tout classifieur f, on a

$$\begin{aligned} \mathbf{R}(f) &= \mathbb{E}\left[\mathbb{P}(Y \neq f(\mathbf{X}) \;|\; \mathbf{X})\right] \\ &= \mathbb{E}\left[\mathbb{1}_{f(\mathbf{X})=0}\eta(\mathbf{X}) + \mathbb{1}_{f(\mathbf{X})=1}(1-\eta(\mathbf{X}))\right] \\ &\geq \mathbb{E}\left[\min\left\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\right\}\right]\,, \end{aligned}$$

avec égalité pour $f = f^\star$. De plus, en remarquant que

$$\mathbf{R}(f) = \mathbb{E}[\eta(\mathbf{X})] + \mathbb{E}\left[(1-2\eta(\mathbf{X}))f(\mathbf{X})\right]\,,$$

on a

$$\begin{aligned}\mathbf{R}(f) - \mathbf{R}(f^\star) &= \mathbb{E}\left[(1-2\eta(\mathbf{X}))(f(\mathbf{X}) - f^\star(\mathbf{X}))\right] \\ &= \mathbb{E}\left[|2\eta(\mathbf{X}) - 1|\, \mathbb{1}_{\{f(\mathbf{X}) \neq f^\star(\mathbf{X})\}}\right] .\end{aligned}$$

■

On voit donc que si la loi du couple $(\mathbf{X}, Y)$ est connue, le problème de classification revient à un simple problème de test bayésien, pour lequel un test optimal (du point de vue du risque de classification) est donné par le test de Bayes. En pratique cependant, la loi du couple est inconnue et il faut *apprendre* à classifier à partir d'observations. On dispose d'un échantillon

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\} ,$$

où les $(\mathbf{X}_i, Y_i)$ sont i.i.d. selon la loi (inconnue) de $(\mathbf{X}, Y)$. Le but est alors de construire, à partir de $\mathcal{D}_n$, un classifieur $\widehat{f}_n$ dont le risque $\mathbf{R}(\widehat{f}_n)$ soit aussi proche que possible du risque de Bayes $\mathbf{R}^\star = \mathbf{R}(f^\star)$. Plus précisément, on souhaite construire à l'aide de $\mathcal{D}_n$ une fonction $\widehat{f}_n$ qui soit telle que, si l'on observe une nouvelle variable explicative distribuée selon $\mathbf{X}$, la probabilité que $\widehat{f}_n(\mathbf{X})$ prédise mal Y, conditionnellement à $\mathcal{D}_n$, soit la plus petite possible. Le risque $\mathbf{R}(\widehat{f}_n)$ est donc en fait une quantité aléatoire puisque la fonction $\widehat{f}_n$ elle-même est aléatoire (elle dépend de $\mathcal{D}_n$). On a

$$\mathbf{R}(\widehat{f}_n) = \mathbb{P}\left(Y \neq \widehat{f}_n(\mathbf{X}) \mid \mathcal{D}_n\right) .$$

Exemple 4.4. Dans le cas où l'ensemble E est un ensemble discret, un classifieur naturel, appelé classifieur par majorité, est construit de la façon suivante : pour tout $x \in E$, on calcule

$$N_0(x) = |\{i \in [\![1, n]\!],\, \mathbf{X}_i = x,\, Y_i = 0\}| ,$$

et

$$N_1(x) = |\{i \in [\![1, n]\!],\, \mathbf{X}_i = x,\, Y_i = 1\}| ,$$

et on pose

$$\widehat{f}_n^{\mathsf{maj}}(x) = \begin{cases} 1 & \text{si } N_1(x) \geq N_0(x), \\ 0 & \text{si } N_0(x) > N_1(x). \end{cases}$$

Autrement dit, on attribue à x le label majoritaire parmi les observations de $\mathcal{D}_n$ pour lesquelles $\mathbf{X}_i = x$.

Définition 4.5. La suite de classifieurs $(\widehat{f}_n)_{n\geq 1}$ est dite consistante si, quelle que soit la loi du couple $(\mathbf{X}, Y)$, on a

$$\mathbf{R}(\widehat{f}_n) \xrightarrow{\mathbb{P}} \mathbf{R}^\star .$$

Cette notion de consistance peut être vue comme une convergence ponctuelle sur l'ensemble des lois de probabilité sur $E \times \{0, 1\}$. On peut vouloir être plus exigeant et demander une convergence uniforme sur l'ensemble de ces lois. Dans la définition ci-dessous, on note $\mathbf{R}_P$ pour souligner qu'il s'agit du risque de classification lorsque la loi de $(\mathbf{X}, Y)$ est P.

Définition 4.6. La suite de classifieurs $(\widehat{f}_n)_{n\geq 1}$ est dite uniformément consistante si

$$\sup_P \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}}\left[\mathbf{R}_P(\widehat{f}_n) - \mathbf{R}_P^\star\right] \underset{n\to\infty}{\longrightarrow} 0 ,$$

où le supremum est pris sur toutes les lois de probabilités sur $E \times \{0,1\}$.

En fait, dans la plupart des cas (plus précisément dès que E est un ensemble infini), la consistance uniforme est impossible à obtenir. Nous allons cependant voir que si E est fini, alors on peut construire un classifieur uniformément consistant.

Une méthode souvent utilisée pour construire un classifieur $\widehat{f}_n$ est la méthode de minimisation du risque empirique. L'idée est d'approcher le risque $\mathbf{R}(f)$ d'un classifieur f par son équivalent empirique

$$\mathbf{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq f(\mathbf{X}_i)\}}\,.$$

Par la loi des grands nombres, $\mathbf{R}_n(f) \xrightarrow{\mathbb{P}} \mathbf{R}(f)$. Étant donné un ensemble $\mathcal{F}$ de classifieurs, souvent appelé dictionnaire, la méthode de minimisation du risque empirique consiste à choisir

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \mathbf{R}_n(f)\,.$$

Remarque 4.5. Le choix de $\mathcal{F}$ est crucial. Prendre $\mathcal{F}$ égal à l'ensemble de tous les classifieurs est souvent un très mauvais choix et conduit au sur-apprentissage. En effet, si E est assez grand pour que, presque sûrement, toutes les observations $\mathbf{X}_i$ soient distinctes, alors le risque empirique est minimisé par le classifieur qui s'ajuste parfaitement aux données, i.e.

$$\widehat{f}_n(x) = \sum_{i=1}^{n} \mathbb{1}_{x=\mathbf{X}_i} Y_i\,.$$

Autrement dit, si $x = \mathbf{X}_i$, le classifieur répond Y_i et si $x \notin \{\mathbf{X}_1,\dots,\mathbf{X}_n\}$, il répond, de façon arbitraire, 0. On a alors $\mathbf{R}_n(\widehat{f}_n) = 0$ mais $\mathbf{R}(\widehat{f}_n)$ peut être grand (plus $\mathcal{F}$ est grand, plus $\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|$ est grand). En fait, il faut choisir $\mathcal{F}$ assez grand pour pouvoir approcher le classifieur de Bayes par des éléments de $\mathcal{F}$ mais assez petit pour que $\mathbf{R}_n(f)$ reste une bonne approximation de $\mathbf{R}(f)$, uniformément sur $\mathcal{F}$. Ce compromis se lit bien sur la décomposition de l'excès de risque :

$$\mathbf{R}(\widehat{f}_n) - \mathbf{R}^\star = \mathbf{R}(\widehat{f}_n) - \inf_{f\in\mathcal{F}} \mathbf{R}(f) + \inf_{f\in\mathcal{F}} \mathbf{R}(f) - \mathbf{R}^\star\,.$$

Le premier terme $\mathbf{R}(\widehat{f}_n) - \inf_{f\in\mathcal{F}} \mathbf{R}(f)$ s'appelle l'erreur stochastique. Le second $\inf_{f\in\mathcal{F}} \mathbf{R}(f) - \mathbf{R}^\star$ l'erreur d'approximation.

Pour $\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \mathbf{R}_n(f)$ un minimiseur sur $\mathcal{F}$ du risque empirique, et pour $f^\diamond \in \arg\min_{f\in\mathcal{F}} \mathbf{R}(f)$ (on suppose pour simplifier que l'infimum $\inf_{f\in\mathcal{F}} \mathbf{R}(f)$ est atteint sur $\mathcal{F}$), on a

$$\mathbf{R}(\widehat{f}_n) - \mathbf{R}(f^\diamond) \leq 2\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|\,. \tag{4.1}$$

En effet,

$$\begin{aligned}
\mathbf{R}(\widehat{f}_n) &\leq \mathbf{R}_n(\widehat{f}_n) + \sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } \widehat{f}_n \in \mathcal{F} \text{ par construction,}\\
&\leq \mathbf{R}_n(f^\diamond) + \sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } \widehat{f}_n \text{ minimise } \mathbf{R}_n \text{ sur } \mathcal{F},\\
&\leq \mathbf{R}(f^\diamond) + 2\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } f^\diamond \in \mathcal{F} \text{ par construction.}
\end{aligned}$$

La quantité $\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|$ est en général difficile à contrôler. Mais si l'on se restreint à des dictionnaires $\mathcal{F}$ finis, alors on peut facilement obtenir des bornes.

Proposition 4.4. *Soit $\mathcal{F} = \{f_1, \dots, f_p\}$ un dictionnaire fini et soit $\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \mathbf{R}_n(f)$ un minimiseur sur $\mathcal{F}$ du risque empirique. Alors, pour tout $\delta \in]0,1[$, avec probabilité au moins $1-\delta$,*

$$\mathbf{R}(\widehat{f}_n) \leq \min_{f\in\mathcal{F}} \mathbf{R}(f) + \sqrt{\frac{2\log\left(\frac{2p}{\delta}\right)}{n}}\,.$$

En particulier,

$$\mathbb{E}\left[\mathbf{R}(\widehat{f}_n)\right] - \min_{f\in\mathcal{F}} \mathbf{R}(f) \leq 2\sqrt{\frac{2\log(2p)}{n}}\,.$$

Démonstration. Par l'inégalité (4.1), une borne union, et l'inégalité de Hoeffding, on a, pour tout $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\mathbf{R}(\widehat{f}_n) - \min_{f\in\mathcal{F}} \mathbf{R}(f) > t\right) &\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| > \frac{t}{2}\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(|\mathbf{R}_n(f_j) - \mathbf{R}(f_j)| > \frac{t}{2}\right) \\
&\leq 2p\exp\left(-\frac{t^2 n}{2}\right)\,.
\end{aligned}$$

En prenant $t = \sqrt{\frac{2\log\left(\frac{2p}{\delta}\right)}{n}}$, on obtient le premier résultat. Pour la deuxième inégalité,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{R}(\widehat{f}_n) - \min_{f\in\mathcal{F}} \mathbf{R}(f)\right] &= \int_0^{+\infty} \mathbb{P}\left(\mathbf{R}(\widehat{f}_n) - \min_{f\in\mathcal{F}} \mathbf{R}(f) > t\right) \mathrm{d}t \\
&\leq \int_0^{+\infty} 2pe^{-\frac{t^2 n}{2}} \wedge 1 \mathrm{d}t \\
&\leq \sqrt{\frac{2\log(2p)}{n}} + 2p\int_{\sqrt{\frac{2\log(2p)}{n}}}^{+\infty} e^{-\frac{t^2 n}{2}}\mathrm{d}t \\
&= \sqrt{\frac{2\log(2p)}{n}} + \frac{2p}{\sqrt{\frac{2\log(2p)}{n}}}\int_{\sqrt{\frac{2\log(2p)}{n}}}^{+\infty} te^{-\frac{t^2 n}{2}}\mathrm{d}t \\
&\leq \sqrt{\frac{2\log(2p)}{n}} + \frac{1}{\sqrt{2n\log(2p)}} \leq 2\sqrt{\frac{2\log(2p)}{n}}\,,
\end{aligned}$$

car $2\log(2p) \geq 1$ pour tout $p \geq 1$. ∎

Proposition 4.5. *Si E est un ensemble fini, le classifieur par majorité $\widehat{f}_n^{maj}$ défini à l'exemple 4.4 satisfait*

$$\sup_P \mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\left[\mathbf{R}_P(\widehat{f}_n^{maj}) - \mathbf{R}_P^\star\right] \leq 2\sqrt{\frac{2(|E|+1)\log(2)}{n}}\,.$$

En particulier, $(\widehat{f}_n^{\mathsf{maj}})$ *est uniformément consistant.*

DÉMONSTRATION. Comme E est fini, l'ensemble $\mathcal{F}$ de tous les classifieurs sur E est lui aussi fini avec $|\mathcal{F}| = 2^{|E|}$. On a alors, pour toute loi P sur $E \times \{0,1\}$, $\min_{f\in\mathcal{F}} \mathbf{R}_P(f) = \mathbf{R}_P^\star$. Il suffit alors de remarquer que $\widehat{f}_n^{\mathsf{maj}}$ est un minimiseur du risque empirique et d'appliquer la Proposition 4.4. ■

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 $\Pi[\cdot \mid X]$ et de la moyenne a posteriori $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]$.

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

La lecture des deux dernières colonnes du tableau suggère une proximité frappante entre la moyenne a posteriori et l'estimateur du maximum de vraisemblance lorsque $n\to+\infty$. Dans ce chapitre, nous allons chercher à étudier la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ en probabilité sous $P_{\theta_0}^{\otimes n}$, avec $\theta_0\in\Theta$ fixé, c'est-à-dire en supposant l'hypothèse fréquentiste

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

On peut en fait se dire que même si le paramètre $\boldsymbol{\theta}$ est aléatoirement distribué selon Π, il y a bien eu un $\theta_0=\boldsymbol{\theta}(\omega)$ fixé choisi selon cette loi. Sous P_{θ_0}, quel est le comportement asymptotique de $\Pi[\cdot \mid \mathbf{X}]$? En particulier, il serait peut-être possible d'utiliser la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ ou un de ses aspects comme estimateur de θ_0. Ainsi dans les exemples ci-dessus (avec $1/\overline{X}_n$ pour le modèle exponentiel), comme $\overline{X}_n\to\theta_0$ en probabilité sous P_{θ_0}, on a

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] \xrightarrow{\mathbb{P}} \theta_0.$$

De plus, on peut également vérifier dans chaque exemple que la variance a posteriori tend vers 0 en probabilité (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 5.1. Dans le cadre bayésien, si l'on note $f(\mathbf{X}) = \int \prod_{i=1}^n p_\theta(X_i)\mathrm{d}\Pi(\theta)$ la densité marginale de $\mathbf{X}$ évaluée en $\mathbf{X}$, on a vu que

$$\mathbb{P}\left(f(\mathbf{X}) = 0\right) = \mathbb{E}[\mathbb{1}_{f(\mathbf{X})=0}] = \int_E \mathbb{1}_{f(x)=0} f(x)\mathrm{d}\mu(x) = 0.$$

Ceci montre que le dénominateur de la formule de Bayes est non nul, presque sûrement sous la loi marginale de $\mathbf{X}$. 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 $\Pi[\cdot \mid \mathbf{X}]$, nous supposerons que le dénominateur de la formule de Bayes est non nul P_{θ_0}-presque sûrement, soit

$$\mathbb{P}_{\theta_0}(f(\mathbf{X}) = 0) = 0\,.$$

de sorte que la formule de Bayes est bien définie P_{θ_0}-presque sûrement. On peut en fait remarquer que l'égalité ci-dessus est de toute façon vérifiée pour Π-presque tout $\theta_0 \in \Theta$. En effet,

$$\begin{aligned}
\int_\Theta \mathbb{P}_{\theta_0}(f(\mathbf{X}) = 0)\mathrm{d}\Pi(\theta_0) &= \int_\Theta \int_E \mathbb{1}_{\{f(x)=0\}} p_{\theta_0}(x)\mathrm{d}\mu(x)\pi(\theta_0)\mathrm{d}\nu(\theta_0) \\
&= \int_E \mathbb{1}_{\{f(x)=0\}} \int_\Theta p_{\theta_0}(x)\pi(\theta_0)\mathrm{d}\nu(\theta_0)\mathrm{d}\mu(x) \\
&= \int_E \mathbb{1}_{\{f(x)=0\}} f(x)\mathrm{d}\mu(x) \\
&= 0\,.
\end{aligned}$$

1. Consistance de lois a posteriori

Cadre. Dans toute la suite de ce chapitre, on considère le cadre d'un modèle $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$ 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 $\Pi[\cdot \mid \mathbf{X}]$, on considère le modèle bayésien

$$\begin{aligned}
\boldsymbol{\theta} &\sim \Pi \\
\mathbf{X} = (X_1, \ldots, X_n) \mid \boldsymbol{\theta} &\sim P_{\boldsymbol{\theta}}^{\otimes n}.
\end{aligned}$$

Une fois $\Pi[\cdot \mid \mathbf{X}]$ formée, on l'étudie sous l'hypothèse fréquentiste

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

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 $P_{\theta_0}^{\otimes n}$ ».

Définition 5.1. On dit que $\Pi[\cdot \mid \mathbf{X}] = \Pi[\cdot \mid X_1, \ldots, X_n]$ est consistante au point $\theta_0 \in \Theta$ si, pour tout $\varepsilon > 0$, sous P_{θ_0},

$$\mathbb{P}\left(\|\boldsymbol{\theta} - \theta_0\| > \varepsilon \mid \mathbf{X}\right) = \Pi\left(\{\theta \in \Theta,\, \|\theta - \theta_0\| > \varepsilon\} \mid \mathbf{X}\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0\,.$$

Remarque 5.2. Pour Z_n une variable aléatoire telle que $0 \leq Z_n \leq 1$, on a

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

et de même $Z_n \xrightarrow{\mathbb{P}} 1$ ssi $\mathbb{E}[Z_n] \to 1$ (exercice). En particulier, pour montrer que l'a posteriori est consistant, il suffit de montrer que

$$\mathbb{E}_{\theta_0}\left[\mathbb{P}\left(\|\boldsymbol{\theta} - \theta_0\| > \varepsilon \mid \mathbf{X}\right)\right] \underset{n\to\infty}{\longrightarrow} 0\,.$$

Exemple 5.3. Voici un exemple de loi a posteriori non-consistante. Soit $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ 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 $[0,1]$. L'a posteriori $\Pi[\cdot \mid \mathbf{X}]$ est donc inconsistant en dehors de $[0,1]$, par exemple en $\theta_0 = 2$ puisque

$$\Pi\left[[3/2, 5/2] \mid \mathbf{X}\right] = 0.$$

1.1. Consistance dans le modèle gaussien avec a priori gaussien.

Proposition 5.1. *Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ avec une loi a priori $\Pi = \mathcal{N}(a, \sigma^2)$, la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ est consistante en tout point $\theta_0 \in \mathbb{R}$.*

Démonstration. La loi a posteriori est donnée par

$$\Pi[\cdot \mid \mathbf{X}] = \mathcal{N}\left(\frac{a\sigma^{-2} + n\overline{X}_n}{n + \sigma^{-2}}, \frac{1}{n + \sigma^{-2}}\right).$$

Notons $m_{\mathbf{X}} = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \frac{a\sigma^{-2} + n\overline{X}_n}{n+\sigma^{-2}}$. Pour tout θ_0 réel et $\varepsilon > 0$, on a

$$\begin{aligned}
\mathbb{P}\left(|\boldsymbol{\theta} - \theta_0| > \varepsilon \mid \mathbf{X}\right) &\leq \mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| + |m_{\mathbf{X}} - \theta_0| > \varepsilon \mid \mathbf{X}\right) \\
&\leq \mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{\varepsilon}{2} \mid \mathbf{X}\right) + \mathbb{1}_{|m_{\mathbf{X}} - \theta_0| > \frac{\varepsilon}{2}},
\end{aligned}$$

où l'on a utilisé l'inégalité triangulaire puis le fait que si $|\boldsymbol{\theta} - m_{\mathbf{X}}| + |m_{\mathbf{X}} - \theta_0| > \varepsilon$, alors au moins l'un des deux termes est strictement supérieur à $\varepsilon/2$. Comme $\overline{X}_n \xrightarrow{\mathbb{P}} \theta_0$ sous P_{θ_0} par la loi des grands nombres, on a $m_{\mathbf{X}} \xrightarrow{\mathbb{P}} \theta_0$ sous P_{θ_0}. Ainsi

$$\mathbb{E}_{\theta_0}\mathbb{1}_{|m_{\mathbf{X}} - \theta_0| > \frac{\varepsilon}{2}} = \mathbb{P}_{\theta_0}\left(|m_{\mathbf{X}} - \theta_0| > \frac{\varepsilon}{2}\right) \underset{n\to\infty}{\longrightarrow} 0\,.$$

D'autre part, d'après l'expression explicite de la loi a posteriori,

$$\begin{aligned}
\mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{\varepsilon}{2} \mid \mathbf{X}\right) &= \mathbb{P}\left(\left|\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n + \sigma^{-2}}\right) - m_{\mathbf{X}}\right| > \frac{\varepsilon}{2} \mid \mathbf{X}\right) \\
&= \mathbb{P}\left(\left|\mathcal{N}\left(0, \frac{1}{n + \sigma^{-2}}\right)\right| > \frac{\varepsilon}{2}\right) \\
&= \mathbb{P}\left(|\mathcal{N}(0,1)| > \frac{\varepsilon}{2}\sqrt{n + \sigma^{-2}}\right) \\
&\underset{n\to\infty}{\longrightarrow} 0\,,
\end{aligned}$$

puisque le lemme de Slutsky donne $\frac{1}{\sqrt{n+\sigma^{-2}}}|\mathcal{N}(0,1)| \xrightarrow{\mathbb{P}} 0$. Donc l'a posteriori est consistant en θ_0. ∎

1.2. Consistance dans le cadre où Θ est fini. Soit $\Theta = \{1, \dots, k\}$. On considère le modèle

$$\mathcal{P} = \{P_1, \dots, P_k\} = \{P_\theta,\ \theta \in \Theta\}, \tag{5.1}$$

où P_j sont des mesures de probabilité sur E. Remarquons que dans le cas fini, le modèle est automatiquement dominé, par exemple par $\mu = P_1 + \cdots + P_k$. On note p_j la densité de P_j par rapport à μ : $\mathrm{d}P_j = p_j \mathrm{d}\mu$. Soit Π une loi a priori sur Θ. Celle-ci est définie par la donnée de

$$\Pi(\{j\}) = \mathbb{P}(\boldsymbol{\theta} = j) = \pi_j, \quad j = 1, \dots, k.$$

Proposition 5.2. *Dans le cadre du modèle discret (5.1), 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 $\Pi[\cdot \mid \mathbf{X}]$ est consistante en tout point $\theta_0 \in \{1, \dots, k\}$.*

Démonstration. Soit $\theta_0 \in \{1, \dots, k\}$. Il suffit de démontrer que

$$\Pi[\{\theta_0\} \mid X_1, \dots, X_n] \xrightarrow{\mathbb{P}} 1$$

sous P_{θ_0}.

Notons $\ell_j(\mathbf{X}) = \prod_{i=1}^n p_j(X_i)$. LA formule de Bayes donne

$$\Pi[\{\theta_0\} \mid \mathbf{X}] = \frac{\pi_{\theta_0} \ell_{\theta_0}(\mathbf{X})}{\sum_{j=1}^k \pi_j \ell_j(\mathbf{X})}.$$

Pour tout $j \neq \theta_0$, on a $\ell_j(\mathbf{X}) \leq \max_{i \neq \theta_0} \ell_i(\mathbf{X})$. Comme $\sum_{i \neq \theta_0} \pi_i = 1 - \pi_{\theta_0}$, on en déduit

$$\Pi[\{\theta_0\} \mid \mathbf{X}] \geq \frac{\pi_{\theta_0} \ell_{\theta_0}(\mathbf{X})}{\pi_{\theta_0} \ell_{\theta_0}(\mathbf{X}) + (1 - \pi_{\theta_0}) \max_{j \neq \theta_0} \ell_j(\mathbf{X})} = \frac{1}{1 + \frac{1-\pi_{\theta_0}}{\pi_{\theta_0}} \frac{\max_{j \neq \theta_0} \ell_j(\mathbf{X})}{\ell_{\theta_0}(\mathbf{X})}}. \tag{5.2}$$

Soit $\varepsilon > 0$. On a

$$\mathbb{P}_{\theta_0}\left(\max_{j \neq \theta_0} \ell_j(\mathbf{X}) \geq \varepsilon \ell_{\theta_0}(\mathbf{X})\right) \leq \sum_{j \neq \theta_0} \mathbb{P}_{\theta_0}\left(\ell_j(\mathbf{X}) \geq \varepsilon \ell_{\theta_0}(\mathbf{X})\right).$$

Pour $j \in [\![1, k]\!] \setminus \{\theta_0\}$, l'inégalité de Markov appliquée avec la fonction $x \mapsto \sqrt{x}$ donne

$$\mathbb{P}_{\theta_0}\left(\ell_j(\mathbf{X}) \geq \varepsilon \ell_{\theta_0}(\mathbf{X})\right) \leq \frac{1}{\sqrt{\varepsilon}} \mathbb{E}_{\theta_0}\left[\sqrt{\frac{\ell_j(\mathbf{X})}{\ell_{\theta_0}(\mathbf{X})}}\right].$$

Or l'espérance dans cette dernière expression s'écrit

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\sqrt{\frac{\ell_j(\mathbf{X})}{\ell_{\theta_0}(\mathbf{X})}}\right] &= \int \left[\frac{\prod_{i=1}^n p_j(x_i)}{\prod_{i=1}^n p_{\theta_0}(x_i)}\right]^{1/2} \prod_{i=1}^n p_{\theta_0}(x_i) \mathrm{d}\mu(x_i) \\
&= \int \sqrt{\prod_{i=1}^n p_j(x_i) \prod_{i=1}^n p_{\theta_0}(x_i)} \prod_{i=1}^n \mathrm{d}\mu(x_i) \\
&= \rho(P_j^{\otimes n}, P_{\theta_0}^{\otimes n}) = \rho(P_j, P_{\theta_0})^n,
\end{aligned}$$

où l'on a utilisé la propriété de l'affinité de Hellinger ρ vue au Chapitre 3. Le modèle étant identifiable, on a $\rho(P_j, P_{\theta_0}) < 1$ pour tout $j \neq \theta_0$ (sinon la distance de Hellinger entre les mesures serait nulle et elles seraient égales), donc $\rho(P_j, P_{\theta_0})^n \xrightarrow[n \to \infty]{} 0$.

Ainsi, pour tout $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0}\left(\max_{j\neq\theta_0} \ell_j(\mathbf{X}) \geq \varepsilon \ell_{\theta_0}(\mathbf{X})\right) \leq \frac{1}{\sqrt{\varepsilon}} \sum_{j\neq\theta_0} \rho(P_j, P_{\theta_0})^n \underset{n\to\infty}{\longrightarrow} 0\,,$$

puisque la somme porte sur un nombre fini de terme $(k-1)$. Autrement dit, sous P_{θ_0},

$$\frac{\max_{j\neq\theta_0} \ell_j(\mathbf{X})}{\ell_{\theta_0}(\mathbf{X})} \xrightarrow{\mathbb{P}} 0\,.$$

et donc la terme de droite dans (5.2) tend vers 1 en probabilité. Comme $\Pi[\{\theta_0\} \,\big|\, \mathbf{X}] \leq 1$, on obtient bien $\Pi[\{\theta_0\} \,\big|\, \mathbf{X}] \xrightarrow{\mathbb{P}} 1$ sous P_{θ_0}. ■

2. Vitesses de convergence

On peut étendre naturellement la notion de consistance en permettant à ε dans la Définition 5.1 de varier, et typiquement de tendre vers 0 avec n.

Définition 5.2. On dit que l'a posteriori $\Pi[\cdot \,\big|\, \mathbf{X}] = \Pi[\cdot \,\big|\, X_1, \ldots, X_n]$ converge à vitesse (au moins) ε_n au point $\theta_0 \in \Theta$ si, sous P_{θ_0},

$$\Pi\left[\{\theta :\ \|\theta - \theta_0\| \leq \varepsilon_n\} \,\big|\, \mathbf{X}\right] \xrightarrow{\mathbb{P}} 1\,.$$

Dans le cadre des modèles paramétriques réguliers, on arrivera typiquement à montrer une convergence à vitesse $M_n/\sqrt{n}$, pour toute suite (M_n) tendant vers l'infini arbitrairement lentement.

Dans certains cas, on peut montrer qu'une vitesse n'est pas améliorable en ordre de grandeur en établissant une borne inférieure à peu près du même ordre. On dira que ζ_n est une borne inférieure pour la vitesse de convergence de $\Pi[\cdot \,\big|\, \mathbf{X}]$ au point $\theta_0 \in \Theta$ si, sous P_{θ_0},

$$\Pi\left[\{\theta :\ \|\theta - \theta_0\| \leq \zeta_n\} \,\big|\, \mathbf{X}\right] \xrightarrow{\mathbb{P}} 0\,.$$

Proposition 5.3. *Dans le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ avec une loi a priori gaussienne $\Pi = \mathcal{N}(a, \sigma^2)$ sur $\boldsymbol{\theta}$, la loi a posteriori $\Pi[\cdot \,\big|\, \mathbf{X}]$ converge en tout point $\theta_0 \in \mathbb{R}$, à vitesse de l'ordre de $1/\sqrt{n}$. Plus précisément, pour tout $\theta_0 \in \mathbb{R}$, et pour (m_n) et (M_n) deux suites telles que $m_n \to 0$ et $M_n \to +\infty$, sous P_{θ_0},*

$$\Pi\left[\left\{\theta :\ \frac{m_n}{\sqrt{n}} \leq \|\theta - \theta_0\| \leq \frac{M_n}{\sqrt{n}}\right\} \,\Big|\, \mathbf{X}\right] \xrightarrow{\mathbb{P}} 1\,.$$

Démonstration. Voir TD. ■

Dans les modèles paramétriques réguliers, la vitesse de convergence sera toujours $1/\sqrt{n}$. Cela résulte du théorème de Bernstein-von Mises.

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, pour des objets beaucoup plus généraux qu'une moyenne empirique. Asymptotiquement, les lois 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, il est d'abord utile de rappeler une notion de proximité pour deux lois, déjà vue au Chapitre 3, la distance en variation totale.

On rappelle la propriété suivante de la distance en variation totale, vue en Proposition 3.8. Soient P, Q deux mesures de probabilité avec $\mathrm{d}P = p\mathrm{d}\mu$ et $\mathrm{d}Q = q\mathrm{d}\mu$. La distance en variation totale entre P et Q vérifie

$$d_{\mathrm{VT}}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|\mathrm{d}\mu(x)\,.$$

Exemple 5.4. Soit $P_n = \mathrm{Unif}[0, 1 + \frac{1}{n}]$ et $P = \mathrm{Unif}[0,1]$. On calcule

$$2d_{\mathrm{VT}}(P_n, P) = \int_0^1 \left|\frac{1}{1+\frac{1}{n}} - 1\right| \mathrm{d}u + \int_1^{1+\frac{1}{n}} \frac{1}{1+\frac{1}{n}}\mathrm{d}u = \frac{2}{n+1}.$$

Ainsi $d_{\mathrm{VT}}(P_n, P) \to 0$ quand $n \to \infty$.

Laplace, au début des années 1800, a remarqué et démontré que dans le modèle binomial $\{\mathcal{B}(n,\theta),\ \theta \in (0,1)\}$, avec une loi a priori uniforme sur $\boldsymbol{\theta}$ (i.e. le modèle considéré par Bayes), la loi a posteriori est une loi $\mathrm{Beta}(1+\mathbf{X}, 1+n-\mathbf{X})$, et que cette loi ressemble étrangement à une loi $\mathcal{N}(\frac{\mathbf{X}}{n}, \frac{\theta_0(1-\theta_0)}{n})$ si $\mathbf{X}$ suit en réalité une loi $\mathcal{B}(n,\theta_0)$. On notera que $\mathbf{X}/n$ 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.

Avant d'énoncer le théorème, on donne une version forte de la notion de modèle régulier.

Définition 5.3. Soit $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, avec $\Theta \subset \mathbb{R}$ ouvert, un modèle dominé avec $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$. On dira que $\mathcal{P}$ est régulier au sens fort si :

— pour tout $x \in E$, la fonction $\theta \mapsto \sqrt{p_\theta(x)}$ est $\mathcal{C}^1$ sur Θ ;

— pour tout $\theta \in \Theta$, il existe $\varepsilon > 0$ tel que

$$\mathbb{E}_\theta\left[\sup_{\eta \in [\theta-\varepsilon, \theta+\varepsilon]} \ell'_\eta(\mathbf{X})^2\right] < \infty\,.$$

On peut vérifier que ces conditions impliquent la notion de régularité donnée dans la première partie du cours (Définition 26). En particulier, elles garantissent l'existence et la continuité de l'information de Fisher $\theta \mapsto \mathbf{I}(\theta)$.

Theorème 5.4 (Théorème de Bernstein-von Mises (BvM))**.** *Soit $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, avec $\Theta \subset \mathbb{R}$ ouvert, un modèle régulier au sens fort. Soit $\theta_0 \in \Theta$. On suppose que $\mathbf{I}(\theta_0) > 0$ et que la loi a priori Π sur Θ vérifie*

— *Π a une densité π par rapport à la mesure de Lebesgue sur $\mathbb{R}$.*

— *$\pi(\theta_0) > 0$ et π est continue au point θ_0.*

On suppose de plus que pour tout $\varepsilon > 0$, il existe une suite de tests (φ_n) telle que

$$\mathbb{P}_{\theta_0}(\varphi_n(\mathbf{X}) = 1) \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{et} \qquad \sup_{\theta,\, |\theta-\theta_0|\geq\varepsilon} \mathbb{P}_\theta(\varphi_n(\mathbf{X}) = 0) \underset{n\to\infty}{\longrightarrow} 0\,. \tag{5.3}$$

Soit $\widehat{\theta}_n(\mathbf{X})$ l'estimateur du maximum de vraisemblance dans ce modèle, supposé unique et consistant. Alors quand $n \to \infty$,

$$d_{\mathrm{VT}}\left(\Pi[\cdot \mid \mathbf{X}], \mathcal{N}\left(\widehat{\theta}_n(\mathbf{X}), \frac{\mathbf{I}(\theta_0)^{-1}}{n}\right)\right) \xrightarrow{\mathbb{P}} 0,$$

sous P_{θ_0}.

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

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

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

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

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.

Démonstration du Théorème 5.4. Nous faisons la preuve dans le modèle gaussien pour une loi a priori gaussienne. Pour une preuve générale, 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 $\Pi = \mathcal{N}(\mu, \sigma^2)$ pour $a \in \mathbb{R}$ fixé. Au vu de l'expression explicite de la loi a posteriori et de l'EMV, et du fait que $\mathbf{I}(\theta) = 1$ pour tout θ dans le modèle gaussien, il s'agit de montrer que sous P_{θ_0},

$$d_{\mathrm{VT}}\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n+\sigma^{-2}}\right), \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right)\right) \xrightarrow{\mathbb{P}} 0\,,$$

avec $m_{\mathbf{X}} = \frac{\mu\sigma^{-2} + n\overline{X}_n}{n+\sigma^{-2}}$. Il y a plusieurs preuves possibles. Celle ci-dessous repose sur une comparaison de distances et un calcul explicite. En utilisant le fait que $d_{\mathrm{VT}}(P,Q) \le h(P,Q)$, où h est la distance de Hellinger (cf. Chapitre 3, Proposition 3.10), puis l'inégalité triangulaire, on a

$$\begin{aligned} d_{\mathrm{VT}}\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n+\sigma^{-2}}\right), \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right)\right) &\le h\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n+\sigma^{-2}}\right), \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right)\right) \\ &\le h\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n+\sigma^{-2}}\right), \mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n}\right)\right) \\ &\quad + h\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n}\right), \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right)\right). \end{aligned}$$

Pour chacun des deux termes ci-dessus, on utilise le Lemme 5.5 ci-dessous pour montrer que l'affinté de Hellinger tend en probabilité vers 1, et donc la distance de Hellinger vers 0. Pour le premier terme, on a, par le Lemme 5.5,

$$\rho\left(\mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n+\sigma^{-2}}\right), \mathcal{N}\left(m_{\mathbf{X}}, \frac{1}{n}\right)\right) = \sqrt{\frac{2\frac{1}{\sqrt{n(n+\sigma^{-2})}}}{\frac{1}{n} + \frac{1}{n+\sigma^{-2}}}} \xrightarrow[n\to\infty]{} 1\,.$$

Pour le deuxième terme, on a, toujours par le Lemme 5.5,

$$\rho\left(\mathcal{N}\left(m_{\mathbf{X}},\frac{1}{n}\right),\mathcal{N}\left(\overline{X}_n,\frac{1}{n}\right)\right)=e^{-\frac{n(m_{\mathbf{X}}-\overline{X}_n)^2}{8}}.$$

Or, sous P_{θ_0},

$$n(m_{\mathbf{X}}-\overline{X}_n)^2=n\left(\frac{\sigma^{-2}(a-\overline{X}_n)}{n+\sigma^{-2}}\right)^2\xrightarrow{\mathbb{P}}0,$$

par le lemme de Slutsky et le fait que $\overline{X}_n\xrightarrow{\mathbb{P}}\theta_0$ sous P_{θ_0}. Ainsi, sous P_{θ_0}, la somme des deux distances converge en probabilité vers 0, ce qu'il fallait démontrer. ■

Lemme 5.5. *Soit ρ l'affinité de Hellinger. Pour tout $a,b\in\mathbb{R}$ et $\sigma,\eta>0$,*

$$\rho\left(\mathcal{N}(a,\sigma^2),\mathcal{N}(b,\sigma^2)\right)=e^{-\frac{(a-b)^2}{8\sigma^2}},$$
$$\rho\left(\mathcal{N}(a,\sigma^2),\mathcal{N}(a,\eta^2)\right)=\sqrt{\frac{2\sigma\eta}{\sigma^2+\eta^2}}.$$

Démonstration. Preuve laissée en exercice. ■

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 $\Pi[\cdot\mid\mathbf{X}]$ dans le modèle $\mathcal{P}$ à partir d'une loi a priori Π et d'observations $\mathbf{X}$. On suppose que la fonction de répartition a posteriori $F_{\mathbf{X}}$ est continue et l'on considère la région de crédibilité $]a_n(\mathbf{X}),b_n(\mathbf{X})]$ de niveau $1-\alpha$ formée par les quantiles de la loi a posteriori

$$\Pi\Big(]-\infty,a_n(\mathbf{X})]\mid\mathbf{X}\Big)=\frac{\alpha}{2},\tag{5.4}$$
$$\Pi\Big(]b_n(\mathbf{X}),+\infty[\mid\mathbf{X}\Big)=\frac{\alpha}{2}.\tag{5.5}$$

Dans la suite, on note $o_{\mathbb{P}}(1)$ toute quantité qui tend vers 0 en probabilité sous $P_{\theta_0}^{\otimes n}$.

Theorème 5.6. *Soit $0<\alpha<1$ et z_α le quantile de niveau $1-\frac{\alpha}{2}$ d'une loi normale standard. Supposons le théorème BvM vérifié. Alors, pour $a_n(\mathbf{X}),b_n(\mathbf{X})$ définis par (5.4)-(5.5), et $\widehat{\theta}_n$ l'EMV,*

$$[a_n(\mathbf{X}),b_n(\mathbf{X})]=\left[\widehat{\theta}_n(\mathbf{X})-\frac{z_\alpha}{\sqrt{n\mathbf{I}(\theta_0)}}(1+o_{\mathbb{P}}(1)),\,\widehat{\theta}_n(\mathbf{X})+\frac{z_\alpha}{\sqrt{n\mathbf{I}(\theta_0)}}(1+o_{\mathbb{P}}(1))\right].$$

Ce résultat donne un développement asymptotique à l'ordre 1 des bornes de l'intervalle de crédibilité $[a_n(\mathbf{X}),b_n(\mathbf{X})]$ 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 l'estimateur du maximum de vraisemblance $\widehat{\theta}_n(\mathbf{X})$. En effet, si l'on suppose les conditions réunies pour que $\widehat{\theta}_n(\mathbf{X})$ soit asymptotiquement efficace au sens où

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{X})-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mathbf{I}(\theta_0)^{-1}),$$

alors l'intervalle

$$I^*(\mathbf{X}) = \left[\widehat{\theta}_n(\mathbf{X}) \pm \frac{z_\alpha}{\sqrt{n\mathbf{I}(\theta_0)}}\right]$$

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

$$\mathbb{P}_{\theta_0}\left(\sqrt{n\mathbf{I}(\theta_0)}\left|\widehat{\theta}_n(\mathbf{X}) - \theta_0\right| \leq z_\alpha\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(|\mathcal{N}(0,1)| \leq z_\alpha\right) = 1-\alpha.$$

En général cependant, l'EMV peut être difficile à déterminer. De plus, l'information de Fisher $\mathbf{I}(\theta_0)$ est généralement inconnue puisqu'elle dépend de θ_0. Une solution standard consiste à remplacer $\mathbf{I}(\theta_0)$ par un estimateur, par exemple $\mathbf{I}(\widehat{\theta}_n(\mathbf{X}))$ (sous les conditions de régularité habituelles, $\theta \mapsto \mathbf{I}(\theta)$ est continue, donc la convergence en probabilité de $\widehat{\theta}_n(\mathbf{X})$ vers θ_0 implique celle de $\mathbf{I}(\widehat{\theta}_n(\mathbf{X}))$ vers $\mathbf{I}(\theta_0)$ et l'on peut appliquer le lemme de Slutsky).

Un des intérêts de l'approche bayésienne est que l'obtention de la région de crédibilité est « automatique » (sous réserve de savoir calculer les quantiles a posteriori, ce qui n'est pas toujours évident). De plus, nous allons voir ci-dessous que l'on peut l'utiliser comme région de confiance.

Démonstration du Théorème 5.6. Soient A et B les ensembles mesurables définis par

$$A =]-\infty, a_n(\mathbf{X})], \qquad B =]b_n(\mathbf{X}), +\infty[.$$

Par définition de $a_n(\mathbf{X})$ et $b_n(\mathbf{X})$, on a

$$\Pi[A \mid \mathbf{X}] = \Pi[B \mid \mathbf{X}] = \frac{\alpha}{2}.$$

Le théorème BvM est vérifié par hypothèse et d'après la définition de la distance en variation totale, on a donc, en notant $\mathcal{B}(\Theta)$ la tribu borélienne sur Θ,

$$\sup_{\Lambda\in\mathcal{B}(\Theta)}\left|\Pi[\Lambda \mid \mathbf{X}] - \mathcal{N}\left(\widehat{\theta}_n(\mathbf{X}), \frac{\mathbf{I}(\theta_0)^{-1}}{n}\right)(\Lambda)\right| = o_{\mathbb{P}}(1).$$

En particulier, en appliquant ceci à $\Lambda = A$, on en déduit que

$$\mathcal{N}\left(\widehat{\theta}_n(\mathbf{X}), \frac{\mathbf{I}(\theta_0)^{-1}}{n}\right)(A) = \frac{\alpha}{2} + o_{\mathbb{P}}(1).$$

En notant Φ la fonction de répartition d'une $\mathcal{N}(0,1)$, cela peut se réécrire

$$\Phi\left(\sqrt{n\mathbf{I}(\theta_0)}(a_n(\mathbf{X}) - \widehat{\theta}_n(\mathbf{X}))\right) = \frac{\alpha}{2} + o_{\mathbb{P}}(1),$$

soit encore

$$\sqrt{n\mathbf{I}(\theta_0)}(a_n(\mathbf{X}) - \widehat{\theta}_n(\mathbf{X})) = \Phi^{-1}\left(\frac{\alpha}{2} + o_{\mathbb{P}}(1)\right).$$

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

$$a_n(\mathbf{X}) = \widehat{\theta}_n(\mathbf{X}) - \frac{z_\alpha}{\sqrt{n\mathbf{I}(\theta_0)}}(1 + o_{\mathbb{P}}(1)),$$

et le résultat pour $b_n(\mathbf{X})$ s'obtient de la même façon. ■

Theorème 5.7 (Confiance asymptotique des régions de crédibilité)**.** *Supposons le théorème BvM vérifié. Alors l'intervalle de crédibilité* $I(\mathbf{X}) = [a_n(\mathbf{X}), b_n(\mathbf{X})]$ *défini par* (5.4)-(5.5) *est un intervalle de confiance asymptotique au niveau* $1-\alpha$*, c'est-à-dire*

$$\mathbb{P}_{\theta_0}\left(\theta_0 \in [a_n(\mathbf{X}), b_n(\mathbf{X})]\right) \underset{n\to\infty}{\longrightarrow} 1-\alpha.$$

Démonstration. Il suffit de montrer que $\mathbb{P}_{\theta_0}(\theta_0 < a_n(\mathbf{X})) \to \alpha/2$ et que $\mathbb{P}_{\theta_0}(\theta_0 > b_n(\mathbf{X})) \to \alpha/2$. Pour cela, on utilise les développements asymptotiques obtenus au Theorème 5.6.

$$\begin{aligned}\mathbb{P}_{\theta_0}(\theta_0 < a_n(\mathbf{X})) &= \mathbb{P}_{\theta_0}\left(\theta_0 < \widehat{\theta}_n(\mathbf{X}) - \frac{z_\alpha}{\sqrt{n\mathbf{I}(\theta_0)}}(1+o_{\mathbb{P}}(1))\right)\\ &= \mathbb{P}_{\theta_0}\left(\sqrt{n\mathbf{I}(\theta_0)}(\widehat{\theta}_n(\mathbf{X})-\theta_0) - o_{\mathbb{P}}(1) > z_\alpha\right).\end{aligned}$$

Comme la quantité à gauche du signe $>$ 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 $\mathbb{P}_{\theta_0}(\theta_0 > b_n(\mathbf{X}))$, ce qui conclut la démonstration. ■

5. Analyse asymptotique des tests bayésiens

Définition 5.4. Un test $\varphi(\mathbf{X})$ est dit consistant si

$$\forall\,\theta\in\Theta,\quad \mathbb{E}_\theta\varphi(\mathbf{X}) \underset{n\to\infty}{\longrightarrow} \begin{cases}0 & \text{si } \theta\in\Theta_0,\\ 1 & \text{si } \theta\in\Theta_1.\end{cases}$$

Cas où $\Theta\subset\mathbb{R}$. Supposons que les sous-ensembles Θ_0 et Θ_1 sont bien séparés au sens où

$$\inf\{|\theta_0-\theta_1|,\ \theta_0\in\Theta_0, \theta_1\in\Theta_1\} =: \rho > 0. \tag{5.6}$$

Proposition 5.8. *Si* Θ_0 *et* Θ_1 *sont bien séparés au sens de* (5.6)*, alors, si la loi a posteriori est consistante, le test de Bayes pour la perte pondérée avec* $a_0, a_1 > 0$ *est consistant.*

Démonstration. Supposons l'hypothèse (5.6) vérifiée et la loi a posteriori consistante. Soit $\theta_0\in\Theta_0$. Par l'hypothèse de consistance de $\Pi[\cdot \mid \mathbf{X}]$, on a pour tout $\varepsilon>0$

$$\Pi\left(\{\theta:\ |\theta-\theta_0|>\varepsilon\} \mid \mathbf{X}\right) \overset{\mathbb{P}_{\theta_0}}{\longrightarrow} 0.$$

Or, par (5.6), on a

$$\Theta_1 \subset \{\theta,\ |\theta-\theta_0|>\rho/2\}.$$

On a donc $\Pi(\Theta_1 \mid \mathbf{X}) \overset{\mathbb{P}_{\theta_0}}{\longrightarrow} 0$. Pour le test de Bayes $\varphi^\star(\mathbf{X}) = \mathbb{1}_{\Pi(\Theta_0|\mathbf{X})\le\frac{a_1}{a_0+a_1}}$, on a donc

$$\mathbb{E}_{\theta_0}\varphi^\star(\mathbf{X}) = \mathbb{P}_{\theta_0}\left(\Pi(\Theta_1\mid\mathbf{X}) \ge \frac{a_0}{a_0+a_1}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Un raisonnement similaire montre que si $\theta\in\Theta_1$, alors $\mathbb{E}_\theta\varphi^\star(X) \underset{n\to\infty}{\longrightarrow} 1$. ■

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 $H_0:\{\theta=0\}$ contre $H_1:\{\theta\neq 0\}$ dans le modèle

gaussien (Chapitre 4), l'hypothèse (5.6) n'est clairement pas vérifiée. On peut néanmoins montrer que, pour $\Pi = (1-\pi_0)\mathcal{N}(0,1) + \pi_0\delta_0$, on a

$$\Pi(\{0\} \mid \mathbf{X}) \xrightarrow{\mathbb{P}} 1 \quad \text{sous } P_\theta, \quad \text{pour } \theta = 0$$
$$\Pi(\{0\} \mid \mathbf{X}) \xrightarrow{\mathbb{P}} 0 \quad \text{sous } P_\theta, \quad \text{pour } \theta \neq 0,$$

ce qui implique que le test est consistant.

Chapitre 6

Simulation de la loi a posteriori (bis) : les méthodes MCMC

Dans ce dernier chapitre, nous faisons un bref tour d'horizon de méthodes MCMC. Nous présentons notamment l'algorithme de Metropolis-Hastings et l'échantillonnage de Gibbs.

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

1. Un bref aperçu sur les chaînes de Markov

Une chaîne de Markov homogène $(X_t)_{t\in\mathbb{N}}$ à espace d'états mesurable $(\Omega, \mathcal{F})$ est un processus aléatoire à valeurs dans Ω dont les transitions se font de la façon suivante : si la chaîne est en $x \in \Omega$, alors, quelle que soit la trajectoire passée, l'état suivant est choisi selon une loi de probabilité fixée $P(x, \cdot)$ sur Ω, qui ne dépend que de x. Autrement dit, pour tout $A \in \mathcal{F}$ et pour tout $t \geq 0$, on a

$$\mathbb{P}(X_{t+1} \in A \mid X_0 \ldots, X_t) = P(X_t, A)\,.$$

La loi de la chaîne est ainsi complètement caractérisée par la loi de X_0 et par la collection de lois $(P(x, \cdot))_{x\in\Omega}$.

L'application

$$\begin{aligned} P : \Omega \times \mathcal{F} &\to [0, 1] \\ (x, A) &\mapsto P(x, A) \end{aligned}$$

s'appelle un noyau de transition. On a en particulier

$$\forall x \in \Omega\,, \quad P(x, \Omega) = 1\,.$$

On suppose de plus que pour tout $A \in \mathcal{F}$, l'application $x \mapsto P(x, A)$ est mesurable.

Si Ω est un ensemble fini, on verra P comme une matrice stochastique (tous les coefficients sont positifs et, sur chaque ligne, la somme des coefficients vaut 1), de taille $|\Omega| \times |\Omega|$, l'entrée $P(x, y)$ correspondant à la probabilité, partant de x d'arrivée en y en un pas.

Exemple 6.1 (Marche aléatoire sur $\mathbb{R}$)**.** Soit $X_0 \sim \mathcal{N}(0, 1)$ et soit $(\xi_i)_{i\geq 1}$ une suite i.i.d. de variables de loi $\mathcal{N}(0, 1)$, indépendante de X_0. Le processus donné par, pour $n \geq 0$,

$$X_{n+1} = X_n + \xi_{n+1}\,,$$

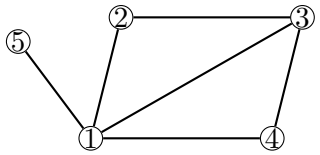
est une chaîne de Markov à valeurs dans $\Omega = \mathbb{R}$ avec noyau de transition $P(x, \cdot) = \mathcal{N}(x, 1)$. Il s'agit d'une marche aléatoire avec sauts gaussiens.

Exemple 6.2 (Marche aléatoire sur un graphe fini)**.** Une marche aléatoire sur un graphe $G = (V, E)$ fini est un processus sur V qui se déplace en sautant, à chaque temps, sur un

voisin choisi uniformément au hasard. Il s'agit d'une chaîne de Markov sur $\Omega = V$, dont la matrice de transition P est donnée par

$$P(u,v) = \begin{cases} \frac{1}{\deg(u)} & \text{si } \{u,v\} \in E, \\ 0 & \text{sinon,} \end{cases}$$

où $\deg(u)$ (le *degré* de u) correspond au nombre de voisins de u dans G. Voici par exemple un graphe à 5 sommets et la matrice de transition correspondante :



$$P = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Si ν_0 est la loi initiale de X_0, alors la loi de X_1 est donnée par

$$\forall A \in \mathcal{F}\,, \quad \nu_1(A) = \nu_0 P(A) = \int_\Omega P(x,A)\nu_0(\mathrm{d}x)\,.$$

Pour $t \geq 1$, la loi de X_t peut être définie par récurrence :

$$\forall A \in \mathcal{F}\,, \quad \nu_t(A) = \nu_{t-1} P(A) = \int_\Omega P(x,A)\nu_{t-1}(\mathrm{d}x)\,.$$

On définit ainsi P^t, le $t^{\text{ième}}$ itéré de P comme $\nu_t P$ pour $\nu_0 = \delta_x$:

$$P^t(x,A) = \int_\Omega P^{t-1}(z,A)P(x,\mathrm{d}z)\,.$$

On a alors $P^t(x,A) = \mathbb{P}(X_t \in A \,\big|\, X_0 = x)$, et, pour ν_0 quelconque,

$$\forall A \in \mathcal{F}\,, \quad \nu_t(A) = \int_\Omega P^t(x,A)\nu_0(\mathrm{d}x)\,.$$

Dans le cas discret fini, P^t est simplement la matrice P à la puissance t. Si l'on considère la loi initiale ν_0 comme un vecteur ligne avec, pour $x \in \Omega$, $\nu_0(x) = \mathbb{P}(X_0 = x)$, alors la loi de la chaîne au temps t est donnée par le vecteur ligne $\nu_t = \nu_0 P^t$:

$$\forall y \in \Omega\,,\ \nu_t(y) = \mathbb{P}(X_t = y) = \sum_{x\in\Omega} \nu_0(x)\mathbb{P}(X_t = y \,\big|\, X_0 = x) = \sum_{x\in\Omega} \nu_0(x)P^t(x,y) = \nu_0 P^t(y)\,.$$

Définition 6.1. On dit que π est une loi stationnaire (ou invariante) pour la chaîne de Markov $(X_t)_{t\in\mathbb{N}}$ si $\pi P = \pi$, i.e. si pour tout $A \in \mathcal{F}$,

$$\int_\Omega P(x,A)\pi(\mathrm{d}x) = \pi(A)\,.$$

Dans le cas discret fini, cela revient à dire que pour tout $y \in \Omega$, $\sum_{x\in\Omega} \pi(x)P(x,y) = \pi(y)$.

Autrement dit, si la loi de X_0 est de loi π et que l'on applique une transition de la chaîne, alors la loi de X_1 est toujours π, et il en est de même de la loi de X_t, pour tout $t \geq 1$.

Remarque 6.3. Une loi stationnaire n'existe pas toujours. Ainsi par exemple, la marche aléatoire simple sur $\mathbb{Z}$ (si l'état courant est $x \in \mathbb{Z}$, alors l'état suivant est $x+1$ ou $x-1$ avec probabilité $1/2$) n'admet pas de probabilité stationnaire. Elle a un comportement « trop diffusif ».

Dans ce qui suit, nous nous restreignons à des espaces d'états finis. On peut étendre la plupart des résultats ci-dessous à des espaces plus généraux, mais cela dépasserait assez largement le cadre de ce cours.

Définition 6.2. Soit Ω un ensemble fini. Un noyau de transition P sur Ω est dit irréductible si pour tous $x, y \in \Omega$, il existe $t \in \mathbb{N}$ tel que $P^t(x,y) > 0$.

Theorème 6.1. *Soit Ω un ensemble fini et P est un noyau de transition sur Ω. Alors P admet une probabilité stationnaire π, et, si P est irréductible, cette probabilité est unique et charge tous les états.*

Démonstration. Soit ν_0 la mesure initiale de la chaîne, et pour $s \geq 1$, soit $\nu_s = \nu_0 P^s$, la loi de la chaîne au temps s. Définissons la mesure

$$\pi_t = \frac{1}{t} \sum_{s=0}^{t-1} \nu_s \,.$$

Par compacité de l'espace des mesures de probabilité sur Ω, il existe une sous-suite de (π_t) qui converge, notons π la limite de cette sous-suite. Alors π est stationnaire. En effet,

$$\pi_t P - \pi_t = \frac{\nu_t - \nu_0}{t} \underset{t \to +\infty}{\longrightarrow} 0 \,.$$

Supposons maintenant que P est irréductible et montrons que π charge tous les états. Comme π est une mesure de probabilité, il existe $x_\star \in \Omega$ tel que $\pi(x_\star) > 0$. Soit y un état quelconque. Par irréductibilité, il existe $t \geq 0$ tel que $P^t(x_\star, y) > 0$. Donc

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P^t(x,y) \geq \pi(x_\star) P^t(x_\star, y) > 0 \,.$$

Montrons maintenant que π est unique. En passant à la transposée, on a ${}^tP\,{}^t\pi = {}^t\pi$. Ainsi, pour montrer que π est unique, il suffit de montrer que le noyau de ${}^tP - I$ est de dimension 1. Une fonction h est dite P-harmonique si, pour tout $x \in \Omega$,

$$\sum_{y \in \Omega} P(x,y) h(y) = h(x) \,.$$

Montrons le résultat intermédiaire suivant : si P est irréductible, toutes les fonctions harmoniques sont constantes. Comme Ω est fini, il existe x_0 tel que $h(x_0) = M = \max_{x \in \Omega} h(x)$. Supposons qu'il existe $z \in \Omega$ avec $P(x_0, z) > 0$ et $h(z) < M$. Alors

$$h(x_0) = \sum_{y \in \Omega} P(x_0, y) h(y) = P(x_0, z) h(z) + \sum_{y \neq z} P(x_0, y) h(y) < M \,,$$

ce qui est absurde. Donc pour tout z tel que $P(x_0, z) > 0$, on a $h(z) = M$. Par irréductibilité, pour tout $y \in \Omega$, il existe un chemin $x_0, \ldots, x_n = y$ avec $P(x_i, x_{i+1}) > 0$ pour tout $i \in [\![0, n-1]\!]$. En répétant l'argument ci-dessus, on obtient $h(x_0) = h(x_1) = \cdots = h(y) = M$, donc h est

constante. Autrement dit, si P est irréductible le noyau de $P-I$ est de dimension 1. Et comme le rang d'une matrice carrée est égale au rang de sa transposée, le noyau de ${}^tP-I$ est aussi de dimension 1. Donc il existe un unique élément de ce noyau dont les coordonnées somment à 1. ■

Une façon simple de trouver une probabilité stationnaire est souvent de chercher une probabilité qui satisfait la condition dite d'équilibre détaillé.

Proposition 6.2. *Soit P un noyau de transition sur Ω fini. Si π est une probabilité sur Ω qui vérifie la condition d'équilibre détaillé*

$$\forall x, y \in \Omega\,, \quad \pi(x)P(x,y) = \pi(y)P(y,x)\,,$$

(on dit que P est réversible par rapport à π), alors π est stationnaire.

Démonstration. En sommant la condition d'équilibre détaillé sur y, on obtient

$$\sum_{y\in\Omega} \pi(y)P(y,x) = \sum_{y\in\Omega} \pi(x)P(x,y) = \pi(x)\sum_{y\in\Omega} P(x,y) = \pi(x).$$

La probabilité π donc bien stationnaire. ■

On a vu que si le noyau P est irréductible, alors il existe donc une unique probabilité stationnaire π qui peut être vue comme un point fixe pour l'action de P. On peut alors montrer que la moyenne de n'importe quelle fonction le long de la trajectoire de la chaîne, converge presque sûrement, quelle que soit la loi initiale, vers l'espérance de cette fonction sous la loi π. C'est le théorème ergodique, qui correspond à un équivalent de la loi forte des grands nombres pour les chaînes de Markov.

Theorème 6.3 (Théorème ergodique)**.** *Soit P est un noyau ergodique sur Ω fini et π sa probabilité stationnaire. Soit $f : \Omega \to \mathbb{R}$. Alors, pour toute mesure initiale ν sur Ω,*

$$\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) \xrightarrow[t\to+\infty]{\text{p.s.}} \mathbb{E}_\pi f = \sum_{x\in\Omega} f(x)\pi(x)\,.$$

Une conséquence importante du théorème ergodique est que si l'on souhaite approcher l'intégrale $\mathbb{E}_\pi f$, il n'est pas nécessaire de savoir simuler selon π. Il suffit de trouver une chaîne de Markov dont π est la mesure stationnaire.

Sous des hypothèses additionnelles, on peut montrer que la loi stationnaire π est la loi limite de la chaîne de Markov : asymptotiquement, la chaîne est distribuée selon π, on dit qu'elle *mélange*. On a alors un moyen de simuler approximativement selon π : on lance la chaîne de Markov, on la laisse évoluer pendant un temps assez long, la loi de X_t sera alors proche de π. Pour garantir cette convergence, l'irréductibilité ne suffit pas. Il faut une propriété plus forte : l'ergodicité.

Définition 6.3. Soit Ω un ensemble fini. Le noyau P est dit ergodique si

$$\exists t \in \mathbb{N}\,, \quad \forall x, y \in \Omega\,, \quad P^t(x,y) > 0\,.$$

Pour quantifier l'écart entre la loi de la chaîne à un certain temps t et la loi stationnaire π, il nous faut une distance entre lois de probabilité. Pour $x \in \Omega$, notons $\mathcal{D}_x(t)$ la distance en variation totale entre la loi de la chaîne au temps t partie de x et la loi stationnaire, i.e.

$$\mathcal{D}_x(t) = d_{\text{VT}}\left(P^t(x,\cdot)\,,\,\pi\right) = \max_{A\subset\Omega}\left(P^t(x,A)-\pi(A)\right) = \sum_{y\in\Omega}\left(P^t(x,y)-\pi(y)\right)_+\,,$$

et

$$\mathcal{D}(t) = \max_{x\in\Omega}\mathcal{D}_x(t)\,.$$

Theorème 6.4. *Si P est un noyau ergodique sur Ω fini, alors*

$$\mathcal{D}(t) \underset{t\to\infty}{\longrightarrow} 0\,.$$

Démonstration. Commençons par remarquer que la suite $\mathcal{D}(t)$ converge car elle est minorée par 0 et est décroissante. En effet, pour tout $x \in \Omega$, par l'inégalité triangulaire ($(a+b)_+ \le a_+ + b_+$), on a

$$\begin{aligned}\mathcal{D}_x(t+1) &= \sum_{y\in\Omega}\left(P^t(z,y)-\pi(y)\right)_+\\ &= \sum_{y\in\Omega}\left(\sum_{z\in\Omega}P(x,z)\left(P^t(z,y)-\pi(y)\right)\right)_+\\ &\le \sum_{z\in\Omega}P(x,z)\mathcal{D}_z(t) \le \mathcal{D}(t)\,.\end{aligned}$$

Notons

$$\overline{\mathcal{D}}(t) = \max_{x,y\in\Omega} d_{\text{VT}}\left(P^t(x,\cdot)\,,\,P^t(y,\cdot)\right)\,.$$

On a $\mathcal{D}(t) \le \overline{\mathcal{D}}(t)$. En effet, par définition de π et l'inégalité triangulaire, on a, pour tout $x \in \Omega$,

$$\begin{aligned}\mathcal{D}_x(t) &= \sum_{z\in\Omega}\left(P^t(x,z)-\sum_{y\in\Omega}\pi(y)P^t(y,z)\right)_+\\ &\le \sum_{y\in\Omega}\pi(y)\sum_{z\in\Omega}\left(P^t(x,z)-P^t(y,z)\right)_+\\ &\le \max_{y\in\Omega} d_{\text{VT}}\left(P^t(x,\cdot)\,,\,P^t(y,\cdot)\right)\,.\end{aligned}$$

En prenant le maximum sur $x \in \Omega$, on obtient l'inégalité voulue. On va montrer que $\overline{\mathcal{D}}(t) \to 0$. Remarquons que $\overline{\mathcal{D}}(\cdot)$ est sous-multiplicative : $\overline{\mathcal{D}}(t+s) \le \overline{\mathcal{D}}(t)\overline{\mathcal{D}}(s)$. En effet, soit $A \subset \Omega$ et soit $B = \{z \in \Omega,\, P^t(x,z) \ge P^t(y,z)\}$. En décomposant selon que la chaîne est en B ou en B^c au temps t, on a

$$\begin{aligned}P^{t+s}(x,A)-P^{t+s}(y,A) &= \sum_{z\in B}\left(P^t(x,z)-P^t(y,z)\right)P^s(z,A) - \sum_{z\in B^c}\left(P^t(y,z)-P^t(x,z)\right)P^s(z,A)\\ &\le d_{\text{VT}}\left(P^t(x,\cdot)\,,\,P^t(y,\cdot)\right)\max_{u,v\in V}\left(P^s(u,A)-P^s(v,A)\right)\,,\end{aligned}$$

où l'on a utilisé le fait que $d_{\mathrm{VT}}\left(P^t(x,\cdot) - P^t(y,\cdot)\right) = P^t(x,B) - P^t(y,B)$. En prenant le maximum sur $A \subset \Omega$, on obtient

$$d_{\mathrm{VT}}\left(P^{t+s}(x,\cdot)\,,\,P^{t+s}(y,\cdot)\right) \leq d_{\mathrm{VT}}\left(P^t(x,\cdot)\,,\,P^t(y,\cdot)\right)\overline{\mathcal{D}}(s)\,,$$

et en prenant le maximum sur $x, y \in \Omega$, on a bien $\overline{\mathcal{D}}(t+s) \leq \overline{\mathcal{D}}(t)\overline{\mathcal{D}}(s)$. Par l'hypothèse d'ergodicité, on peut trouver $t_\star$ tel que tous les coefficients de la matrice $P^{t_\star}$ soient strictement positifs. Ainsi, en utilisant la Proposition 3.8,

$$\overline{\mathcal{D}}(t_\star) = \max_{x,y}\left\{1 - \sum_{z\in\Omega} P^{t_\star}(x,z) \wedge P^{t_\star}(y,z)\right\} < 1\,,$$

et $\overline{\mathcal{D}}(kt_\star) \leq \overline{\mathcal{D}}(t_\star)^k \underset{k\to\infty}{\to} 0$. Donc $\overline{\mathcal{D}}(t) \to 0$ et il en est de même de $\mathcal{D}(t)$. ■

2. Algorithmes MCMC

Le cadre est le suivant. Supposons que l'on veuille soit simuler (disons approximativement) suivant une loi de densité π, ou bien que l'on veuille évaluer une intégrale du type $\mathbf{I} = \int \phi(x)\pi(x)\mathrm{d}\mu(x)$ (avec μ typiquement la mesure de Lebesgue ou la mesure de comptage), comme c'est le cas en statistiques bayésiennes pour π la densité a posteriori et $\mathbf{I}$ la moyenne a posteriori par exemple. On aimerait construire une chaîne de Markov (X_t) de densité stationnaire π, car alors, d'après les deux faits ci-dessus, la loi de X_t avec t grand sera proche d'une loi de densité π, tandis que la moyenne $\frac{1}{t}\sum_{i=0}^{t-1}\phi(X_i)$ approchera l'intégrale cherchée, par le théorème ergodique.

2.1. L'algorithme de Metropolis-Hastings.

Soit $Q(x,\cdot)$ une collection de lois et supposons que l'on sait simuler rapidement suivant ces lois. Par exemple, si $Q(x,\cdot)$ correspond à une loi $\mathcal{N}(x,1)$, alors on sait bien simuler selon cette loi (c'est ce qu'on appelle *Random walk Metropolis-Hastings*).

Soit π la loi suivant laquelle on veut simuler, ou pour laquelle on veut calculer $\int f(x)\pi(\mathrm{d}x)$.

Pour simplifier les choses, nous allons à nouveau supposer que Ω est un ensemble fini.

Soit Q une matrice de transition irréductible sur Ω et tel que, pour tous $x, y \in \Omega$, $Q(x,y) > 0$ ssi $Q(y,x) > 0$. On définit, à partir de Q, un noyau P dont la loi stationnaire est π, de la façon suivante : si l'état courant est $x \in \Omega$, on génère $y \in \Omega$ selon $Q(x,\cdot)$, et l'on accepte cette transition en y avec probabilité :

$$r(x,y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)} \wedge 1\,.$$

Le rapport $r(x,y)$ est appelé rapport de Metropolis-Hastings. Le noyau de transition P de cette nouvelle chaîne est donné par

$$P(x,y) = \begin{cases} Q(x,y)r(x,y)\,, & \text{si } y \neq x\ , \\ 1 - \sum_{z\neq x} Q(x,z)r(x,z)\,, & \text{si } y = x\ . \end{cases} \tag{6.1}$$

Theorème 6.5. *Le noyau P défini par (6.1) est réversible pour π, donc π est stationnaire pour P.*

DÉMONSTRATION. Soient x, y dans Ω avec $x \neq y$. Par symétrie on peut toujours supposer $\pi(y)Q(y,x) \leq \pi(x)Q(x,y)$, quitte à échanger les rôles de x et y (la condition d'équilibre détaillé ne change pas si on permute x et y). Dans ce cas notons que

$$r(x,y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, \qquad r(y,x) = 1.$$

Comme $x \neq y$, pour passer de x à y avec la chaîne définie par l'algorithme, il faut deux choses : générer y avec probabilité $Q(x,y)$ et accepter le mouvement de x à y avec probabilité $r(x,y)$. Ainsi

$$P(x,y) = Q(x,y)r(x,y) = \frac{\pi(y)Q(y,x)}{\pi(x)}.$$

On en déduit que $\pi(x)P(x,y) = \pi(y)Q(y,x)$. Mais $Q(y,x) = P(y,x)$ puisque $r(y,x) = 1$. Ainsi les conditions d'équilibre détaillé sont vérifiées pour $x \neq y$, et elles sont immédiates pour $x = y$. ■

Exemple 6.4. Soit Q la matrice de transition de la marche aléatoire simple sur un graphe fini connexe $G = (V, E)$. La distribution stationnaire de cette chaîne est $\frac{\deg(\cdot)}{2|E|}$. On souhaite obtenir un échantillon issu de la loi uniforme sur V, $\pi = \mathrm{Unif}(V)$. L'algorithme de Metropolis-Hastings consiste alors à définir la chaîne de noyau P donnée par : si l'état courant est x, on choisit uniformément un voisin y de x et l'on accepte la transition de x à y avec probabilité

$$r(x,y) = \frac{\deg(x)}{\deg(y)} \wedge 1\,.$$

La loi stationnaire de P est la loi uniforme sur V.

Application aux statistiques bayésiennes. Dans un cadre bayésien, on cherche typiquement à simuler suivant la loi a posteriori, ou à calculer des intégrales de type $\int \phi(\theta)\mathrm{d}\Pi(\theta \mid \mathbf{X})$. La loi cible est donc généralement $\Pi[\cdot \mid \mathbf{X}]$. Pour utiliser l'algorithme de Metropolis-Hastings, il faut savoir simuler suivant $Q(x,\cdot)$ pour tout x. Comme on a le choix du noyau Q, on peut choisir un noyau selon lequel on peut simuler. Mais il faut aussi pouvoir calculer le quotient dans la probabilité d'acceptation $r(x,y)$. C'est en principe un problème, car $\pi(\theta \mid \mathbf{X})$ est typiquement difficile à évaluer, notamment parce que son expression contient le dénominateur $\int p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\theta$. Le point remarquable ici est que cette quantité se simplifie et que l'on a seulement besoin de connaître la loi a posteriori à constante près. En effet, ici

$$\begin{aligned} r(x,y) &= \frac{\pi(y \mid \mathbf{X})Q(y,x)}{\pi(x \mid \mathbf{X})Q(x,y)} \wedge 1 \\ &= \frac{\frac{p_y(\mathbf{X}))\pi(y)}{\int p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\theta}}{\frac{p_x(\mathbf{X})\pi(x)}{\int p_\theta(\mathbf{X})\pi(\theta)\mathrm{d}\theta}} \frac{Q(y,x)}{Q(x,y)} \wedge 1 \\ &= \frac{p_y(\mathbf{X})\pi(y)}{p_x(\mathbf{X})\pi(x)} \frac{Q(y,x)}{Q(x,y)} \wedge 1\,. \end{aligned}$$

Cette expression se calcule directement, du moins si l'expression de la densité a priori π n'est pas trop complexe. On remarque également qu'à nouveau ici, π n'a besoin d'être connue qu'à constante multiplicative près.

2.2. L'algorithme de Gibbs. On souhaite simuler suivant la loi d'un vecteur de Ω^d, de densité $\pi : \Omega^d \to \mathbb{R}_+$. Pour $x \in \Omega^d$ et $\ell \in [\![1, d]\!]$, on note

$$x^{(\ell)} = (x_1, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_d) \in \Omega^{d-1}\,.$$

On suppose que l'on sait facilement simuler suivant les densités conditionnelles $\pi(\cdot \mid x^{(\ell)})$, pour tout $\ell \in [\![1, d]\!]$. L'idéee de l'algorithme de Gibbs est de rejouer une par une les coordonnée du vecteur, selon la loi conditionnelle sachant toutes les autres. On présente ici deux versions de l'algorithme, qui diffèrent en la façon de choisir les coordonnées que l'on rejoue.

2.2.1. *Gibbs avec balayage aléatoire.* Pour $t \in \mathbb{N}$, partant de $X_t = x \in \Omega^d$, on génère X_{t+1} de la façon suivante :
— on tire une coordonnée $\ell \in [\![1, d]\!]$ uniformément au hasard ;
— on génère y_ℓ selon la densité $\pi(\cdot \mid x^{(\ell)})$;
— on pose $X_{t+1} = (x_1, \dots, x_{\ell-1}, y_\ell, x_{\ell+1}, \dots, x_d)$.

Proposition 6.6. *La chaîne $(X_t)_{t\in\mathbb{N}}$ est réversible pour π.*

Démonstration. Soient $x, y \in \Omega^d$. Si y diffère de x plus de deux coordonnées, alors $P(x, y) = P(y, x) = 0$. Sinon, il existe ℓ et $y_\ell \in \Omega$, tel que $y = (x_1, \dots, x_{\ell-1}, y_\ell, x_{\ell+1}, \dots, x_d)$. Dans ce cas, on a

$$P(x, y) = \frac{1}{d}\pi(y_\ell \mid x^{(\ell)}) = \frac{1}{d}\frac{\pi(y)}{\pi(x^{(\ell)})}\,.$$

De même, $P(y, x) = \frac{1}{d}\frac{\pi(x)}{\pi(x^{(\ell)})}$. Ainsi pour tous $x, y \in \Omega^d$, on a $\pi(x)P(x, y) = \pi(y)P(y, x)$ et la chaîne est bien réversible pour π. ∎

2.2.2. *Gibbs avec balayage déterministe.* Pour $t \in \mathbb{N}$, partant de $X_t = x \in \Omega^d$, on génère X_{t+1} de la façon suivante :
— pour $\ell = 1, \dots, d$, on génère y_ℓ selon la densité $\pi(\cdot \mid y_1, \dots, y_{\ell-1}, x_{\ell+1}, \dots, x_d)$;
— on pose $X_{t+1} = (y_1, \dots, y_d)$.

Proposition 6.7. *La mesure π est stationnaire pour la chaîne $(X_t)_{t\in\mathbb{N}}$.*

DÉMONSTRATION. Notons que le noyau P peut s'écrire $P = Q_1 \dots Q_d$, où Q_ℓ correspond au changement de la $\ell^{\text{ième}}$ coordonnée. Pour tout $\ell \in [\![1, d]\!]$, et pour tout $y \in \Omega^d$, on a

$$\begin{aligned}
\pi Q_\ell(y) &= \sum_{x \in \Omega^d} \pi(x) Q_\ell(x, y) \\
&= \sum_{x_\ell \in \Omega} \pi(y^{(\ell)} x_\ell) Q_\ell(y^{(\ell)} x_\ell, y) \\
&= \sum_{x_\ell \in \Omega} \pi(y^{(\ell)} x_\ell) \pi(y_\ell \mid y^{(\ell)}) \\
&= \sum_{x_\ell \in \Omega} \pi(y^{(\ell)}) \pi(x_\ell \mid y^{(\ell)}) \frac{\pi(y)}{\pi(y^{(\ell)})} \\
&= \pi(y) \sum_{x_\ell \in \Omega} \pi(x_\ell \mid y^{(\ell)}) = \pi(y) \,.
\end{aligned}$$

Ainsi π est stationnaire pour chacun des noyau Q_ℓ, donc pour P. ■

Exemple 6.5. Soit (X, Y) un couple de variables aléatoires de densité sur $\mathbb{R}^2$ donnée par

$$h(x, y) = C \exp\left(-\frac{y^2}{2} - \frac{x^2(1 + y + y^2)}{2}\right) .$$

La loi conditionnelle de X sachant $Y = y$ a pour densité

$$f(x \mid y) \propto \exp\left(-\frac{x^2(1 + y + y^2)}{2}\right) .$$

Ainsi, $\mathcal{L}(X \mid Y) = \mathcal{N}\left(0, \frac{1}{1+Y+Y^2}\right)$. De même, la densité de Y sachant $X = x$ est

$$g(y \mid x) \propto \exp\left(-\frac{1 + x^2}{2}\left(y^2 + \frac{x^2 y}{1 + x^2}\right)\right) \propto \exp\left(-\frac{1 + x^2}{2}\left(y + \frac{x^2}{2(1 + x^2)}\right)^2\right) .$$

Ainsi, $\mathcal{L}(Y \mid X) = \mathcal{N}\left(-\frac{X^2}{2(1+X^2)}, \frac{1}{1+X^2}\right)$. Pour simuler selon la loi de (X, Y), l'algorithme de Gibbs consiste à considérer la chaîne de Markov suivante : on part de $(x_0, y_0) = (0, 0)$, puis à chaque temps $t \geq 0$, conditionnellement à $(X_t, Y_t) = (x_t, y_t)$, on génère (X_{t+1}, Y_{t+1}) selon :

— $X_{t+1} \sim \mathcal{N}\left(0, \frac{1}{1+y_t+y_t^2}\right)$;
— $Y_{t+1} \sim \mathcal{N}\left(-\frac{x_{t+1}^2}{2(1+x_{t+1}^2)}, \frac{1}{1+x_{t+1}^2}\right)$.

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 de la section 3.4, si l'on sait simuler suivant les lois $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \mathbf{X})$ et $\mathcal{L}(\boldsymbol{\alpha} \mid \boldsymbol{\theta}, \mathbf{X})$, l'algorithme de Gibbs permet de simuler approximativement selon $\mathcal{L}((\boldsymbol{\theta}, \boldsymbol{\alpha}) \mid \mathbf{X})$.