D-SYS-315 – Introduction à la statistique Bayésienne

F. Simatos
6 janvier 2026

Institut Supérieur de l'Aéronautique et de l'Espace

SUPAERO

D-SYS-315 – Introduction à la statistique Bayésienne

F. Simatos
6 janvier 2026

Table des matières

Bien que la connaissance débute avec l'expérimentation, il ne s'ensuit pas que la connaissance soit entièrement déduite de l'expérimentation.

Kant

On dit souvent qu'il faut expérimenter sans idée préconçue. Cela n'est pas possible ; non seulement ce serait rendre toute expérience stérile, mais on le voudrait qu'on ne le pourrait pas. Chacun porte en soi sa conception du monde dont il ne peut se défaire si aisément.

Poincaré

1.1 Généralités

On se place dans le cadre de l'estimation paramétrique classique, i.e., on a un modèle paramétrique $\{f_\theta : \theta \in \Theta\}$ avec f_θ une densité de probabilité (par défaut, sur $\mathbb{R}$ et $\Theta \subset \mathbb{R}$). On a un échantillon $x = (x_1, \cdots, x_n)$ supposé, sauf mention explicite du contraire, i.i.d. et tirée selon une "vraie" densité f_{θ_0}. Le but est de faire une "bonne" estimation de θ_0 à partir de l'échantillon x.

Dans le cadre fréquentiste classique, que vous avez par exemple vu en tronc commun de première année, cette estimation prend la valeur d'une variable aléatoire $\hat{\theta} \in \Theta$ qui vit dans le même espace que θ_0. L'estimateur le plus connu, et sur lequel nous reviendrons régulièrement dans le cours, est l'estimateur du maximum de vraisemblance qui, comme son nom l'indique, sélectionne le paramètre θ qui maximise la vraisemblance de l'échantillon :

$$\hat{\theta}_n^{\mathrm{MV}} \in \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(x_i).$$

On remarque ainsi en particulier que dans ce cadre fréquentiste, aucune estimation n'est proposée en l'absence d'observation. En pratique en revanche, face à un problème d'estimation et même en l'absence d'observation on est amenés à proposer un estimateur. C'est ce que l'on appelle la loi a priori en statistique bayésienne, qui sera $\pi(\theta)$, et qui représente le degré de confiance, l'état des connaissances que l'on a sur le paramètre à estimer. L'approche bayésienne consiste à mettre à jour cette loi a priori en fonction des observations : on obtient ainsi la loi a posteriori, notée $\pi(\theta \mid x)$, qui est la loi de θ conditionnée par les observations. On calcule cette loi à l'aide de la formule de Bayes, que l'on note

$$\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta)$$

où, dans le contexte bayésien, on note $f(x \mid \theta) = f_\theta(x)$ pour mettre en avant que f_θ est la loi d'échantillonnage conditionnée par le paramètre θ, qui est effectivement rendu aléatoire via la loi a priori π. Une différence majeure entre l'approche fréquentiste et bayésienne est donc qu'en statistique bayésienne, l'estimation du paramètre n'est pas un nombre (ou un vecteur), mais une distribution de probabilités sur l'espace des paramètres : c'est la loi a posteriori, qui résulte de la loi a priori et des observations. Nous verrons qu'il y a en fait d'autres motivations

ou justifications de cette approche, et que le choix de la loi a priori est un des enjeux majeurs de la statistique bayésienne. Pour illustrer ces concepts on commence par considérer deux exemples historiques.

1.2 Exemples

1.2.1 Exemples historiques

EXEMPLE 1.1 Le problème suivant a été posé par Bayes en 1763. Une boule de billard W roule sur une ligne de longueur un, avec une probabilité uniforme de s'arrêter n'importe où. Supposons qu'elle s'arrête en p. Une deuxième boule O roule alors n fois dans les mêmes conditions, et on note X le nombre de fois où la boule O s'arrête à la gauche de W. Avant d'avoir lancé O, quelle estimation de p pouvez-vous faire ? Comment cette estimation est-elle modifiée après avoir lancé O et connaissant X ?

On pourrait adopter un raisonnement fréquentiste en raisonnant conditionnellement à p et considérer X/n comme estimateur, qui converge par le théorème central limite. Ce raisonnement, justifié, ignore le fait que p elle-même est aléatoire et même suit une loi uniforme. On peut essayer d'exploiter cette information pour calculer la nouvelle loi de p mise à jour avec les observations : on verra que cette loi s'appelle la loi a posteriori. ■

EXEMPLE 1.2 Le problème suivant a été posé par Laplace en 1773. Une urne contient un nombre n de boules noires ou blanches. Si la première boule sortie de l'urne est blanche, quelle est la probabilité que la proportion p de boules blanches soit p_0 ?

Ici la situation est un peu différente de l'exemple précédent, puisqu'on pose explicitement une question en terme de probabilité concernant la proportion de boules blanches : là on sort forcément du cadre fréquentiste habituel, et on voit qu'il faut nécessairement faire une hypothèse de base sur la loi de p. Par exemple, Laplace a supposé que p était uniformément réparti dans $\{2/n, \dots, (n-1)/n\}$ et la règle de Bayes nous donne alors comme loi a posteriori

$$\mathbb{P}(p = p_0 \mid \text{données}) = \frac{2np_0}{n(n-1)-2}$$

ce qui représente la loi biaisée. ■

1.2.2 Exemple fil rouge : le modèle gaussien à variance connue

Dans tout le cours, on considèrera un exemple dit fil rouge qui servira à illustrer les notions introduites : il s'agit du modèle d'échantillonage gaussien à variance σ^2 connue, avec une loi a priori elle aussi gaussienne de paramètre (μ_0, σ_0^2) :

$$\pi(\theta) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-(\theta-\mu_0)^2/(2\sigma_0^2)} \tag{1.1}$$

et

$$f(x \mid \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\theta)^2\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|x-\theta\mathbf{1}_n\|^2}{2\sigma^2}\right) \tag{1.2}$$

avec $\mathbf{1}_n = (1, \dots, 1)$ le vecteur des 1 en dimension n. Lorsque l'on sera amené à manipuler des variables gaussiennes, on utilisera les deux résultats suivants.

Lemme 1.3. *Si X est un vecteur gaussien de moyenne μ et de matrice de variance-covariance Σ, alors $MX+A$ est aussi un vecteur gaussien de moyenne $M\mu+A$ et de matrice de variance-covariance $M\Sigma M^T$.*

Théorème 1.4. *Soit $X=(X_1,X_2)$ un vecteur gaussien tel que X_1 est absolument continu. Alors X_2 conditionnellement à X_1 est un vecteur gaussien de moyenne*

$$\mathbb{E}(X_2 \mid X_1) = \mathbb{E}(X_2) + \mathbb{C}\mathrm{ov}(X_2,X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1}(X_1 - \mathbb{E}(X_1))$$

et de matrice de variance-covariance

$$\Sigma_{X_1} = \mathbb{V}\mathrm{ar}(X_2) - \mathbb{C}\mathrm{ov}(X_2,X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1}\mathbb{C}\mathrm{ov}(X_1,X_2).$$

Pour x un vecteur on notera $|x| = \sum|x_i|$ sa norme L_1 et $\|x\|$ sa norme L_2, i.e., $\|x\|^2 = \sum x_i^2$. On déduit en particulier du second résultat la loi a posteriori.

Lemme 1.5. *Dans l'exemple fil rouge, la loi a posteriori est la loi*

$$\mathscr{N}\left(p\bar{x}+q\mu_0, \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right) \quad \textit{avec} \quad p = 1-q = \frac{\sigma_0^2}{\sigma^2/n+\sigma_0^2}. \tag{1.3}$$

Démonstration. On fait d'abord le calcul avec une observation. On a $\theta \sim \mathscr{N}(\mu_0,\sigma_0^2)$ et $x \mid \theta \sim \mathscr{N}(\theta,\sigma^2)$ ce que l'on peut écrire sous la forme $x = \theta+\sigma\varepsilon$ et $\theta = \mu_0+\sigma_0\delta$ avec $\varepsilon,\delta \sim \mathscr{N}(0,1)$, soit

$$\begin{cases} \theta = \mu_0+\sigma_0\delta, \\ x = \mu_0+\sigma_0\delta+\sigma\varepsilon \end{cases}$$

Donc (x,θ), en tant que transformation affine du vecteur gaussien (δ,ε), est un vecteur gaussien. Par ailleurs, x admet une densité (on est en dimension 1), et donc le théorème précédent nous assure que $\theta \mid x$ suit une loi gaussienne de moyenne

$$\mu_0 + \mathbb{C}\mathrm{ov}(\theta,x)\mathbb{V}\mathrm{ar}(x)^{-1}(x-\mu_0)$$

et de variance

$$\sigma_0^2 - \mathbb{C}\mathrm{ov}(\theta,x)^2\mathbb{V}\mathrm{ar}(x)^{-1}.$$

On calcule

$$\mathbb{V}\mathrm{ar}(x) = \mathbb{V}\mathrm{ar}(\mu_0+\sigma_0\delta+\sigma\varepsilon) = \sigma^2+\sigma_0^2$$

et

$$\mathbb{C}\mathrm{ov}(\theta,x) = \mathbb{E}(\sigma_0\delta(\sigma_0\delta+\sigma\varepsilon)) = \sigma_0^2$$

ce qui donne le résultat pour $n=1$. Pour $n \geq 1$, on se ramène au cas $n=1$ via les calculs

suivants. La vraisemblance est donnée par

$$\begin{aligned} f(x \mid \theta) &\propto_\theta \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\theta)^2\right) \\ &\propto_\theta \exp\left(-\frac{1}{2\sigma^2}\sum_i (\theta^2-2\theta x_i)\right) \\ &\propto_\theta \exp\left(-\frac{n\theta^2-2n\theta\bar{x}}{2\sigma^2}\right) \\ &\propto_\theta \exp\left(-\frac{(\bar{x}-\theta)^2}{2(\sigma^2/n)}\right) \\ &\propto_\theta f'(\bar{x} \mid \theta) \end{aligned}$$

avec $f'(\cdot \mid \theta)$ la densité gaussienne de moyenne θ et de variance σ^2/n. Donc $\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta) \propto \pi(\theta) f'(\bar{x} \mid \theta) \propto \pi'(\theta \mid \bar{x})$ avec π' définie comme π mais avec une variance σ^2/n. On obtient donc la même formule que avec une observation, mais avec σ^2/n à la place de σ^2 et $\bar{x}$ à la place de x. □

Ce résultat permet dès à présent d'illustrer plusieurs idées que l'on reverra pendant le cours :

1. la moyenne a posteriori est une combinaison linéaire de l'information a priori via μ_0 et de l'information apportée par les données *via* le maximum de vraisemblance $\bar{x}$. En outre, lorsque la taille de l'échantillon augmente, l'effet de la loi a priori s'estompe et l'estimation se base principalement sur les données ;
2. si l'on centre et que l'on met à l'échelle, la loi a posteriori suit une loi gaussienne.

1.3 Motivations et justifications

1.3.1 Théorème de De Finetti (*)

Dans l'approche bayésienne, les observations x ne sont pas i.i.d., mais ne le sont que conditionnellement au paramètre θ vu comme un paramètre aléatoire. En fait, le théorème de De Finetti montre que cette situation est caractéristique de suites dites échangeables.

Définition 1.6. Une suite $x = (x_1, x_2, \ldots)$ est dite échangeable si pour tout n et pour toute permutation $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$, $(x_1,\ldots,x_n)$ et $(x_{\sigma(1)},\cdots,x_{\sigma(n)})$ ont la même loi.

Evidemment, une suite i.i.d. est échangeable, mais la réciproque n'est pas vraie. Un contre-exemple simple est donnée par la suite $(x_0 + x_k, k \geq 1)$ avec les $(x_k, k \geq 0)$ i.i.d.. Néanmoins, le théorème de De Finetti montre qu'une suite échangeable n'est qu'un mélange de suites i.i.d.

Théorème 1.7 (De Finetti)**.** *Une suite de variables aléatoires $(x_1, x_2, \ldots)$ de densité f est échangeable si et seulement si il existe une vraisemblance $f(x \mid \theta)$ et une mesure de probabilités P sur Θ telles que pour tout n,*

$$f(x_1, \ldots, x_n) = \int f(x_1 \mid \theta) \cdots f(x_n \mid \theta) P(\mathrm{d}\theta).$$

L'exemple suivant illustre les relations entre le théorème de De Finetti et l'approche bayésienne.

EXEMPLE 1.8 Historiquement les moteurs de recherche ont utilisé des modèles de "sacs de mots" pour modéliser ces documents, dans lesquels l'ordre des mots ne compte pas. Si l'on voit un mot et qu'il est français, alors on s'attend à ce que le reste du document soit en français. Cela montre que les mots ne sont pas i.i.d., mais qu'ils le sont possiblement conditionnellement à un paramètre, par exemple ici la langue. ■

1.3.2 Justification axiomatique (*)

Principe de conditionnement

Supposons qu'un échantillon doive être analysé par un laboratoire qui fournit une mesure avec une grande précision, correspondant à une variance de 0,1. La moitié du temps, ce laboratoire est en panne et on doit donc faire analyser l'échantillon par un autre laboratoire, moins précis, correspondant à une variance de 10. Etant donné le bon laboratoire n'était pas en panne, quelle est la variance de la mesure ? La réponse fréquentiste correspond à moyenniser les variances :

$$0{,}5 \times 0{,}1 + 0{,}5 \times 10 = 5{,}05.$$

En effet, ce serait la variance si on avait répété l'expérience de nombreuses fois. Néanmoins, un bayésien prendra l'observation en compte, et estimera la variance comme valant 0,1. L'inférence fréquentiste viole donc le principe suivant.

Principe de conditionnement. *Si deux expériences E_1 et E_2 sur le paramètre θ sont possibles et si on choisit une de ces expériences au hasard avec probabilité p, alors l'inférence sur θ ne doit dépendre que de l'expérience choisie.*

Principe d'exhaustivité

Définition 1.9. T est une statistique exhaustive si, alors que $x \sim f(x \mid \theta)$, la loi de x conditionné à $T(x)$ ne dépend pas de θ.

En d'autres termes, une statistique est exhaustive si elle contient toute l'information apportée sur x par θ. L'importance théorique des statistiques exhaustives découle en grande partie du théorème de Rao–Blackwell.

Théorème 1.10 (Théorème de Rao–Blackwell)**.** *Si T est une statistique exhaustive, alors pour tout estimateur $\hat{\theta}$ l'erreur quadratique moyenne est améliorée en considérant comme estimateur l'espérance de $\hat{\theta}$ conditionnellement à T, i.e.,*

$$\mathbb{E}_\theta\left[\left(\theta - \mathbb{E}_\theta(\hat{\theta} \mid T)\right)^2\right] \leq \mathbb{E}_\theta\left[\left(\theta - \hat{\theta}\right)^2\right].$$

Démonstration. En considérant $\hat{\theta} - \theta$ on peut considérer sans perte de généralité que $\theta = 0$. On a

$$\mathbb{E}_\theta\left[\hat{\theta}\mathbb{E}_\theta(\hat{\theta} \mid T)\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta(\hat{\theta} \mid T)^2\right]$$

et donc

$$0 \leq \mathbb{E}_\theta\left[\left(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta} \mid T)\right)^2\right] = \mathbb{E}_\theta\left(\hat{\theta}^2\right) - \mathbb{E}_\theta\left[\mathbb{E}_\theta(\hat{\theta} \mid T)^2\right]$$

ce qui prouve le résultat. □

Principe d'exhaustivité. *Deux observations x et y donnant la même valeur d'une statistique exhaustive T doivent conduire à la même inférence sur θ.*

Théorème de Birnbaum et principe de vraisemblance

Principe de vraisemblance. *Si deux expériences d'inférence sur un même paramètre θ mènent à la même fonction de vraisemblance à une constante multiplicative près, alors l'inférence devrait être la même.*

Théorème 1.11 (Birnbaum)**.** *Le principe de vraisemblance est équivalent à la conjonction des principes d'exhaustivité et de conditionnement.*

On a vu sur l'exemple ci-dessus que le principe de conditionnement était violé par la statistique fréquentiste. Un autre exemple montre que le principe de vraisemblance l'est aussi. Soit θ le biais d'une pièce de monnaie, que l'on estime par deux expériences différentes :

Expérience 1 **:** on lance la pièce un nombre n de fois déterminé en avance, et on observe le nombre X de succès ;

Expérience 2 **:** on lance la pièce un nombre N de fois, aléatoire, jusqu'à observer un nombre $x \geq 1$ de succès, déterminé en avance.

La vraisemblance de ces expériences est donnée par

$$\ell_1(\theta \mid x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \text{ et } \ell_2(\theta \mid n) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}.$$

Pour la deuxième probabilité, notons N_k pour $k = 1, \ldots, n$ le nombre d'échecs entre le $(k-1)$-ième et k-ième succès : alors

$$\mathbb{P}_\theta(N_1 = n_1, \ldots, N_n = n_n) = \theta^n(1-\theta)^{n_1+\cdots+n_n}$$

donc toutes les séries de longueur n avec x succès sont équiprobables, et donc

$$\ell_2(\theta \mid n) = \#\text{séries} \times \theta^x(1-\theta)^{n-x} = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}.$$

Si ces expériences ont la même vraisemblance, leur inférence peut néanmoins varier, tout du moins lorsque l'on fait de l'estimation par intervalle de confiance ou des tests. En effet, l'estimation ponctuelle par maximum de vraisemblance est nécessairement la même dans les deux cas, c'est la moyenne empirique. Par contre, considérons le test $H_0 : \theta \leq 0{,}5$ et $H_1 : \theta > 0{,}5$. Dans les deux cas, les tests sont de la forme

$$\text{Accepter } H_0 \Longleftrightarrow \frac{x}{n} > c_i$$

et plus précisément, pour l'expérience 1,

$$\text{Accepter } H_0 \Longleftrightarrow X > nc_1$$

avec n fixe, et pour l'expérience 2

$$\text{Accepter } H_0 \Longleftrightarrow N < x/c_2$$

avec x fixe. Néanmoins, le seuil c_i va dépendre de l'expérience. Pour l'expérience 1, ce seuil satisfait (à un risque de première espèce α fixé)

$$\mathbb{P}_{1/2}(\bar{x} > c_1) = \mathbb{P}(\mathscr{B}in(n, 1/2) > nc_1) = \sum_{k=nc_1}^{n} \binom{n}{k} 2^{-n} \approx \alpha.$$

Pour l'expérience 2, c_2 est fixé par

$$\mathbb{P}_{1/2}(n < x/c_2) = \sum_{m=x}^{x/c_2} \binom{m-1}{x-1} 2^{-m} \approx \alpha.$$

En général les résolutions de ces deux équations amènent à des solutions différentes, et donc l'inférence sera potentiellement différente pour deux expériences avec des vraisemblances identiques (à une constante multiplicative près). Attention : cela ne veut pas dire que l'inférence fréquentiste ne satisfait jamais le principe de vraisemblance : par exemple, l'estimation ponctuelle par maximisation de la vraisemblance vérifie, par construction, le principe de vraisemblance. L'exemple ci-dessus n'est qu'un contre-exemple particulier. Comme on le verra, l'approche bayésienne quant à elle satisfait ces trois principes puisque l'inférence est entièrement faite à partir de la fonction de vraisemblance (et de la loi a priori).

2.1 Entropie

Définition 2.1. Etant donné une distribution f, on définit son entropie différentielle

$$H(f) = -\int f(x)\log f(x)\mathrm{d}x \in [-\infty, +\infty].$$

Si f est la densité de la loi uniforme sur $[0,a]$, on voit que

$$H(f) = \int_0^a \frac{1}{a}\log a\mathrm{d}x = \log a,$$

ce qui montre que l'entropie différentielle peut être positive ou négative. Si on considère

$$f(x) = \frac{c}{x(\log x)^2}\mathbb{1}\left(x \geq x_0\right)$$

avec x_0 assez grand et c la constante de normalisation, on a

$$H(f) = -\log c + c\int_{x_0}^{\infty}\frac{1}{x\log x}\mathrm{d}x + c\int_{x_0}^{\infty}\frac{\log(\log x)}{x(\log x)^2}\mathrm{d}x$$

qui vaut $+\infty$, puisque $1/(x\log x)$ n'est pas intégrable. On peut aussi construire une densité avec une entropie différentielle négative, mais c'est beaucoup plus compliqué (en tout cas, je ne connais pas d'exemple simple). Cf. par exemple l'exemple construit la page web accessible via ce lien [1]. A noter que cet exemple fournit un exemple de densité sur $[0,1]$ avec une entropie qui vaut $-\infty$, alors que le Lemme 3.1 ci-dessous implique que l'entropie est nécessairement bornée supérieurement sur $[0,1]$.

Pour une loi discrète, l'entropie est (un peu) mieux définie car

$$h(p) = -\sum_x p(x)\log p(x) \in [0,\infty].$$

Notamment, elle est positive, mais l'exemple construit en continu est facilement adaptable pour construire une loi discrète d'entropie infinie.

1. L'adresse url explicite : https://stats.stackexchange.com/questions/354643/can-the-differential-entropy-be-negative-infinity

L'entropie est une mesure de la quantité d'information contenue dans une loi de probabilité : une entropie (discrète) faible correspond à une loi de probabilité très resserrée et donc avec beaucoup d'information. En effet, on a par exemple $h(p) = 0$ si et seulement si p est une mesure de Dirac.

L'entropie différentielle s'approche par l'entropie discrète via une constante additive. En effet, soit f une densité et $\varepsilon > 0$. Pour chaque $i \in \mathbb{Z}$, il existe x_k avec $f(x_k) = \frac{1}{\varepsilon}\int_{k\varepsilon}^{(k+1)\varepsilon} f$ par le théorème des valeurs intermédiaires. On considère alors la loi de probabilité p_ε sur $\{x_k\}$ définie par $p_\varepsilon(x_k) = \int_{k\varepsilon}^{(k+1)\varepsilon} f = \varepsilon f(x_i)$. On a alors

$$h(p_\varepsilon) = -\sum_k \varepsilon f(x_i)\log(\varepsilon f(x_i)) = -\varepsilon\sum_i f(x_i)\log f(x_i) - \log\varepsilon$$

Si $f\log f$ est Riemann intégrable, le premier terme converge vers $H(f)$, alors que le second terme diverge. Cela montre que l'entropie d'une discrétisation à n-bits d'une variable continue X est $h(X)+n$, cf. [2] pour plus de détails.

2.2 Divergence de Kullback–Leibler et entropie croisée

Définition 2.2. La divergence de Kullback–Leibler entre deux densités de probabilité f et g est donnée par

$$D(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x.$$

Par convention, on a $0/0 = 0$ et $D(f,g) = +\infty$ si l'ensemble $\{x : f(x) \neq 0, g(x) = 0\}$ n'est pas de mesure nulle.

Proposition 2.3. *$D(f,g) \geq 0$ avec égalité si et seulement si $f = g$ (p.p.).*

Démonstration. C'est l'inégalité de Jensen : pour toute variable aléatoire X et fonction φ convexe, on a sous réserve d'intégrabilité $\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$ avec égalité si et seulement si φ est linéaire ou X est une constante. Donc si $X \sim f$ on a $D(f,g) = \mathbb{E}(-\log(g(X)/f(X)))$ et puisque $-\log$ est convexe, l'inégalité de Jensen donne $D(f,g) \geq -\log\mathbb{E}(g(X)/f(X)) = 0$ avec égalité, puisque log n'est pas linéaire, si et seulement si $f(X)/g(X)$ est une constante, i.e., si $f = g$ (f-p.p.). □

Définition 2.4. L'entropie croisée entre f et g est définie par

$$\mathrm{CE}(f,g) = -\int f(x)\log g(x)\mathrm{d}x.$$

Ainsi, $\mathrm{CE}(f,g) = D(f,g) + H(f)$ et donc minimiser l'entropie croisée revient à minimiser la divergence de Kullback–Leibler, ce qui par le résultat précédent est obtenu pour $g = f$, i.e.,

$$f = \arg\min_g \mathrm{CE}(f,g) = \arg\min_g D(f,g). \tag{2.1}$$

2.3 Score et information de Fisher

Définition 2.5. Soit $\{f_\theta : \theta \in \Theta\}$ un modèle paramétrique avec $\theta \mapsto f_\theta(x)$ dérivable pour tout x et dans $L_2(f_\theta)$. On définit le score de ce modèle paramétrique par

$$\ell(x \mid \theta) = \partial_\theta \log f(x \mid \theta) = \frac{\partial_\theta f(x \mid \theta)}{f(x \mid \theta)}$$

et l'information de Fisher par

$$I(\theta) = \mathbb{E}_\theta(\partial_\theta \log f_\theta(x)^2) = \int \partial_\theta \log f_\theta(x)^2 f_\theta(x) \mathrm{d}x.$$

Lemme 2.6. *Lorsque le modèle est régulier, on a* $I(\theta) = -\mathbb{E}_\theta \partial_\theta^2 \log f_\theta(x_1) = \mathbb{E}(\partial_\theta \log f_\theta(x)^2)$.

Démonstration. En effet,

$$\begin{aligned} -\mathbb{E}_\theta \partial_\theta^2 \log f_\theta(x) &= -\int \partial_\theta^2 \log f_\theta(x) f_\theta(x) \mathrm{d}x \\ &= -\int \left(\frac{\partial_\theta^2 f_\theta(x)}{f_\theta(x)} - \left(\frac{\partial_\theta f_\theta(x)}{f_\theta(x)} \right)^2 \right) f_\theta(x) \mathrm{d}x \\ &= -\int \partial_\theta^2 f_\theta(x) \mathrm{d}x + I(\theta) \end{aligned}$$

et $\int \partial_\theta^2 f_\theta(x) \mathrm{d}x = \partial_\theta^2 \int f_\theta(x) \mathrm{d}x = 0$. □

Le score peut être interprété comme une mesure de la vitesse à laquelle la densité change lorsque l'on change légèrement le paramètre θ. L'information de Fisher, qui n'est autre que la moyenne du score au carré, permet donc d'obtenir une version moyennisée de cette mesure. Donc si l'information de Fisher est grande, cela veut dire que la distribution va changer rapidement lorsque l'on bouge le paramètre, et que donc la distribution avec le paramètre θ_0 va être "bien différente" et donc peut être "bien distinguée" des distributions avec des paramètres relativement éloignés de θ_0. Cela veut dire que l'on devrait être capable d'estimer θ_0 correctement à partir des données. D'un autre côté, si l'information de Fisher est petite, cela veut dire que la distribution est très similaire à d'autres distributions qui ont des paramètres distincts et donc qu'il sera plus difficile de discriminer entre les deux, et donc l'estimation sera moins bonne. On retrouve cette interprétation dans la borne de Cramer-Rao, qui dit que si l'information de Fischer est petite, alors la variance d'un estimateur non-biaisé est forcément grande.

Dans la suite on aura besoin du résultat technique suivant, qui explicite l'information de Fisher lorsque l'on change de variable. Pour cela on considère $h : \Theta \to \Theta'$ un $\mathscr{C}^1$-difféomorphisme et $I_h : \Theta' \to \mathbb{R}$ l'information de Fisher associé au modèle paramétrique $\{f_{h^{-1}(\theta')} : \theta' \in \Theta'\}$, ce qui correspond à faire le changement de variable $\theta' = h(\theta)$.

Proposition 2.7. *On a* $I(\theta) = I_h(h(\theta))(h'(\theta))^2$.

Démonstration. Par définition, $I_h(\theta')$ est l'information de Fisher associé au modèle d'échantillonnage $g(x \mid \theta') = f(x \mid h^{-1}(\theta'))$ avec $\theta' \in h(\Theta)$. Le vecteur du score associé est

$$\ell_h(x \mid \theta') = \ell(x \mid h^{-1}(\theta'))(h^{-1})'(\theta') = \frac{\ell(x \mid h^{-1}(\theta'))}{h' \circ h^{-1}(\theta')}$$

et donc pour $\theta' = h(\theta)$, on a

$$\begin{aligned} I_h(h(\theta)) &= \int [\ell_h(x \mid \theta')]^2 g(x \mid \theta') \mathrm{d}x \\ &= \int \left(\frac{\ell(x \mid h^{-1}(\theta'))}{h' \circ h^{-1}(\theta')} \right)^2 f(x \mid h^{-1}(\theta')) \mathrm{d}x \\ &= \frac{1}{h'(\theta)^2} I(\theta) \end{aligned}$$

ce qui prouve le résultat. □

3.1 Approximations paramétriques

Consiste à restreindre π à une certaine famille paramétrique et à calibrer les paramètres en fonction des informations disponibles type moyenne/médiane. Une autre manière de "calibrer" est de rajouter une couche d'aléa et supposer que les hyperparamètres sont eux-mêmes tirés selon une distribution de probabilité : on parle alors de modèles hiérarchiques, qui ne seront pas abordés dans ce cours, cf. par exemple [8, Chapitre 10].

3.2 Maximum d'entropie

3.2.1 Idée générale

Il arrive que l'on ait de l'information sur la loi a priori sous la forme de moments $\int g(\theta)\pi(\theta)\mathrm{d}\theta$. Dans ce cas, on peut utiliser la méthode du maximum d'entropie, qui repose sur le principe d'incertitude de Laplace, qui postule d'utiliser la loi uniforme lorsqu'aucune information n'est disponible. Puisqu'une loi avec une entropie élevée contient peu d'information, le résultat suivant fournit une justification de ce principe.

Lemme 3.1. *Soit A un ensemble de mesure de Lebesgue $m(A)$ finie. Alors la mesure uniforme sur A est l'unique solution du problème d'optimisation*

$$\begin{aligned} &\arg\max H(f) \\ &\text{t.q. } f \in \mathscr{P} \end{aligned}$$

Démonstration. On a

$$H(f) = \mathrm{CE}(f,u) - D(f,u) \leq \mathrm{CE}(f,u) = -\int_A f(x)\log 1/\lambda(A)\mathrm{d}x = \log\lambda(A)$$

et puisque $\log\lambda(A) = H(u)$ on obtient bien le résultat. □

Ainsi, lorsque de l'information sur la loi a priori est connue sous la forme de moments, il

est naturel de considérer le problème suivant :

$$\begin{array}{l} \arg\max H(f) \\ \text{t.q. } f \in \mathscr{P} \\ \text{et } \int g_k(x) f(x) \mathrm{d}x = 0, k = 1, \ldots, m, \end{array} \tag{3.1}$$

avec des fonctions $g_k : \mathbb{R}^d \to \mathbb{R}$. Les contraintes de ce problème d'optimisation reflètent alors la connaissance que l'on a du problème.

3.2.2 Cas d'un espace borné

On se restreint dans un premier temps dans le cas où l'espace Θ des paramètres (sur lequel sont définies les distributions $f \in \mathscr{P}$ que l'on cherche) est de mesure de Lebesgue finie (ou, dans le cas discret, fini). Dans ce cas, on a le résultat suivant, formulé dans $\mathbb{R}$ et qui reste probablement vrai dans $\mathbb{R}^d$.

Lemme 3.2. *Soit $A \subset \mathbb{R}$ un ensemble de mesure de Lebesgue $m(A)$ finie. Alors la mesure*

$$\pi^*(x) \propto \exp\left(\sum_{k=1}^{m} \lambda_k g_k(x)\right).$$

sur A est l'unique solution du problème d'optimisation (3.1).

Démonstration. Dans le cas discret et fini, on peut prouver ce résultat à l'aide des multiplicateurs de Lagrange. En effet, on cherche $p = (p_1, \ldots, p_n)$ et

$$L(p, \lambda) = \sum_{i=1}^{n} p_i \log(p_i / p_0(i)) + \sum_{k=1}^{m} \sum_{i=1}^{n} \lambda_k g_k(i) p_i$$

et donc

$$\partial_{p_i} L = \log(p_i / p_0(i)) + 1 + \sum_{k=1}^{m} \lambda_k g_k(i)$$

et donc

$$\partial_{p_i} L = 0 \Longleftrightarrow p_i = p_0(i) \exp\left(-1 - \sum_{k=1}^{m} \lambda_k g_k(i)\right).$$

Dans le cas continu, cf. par exemple [7]. □

REMARQUE 3.3 Un problème majeur de cette approche dans le continu est qu'elle n'est pas stable par reparamétrisation : si θ suit la loi uniforme sur $[0,1]$, alors θ^2 est de densité $1/(2\sqrt{t})$ et ne suit en particulier pas une loi uniforme. Ce problème n'est pas présent en discret, mais d'une certaine manière parce que l'espace d'état change. ■

On peut aussi prendre une mesure de référence f_0 et considérer plutôt le problème

$$\begin{array}{l} \arg\min D(f||f_0) \\ \text{t.q. } f \in \mathscr{P} \\ \text{et } \int g_k(x) f(x) \mathrm{d}x = 0, k = 1, \ldots, m \end{array} . \tag{3.2}$$

On a alors la généralisation suivante du Lemme 3.2.

Lemme 3.4. *Soit $A \subset \mathbb{R}$ un ensemble de mesure de Lebesgue $m(A)$ finie. Alors la mesure*

$$\pi^*(x) \propto f_0(c) \exp\left(\sum_{k=1}^{m} \lambda_k g_k(x)\right).$$

sur A est l'unique solution du problème d'optimisation (3.2).

3.2.3 Cas d'un espace non-borné

Dans le cas d'un espace non-borné (e.g., $\Theta = [0,\infty)$ ou $\Theta = \mathbb{R}$), certaines complications techniques apparaissent.

Déjà, comme il existe f définie sur $[a,\infty)$ avec $H(f) = \infty$, on voit que les solutions au problème $\arg\max H(f)$ perdent leur pertinence sur $\mathbb{R}$ et même sur $\mathbb{R}_+$: il existe une infinité de solutions, et toutes ces solutions ont une entropie infinie. Les résultats suivants montrent que les solutions au problème du maximum d'entropie restent pertinentes sur $\mathbb{R}_+$ si on rajoute une contrainte de premier moment, et sur $\mathbb{R}$ si on rajoute une contrainte de second moment.

Lemme 3.5. *Soit $\mu > 0$. Le problème $\arg\max H(f)$ sous la contrainte $\int x f(x)\mathrm{d}x = \mu$ a une unique solution parmi les densités f sur $\mathbb{R}_+$, à savoir la loi exponentielle de paramètre λ.*

Démonstration. Cf. [1, Theorem 3.3]. □

Lemme 3.6. *Soit $\sigma > 0$. Le problème $\arg\max H(f)$ sous la contrainte $\int x^2 f(x)\mathrm{d}x - (\int x f(x)\mathrm{d}x)^2 = \sigma^2$ a une unique solution parmi les densités f sur $\mathbb{R}$, à savoir la loi normale de variance σ^2.*

Démonstration. Cf. [1, Theorem 3.2]. □

REMARQUE 3.7 Il est facile de voir que l'entropie est invariante par translation. En particulier, toutes les lois gaussiennes avec la même variance ont la même entropie, ce qui explique pourquoi la moyenne de la loi gaussienne qui maximise l'entropie n'est pas spécifiée dans le résultat précédent. ■

Ces résultats sont cohérents avec le Lemme 3.2, dans le sens où ils peuvent s'obtenir en passant à la limite à partir de la densité maximisant l'entropie sur un intervalle $[-a,b]$ et en faisant tendre b et/ou a vers l'infini. Néanmoins, ils ne répondent pas à toutes les questions, par exemple la loi maximisant l'entropie sur $\mathbb{R}_+$ avec une variance connue, ou bien la loi maximisant l'entropie sur $\mathbb{R}$ avec une moyenne connue.

On notera quand même que toute distribution avec un premier moment fini a une entropie finie (cf. par exemple [6]). Néanmoins, l'entropie n'est pas bornée, i.e., pour toute densité f sur $\mathbb{R}$ de moyenne finie on a $H(f) < \infty$ mais $\sup H(f) = \infty$ où le supremum est pris sur toutes les densités sur $\mathbb{R}$ de moyenne finie. En particulier, le problème du maximum d'entropie sur $\mathbb{R}$ sous contrainte de premier moment n'a pas de solution. En fait, le lemme suivant montre entre autre que le supremum reste infini même si on se restreint aux distributions de moyenne μ fixée (considérer la loi de Laplace de moyenne μ, puis faire tendre $\lambda \to \infty$). Ainsi, le principle

du maximum d'entropie devient pertinent sur $\mathbb{R}_+$ si une condition de premier moment est imposée, et sur $\mathbb{R}$ si une condition de second moment est imposée.

Le résultat principal du lemme suivant est d'expliciter la densité d'entropie maximale sous contrainte de premier moment, et de moment centré.

Lemme 3.8. *Soit $\lambda > 0$ et $\mu \in \mathbb{R}$. Pour toute distribution f de moyenne μ et de moment centré λ, i.e., pour toute distribution f avec*

$$\int_{\mathbb{R}} x f(x)\mathrm{d}x = \mu \quad et \quad \int_{\mathbb{R}} |x-\mu|\, f(x)\mathrm{d}x = \lambda,$$

on a $H(f) \leq 1+\log(2\lambda)$, avec égalité si et seulement si f est la densité de Laplace de moyenne μ et de variance λ^2, i.e., $f(x) = 1/(2\lambda)e^{-|x-\mu|/\lambda}$.

Démonstration. Cf. [1, Theorem 5.3]. □

3.3 Lois conjuguées et familles exponentielles

3.3.1 Résultats généraux

Définition 3.9. La famille $\{f(\cdot \mid \theta)\}$ est dite exponentielle de dimension k si elle peut s'écrire

$$f(x \mid \theta) = C(\theta)h(x)\exp\left(R(\theta)\cdot T(x)\right)$$

avec R et T à valeurs dans $\mathbb{R}^k$, et $\cdot$ le produit scalaire. Si $R = T = \mathrm{Id}$ on parle de famille naturelle.

A noter que $C(\theta)$ peut être vue comme une constante de normalisation. Par ailleurs, un changement de variable de x en $z = T(x)$ et une reparamétrisation $\eta = R(\theta)$ amène à la forme naturelle.

EXEMPLE 3.10 La loi exponentielle $f(x \mid \theta) = \theta e^{-\theta x}\mathbb{1}(x > 0)$ constitue une famille exponentielle avec $C(\theta) = \theta$, $h(x) = \mathbb{1}(x > 0)$, $R(\theta) = -\theta$ et $T(x) = x$. Ainsi, elle est presque naturelle. ■

EXEMPLE 3.11 Le modèle gaussien à variance connue $f(x \mid \theta) = (2\pi\sigma^2)^{-1/2}e^{-(x-\theta)^2/(2\sigma^2)}$ constitue une famille exponentielle avec $C(\theta) = (2\pi\sigma^2)^{-1/2}e^{-\theta^2/(2\sigma^2)}$, $h(x) = e^{-x^2/(2\sigma^2)}$, $T(x) = x$ et $R(\theta) = -\theta^2/(2\sigma^2)$. Ainsi, on obtient la forme naturelle en paramétrant par $\eta = -\theta^2/(2\sigma^2)$. ■

EXEMPLE 3.12 Le modèle gaussien à moyenne connue $f(x \mid \theta) = (2\pi\theta)^{-1/2}e^{-(x-\mu)^2/(2\theta)}$ constitue une famille exponentielle avec $C(\theta) = (2\pi\theta)^{-1/2}$, $h(x) = 1$, $T(x) = -(x-\mu)^2/2$ et $R(\theta) = 1/\theta$. Ainsi, on obtient la forme naturelle en paramétrant par $\eta = 1/\theta$ et $z = (x-\mu)^2/2$. ■

EXEMPLE 3.13 Le modèle gaussien à moyenne et variance inconnues. ■

EXEMPLE 3.14 La loi de Poisson : $h(x) = 1/x!$, $C(\theta) = e^{-\theta}$, $T(x) = x$ et $R(\theta) = \log\theta$. ■

EXEMPLE 3.15 La loi de Dirichlet, et donc la loi beta et donc la loi uniforme : la densité est $\frac{\prod_{i=1}^n \Gamma(\theta_i)}{\Gamma(|\theta|)} \prod_{i=1}^n x_i^{\theta_i - 1}$, ce qui correspond à $T(x) = (\log x_i)$, $R(\theta) = \theta - \mathbf{1}_n$. ■

REMARQUE 3.16 T est une statistique exhaustive si, alors que $x \sim f(x \mid \theta)$, la loi de x conditionné à $T(x)$ ne dépend pas de θ. Le théorème de factorisation garantit que si T est une statistique exhaustive, alors $f(x \mid \theta)$ peut s'écrire $f(x \mid \theta) = g(T(x) \mid \theta)h(x \mid T(x))$ avec g la densité de T. Le Lemme de Pitman-Koopman dit par ailleurs que si une famille de lois $f(\cdot \mid \theta)$ à support constant est telle que, à partir d'une taille d'échantillon suffisamment grande, il existe une statistique exhaustive de taille fixe, alors la famille est exponentielle. ■

Définition 3.17. Une famille de lois a priori $\{\pi_\alpha\}$ est dite conjuguée pour le modèle paramétrique $\{f_\theta\}$ si pour tout α, la loi a posteriori partant de la loi a priori π_α fait partie de la même famille : $\pi_\alpha(\cdot \mid x) = \pi_\beta$ pour un certain $\beta = \beta(x)$.

Proposition 3.18. *Soit $f(x \mid \theta) = h(x)e^{\theta \cdot x - \psi(\theta)}$ la loi générique d'une famille exponentielle. Une famille conjuguée pour $f(x \mid \theta)$ est donnée par*

$$\pi(\theta \mid \mu, \lambda) \propto e^{\theta \cdot \mu - \lambda \psi(\theta)}.$$

La loi a posteriori correspondante est $\pi(\theta \mid \mu + x, \lambda + 1)$.

Démonstration. On a

$$\pi(\theta \mid x) \propto e^{\theta \cdot x - \psi(\theta) + \theta \cdot \mu - \lambda \psi(\theta)} \propto \pi(\theta \mid \mu + x, \lambda + 1).$$

□

Il existe des familles conjuguées en dehors des familles exponentielles comme le montre le résultat suivant.

Lemme 3.19. *La loi de Pareto $\mathscr{P}a(\alpha, \theta)$ est la loi sur $\mathbb{R}$ de densité donnée par $\alpha\theta^\alpha x^{-\alpha-1}\mathbb{1}(x \geq \theta)$. La famille de lois a priori $\pi(\theta \mid a, b) \propto \theta^{a+1}\mathbb{1}(0 \leq \theta \leq b)$ est conjuguée par les lois de Pareto.*

Démonstration. Si $\pi = \pi(\cdot \mid a, b)$ et $f(x \mid \theta) \sim \mathscr{P}a(\alpha, \theta)$, alors

$$\pi(\theta \mid x) \propto \theta^{a+1}\mathbb{1}(\theta \leq b) \times \theta^\alpha \mathbb{1}(\theta \leq x) \propto \pi(\theta \mid a + \alpha, \min(b, x)).$$

□

Lemme 3.20. *Soit $\mathscr{F}$ la famille conjuguée naturelle d'une famille exponentielle. Alors l'ensemble des mélanges de N lois conjuguées*

$$\hat{\mathscr{F}}_N = \left\{ \sum_{i=1}^{N} \omega_i \pi(\theta \mid \mu_i, \lambda_i) : \omega \geq 0, |\omega| = 1 \right\}$$

est aussi une famille conjuguée. De plus, si

$$\pi(\theta) = \sum_{i=1}^{N} \omega_i \pi(\theta \mid \mu_i, \lambda_i)$$

alors la loi a posteriori est un mélange

$$\pi(\theta \mid x) = \sum_{i=1}^{N} \omega_i'(x) \pi(\theta \mid \mu_i + x, \lambda_i + 1)$$

avec

$$\omega_i'(x) = \frac{\omega_i K(\mu_i, \lambda_i)/K(\mu_i + x, \lambda_i + 1)}{\sum_{j=1}^{N} \omega_j K(\mu_j, \lambda_j)/K(\mu_j + x, \lambda_j + 1)}$$

avec K la constante de normalisation de $\pi(\theta \mid \lambda, \mu)$, i.e., $\pi(\theta \mid \lambda, \mu) = K(\mu, \lambda) e^{\theta \cdot \mu - \lambda \psi(\theta)}$.

Démonstration. Soit

$$\begin{aligned} f_i(x) &= \int \pi(\theta \mid \lambda_i, \mu_i) f(x \mid \theta) \mathrm{d}\theta \\ &= K(\lambda_i, \mu_i) \int e^{\theta \mu_i - \lambda_i \psi(\theta)} h(x) e^{\theta \mu - \lambda \psi(\theta)} \mathrm{d}\theta \\ &= K(\lambda_i, \mu_i)/K(\lambda_i + 1, \mu_i + x) \end{aligned}$$

de telle sorte que

$$\pi(\theta \mid \lambda_i, \mu_i) f(x \mid \theta) = f_i(x) \pi_i(\theta \mid \lambda_i + 1, \mu_i + x).$$

On a donc

$$\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta) = \sum_i \omega_i \pi(\theta \mid \lambda_i, \mu_i) f(x \mid \theta) = \sum_i \omega_i f_i(x) \pi_i(\theta \mid \lambda_i + 1, \mu_i + x)$$

ce qui donne le résultat avec $\omega_i'(x) \propto_i \omega_i f_i(x)$. □

L'intérêt de ce résultat est que les combinaisons linéaires sont denses dans l'ensemble des distributions : en théorie, on peut donc approcher n'importe quelle loi a priori par un mélange de lois conjuguées, cf. par exemple [8, Théorème 3.24]. On étudie maintenant plus en détail le cas gaussien.

3.3.2 Modèle gaussien à variance connue

Il s'agit de l'exemple du fil rouge. En fait, le Lemme 1.5 nous dit que pour le modèle paramétrique

$$f(x \mid \theta) \propto_x \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right),$$

la famille de lois normales forme une famille conjuguée.

3.3.3 Modèle gaussien à moyenne connue

On considère maintenant le cas où la moyenne μ est connue, et l'on cherche à estimer la variance $\theta = \sigma^2$. Le modèle paramétrique est alors

$$f(x \mid \theta) = \frac{1}{\sqrt{(2\pi\theta)^n}} \exp\left(-\frac{1}{2\theta} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

Définition 3.21. La loi Gamma de paramètre $\alpha, \beta > 0$, notée $\mathscr{G}(\alpha, \beta)$, est la loi sur $\mathbb{R}$ de densité

$$g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}(x > 0).$$

Pour $\alpha = 1$ on retrouve donc la loi exponentielle.

Définition 3.22. La loi Inverse Gamma de paramètre $\alpha, \beta > 0$, notée $\mathscr{IG}(\alpha, \beta)$, est la loi de $1/Z$ avec $Z \sim \mathscr{G}(\alpha, \beta)$.

En utilisant la formule de changement de variable

$$f_{\varphi(X)}(y) = f_X(\varphi^{-1}(y)) \left| (\varphi^{-1})'(y) \right|$$

avec $\varphi(x) = \varphi^{-1}(x) = 1/x$, on obtient donc que $\mathscr{IG}(\alpha, \beta)$ est la loi de densité

$$g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} \mathbb{1}(x > 0)$$

Lemme 3.23. *Les lois inverse gamma forment une famille de lois conjuguées pour le modèle paramétrique gaussien à moyenne connue.*

Démonstration. Si $\pi \sim \mathscr{IG}(\alpha, \beta)$, on a alors

$$\pi(\theta \mid x) \propto \frac{1}{\theta^{n/2+\alpha+1}} \exp\left(-\frac{1}{2\theta} \sum_{i=1}^{n} (x_i - \mu)^2 - \beta/\theta \right)$$

et l'on reconnaît la loi $\mathscr{IG}(\alpha + n/2, \beta + \sum (x_i - \mu)^2/2)$ □

Lemme 3.24. *Les lois gamma forment une famille de lois conjuguées pour le modèle paramétrique gaussien à moyenne connue et paramétrée par la précision $\tau = 1/\sigma^2$.*

3.4 Lois a priori impropres

Les lois a priori impropres apparaissent naturellement lorsque l'on considère des modèles paramétriques invariants, par exemple par translation. C'est le cas notamment de l'exemple fil rouge, pour lequel

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)} = f(x-\theta \mid 0)$$

Dans ce cas, il est naturel de chercher une loi a priori π qui respecte cette invariance, i.e., telle que $\pi(\theta - a) = \pi(\theta)$ pour tout $a \in \mathbb{R}$. La solution de cette équation est bien évidemment la fonction constante $\pi(\theta) = \pi(0)$, qui présente le problème de ne pas être intégrable (sauf pour $\pi \equiv 0$) : on dit que π est une loi a priori **impropre**. Ainsi, une loi a priori impropre ne peut pas être interprétée comme une distribution de probabilité.

Néanmoins, le fait que π ne soit pas intégrable ne présente pas forcément de problème puisque cela n'empêche pas forcément la distribution a posteriori d'être bien définie. En effet, pour que la mesure a posteriori $\pi(\theta \mid x)$ puisse être proportionnelle à $\pi(\theta) f(x \mid \theta)$, tout ce qu'il faut est que cette fonction soit intégrable, i.e., $\int \pi(\theta) f(x \mid \theta) \mathrm{d}\theta < \infty$. Si c'est le cas, on peut alors effectivement définir

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{\int \pi(\theta) f(x \mid \theta)}.$$

C'est bien le cas dans l'exemple du fil rouge, puisque l'on a

$$\int \pi(\theta) f(x \mid \theta) \mathrm{d}\theta = \pi(0) \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)} \mathrm{d}\theta = \pi(0)$$

ce qui donne donc

$$\pi(\theta \mid x) \propto f(x \mid \theta).$$

On remarque en particulier que la valeur de la constante n'a pas d'importance.

Il existe de nombreuses raisons de considérer des lois impropres (cf. [8, Chapitre 1.5]), parmi lesquelles :

1. dans le cas précédent, la mesure de Lebesgue est une généralisation naturelle de la loi uniforme, justifiée comme on l'a vu par la méthode du maximum d'entropie et défendue par Laplace dans son principe de la raison insuffisante. Dit autrement, en l'absence de toute information il est naturel de vouloir considérer la mesure uniforme sur $\mathbb{R}$, même si celle-ci n'est pas intégrable ;
2. dans de nombreux cas, les mesures impropres apparaissent comme limite de mesures propres. Elles peuvent donc être interprétées comme un cas extrême où la précision de l'information a priori a complètement disparu. Dans l'exemple du fil rouge, on peut ainsi voir la mesure de Lebesgue comme limite d'un loi normale centrée et de variance $\sigma^2 \to \infty$;
3. comme on le verra plus loin, elles apparaissent naturellement dans le cadre des lois non-informatives telles que la loi de Jeffreys.

Un autre exemple de loi a priori impropre qui apparaît par des considérations d'invariance est la mesure $\pi(\sigma) = 1/\sigma$ sur $\mathbb{R}_+$, motivée par des lois qui satisfont une relation d'invariance d'échelle

$$f(x \mid \theta) = \frac{1}{\theta} f\left(\frac{x}{\theta} \mid 1\right). \tag{3.3}$$

C'est par exemple le cas du modèle gaussien centré à variance inconnue, où

$$f(x \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)} = \frac{1}{\sigma} f\left(\frac{x}{\sigma} \mid 1\right).$$

Comme dans le cas de l'invariance par translation, la relation d'invariance d'échelle (3.3) suggère de prendre une loi a priori qui satisfait l'équation fonctionnelle

$$\pi(\theta) = \frac{1}{A}\pi\left(\frac{\theta}{A}\right)$$

dont la solution est donnée par $\pi(\theta) = A/\theta$.

EXEMPLE 3.25 Dans le cas du modèle gaussien centré, cela donne alors

$$\pi(\theta \mid x) \propto \frac{1}{\theta} \times \frac{1}{\theta^{n/2}} e^{-\|x\|^2/(2\theta)}.$$

Le symbole de proportionnalité est alors bien définie puisque pour tout $n \geq 1$ on a bien

$$\int \frac{1}{\theta^{n/2+1}} e^{-\|x\|^2/(2\theta)} \mathrm{d}\theta = \int \eta^{n/2-1} e^{-\eta\|x\|^2/2} \mathrm{d}\eta < \infty.$$

■

Dans l'exemple ci-dessus, dans le cas $n = 1$ l'intégrabilité n'est en fait obtenue que pour $x \neq 0$, mais comme cet évènement est de mesure nulle il semble raisonnable de le négliger. Dans le cas discret l'exemple de la loi binomiale montre que la situation peut être plus subtile.

3.5 Zellner

TODO

3.6 Lois a priori non-informatives

3.6.1 Lois invariantes et a priori de Laplace

Lorsqu'aucune information n'est disponible sur la loi a priori, il est raisonnable de la choisir à partir de la seule information disponible, à savoir le modèle d'échantillonnage $\{f(\cdot \mid \theta) : \theta \in \Theta\}$: on parle dans ce cas de lois a priori *non-informatives*. Les lois impropres découlant de principes d'invariance rencontrées dans la section 3.4 rentrent elles aussi dans le cadre des lois non-informatives, puisqu'elles n'ont été choisies qu'à partir d'information disponible sur le modèle d'échantillonnage.

La loi uniforme préconisée par Laplace, ou la mesure de Lebesgue dans le cas d'un espace non compact, peut elle aussi être considérée comme une loi non-informative. Néanmoins, une de ses limitations majeures, déjà évoquées dans la section 3.2 sur le maximum d'entropie

TODO
compléter

3.6.2 La loi a priori de Jeffreys

Définition 3.26. L'a priori de Jeffreys est défini par

$$\pi^*(\theta) \propto \sqrt{I(\theta)}.$$

En fait, la loi de Jeffreys est fréquemment impropre et donc la constante qui apparaît n'est pas importante. Deux manières qui n'ont a priori rien à voir justifient cette loi a priori.

Invariance

La première justification vient de considérations d'invariance. On rappelle que $I(\theta)$ mesure la capacité du modèle à discriminer entre θ et $\theta \pm \mathrm{d}\theta$: si $I(\theta)$ est grande, cela veut dire que la distribution avec le paramètre θ va être "bien différente" et donc peut être "bien distinguée" des distributions avec des paramètres relativement éloignés de θ_0. Il est donc naturel de prendre une distribution a priori de la forme $\pi(\theta) \propto \varphi(I(\theta))$ avec φ croissante, puisque cela revient à privilégier des valeurs de θ qui pourront être bien estimées.

Supposons maintenant que l'on veuille en outre imposer une contrainte d'invariance par reparamétrisation $h : \Theta \to \Theta'$. Si π était la loi d'une variable aléatoire X, on voudrait donc choisir la loi de $h(X)$ comme loi a priori π_h, i.e., on voudrait prendre $\pi_h(y) = \pi(x)/h'(x) = \varphi(I(x))/h'(x)$ puisque

$$f_{h(X)}(y) = f_X(h^{-1}(y))|(h^{-1})'(y)| = \frac{f_X(h^{-1}(y))}{|h'(h^{-1}(y))|}.$$

D'un autre côté, on voudrait que la relation $\pi = \varphi(I)$, i.e., on voudrait $\pi_h = \varphi(I_h)$. Comme $I_h(h(\theta)) = I(\theta)/h'(\theta)^2$ d'après la Proposition 2.7, on voudrait donc

$$\frac{\varphi(I)}{h'} = \pi_h = \varphi(I_h) = \varphi(I/h'^2),$$

i.e., $\varphi(I/h'^2) = \varphi(I)/h'$ dont la solution est $\varphi(x) = \sqrt{x}$.

Loi a priori de référence

Pour une loi a priori π on définit $G(\pi)$ la moyenne de la divergence de Kullback–Leibler entre la loi a priori et la loi a posteriori :

$$G(\pi) = \mathbb{E}\left[D\left(\pi(\cdot \mid x)||\pi\right)\right].$$

Ici l'espérance est prise par rapport à la loi moyennisée (par rapport à la loi a priori) des observations, i.e., on a de manière explicite

$$\mathbb{E}\left[D\left(\pi(\cdot \mid x)||\pi\right)\right] = \int \mathrm{d}x f(x) \int \mathrm{d}\theta \pi(\theta \mid x) \log\left(\frac{\pi(\theta \mid x)}{\pi(\theta)}\right) \tag{3.4}$$

avec $f(x) = \int \pi(\theta) f_\theta(x) \mathrm{d}x$. Il est naturel de chercher à maximiser cette distance, afin de maximiser l'information apportée par les observations : on parle alors de loi a priori de référence (*reference prior*). Le Théorème 4.15 de Bernstein–von Mises ci-dessous fournira alors une nouvelle justification de la loi a priori de Jeffreys, puisqu'il implique que cette loi a priori est approximativement une loi a priori de référence.

EXEMPLE 3.27 (FIL ROUGE) On a dans ce cas

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)}$$

et donc le score est défini par

$$\ell(\theta; x) = \partial_\theta \log f(x \mid \theta) = -\frac{x-\theta}{\sigma^2}$$

et donc l'information de Fisher par

$$I(\theta) = \int \ell(\theta, x)^2 f(x \mid \theta) \mathrm{d}x = \frac{1}{\sigma^3}$$

qui est une constante : on retrouve donc bien la loi impropre suggérée par l'invariance d'échelle. ■

EXEMPLE 3.28 On considère maintenant le cas du modèle gaussien centré à variance inconnue :

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/(2\theta)}.$$

Alors

$$\ell(\theta; x) = \partial_\theta \log f(x \mid \theta) = -\frac{1}{\theta} + \frac{x^2}{2\theta^2}$$

et donc

$$I(\theta) = \mathbb{E}_\theta \left[\left(-\frac{1}{\theta} + \frac{x^2}{2\theta^2} \right)^2 \right] = \mathbb{E}_1 \left[\left(-\frac{1}{\theta} + \frac{x^2}{2\theta} \right)^2 \right] = \frac{1}{\theta^2} \mathbb{E}_1 \left[(x-1)^2 \right] = \frac{1}{2\theta^2}$$

ce qui donne $\pi(\theta) \propto 1/\sigma$, encore une fois, comme suggéré par la relation d'invariance. ■

Si l'on veut pouvoir regarder le modèle gaussien avec moyenne et variance inconnues, il faut pouvoir généraliser l'information de Fisher et l'a priori de Jeffreys en plus grande dimension. Lorsque $\theta \in \mathbb{R}^d$ avec $d \geq 1$, l'information de Fisher $I(\theta)$ est une matrice, définie comme

$$-\left(\mathbb{E}_\theta \left(\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta_i \partial \theta_j} \right) \right)_{1 \leq i,j \leq d}$$

L'a priori de Jeffreys est alors donnée par $\sqrt{\det(I(\theta))}$.

EXEMPLE 3.29 Dans le cas du modèle gaussien avec moyenne et variance inconnues, on a alors $\theta = (\mu, \sigma^2)$ et

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x-\theta_1)^2/(2\theta_2)}$$

et donc

$$\log f(x \mid \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta_2 - \frac{(x-\theta_1)^2}{2\theta_2}$$

et donc

$$\begin{cases} \dfrac{\partial^2 \log f(x \mid \theta)}{\partial \theta_1^2} = -\dfrac{1}{\theta_2}, \\ \dfrac{\partial^2 \log f(x \mid \theta)}{\partial \theta_2^2} = \dfrac{1}{2\theta_2^2} - \dfrac{(x-\theta_1)^2}{\theta_2^3}, \\ \dfrac{\partial^2 \log f(x \mid \theta)}{\partial \theta_1 \partial \theta_2} = -\dfrac{\theta_1 - x}{\theta_2^2}, \end{cases}$$

et donc

$$I(\theta) = -\begin{pmatrix} -1/\theta_2 & \mathbb{E}_\theta(x-\theta_1)/\theta_2^2 \\ \mathbb{E}_\theta(x-\theta_1)/\theta_2^2 & 1/(2\theta_2^2) - \mathbb{E}_\theta(x-\theta_1)^2/\theta_2^2 \end{pmatrix} = \begin{pmatrix} 1/\theta_2 & 0 \\ 0 & 1/(2\theta_2^2) \end{pmatrix}$$

et donc l'a priori de Jeffreys est donné par $\sqrt{1/\theta_2^3} = 1/\sigma^{3/2}$. ■

On s'intéresse au comportement des estimateurs bayésiens dans le régime asymptotique $n \to \infty$.

4.1 Régularité et différentiabilité en moyenne quadratique

On va examiner des conditions techniques sous lesquelles les estimateurs convergent. Ces conditions impliquent des conditions de régularité que l'on énonce ici dans le cas à densité sur $\mathbb{R}$, i.e., on considère un modèle paramétrique $\{f_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}$. On généralise au cas vectoriel en remplaçant principalement les dérivées par des gradients.

Définition 4.1. Le modèle paramétrique $\{f_\theta, \theta \in \Theta\}$ est dit régulier si :

1. Θ est un ouvert de $\mathbb{R}^k$;
2. pour tout x, $\theta \mapsto f_\theta(x)$ est continûment dérivable, de dérivée $\partial_\theta f_\theta$;
3. la fonction du score $\ell_\theta = \partial_\theta f_\theta / f_\theta$ satisfait $\int \ell_\theta(x)^2 f_\theta(x) \mathrm{d}x < \infty$;
4. l'information de Fisher $I(\theta) = \int \ell_\theta(x)^2 f_\theta(x) \mathrm{d}x$ est strictement positive et continue en θ.

Lorsque le modèle est régulier dans le sens ci-dessus, alors il est aussi différentiable en moyenne quadratique, cf. par exemple [11, Lemme 7.6].

Définition 4.2. Un modèle paramétrique $\{f_\theta, \theta \in \Theta\}$ est dit différentiable en moyenne quadratique s'il existe une fonction mesurable $\dot{\ell}_\theta$ telle que, lorsque $\theta \to \theta_0$,

$$\int \left[\sqrt{f_\theta(x)} - \sqrt{f_{\theta_0}(x)} - \frac{1}{2}(\theta - \theta_0) \dot{\ell}_{\theta_0}(x) \sqrt{f_{\theta_0}(x)} \right]^2 \mathrm{d}x = o\left((\theta - \theta_0)^2\right).$$

Un des intérêts de la notion plus générale de différentiabilité en moyenne quadratique est de s'affranchir de l'hypothèse de continuité de $\theta \mapsto f_\theta(x)$, qui n'est par exemple pas satisfaite dès lors que les lois n'ont pas toutes le même support (cf. par exemple la loi uniforme, qui n'est d'ailleurs pas non plus différentiable en moyenne quadratique, cf. [11, Exemple 7.9]). A

noter par ailleurs que si le modèle est régulier, alors on a en fait $\dot{\ell}_\theta = \ell_\theta$ puisque

$$\partial_\theta \sqrt{f_\theta} = \frac{\partial_\theta f_\theta}{2\sqrt{f_\theta}} = \frac{\partial_\theta f_\theta}{2f_\theta}\sqrt{f_\theta}.$$

Mais, encore une fois, la différentiabilité quadratique ne nécessite pas la différentiabilité de f_θ.

4.2 Comportement asymptotique de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance joue un rôle important dans l'étude du comportement asymptotique des estimateurs bayésiens, et on commence donc pas son étude. On définit

$$L_n(\theta) = \frac{1}{n}\log f_\theta(x)$$

la log-vraisemblance pondérée. La loi des grands nombres donne alors le résultat suivant.

Lemme 4.3. *Si les observations sont i.i.d. tirées selon la vraie loi f_{θ_0}, alors pour tout $\theta \in \Theta$ on a $L_n(\theta) \xrightarrow{\text{p.s.}} L(\theta) := \mathbb{E}_{\theta_0}(\log f_\theta(x_1)) = -\mathrm{CE}(f_{\theta_0}, f_\theta)$.*

Ainsi, le résultat d'optimisation (2.1) donne $\theta_0 = \arg\max_\theta L(\theta)$ et puisque $\hat{\theta}^{\mathrm{MV}} = \arg\max L_n$, la convergence $\hat{\theta}^{\mathrm{MV}} \to \theta_0$ est naturelle. Néanmoins, la convergence ponctuelle $L_n(\theta) \xrightarrow{\text{p.s.}} L(\theta)$ n'est pas suffisante pour assurer que le maximiseur de L_n converge vers le maximiseur de L. Il faut typiquement une convergence fonctionnelle plus forte et une unicité du maximiseur. On a par exemple le résultat suivant.

Théorème 4.4 ([11, Théorème 5.7])**.** *Soit M_n des fonctions aléatoires et M une fonction déterministe. Si pour tout $\varepsilon > 0$ on a*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0 \quad \textit{et} \quad \sup_{\theta : d(\theta,\theta_0)\geq\varepsilon} M(\theta) < M(\theta_0) \tag{4.1}$$

alors $\hat{\theta}_n^{\mathrm{MV}} \xrightarrow{\mathbb{P}} \theta_0$ avec $\hat{\theta}_n^{\mathrm{MV}} = \arg\max M_n$.

Démonstration. Puisque $\|M_n - M\|_\infty \xrightarrow{\mathbb{P}} 0$ on a en particulier $M_n(\theta_0) \xrightarrow{\mathbb{P}} M(\theta_0)$ et donc, puisque θ_0 maximise M, on a

$$0 \leq M(\theta_0) - M(\hat{\theta}_n^{\mathrm{MV}}) \leq M_n(\theta_0) - M(\hat{\theta}_n^{\mathrm{MV}}) + o_{\mathbb{P}}(1).$$

Puisque $\hat{\theta}^{\mathrm{MV}} \in \arg\max M_n$ cela donne

$$0 \leq M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n^{\mathrm{MV}}) - M(\hat{\theta}_n^{\mathrm{MV}}) + o_{\mathbb{P}}(1) \leq \|M_n - M\|_\infty + o_{\mathbb{P}}(1)$$

et donc $M(\theta_0) - M(\hat{\theta}_n) \xrightarrow{\mathbb{P}} 0$. Par ailleurs, la deuxième condition implique que si $M(\theta)$ est proche de $M(\theta_0)$, alors nécessairement θ est proche de θ_0. Ainsi, la convergence $M(\hat{\theta}_n) \xrightarrow{\mathbb{P}} M(\theta_0)$ implique que $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$. □

REMARQUE 4.5 Sous les hypothèses (4.1), on a en fait $\hat{\theta}_n \to \theta_0$ pour tout estimateur $\hat{\theta}_n$ qui satisfait $\hat{\theta}_n \geq M_n(\theta_0) - o_{\mathbb{P}}(1)$. ■

La deuxième hypothèse du théorème ci-dessus implique en particulier que M admet un unique maximiseur, par ailleurs bien séparé. Concernant l'estimateur du maximum de vraisemblance, cette hypothèse est vérifiée si le modèle est identifiable par la Proposition 2.3 et il s'agit donc de vérifier la première condition dans (4.1). Pour comprendre cette convergence uniforme, on fait un détour via le théorème de Glivenko–Cantelli.

4.2.1 Théorème et classe de Glivenko–Cantelli (*)

Si on définit $\mathscr{F} = \{f_\theta : \theta \in \Theta\}$, $\mathbb{P}_n f = \frac{1}{n}\sum_{k=1}^n f(x_i)$ la moyenne sous la mesure empirique et $\mathbb{P}f = \mathbb{E}_{\theta_0} f(x)$ la moyenne sous la vraie valeur du paramètre, on peut donc réécrire la condition de convergence uniforme qui apparaît dans (4.1)

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

en

$$\sup\{|\mathbb{P}_n f - \mathbb{P}f| : f \in \mathscr{F}\} \xrightarrow{\mathbb{P}} 0. \tag{4.2}$$

Cela mène à la définition générale suivante.

Définition 4.6. Une classe de fonctions mesurables $\mathscr{F}$ est dite $\mathbb{P}$-Glivenko-Cantelli si la convergence (4.2) est satisfaite.

Lorsque $\mathscr{F}$ est l'ensemble des fonctions indicatrices d'intervalles de la forme $(-\infty, t]$ on obtient le théorème de Glivenko–Cantelli que certains d'entre vous ont déjà vu puisqu'il est à la base du test de Kolmogorov–Smirnov. En effet, pour $\mathscr{F} = \{x \mapsto \mathbb{1}(x \leq t) : t \in \mathbb{R}\}$ on a

$$\sup\{|\mathbb{P}_n f - \mathbb{P}f| : f \in \mathscr{F}\} = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$$

avec $F_n(t) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}(x_k \leq t)$ la fonction de répartition empirique et $F(t) = \mathbb{P}_{\theta_0}(x \leq t)$ la vraie fonction de répartition.

Théorème 4.7 (Théorème de Glivenko–Cantelli)**.** *Si $x_1, x_2, \ldots$ sont i.i.d. de fonction de répartition F, alors $\|F_n - F\|_\infty \xrightarrow{\mathbb{P}} 0$.*

Démonstration. Pour $\varepsilon > 0$ donné, il existe une partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ telle que $F(t_i-) - F(t_{i-1}) < \varepsilon$. Pour $t_{i-1} \leq t < t_i$ on a

$$F_n(t_{i-1}) \leq F_n(t) \leq F_n(t_i-) \text{ et } F(t_i-) - \varepsilon \leq F(t) \leq \varepsilon + F(t_{i-1})$$

ce qui donne

$$F_n(t_{i-1}) - F(t_{i-1}) - \varepsilon \leq F_n(t) - F(t) \leq F_n(t_i-) - F(t_i-) + \varepsilon$$

et donc

$$\limsup_n \|F_n - F\|_\infty \leq \varepsilon$$

puisque $F_n(t_i) \xrightarrow{\text{p.s.}} F(t_i)$ et $F_n(t_i-) \xrightarrow{\text{p.s.}} F(t_i-)$ pour chaque i par la loi forte des grands nombres. □

L'idée de la preuve ci-dessus est de trouver des "enveloppes" qui entourent la limite attendue F.

Définition 4.8. Un crochet $[\ell, u]$ est l'ensemble des fonctions $\{f : \ell \leq f \leq u\}$. Un crochet $[\ell, u]$ est un ε-crochet si $\mathbb{P}(u - \ell) \leq \varepsilon$. Le nombre crochetant $N(\varepsilon, \mathscr{F})$ est le nombre minimal de ε-crochets nécessaires pour couvrir $\mathscr{F}$.

Théorème 4.9 ([11, Théorème 19.4])**.** *Si $N(\varepsilon, \mathscr{F}) < \infty$ pour tout $\varepsilon > 0$, alors $\mathscr{F}$ est $\mathbb{P}$-Glivenko–Cantelli.*

4.2.2 Retour sur l'estimateur du maximum de vraisemblance

On utilise maintenant les résultats ci-dessus pour établir des résultats de convergence de l'estimateur du maximum de vraisemblance.

Théorème 4.10 (Consistance du MV, [11, Exemple 19.8])**.** *Si Θ est compact, $\sup_\theta f_\theta$ est intégrable et $\theta \mapsto f_\theta(x)$ est continue pour chaque x, alors $\{\log f_\theta : \theta \in \Theta\}$ est $\mathbb{P}$-Glivenko–Cantelli, i.e., $L_n \to L$ uniformément en probabilités.*
En particulier, si les hypothèses ci-dessus sont satisfaites, L est continue et le modèle est identifiable alors $\hat{\theta}_n^{\mathrm{MV}} \xrightarrow{\mathbb{P}} \theta_0$.

Démonstration. On montre que la première partie des hypothèses implique que le nombre crochetant est fini. Pour $B \subset \Theta$ une boule ouverte, on définit $f_B(x) = \inf_{\theta \in B} f_\theta(x)$ et $f^B(x) = \sup_{\theta \in B} f_\theta(x)$. Si B_m est une suite de boules avec un centre commun θ et un rayon décroissant vers 0, alors on a $f^{B_m} - f_{B_m} \downarrow 0$ par la continuité supposée, pour chaque x et donc dans L_1 par convergence dominée (en utilisant l'hypothèse $\sup_\theta f_\theta$ intégrable). Ainsi, pour $\varepsilon > 0$ fixé, pour chaque θ il existe une boule ouverte B centrée sur θ telle que le crochet $[f_B, f^B]$ a une taille au plus ε. Par la compacité de Θ, on peut extraire des B que l'on vient de construire un recouvrement fini. Les crochets correspondant recouvrent $\mathscr{F}$ et donc le nombre crochetant est fini.

Ainsi, pour avoir la convergence de l'estimateur du maximum de vraisemblance il ne reste qu'à montrer la deuxième condition dans (4.1), ce qui suit directement de la continuité et du fait que θ_0 est l'unique maximiseur de la log-vraisemblance lorsque le modèle est identifiable. □

EXEMPLE 4.11 (FIL ROUGE) Pour $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)}$, $\theta \mapsto f_\theta(x)$ est bien intégrable pour tout $x \in \mathbb{R}$. Par contre, $\sup_{\theta \in \mathbb{R}} f_\theta(x) = (2\pi\sigma^2)^{-1/2}$ n'est pas intégrable mais si on se restreint à $\Theta = [\theta_*, \theta_0]$ on a alors

$$\sup_{\theta_* \leq \theta \leq \theta_0} f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \begin{cases} 1 & \text{si } \theta_* \leq x \leq \theta_0, \\ e^{-(x-\theta_0)^2/(2\sigma^2)} & \text{si } \theta_* \leq x \leq \theta_0, \end{cases}$$

avec $\theta_0 = \theta_*$ ou $\theta_0 = \theta_0$ en fonction de x. Ainsi, en se restreignant à un compact on a bien que $\sup_\theta f_\theta$ est intégrable. ■

On rappelle maintenant un résultat de normalité asymptotique. Dans le reste de ces notes, N désigne une variable aléatoire normale standard.

Théorème 4.12 (Normalité asymptotique du MV, [11, Théorème 5.39]). *Si les conditions suivantes sont satisfaites :*

- *le modèle est identifiable et différentiable en moyenne quadratique ;*
- *il existe une fonction mesurable ℓ avec $\mathbb{P}\ell^2 < \infty$ telle que pour tous θ_1, θ_2 dans un voisinage de θ_0 :*

$$\left|\log f_{\theta_1}(x) - \log f_{\theta_2}(x)\right| \leq \ell(x)\,\|\theta_1 - \theta_2\|\,;$$

- *$I(\theta_0) > 0$;*
- *$\hat{\theta}_n^{\mathrm{MV}}$ est consistent ;*

alors $\sqrt{n}(\hat{\theta}_n^{\mathrm{MV}} - \theta_0) \xrightarrow{\mathrm{L}} I(\theta)^{-1/2} N$.

Eléments de démonstration. On présente l'idée de la preuve lorsque f_θ est deux fois continûment dérivable, et on note L_n' et L_n'' et L' et L'' les dérivées première et seconde de L_n et L, respectivement. Puisque $L_n'(\hat{\theta}^{\mathrm{MV}}) = 0$, on obtient par l'expansion de Taylor

$$0 = L_n'(\hat{\theta}^{\mathrm{MV}}) = L_n'(\theta_0) + (\hat{\theta}^{\mathrm{MV}} - \theta_0) L_n''(\theta_1)$$

pour un certain θ_1 entre θ_0 et $\hat{\theta}_n^{\mathrm{MV}}$, et donc

$$\sqrt{n}(\theta_0 - \hat{\theta}) = \frac{\sqrt{n} L_n'(\theta_0)}{L_n''(\theta_1)}$$

On a

$$\sqrt{n} L_n'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \partial_\theta \log f_\theta(x_i) \mid_{\theta=\theta_0}$$

et puisque $\mathbb{E}_{\theta_0}(\partial_\theta \log f_\theta(x_1) \mid_{\theta=\theta_0}) = L'(\theta_0) = 0$ puisque θ_0 maximise L, on obtient donc que $\sqrt{n} L_n'(\theta) \xrightarrow{\mathrm{L}} \sqrt{\mathbb{E}_{\theta_0}(\partial_\theta \log f_{\theta_0}(x)^2)} N = I(\theta_0)^{1/2} N$. Par ailleurs,

$$L_n''(\theta_1) = \frac{1}{n} \sum_{i=1}^{n} \partial_\theta^2 \log f_{\theta_1}(x_i) \to \mathbb{E}_{\theta_0} \partial_\theta^2 \log f_\theta(x) = -I(\theta_0)$$

en utilisant pour la convergence le fait que $\hat{\theta}_n^{\mathrm{MV}} \xrightarrow{\mathbb{P}} \theta_0$ (et donc $\theta_1 \to \theta_0$), ce qui donne le résultat. □

4.3 Comportement asymptotique des estimateurs bayésiens

Ibragimov et Has′minskiĭ [5, Théorème *II*.2.1] présentent un résultat très général garantissant qu'une suite d'estimateurs bayésiens est consistante et asymptotiquement normale. Nous nous restreindrons ici à une discussion sur l'estimateur du maximum a posteriori, sur la normalité asymptotique de la densité a posteriori (théorème de Bernstein-von Mises) et à une discussion rapide sur la moyenne a posteriori.

4.3.1 Comportement asymptotique de la densité a posteriori

La consistance est valable en tout généralité dès lors que les observations sont à valeurs dans un espace euclidien.

Théorème 4.13 (Consistance de la densité a posteriori, [11, Théorème 10.10])**.** *Si le modèle est identifiable, alors pour toute loi a priori π sur Θ la suite des lois a posteriori est consistante pour π-presque tout θ, i.e., pour π-presque tout θ et toute fonction continue bornée f on a*

$$\int f(\eta)\pi(\eta \mid x)\mathrm{d}\eta \xrightarrow{\mathbb{P}} f(\theta).$$

La limitation de ce résultat est que la convergence n'est garantie que pour π-presque tout θ. On considère maintenant un exemple plus simple.

Théorème 4.14. *Si Θ est compact et A est un voisinage de θ_0 avec $\pi(A) > 0$, alors $\pi(A \mid x) \xrightarrow{\text{p.s.}} 1$.*

Démonstration. On se ramène au cas discret en recouvrant A d'un nombre fini (puisque A est compact) de boules dont une seule contient θ_0. Il suffit de montrer que $\mathbb{P}(\theta = \theta_0 \mid x) \to 1$ lorsque Θ est fini et $\pi(\theta_0) = \mathbb{P}(\theta = \theta_0) > 0$. On a

$$\log\left(\frac{\pi(\theta \mid x)}{\pi(\theta_0 \mid x)}\right) = \log\left(\frac{\pi(\theta)}{\pi(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{f(x_i \mid \theta)}{f(x_i \mid \theta_0)}\right)$$

et donc

$$\frac{1}{n}\log\left(\frac{\pi(\theta \mid x)}{\pi(\theta_0 \mid x)}\right) \xrightarrow{\text{p.s.}} -D(f_{\theta_0}, f_\theta).$$

Puisque $D(f_{\theta_0}, f_\theta) < 0$ pour $\theta \neq \theta_0$, cela implique que $\pi(\theta \mid x)/\pi(\theta_0 \mid x) \xrightarrow{\text{p.s.}} 0$ pour $\theta \neq \theta_0$ et par suite, puisque Θ est fini, que $\pi(\theta_0 \mid x) \xrightarrow{\text{p.s.}} 1$. □

Il suit directement du résultat précédent que si $\pi(A) > 0$ pour tout voisinage de θ_0, alors $\pi \to \delta_{\theta_0}$. Si en revanche $\pi(A) = 0$ pour un voisinage de θ_0, i.e., la loi a priori ne met pas de masse autour de θ_0, alors on peut en fait prouver avec exactement les mêmes arguments que $\pi \to \delta_{\theta^*}$ avec $\theta^* = \arg\min D(f_{\theta_0}, f_\theta)$ où θ est dans le support de π. En d'autres termes, la loi a posteriori se concentre autour du paramètre qui minimise la distance de Kullback–Leibler, cf [3].

Concernant la normalité asymptotique, il s'agit du Théorème de Bernstein–von Mises.

Encore une fois, l'intuition est assez claire : on a

$$\begin{aligned} f(x \mid \theta) &= \prod_{i=1}^{n} f(x_i \mid \theta) \\ &= \exp\left(\sum_{i=1}^{n} \log f(x_i \mid \theta)\right) \\ &= f(x \mid \hat{\theta}^{\mathrm{MV}}) \exp\left(n(L_n(\theta) - L_n(\hat{\theta}^{\mathrm{MV}}))\right) \\ &\approx f(x \mid \hat{\theta}^{\mathrm{MV}}) \exp\left(\frac{1}{2} n(\theta - \hat{\theta}^{\mathrm{MV}})^2 L_n''(\hat{\theta}^{\mathrm{MV}})\right) \\ &\approx f(x \mid \hat{\theta}^{\mathrm{MV}}) \exp\left(-\frac{1}{2} n(\theta - \hat{\theta}^{\mathrm{MV}})^2 I(\theta_0)\right) \end{aligned}$$

(en utilisant $L_n''(\hat{\theta}^{\mathrm{MV}}) \to -I(\theta_0)$ pour la dernière égalité, cf. la preuve du Théorème 4.12) et donc

$$\pi(\theta \mid x) = \pi(\theta) \frac{f(x \mid \theta)}{f(x)} \propto \pi(\theta) \exp\left(-\frac{1}{2} n(\theta - \hat{\theta}^{\mathrm{MV}})^2 I(\theta_0)\right)$$

Or si $\bar{f}(\theta \mid x)$ est la densité de $\sqrt{n}(\theta - \hat{\theta}_n^{\mathrm{MV}})$ conditionnellement à x[1], on a

$$\bar{f}(\theta \mid x) = \frac{1}{\sqrt{n}} \pi\left(\hat{\theta}^{\mathrm{MV}} + \frac{\theta}{\sqrt{n}} \mid x\right) \propto \pi\left(\hat{\theta}^{\mathrm{MV}} + \frac{\theta}{\sqrt{n}}\right) \exp\left(-\frac{1}{2} \theta^2 I(\theta_0)\right)$$

Cela suggère donc que $\sqrt{n}(\theta - \hat{\theta}^{\mathrm{MV}})$ suit asymptotiquement une loi normale, ce qui est bien le cas. Il existe plusieurs versions de ce résultat appelé théorème de Bernstein-von Mises, dont la suivante.

TODO
Vérifier le résultat suivant (que veut-il dire ?)

Théorème 4.15 ([11, Théorème 10.1])**.** *On suppose que :*

- *l'échantillon $x = (x_1, \dots, x_n)$ est constitué de n observations i.i.d. de loi f_θ ;*
- *le modèle paramétrique $\{f_\theta, \theta \in \Theta\}$ est différentiable en moyenne quadratique en θ_0 avec $I(\theta_0) > 0$;*
- *π est absolument continue dans un voisinage de θ_0 avec $\pi(\theta_0) > 0$;*
- *pour tout $\varepsilon > 0$ il existe une suite de tests ϕ_n tels que $\mathbb{P}_{\theta_0} \phi_n \to 0$ et $\sup_{|\theta - \theta_0| \geq \varepsilon} \mathbb{P}_n(1 - \phi_n) \to 0$ pour tout $\varepsilon > 0$.*

Alors la loi a posteriori de $\sqrt{n}(\theta - \hat{\theta}_n^{\mathrm{MV}})$ sachant x converge en variation totale en probabilité vers une loi gaussienne centrée de variance $I(\theta)^{-1}$, i.e.,

$$\mathbb{P}_{\theta_0}\left(\int \left| \frac{1}{\sqrt{n}} \pi\left(\hat{\theta}^{\mathrm{MV}} + \frac{t}{\sqrt{n}} \mid x\right) - \frac{I(\theta_0)^{1/2}}{\sqrt{2\pi}} e^{-t^2 I(\theta_0)/2} \right| \mathrm{d}t \geq \varepsilon\right) \xrightarrow[n \to \infty]{} 0.$$

Dans ce résultat, on compare le paramètre à un estimateur du maximum de vraisemblance, et donc sous des hypothèses garantissant la consistance de l'estimateur du maximum de

1. $\hat{\theta}^{\mathrm{MV}}$ étant une fonction mesurable de x, on a n'a pas besoin de préciser la structure de dépendance entre θ et $\hat{\theta}_n^{\mathrm{MV}}$.

vraisemblance on obtient que la loi de θ conditionnelle aux observations suit une loi gaussienne. On peut aussi avoir des résultats de convergence uniforme sur les compacts de la densité a posteriori. Par exemple, sous certaines hypothèses techniques Schervish [10, Théorème 7.89] obtient le résultat suivant :

$$\mathbb{P}_{\theta_0}\left(\sup_{t\in B}\left|\pi'(t\mid x)-\phi(t)\right|\geq\varepsilon\right)\underset{n\to\infty}{\longrightarrow}0$$

pour tout $\varepsilon>0$ et tout compact B, où ϕ est la densité de la loi normale standard et π' la densité de $\sqrt{n}(\theta-\hat{\theta}_n^{\mathrm{MV}})$.

4.3.2 Comportement asymptotique du MAP

Par définition, on a

$$\hat{\theta}_n^{\mathrm{MAP}}\in\arg\max_\theta M_n(\theta)\quad\text{avec}\quad M_n(\theta)=L_n(\theta)+\frac{1}{n}\log\pi(\theta).$$

TODO
Lien MAP $\leftrightarrow$ optim avec pénalité

On a donc $M_n(\theta)\xrightarrow{\text{p.s.}}L(\theta)$ mais en général, même si $L_n\xrightarrow{\text{p.s.}}L$ uniformément, à cause du terme $\frac{1}{n}\log\pi(\theta)$ on n'aura pas convergence uniforme de M_n vers L (sauf si $\sup|\log\pi|<\infty$, i.e., $\sup\pi<\infty$ et $\inf\{\pi(\theta):\theta\text{ t.q. }\pi(\theta)>0\}>0$). Pour éviter la deuxième condition qui est très forte, on se ramène à la maximisation de la vraisemblance (et non de la log-vraisemblance) ce qui nécessite une hypothèse supplémentaire que L est bornée. Ces hypothèses doivent probablement pouvoir être allégées mais je n'ai pas trouvé de référence[2].

TODO
Trouver de meilleures hypothèses, par exemple $\theta\in L^1$?

Théorème 4.16 (Consistance du MAP)**.** *Si les hypothèses du Théorème 4.10 sont satisfaites, que L et π sont bornées et que $\pi(\theta_0)>0$, alors $\hat{\theta}_n^{\mathrm{MAP}}\xrightarrow{\mathbb{P}}\theta_0$.*

Démonstration. Sous les hypothèses du Théorème 4.10, on a $\|L_n-L\|_\infty\xrightarrow{\mathbb{P}}0$ et donc $\|e^{L_n}-e^L\|_\infty\xrightarrow{\mathbb{P}}0$ aussi, puisque

$$\left|e^{L_n(\theta)}-e^{L(\theta)}\right|\leq e^{L(\theta)}\left(e^{\|L_n-L\|_\infty}-1\right),$$

et donc $M_n=(\pi(\theta))^{1/n}e^{L_n(\theta)}\xrightarrow{\mathbb{P}}e^{L(\theta)}$ uniformément. Puisque $\hat{\theta}^{\mathrm{MAP}}=\arg\max M_n$ on peut donc appliquer le Théorème 4.4. □

TODO
Préciser les hypothèses et la conclusion.

2. Par exemple, en utilisant la remarque (4.5), on pourrait s'affranchir de l'hypothèse L bornée si l'on pouvait montrer que $\frac{1}{n}\log\pi(\hat{\theta}_n^{\mathrm{MAP}})\to0$, ce qui est consistent avec l'hypothèse $\pi(\theta_0)>0$ et le comportement attendu $\hat{\theta}_n^{\mathrm{MV}}$.

Théorème 4.17 (Normalité asymptotique du MAP)**.** *Sous certaines hypothèses, on a* $\sqrt{n}(\hat{\theta}^{\mathrm{MAP}} - \theta_0)$ *converge vers une loi normale.*

Retour sur les lois a priori de référence

On justifie maintenant la remarque de la Section 3.6.2 selon laquelle la loi a priori de Jeffreys maximise approximativement la distance de Kullback–Leibler entre la loi a priori et la loi a posteriori. On rappelle que le but est de maximiser, en π, la quantité $\mathbb{E}[D(\pi(\cdot \mid x)||\pi)]$. Le comportement asymptotique de la loi a posteriori $\pi(\cdot \mid x)$ dépend de la loi des x : le théorème de Bernstein–Van Mises dit que, lorsque l'échantillon est tiré selon f_θ pour un certain paramètre θ fixé, la loi a posteriori correctement normalisée converge vers une gaussienne centrée de variance $1/I(\theta)$. Formellement, sous $\mathbb{P}_\theta$ (i.e., lorsque x est tiré selon f_θ pour $\theta \in \Theta$ fixé), alors le théorème de Bernstein–Van Mises se traduit de la manière suivante :

$$\pi(\theta' \mid x) \approx \sqrt{\frac{nI(\theta)}{2\pi}} e^{-nI(\theta)(\theta'-\hat{\theta})^2/2}.$$

Attention : dans cette expression, θ' est la variable muette, et θ est fixé et représente le paramètre de la loi selon laquelle l'échantillon x est tiré. On va maintenant utiliser cette approximation pour étudier $\mathbb{E}[D(\pi(\cdot \mid x)||\pi)]$.

L'idée est de conditionner par rapport à θ afin de pouvoir utiliser l'approximation précédente. On va donc écrire

$$\mathbb{E}[D(\pi(\cdot \mid x)||\pi) \mid \theta] = \mathbb{E}_\theta\left[\int \pi(\theta' \mid x) \log\left(\frac{\pi(\theta' \mid x)}{\pi(\theta')}\right) \mathrm{d}\theta'\right].$$

Sous $\mathbb{P}_\theta$, on peut donc utiliser l'approximation de la loi a posteriori, ce qui donne

$$\int \pi(\theta' \mid x) \log\left(\frac{\pi(\theta' \mid x)}{\pi(\theta')}\right) \mathrm{d}\theta' \approx \int \sqrt{\frac{nI(\theta)}{2\pi}} e^{-nI(\theta)(\theta'-\theta)^2/2} \log\left(\frac{\sqrt{\frac{nI(\theta)}{2\pi}} e^{-nI(\theta)(\theta'-\theta)^2/2}}{\pi(\theta')}\right) \mathrm{d}\theta'.$$

Si on change de variable puis qu'on utilise l'approximation $\pi(\theta + u/\sqrt{n}) \approx \pi(\theta)$, on obtient

$$\begin{aligned}\int \pi(\theta' \mid x) \log\left(\frac{\pi(\theta' \mid x)}{\pi(\theta')}\right) \mathrm{d}\theta' &\approx \int \sqrt{\frac{I(\theta)}{2\pi}} e^{-I(\theta)u^2/2} \log\left(\frac{\sqrt{\frac{nI(\theta)}{2\pi}} e^{-I(\theta)u^2/2}}{\pi(\theta + u/\sqrt{n})}\right) \mathrm{d}u \\ &\approx \int \sqrt{\frac{I(\theta)}{2\pi}} e^{-I(\theta)u^2/2} \log\left(\frac{\sqrt{\frac{nI(\theta)}{2\pi}} e^{-I(\theta)u^2/2}}{\pi(\theta)}\right) \mathrm{d}u.\end{aligned}$$

Il reste donc à faire des intégrales par rapport à la densité normale, et on obtient

$$\int \pi(\theta' \mid x) \log\left(\frac{\pi(\theta' \mid x)}{\pi(\theta')}\right) \mathrm{d}\theta' \approx \frac{1}{2}\log\left(\frac{n}{2\pi}\right) + \log\left(\frac{\sqrt{I(\theta)}}{\pi(\theta)}\right) \mathrm{d}u - \frac{1}{2}.$$

La quantité que l'on cherche à calculer, $\mathbb{E}[D(\pi(\cdot \mid x)||\pi)]$, est obtenue en intégrant cette dernière quantité par rapport à π, et on obtient donc

$$\mathbb{E}[D(\pi(\cdot \mid x)||\pi)] \approx \int \log\left(\frac{\sqrt{I(\theta)}}{\pi(\theta)}\right) \pi(\theta) \mathrm{d}\theta + \frac{1}{2}\log\left(\frac{n}{2\pi}\right) - \frac{1}{2}.$$

Si $\sqrt{I}$ était une densité, on aurait une divergence de Kulback qui serait minimisée pour $\pi = \sqrt{I}$. Les mêmes arguments montrent que si $\sqrt{I}$ et π non intégrables (on s'autorise donc des lois impropres), alors c'est toujours $\sqrt{I}$ qui minimise cette quantité. On retrouve ainsi l'a priori de Jeffreys.

4.3.3 Moyenne a posteriori (*)

Lorsqu'une loi converge, il est naturel de s'attendre à ce que la moyenne associée converge aussi. Ainsi, sous des hypothèses supplémentaires, un corollaire du théorème de Bernstein-von Mises est que la moyenne a posteriori est un estimateur consistant et asymptotiquement normal.

On s'intéresse maintenant à l'estimateur bayésien $\hat{\theta}_n = \mathbb{E}(\theta \mid x_n)$ donné par la moyenne a posteriori. Puisque

$$\hat{\theta}_n = \int \eta \pi(\eta \mid x) \mathrm{d}\eta,$$

on est presque dans le cadre d'application du Théorème 4.13, sauf qu'on voudrait l'appliquer à $f(\eta) = \eta$ qui n'est pas bornée. En fait, le cœur de la preuve du Théorème 4.13 consiste à prouver que θ est mesurable par rapport à $x_\infty = (x_1, x_2, \ldots)$, i.e., qu'il existe une fonction mesurable h telle que

$$\theta = h(x_\infty).$$

TODO
Essayer de comprendre pourquoi c'est vrai...

Une fois ce résultat admis, on peut alors prouver la consistance de la moyenne a posteriori grâce à un argument de martingale. En effet, la suite $(\hat{\theta}_n, n \geq 1)$ vérifie une propriété bien particulière, à savoir

$$\mathbb{E}\left(\hat{\theta}_{n+1} \mid x_n\right) = \mathbb{E}\left[\mathbb{E}\left(\theta \mid x_{n+1}\right) \mid x_n\right] = \mathbb{E}\left(\theta \mid x_n\right) = \hat{\theta}_n.$$

La deuxième égalité vient de la propriété de la tour, à savoir

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

On a donc l'égalité remarquable $\mathbb{E}\left(\hat{\theta}_{n+1} \mid x_n\right) = \hat{\theta}_n$ qui est caractéristique d'une structure de martingale. Dans le cas présent, où $\hat{\theta}_n = \mathbb{E}(\theta \mid x_n)$, on a le résultat fondamental suivant.

Théorème 4.18. *Si $\mathbb{E}(|\theta|) < \infty$ et le modèle est identifiable, alors $\hat{\theta}_n$ converge presque sûrement et dans L^1 vers $\mathbb{E}(\theta \mid x_1, x_2, \ldots)$.*

Puisque θ est mesurable par rapport à x_∞ on obtient bien la consistance de $\hat{\theta}_n$. Concernant la normalité asymptotique on se contentera à nouveau de calculs heuristiques. On a

$$\hat{\theta} = \int \theta \pi(\theta \mid x) \mathrm{d}\theta$$

et donc le changement de variable $\eta = \sqrt{n}(\theta - \hat{\theta}^{\mathrm{MV}})$ donne

$$\hat{\theta} = \frac{1}{\sqrt{n}} \int \left(\hat{\theta}^{\mathrm{MV}} + \frac{\eta}{\sqrt{n}}\right) \pi\left(\hat{\theta}^{\mathrm{MV}} + \frac{\eta}{\sqrt{n}} \mid x\right) \mathrm{d}\eta$$

et donc

$$\sqrt{n}\left(\hat{\theta}-\hat{\theta}^{\mathrm{MV}}\right)=\int \frac{\eta}{\sqrt{n}}\pi\left(\hat{\theta}^{\mathrm{MV}}+\frac{\eta}{\sqrt{n}}\mid x\right)\mathrm{d}\eta.$$

Le Théorème de Bernstein–von Mises suggère donc que $\sqrt{n}(\hat{\theta}-\hat{\theta}^{\mathrm{MV}})\xrightarrow{\mathrm{L}}0$.

TODO
Préciser les hypothèses.

Théorème 4.19. *$\sqrt{n}(\hat{\theta}-\hat{\theta}^{\mathrm{MV}})\xrightarrow{\mathrm{L}}0$ et donc $\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{\mathrm{L}}\mathcal{N}(0,1/I(\theta_0))$ si l'estimateur du maximum de vraisemblance converge.*

4.4 Inégalité de van Trees (*)

On présente dans cette section l'inégalité de van Trees, qui peut servir à borner inférieurement le risque minimax (puisque $\max_\theta R_f(\delta,\theta)\geq\int R_f(\delta,\theta)\pi(\theta)\mathrm{d}\theta$). Cf. [4, 12, 13, 14] pour des généralisations (notamment les bornes de Weiss–Weinstein) et le lien avec l'inégalité de Fréchet–Darmois–Cramer–Rao.

Théorème 4.20. *Si le modèle est régulier, que $I(\theta)>0$ et que $\hat{g}$ est un estimateur de carré intégable et sans biais de $g(\theta)$, alors pour tout $\theta\in\Theta$ on a $\mathbb{V}\mathrm{ar}_\theta\hat{g}\geq\frac{g'(\theta)^2}{nI(\theta)}$.*

Démonstration. Puisque $\theta=\mathbb{E}_\theta\hat{g}=\int g(x)f_\theta(x)\mathrm{d}x$ on obtient en dérivant

$$\begin{aligned}1&=\int g(x)\partial_\theta f_\theta(x)\mathrm{d}x\\&=\int(g(x)-\theta)\partial_\theta f_\theta(x)\mathrm{d}x\\&\leq\left\{\int(g(x)-\theta)^2f_\theta(x)\mathrm{d}x\right\}\times\left\{\int\frac{\partial_\theta f_\theta(x)^2}{f_\theta(x)}\mathrm{d}x\right\}\end{aligned}$$

en utilisant $\int\partial_\theta f_\theta(x)\mathrm{d}x=0$ pour la deuxième égalité et Cauchy–Schwarz dans $L_2(f_\theta)$ pour la dernière inégalité, ce qui donne le résultat. □

Théorème 4.21 (Inégalité de van Trees)**.** *On suppose que le modèle est différentiable en moyenne quadratique en tout θ, et que $\theta \mapsto f(x \mid \theta)$ est C^1 pour tout x. On suppose que π est dérivable sur $[a,b]$, nulle aux bords, et on note*

$$J = \mathbb{E}[\log \pi(\theta)'^2] = \int \frac{\pi'(\theta)^2}{\pi(\theta)} \mathrm{d}\theta.$$

On suppose en outre que g est C^1 sur $[a,b]$ et telle que $\mathbb{E}(|g'(\theta)|) < \infty$. Alors pour tout estimateur T on a

$$\mathbb{E}\left[(T - g(\theta))^2\right] \geq \frac{[\mathbb{E}(g'(\theta))]^2}{\mathbb{E}(I(\theta)) + J}$$

avec $I(\theta) = \mathbb{E}[\log f(x \mid \theta)'^2 \mid \theta]$ l'information de Fisher.

Démonstration. On a

$$\mathbb{E}(g'(\theta)) = \int g'(\theta)\pi(\theta)\mathrm{d}\theta = \int \left\{ \int g'(\theta)\pi(\theta) f(x \mid \theta)\mathrm{d}\theta \right\} \mathrm{d}x.$$

Par intégration par parties,

$$\int g'(\theta)\pi(\theta) f(x \mid \theta)\mathrm{d}\theta = [(g(\theta) - T(x))\pi(\theta) f(x \mid \theta)]_a^b - \int (g(\theta) - T(x))\, \partial_\theta(\pi(\theta) f(x \mid \theta))\mathrm{d}\theta$$

Par hypothèse, les termes de bord sont nuls et donc

$$\begin{aligned}\mathbb{E}(g'(\theta)) &= -\int (g(\theta) - T(x)) \left(\pi'(\theta) f(x \mid \theta) + \pi(\theta)\partial_\theta f(x \mid \theta)\right) \mathrm{d}\theta\mathrm{d}x \\ &= -\int (g(\theta) - T(x)) \left(\frac{\pi'(\theta)}{\pi(\theta)} + \frac{\partial_\theta f(x \mid \theta)}{f(x \mid \theta)}\right) \pi(\theta) f(x \mid \theta)\mathrm{d}\theta\mathrm{d}x\end{aligned}$$

puis par Cauchy–Schwarz,

$$\left[\mathbb{E}(g'(\theta))\right]^2 \leq \left\{ \int (g(\theta) - T(x))^2 \pi(\theta) f(x \mid \theta)\mathrm{d}\theta\mathrm{d}x \right\} \times \left\{ \int \left(\frac{\pi'(\theta)}{\pi(\theta)} + \frac{\partial_\theta f(x \mid \theta)}{f(x \mid \theta)}\right)^2 \pi(\theta) f(x \mid \theta)\mathrm{d}\theta\mathrm{d}x \right\}$$

Le premier terme donne le terme $\mathbb{E}[(T - g(\theta))^2]$. Quant au second terme, on obtient en développant le carré

$$\int \left(\frac{\pi'(\theta)}{\pi(\theta)} + \frac{\partial_\theta f(x \mid \theta)}{f(x \mid \theta)}\right)^2 \pi(\theta) f(x \mid \theta)\mathrm{d}\theta\mathrm{d}x = J + I(\theta) + \int \pi'(\theta)\partial_\theta f(x \mid \theta)\mathrm{d}\theta\mathrm{d}x$$

et puisque $\int \partial_\theta f(x \mid \theta)\mathrm{d}x = 0$ on obtient bien le résultat voulu. □

5.1 Généralités

Une manière de voir apparaître les estimateurs bayésiens est via la théorie de la décision, dont nous présentons certains éléments dans ce chapitre. En pratique, les observations sont là pour aider un décideur à prendre une décision, vendre un stock, etc. Une décision est donc représentée par une fonction $\delta : \mathscr{O} \to D$ avec $\mathscr{O}$ l'ensemble des observations possibles et D l'ensemble des actions possibles, et l'on appellera une telle fonction une règle de décision (la décision étant plus l'action). Par simplicité, on se restreint dans ce cours au cas où la décision à prendre est une estimation, i.e., $D = \Theta$, et on notera donc $\mathscr{D} = \{\delta : \mathscr{O} \to \Theta \text{ mesurable}\}$ l'ensemble des règles de décision, i.e., estimateurs, possibles. Un exemple simple où $D \neq \Theta$ est le cas où $D = \{0,1\}$: cela correspond à une prise de décision, et sera traité au Chapitre 6.

La théorie de la décision repose sur deux éléments :

- une **fonction de coût** ou **fonction de perte** $L : \Theta \times D \to \mathbb{R}$ qui à une valeur de paramètre $\theta \in \Theta$ et une décision $d \in D$ associe un coût $L(\theta, d) \in \mathbb{R}$;
- une **règle de décision** $\delta : \mathscr{O} \to D$ qui à des observations $x \in \mathscr{O}$ associe une décision/observation $\delta(x) \in D = \Theta$.

Les fonctions de coût les plus classiques correspondent à des distances L_p, par exemple la *perte quadratique* $L(\theta, d) = (\theta - d)^2$ ou encore la *perte absolue* $L(\theta, d) = |\theta - d|$. Dans le cas où Θ est fini ou dénombrable, on peut aussi considérer la fonction de coût binaire $L(\theta, d) = \mathbb{1}(\theta \neq d)$, i.e., $L(\theta, d) = 0$ si $\theta = d$ et 1 sinon.

5.2 Critères fréquentistes

Dans le cadre de la théorie de la décision, la fonction de coût L est supposée fixée une fois pour toute, et on s'intéresse à ce qui se passe lorsque le paramètre $\theta \in \Theta$ varie ainsi que la règle de décision δ. Ainsi, on considère le **risque** ou **risque fréquentiste** $R_f(\theta, \delta)$ associé au paramètre θ et à la règle de décision δ défini par le coût moyen lorsque le paramètre vaut θ :

$$R_f(\theta, \delta) = \mathbb{E}_\theta \left[L(\theta, \delta(\cdot)) \right] = \int L(\theta, \delta(x)) f(x \mid \theta) \mathrm{d}x.$$

Le risque fréquentiste associé à une règle de décision données est donc une fonction du paramètre inconnu θ : comparer les risques de deux règles de décision revient donc à comparer

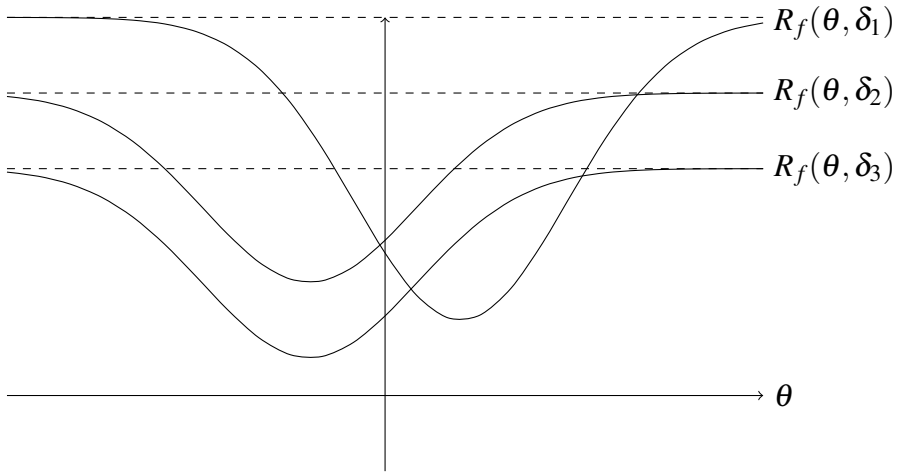

FIGURE 5.1 – Un exemple de trois risques fréquentistes : δ_3 est uniformément meilleur que δ_2, on dit qu'elle lui est préférable et donc δ_2 est inadmissible. Par contre, il n'est a priori pas évident de comparer δ_1 à δ_2 et δ_3.

des fonctions, problème que l'on peut aborder de multiples manières comme il sera discuté dans le reste de ce chapitre.

EXEMPLE 5.1 (FIL ROUGE) On prendra une fonction de coût quadratique : $L(\theta,\delta) = (\theta - \delta)^2$ et une fonction linéaire en les observations comme règle de décision/estimateur, i.e., $\delta = a^T x$ pour un certain vecteur a. Si $a = (1/n,\dots,1/n)$ on a alors $\delta = \bar{x}$, la moyenne empirique, et par ailleurs, l'estimateur est non biaisé si et seulement si $A := a^T \mathbf{1}_n = \sum a_i = 1$. On a alors

$$\begin{aligned} R_f(\theta,\delta) &= \mathbb{E}_\theta\left[\left(\theta - a^T X\right)^2\right] \\ &= \mathbb{E}_\theta\left[\left(\theta - \theta a^T \mathbf{1}_n - (a^T X - \theta a^T \mathbf{1}_n)\right)^2\right] \\ &= (1-A)^2\theta^2 + \mathbb{V}\mathrm{ar}(a^T X) \\ &= (1-A)^2\theta^2 + \|a\|^2\sigma^2. \end{aligned}$$

Si δ est non biaisé, on a donc un risque indépendant de θ et qui augmente avec la variance σ^2 du modéle d'échantillonnage. Pour $a = n^{-1}\mathbf{1}_n$ on a un risque qui diminue en $1/n^2$, ce qui correspond au risque minimal dans cette classe d'estimateurs. ■

On étudie maintenant la performance fréquentiste d'une règle de décision via deux critères : l'admissibilité et la minimaxité.

5.2.1 Admissibilité

Définition 5.2. Une règle de décision δ est **admissible** si aucune autre règle n'a un risque inférieur pour tout θ, i.e., si $R_f(\theta,\delta') \leq R_f(\theta,\delta)$ pour tout θ implique $\delta' = \delta$.

Définition 5.3. On dit qu'une règle de décision δ_1 est **préférable** à une autre règle de décision δ_2 si les deux conditions suivantes sont satisfaites :

1. le risque associé à δ_1 est toujours au moins aussi bon que celui associé à δ_2, i.e., $R_f(\theta,\delta_1) \leq R_f(\theta,\delta_2)$ pour tout $\theta \in \Theta$;
2. le risque associé à δ_1 est strictement meilleur que celui associé à δ_2 pour au moins une valeur de paramètre θ, i.e., il existe θ_0 tel que $R_f(\theta_0,\delta_1) < R_f(\theta_0,\delta_2)$.

REMARQUE 5.4 On vérifie qu'une règle de décision est admissible si aucune autre règle ne lui est préférable. ■

Par exemple, sur la Figure 5.1 δ_2 n'est pas admissible puisque δ_3 lui est préférable. Le premier exemple suivant montre que l'admissibilité n'induit pas un ordre total. Le deuxième exemple montre que la notion d'inadmissibilité n'est pas très forte, notamment du fait que l'on ne peut pas toujours comparer deux règles. Néanmoins, si une règle n'est pas admissible alors il existe une règle dont le risque est moindre quelle que soit la vraie valeur du paramètre θ et on a donc tout intérêt à prendre cette autre règle.

EXEMPLE 5.5 (FIL ROUGE) Si on reprend le risque intégré $R_f(\theta,\delta) = (1-A)^2\theta^2 + \|a\|^2\sigma^2$ obtenu dans l'exemple 2.7, on voit qu'en fonction des valeurs de A et $\|a\|$ les règles de décision peuvent être ou non comparables. Si on se restreint aux règles de décision avec $A = 1$, i.e., les règles sans biais, alors on peut comparer deux règles a_1 et a_2 et a_1 sera préférable à a_2 si $\|a_1\| \leq \|a_2\|$. ■

EXEMPLE 5.6 Soit $\hat{\theta} = \theta_0$ pour un certain $\theta_0 \in \Theta$ fixé, et on considère le coût quadratique $L(\theta,d) = (\theta - d)^2$. Soit T un autre estimateur avec $R_f(\theta,T) \leq R_f(\theta,\hat{\theta})$ pour tout θ. En particulier, pour $\theta = \theta_0$ on a $R_f(\theta_0,T) \leq R_f(\theta_0,\hat{\theta}) = 0$ et donc $\mathbb{E}_{\theta_0}((\theta_0 - T)^2) = 0$, i.e., $T = \theta_0$ $\mathbb{P}_{\theta_0}$-presque sûrement. Ainsi, $\hat{\theta}$ est admissible. ■

En général, si on restreint la classe d'estimateurs considérés, on peut établir des résultats d'admissibilité, par exemple dans la classe des estimateurs non biaisés comme illustré dans l'exemple 5.5. Néanmoins, ça n'est pas toujours le cas comme le montre l'exemple suivant, proposé par Stein en 1955.

EXEMPLE 5.7 On cherche à estimer un paramètre $\theta \in \mathbb{R}^p$ à l'aide d'une observation $X \sim \mathcal{N}(\theta, I_p)$. L'estimateur "ordinaire" est simplement donné par $\hat{\theta} = X$ et présente un risque $R_f(\theta,\hat{\theta}) = p$. En revanche, l'estimateur

$$\hat{\theta}' = \left(1 - \frac{p-2}{\|X\|^2}\right) X$$

présente lui un risque

$$R_f(\theta,\hat{\theta}') = p - (p-2)\mathbb{E}\left(\frac{1}{\|X\|}\right)$$

qui est strictement meilleur dès lors que $p \geq 3$. Ce résultat est particulièrement surprenant, puisqu'il dit que pour estimer les moyennes des X_i qui sont pourtant **indépendants**, il est préférable d'utiliser un estimateur qui les combine ! Cf. par exemple [9] pour plus de détails. ■

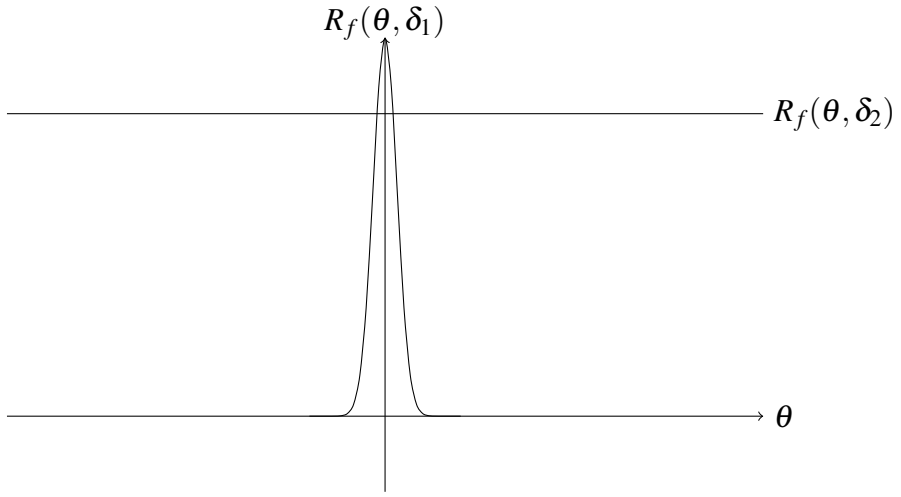

FIGURE 5.2 – Sur cet exemple, δ_2 présente un risque minimax inférieur à δ_1 alors que son risque fréquentiste est plus élevé pour presque toutes les valeurs de θ.

5.2.2 Minimaxité

Un moyen un peu "brutal" de comparer le risque de deux règles de décision est de simplement comparer leur risque maximal : cela conduit à la notion de minimaxité. Par exemple, sur la Figure 5.1, on a $\sup_{\theta\in\Theta} R_f(\theta,\delta_1) \geq \sup_{\theta\in\Theta} R_f(\theta,\delta_2)$ ce qui pourra amener à considérer δ_2 plutôt que δ_1. Dans ce cas, δ_2 est appelé estimateur minimax. Le critère minimax est évidemment très frustre comme l'illustre la Figure 5.2.

Définition 5.8. On appelle **risque minimax** associé à la fonction de coût L la valeur

$$\overline{R} = \inf_{\delta\in\mathscr{D}} \sup_{\theta\in\Theta} R_f(\theta,\delta)$$

et **estimateur minimax** tout estimateur δ_0 avec

$$\sup_{\theta\in\Theta} R_f(\theta,\delta_0) = \overline{R}.$$

L'interprétation de $\overline{R}$ est la suivante : je choisis une règle de décision, et la Nature choisit le pire paramètre. Je cherche donc à choisir le meilleur paramètre avec ces règles.

EXEMPLE 5.9 (FIL ROUGE) Pour le risque $R_f(\theta,\delta) = (1-A)^2\theta^2 + \|a\|^2\sigma^2$, le risque minimax est infini si l'estimateur est biaisé et que Θ n'est pas borné. Si par exemple $\Theta = [0,\theta_0]$, alors on a $\sup_\theta R_f(\theta,\delta) = R_f(\theta_0,\delta)$ et tout a qui minimise $(1-A)^2\theta_0^2 + \|a\|^2\sigma^2$ est minimax. ■

Sous certaines hypothèses générales, l'existence d'un estimateur minimax est assurée par le théorème suivant.

Théorème 5.10 ([8, Théorème 2.20]). *Si $\mathscr{D} \subset \mathbb{R}^k$ est convexe et compact et si $d \in D \mapsto L(\theta, d)$ est continue et convexe pour chaque valeur de θ, alors il existe un estimateur minimax.*

5.2.3 Risque de Bayes

Le critère minimax est extrêmement conservatif puisqu'il ne considère que les valeurs extrêmes du risque. Le risque intégré introduit ci-dessous pallie ce problème en prenant en compte tous les risques possibles, mais en pondérant les valeurs du paramètre. La terminologie en lien avec Bayes (risque de Bayes, règle de décision bayésienne, estimateur bayésien) sera justifiée dans la prochaine section.

Définition 5.11. Pour toute règle de décision $\delta \in \mathscr{D}$, le **risque de Bayes** ou **risque intégré** de δ par rapport à une loi a priori π est défini par

$$R_B(\delta) = \int R_f(\theta, \delta)\pi(\theta)\mathrm{d}\theta = \int L(\theta, \delta(x)) f(x \mid \theta)\pi(\theta)\mathrm{d}\theta\mathrm{d}x = \mathbb{E}\left[L(\theta, \delta(\cdot))\right].$$

Le **risque de Bayes selon π** est la plus petite valeur possible du risque de Bayes :

$$\underline{R} = \inf_{\delta \in \mathscr{D}} R_B(\delta).$$

Une règle de décision δ_π est une **règle de décision bayésienne** pour l'a priori π si

$$R_B(\delta_\pi) = \underline{R} = \inf_{\delta \in \mathscr{D}} R_B(\delta).$$

Si δ_π est une règle bayésienne, l'estimateur $\delta_\pi(x)$ associé est appelé **estimateur bayésien**.

EXEMPLE 5.12 (FIL ROUGE) Le risque de Bayes est donné par

$$R_B(\delta) = \mathbb{E}\left((1-A)^2\theta^2 + \|a\|^2\sigma^2\right)$$

où $\theta \sim \mathscr{N}(\mu_0, \sigma_0^2)$, ce qui donne

$$R_B(\delta) = (1-A)^2\mathbb{E}\left(\theta^2\right) + \|a\|^2\sigma^2 = (1-A)^2\left(\sigma_0^2 + \mu_0^2\right) + \|a\|^2\sigma^2.$$

Si l'on se restreint à la classe des estimateurs non biaisés, les estimateurs bayésiens et minimax coïncident donc puisqu'ils correspondent tous les deux à minimiser $\|a\|^2$ sous la contrainte $a^T\mathbf{1}_n = 1$. Dans le cas général, le risque minimax est infini alors que le risque de Bayes reste fini. ■

EXEMPLE 5.13 On considère $\pi(\theta) \sim \mathscr{E}(1)$, $f(x \mid \theta) \sim \mathscr{P}(\theta)$ et les estimateurs linéaires $\delta_c(x) = cx$. On considère d'abord le coût quadratique $L(\theta, d) = (\theta - d)^2$: on a alors

$$R_f(\theta, \delta_c) = \mathbb{E}\left[(cx-\theta)^2 \mid \theta\right] = c^2\mathbb{V}\mathrm{ar}(x \mid \theta) + (c-1)^2\theta^2 = c^2\theta + (c-1)^2\theta^2.$$

Alors δ_γ est préférable à δ_c si et seulement si

$$\begin{aligned}\forall \theta > 0,\ \gamma^2\theta + (\gamma-1)^2\theta^2 &< c^2\theta + (c-1)^2\theta^2 \\ &\iff \forall \theta > 0,\ \gamma^2 + (\gamma-1)^2\theta < c^2 + (c-1)^2\theta \\ &\iff \gamma < c \ \text{ et } \ (\gamma-1)^2 < (c-1)^2.\end{aligned}$$

Ainsi, δ_c pour $c > 1$ n'est pas admissible, puisque tout δ_γ avec $\gamma \in (c_*, c)$ avec $c_* < c$ et $(1-c_*)^2 = (1-c)^2$ lui est préférable (et on vérifie directement que $\gamma = 1$ lui est préférable, puisque pour $c > 1$ on a $R_f(\theta, \delta_c) \geq \theta = R_f(\theta, \delta_1)$). Par contre, les arguments ci-dessus montrent aussi que δ_c avec $c \leq 1$ est admissible : c^2 représente la pente à l'origine et $(c-1)^2$ le taux d'accroissement pour θ grand, et donc pour $0 \leq c \leq c' \leq 1$ on a $c^2 \leq (c')^2$ mais $(1-c)^2 \geq (1-c')^2$ et donc les courbes se croisent nécessairement. Le meilleur estimateur, qui minimise le risque intégré

$$R_B(\delta_c) = c^2 + 2(1-c)^2,$$

est obtenu pour c solution de $2c + 4(c-1) = 0$, i.e., $c = 2/3$. Si on prend la fonction de coût $L'(\theta, d) = (1 - d/\theta)^2 = \theta^{-2} L(\theta, d)$ on obtient

$$R'_f(\delta_c, \theta) = c^2\theta^{-1} + (c-1)^2.$$

Le raisonnement pour l'admissibilité reste le même, par contre le risque intégré vaut maintenant $+\infty$. ■

5.3 Approche bayésienne

Un fréquentiste est quelqu'un content de regarder d'autres données qu'il aurait pu avoir mais qu'il n'a pas eu.

Michael Jordan

La volonté de minimiser le risque fréquentiste est un peu paradoxal, puisqu'elle repose sur un comportement moyen des données, et non pas les données à disposition. A l'inverse, l'approche bayésienne cherche à minimiser le risque a posteriori.

Définition 5.14. Le risque a posteriori $\rho(\delta, \pi \mid x)$ est le risque intégré par rapport à la loi a posteriori :

$$\rho(\delta, \pi \mid x) = \mathbb{E}\left[L(\theta, \delta(x)) \mid x\right] = \int L(\theta, \delta(x))\pi(\theta \mid x)\mathrm{d}\theta.$$

De manière assez élégante, cette approche rejoint l'approche fréquentiste via le risque de Bayes ou risque intégré, qui consiste à intégrer le risque fréquentiste par rapport à la loi a priori. Le résultat suivant justifie cette terminologie, et montre notamment qu'un estimateur bayésien correspond à l'intégration de la fonction de coût par rapport à la loi a posteriori.

Théorème 5.15. *δ_π est une règle de décision bayésienne pour l'a priori π si et seulement si $\delta_\pi(x)$ minimise le risque a posteriori pour toute observation $x \in \mathscr{O}$, i.e., on a*

$$\delta_\pi(x) \in \arg\min_{d \in \Theta} \ \mathbb{E}\left[L(\theta, d) \mid x\right].$$

REMARQUE 5.16 En fait, on a rigoureusement $\mathbb{P}(\delta_\pi(x) \in \arg\min \cdots) = 1$, mais ici comme dans le reste du cours on passera sous silence ces considérations issues de la théorie de la mesure. ■

Démonstration du Théorème 5.15. Par le théorème de l'espérance totale, on a

$$R_B(\delta) = \mathbb{E}[L(\theta, \delta(\cdot))] = \mathbb{E}[\rho(\delta, \pi \mid x)] = \mathbb{E}[\mathbb{E}[L(\theta, \delta(x)) \mid x]].$$

Si $\delta_\pi(x)$ minimise $\mathbb{E}[L(\theta, d) \mid x]$ en d pour chaque x, alors $R_B(\delta_\pi) \leq R_B(\delta)$ pour toute règle de décision δ au vu de l'égalité précédente et δ_π est donc une règle de décision bayésienne. Réciproquement, soit δ_π une règle de décision bayésienne et $\delta \in \mathscr{D}$ une autre règle de décision avec $\delta(x) \in \arg\min_{d\in\Theta} \mathbb{E}[L(\theta, d) \mid x]$. Alors

$$0 \leq R_B(\delta) - \underline{R} = R_B(\delta) - R_B(\delta_\pi) = \mathbb{E}[\mathbb{E}[L(\theta, \delta(x)) \mid x] - \mathbb{E}[L(\theta, \delta_\pi(x)) \mid x]].$$

Par définition de δ_π on a $\mathbb{E}[L(\theta, \delta(x)) \mid x] - \mathbb{E}[L(\theta, \delta_\pi(x)) \mid x] \leq 0$ ce qui implique que $\mathbb{E}[L(\theta, \delta(x)) \mid x] = \mathbb{E}[L(\theta, \delta_\pi(x)) \mid x]$ pour presque tout x et donc $\delta_\pi(x) \in \arg\min_d \mathbb{E}[L(\theta, d) \mid x]$. □

EXEMPLE 5.17 (FIL ROUGE) On considère l'exemple du fil rouge avec une fonction de coût quadratique et des estimateurs de la forme $a^\top x$. Montrez que

$$\mathbb{E}(L(\theta, d) \mid x) = \mathbb{V}\mathrm{ar}(\theta \mid x) + [\mathbb{E}(\theta \mid x) - d]^2$$

et déduisez-en le risque a posteriori et l'estimateur bayésien.
Dans l'exemple fil rouge, on a alors

$$\rho(\delta, \pi \mid x) = \mathbb{E}\left[\left(\theta - a^T x\right)^2 \mid x\right] = \left(\mathbb{E}(\theta \mid x) - a^T x\right)^2 + \mathbb{V}\mathrm{ar}(\theta \mid x)$$

et puisque

$$\theta \sim \mathscr{N}\left(p\bar{x} + q\mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right) \quad \text{avec} \quad p = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}$$

par (1.3), on obtient

$$\rho(\delta, \pi \mid x) = \left(p\bar{x} + q\mu_0 - a^T x\right)^2 + \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

En particulier, puisque $p \to 1$, on voit que si l'on veut que le risque a posteriori disparaisse lorsque $n \to \infty$, il faut choisir une suite d'estimateurs a_n telle que $a_n^T x \to \mu_0$, par exemple $a = n^{-1}\mathbf{1}_n$ qui mène au risque a posteriori

$$\rho(\delta, \pi \mid x) = \left(\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\right)^2 (\mu_0 - \bar{x})^2 + \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}. \tag{5.1}$$

Pour un risque quadratique on a par définition

$$L(\theta, d \mid x) = \int (\theta - d)^2 f(\theta \mid x) \mathrm{d}\theta = \mathbb{V}\mathrm{ar}(\theta \mid x) + [\mathbb{E}(\theta \mid x) - d]^2$$

et donc

$$\delta_\pi(x) = \mathbb{E}(\theta \mid x) = p\bar{x} + q\mu_0$$

en utilisant (1.3) pour la dernière égalité. ■

5.4 Admissibilité et estimateurs bayésiens (*)

Proposition 5.18. *Une règle de décision bayésienne δ_π est admissible si au moins une des conditions suivantes est satisfaite :*

a) δ_π est l'unique estimateur bayésien ;

b) π admet une densité strictement positive sur Θ et $\theta \mapsto R_f(\theta, \delta)$ est continue pour tout $\delta \in \mathscr{D}$;

c) Θ est fini et $\pi(\{\theta\}) > 0$ pour tout $\theta \in \Theta$.

Démonstration de a). Soit δ une règle de décision telle que $R_f(\theta, \delta) \leq R_f(\theta, \delta_\pi)$ pour tout $\theta \in \Theta$. Puisque $R_B(\delta) = \int R_f(\theta, \delta)\pi(\theta)\mathrm{d}\theta$, en intégrant sous π on obtient $R_B(\delta) \leq R_B(\delta_\pi) = \inf_\delta R(\Pi, \delta)$, donc δ' est aussi une règle de Bayes pour π. Donc $\delta' = \delta$ par hypothèse, et donc $R_f(\theta, \delta) = R_f(\theta, \delta')$ ce qui montre que δ est admissible. □

Démonstration de b). Soit δ une règle de décision telle que $R_f(\theta, \delta) \leq R_f(\theta, \delta_\pi)$ pour tout $\theta \in \Theta$. Alors $\mathbb{P}(R_f(\theta, \delta) = R_f(\theta, \delta_\pi)) = 1$ puisque

$$0 \leq \int_\theta \big(R_f(\theta, \delta_\pi) - R_f(\theta, \delta)\big)\, \pi(\theta)\mathrm{d}\theta = R_B(\delta_\pi) - R_B(\delta) \leq 0$$

où la dernière inégalité provient de la définition de δ_π. On a donc que $R_f(\theta, \delta_\pi)\pi(\theta) = R_f(\theta, \delta)\pi(\theta)$ presque partout, et les hypothèses sur π et R_f impliquent $R_f(\theta, \delta_\pi) = R_f(\theta, \delta)$ pour tout θ. □

Démonstration de c). Le raisonnement est le même que précédemment. □

5.5 Minimaxité et estimateurs bayésiens (*)

Proposition 5.19. *Le risque de Bayes est toujours plus petit que le risque minimax, i.e., $\underline{R} \leq \overline{R}$.*

Démonstration. Pour toute règle de décision $\delta \in \mathscr{D}$ on a

$$\underline{R} \leq R_B(\delta) = \int R_f(\theta, \delta)\pi(\theta)\mathrm{d}\theta \leq \sup_\theta R_f(\theta, \delta)$$

si bien que $\underline{R} \leq \sup_\theta R_f(\theta, \delta)$ pour toute règle de décision δ, et donc $\underline{R} \leq \inf_\delta \sup_\theta R_f(\theta, \delta) = \overline{R}$. □

Définition 5.20. On dit que le problème d'estimation admet une valeur si $\overline{R} = \underline{R}$.

Théorème 5.21. *Soit δ_π une règle de décision bayésienne pour une distribution a priori π. Si*

$$R_B(\delta_\pi) = \sup_{\theta\in\Theta} R_f(\theta, \delta_\pi) \tag{5.2}$$

alors :

1. *δ_π est une règle minimax et le problème d'estimation admet une valeur ;*
2. *$\underline{R}(\pi) = \sup_{\pi'} \underline{R}(\pi')$, i.e., on dit que π est une loi la moins favorable ;*
3. *toute règle minimax est une règle bayésienne pour π. En particulier, si δ_π est l'unique règle de Bayes pour π, alors δ_π est l'unique règle minimax.*

Démonstration. Pour toute règle de décision δ on a (par hypothèse pour la première égalité, par définition de δ_π pour la première inégalité, par définition de R_B pour la seconde égalité et par croissance de l'espérance pour la seconde inégalité)

$$\sup_\theta R_f(\theta, \delta_\pi) = R_B(\delta_\pi) \leq R_B(\delta) = \mathbb{E}_\pi[R_f(\theta, \delta)] \leq \sup_\theta R_f(\theta, \delta)$$

ce qui montre bien que δ_π est minimax. De plus, si δ est une autre règle minimax, alors on a égalité dans les inégalités ci-dessous et donc on a en particulier $R_B(\delta) = R_B(\delta_\pi)$ et donc δ est une règle bayésienne.

Pour démontrer le deuxième point, on note que pour toute loi a priori π'

$$\underline{R}(\pi') = \inf_\delta R_B(\delta, \pi') \leq \inf_\delta \sup_\theta R_f(\theta, \delta) = \underline{R}(\pi)$$

où la dernière égalité vient du premier point. □

La condition 5.2 peut paraître assez forte, et effectivement elle l'est. Elle est satisfaite en particulier si $R_f(\theta, \delta_\pi)$ ne dépend pas de θ ce qui est par exemple le cas dans l'exemple suivant.

EXEMPLE 5.22 (FIL ROUGE) Comme on l'a déjà vu dans l'exemple 5.1, dans le cas non biaisé le risque fréquentiste ne dépend pas de θ et la règle de décision bayésienne est donc minimax, ce que l'on avait déjà vu dans l'exemple 5.12. ■

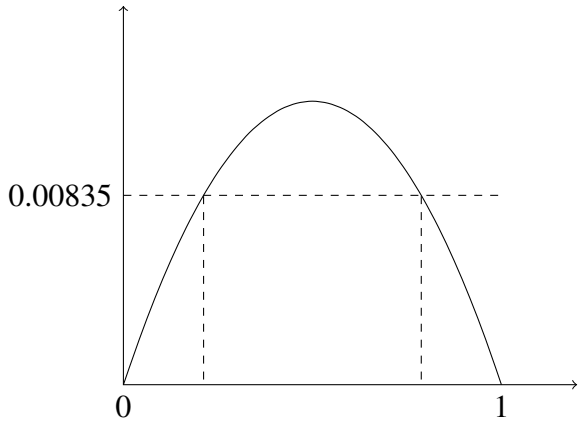

FIGURE 5.3 – Comparaison du risque minimax et du risque bayésien dans l'exemple **??**.

5.6 Estimateurs ponctuels classiques et lien avec la théorie de la décision

Etant donné la distribution a posteriori $\pi(\theta \mid x)$, de nombreux choix sont possibles si l'on souhaite faire une estimation ponctuelle : le mode, la moyenne, la médiane, etc. En fait, chaque estimateur provient d'une fonction de coût différente.

Proposition 5.23. *L'estimateur de Bayes est donné par :*

- *le mode* $\arg\max_{\theta\in\Theta}\pi(\theta \mid x)$ *dans le cas d'une fonction de coût binaire :* $L(\theta,d) = \mathbb{1}(d \neq \theta)$ *;*
- *l'espérance conditionnelle* $\mathbb{E}(\theta \mid x)$ *dans le cas d'une fonction de coût quadratique :* $L(\theta,d) = (\theta-d)^2$ *;*
- *une médiane de* $\pi(\theta \mid x)$ *dans le cas d'une fonction de coût* L_1 *:* $L(\theta,d) = |\theta-d|$.

Une fonction de coût binaire n'est bien définie que si Θ est fini, sinon dans le cas continu on aurait toujours un risque fréquentiste qui vaut 1. Par contre, le mode est toujours bien défini.

Définition 5.24. On dit que δ est un estimateur du maximum a posteriori (MAP) si $\delta(x) \in \arg\max_{\theta\in\Theta}\pi(\theta \mid x)$ pour tout $x \in \mathscr{O}$.

Lemme 5.25. *On a* $\arg\min_{m\in\mathbb{R}}\mathbb{E}((X-m)^2) = \mathbb{E}(X)$.

Démonstration. On a $\mathbb{E}((X-m)^2) = \mathbb{E}(X^2) - 2m\mathbb{E}(X) + m^2$. □

Lemme 5.26. *Soit X à densité : alors*

$$\arg\min_{m\in\mathbb{R}}\mathbb{E}(|X-m|) = \left\{x : \mathbb{P}(X<x) = \frac{1}{2}\right\}.$$

Démonstration. On a

$$\begin{aligned}\mathbb{E}(|X-m|) &= \int_{-\infty}^{m}(m-x)f(x)\mathrm{d}x + \int_{m}^{\infty}(x-m)f(x)\mathrm{d}x \\ &= m\mathbb{P}(X<m) - \int_{-\infty}^{m}xf(x)\mathrm{d}x + \int_{m}^{\infty}xf(x)\mathrm{d}x - m\mathbb{P}(X>m).\end{aligned}$$

Si on dérive par rapport à m, on obtient donc

$$\frac{\mathrm{d}}{\mathrm{d}m}\mathbb{E}(|X-m|) = \mathbb{P}(X<m) + mf(m) - mf(m) - mf(m) + mf(m) - \mathbb{P}(X>m)$$

ce qui donne le résultat. □

Démonstration pour $L(\theta,d) = \mathbb{1}(\theta \neq d)$. Dans ce cas, le risque fréquentiste est

$$R_f(\theta,\delta) = \mathbb{E}_\theta(L(\theta,\delta)) = \mathbb{P}_\theta(\theta \neq \delta);$$

le risque intégré est

$$R_B(\delta) = \mathbb{E}(L(\theta,\delta)) = \mathbb{P}(\theta \neq \delta);$$

et donc les estimateurs bayésiens maximisent $d \mapsto \mathbb{P}(\theta = d \mid x) = \pi(d \mid x)$. □

Démonstration pour $L(\theta,d)=(\theta-d)^2$. Dans ce cas, le risque fréquentiste est

$$R_f(\theta,\delta)=\mathbb{E}_\theta(L(\theta,\delta))=\mathbb{E}_\theta\left[(\theta-\delta)^2\right];$$

le risque intégré est

$$R_B(\theta,\delta)=\mathbb{E}(L(\theta,\delta))=\mathbb{E}\left[(\theta-\delta)^2\right];$$

et donc les estimateurs bayésiens minimisent $d\mapsto\mathbb{E}[(\theta-d)^2\mid x]$ qui est l'espérance conditionnelle par le Lemme 5.25 (ou par définition de l'espérance conditionnelle comme projection L_2). □

Démonstration pour $L(\theta,d)=|\theta-d|$. Dans ce cas, le risque fréquentiste est

$$R_f(\theta,\delta)=\mathbb{E}_\theta(L(\theta,\delta))=\mathbb{E}_\theta\,|\theta-\delta|;$$

le risque intégré est

$$R_B(\theta,\delta)=\mathbb{E}(L(\theta,\delta))=\mathbb{E}\,|\theta-\delta|;$$

et donc les estimateurs bayésiens minimisent $d\mapsto\mathbb{E}[|\theta-d|\mid x]$ qui est une médiane par le Lemme 5.26. □

EXEMPLE 5.27 (FIL ROUGE) Dans le cas gaussien, tous ces estimateurs coïncident. ■

EXEMPLE 5.28 On considère un modèle gaussien à variance σ^2 connue :

$$f(x\mid\theta)=\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\|x-\theta\mathbf{1}_n\|^2/(2\sigma^2)}.$$

Comme loi a priori on prend la loi de Cauchy $\pi(\theta)\propto 1/(\theta^2+1)$. Montrez que le MAP est la racine d'un polynôme du troisième degré que vous identifierez, et que la moyenne a posteriori peut s'écrire

$$\frac{\int\frac{\theta}{\theta^2+1}e^{-(\theta-a)^2/2}\mathrm{d}\theta}{\int\frac{1}{(\theta^2+1)}e^{-(\theta-a)^2/2}\mathrm{d}\theta}$$

pour un paramètre a que vous identifierez.
On considère un modèle gaussien à variance σ^2 connue :

$$f(x\mid\theta)=\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\|x-\theta\mathbf{1}_n\|^2/(2\sigma^2)}.$$

Si on prend comme a priori la loi de Cauchy $\pi(\theta)\propto 1/(\theta^2+1)$, on obtient

$$\pi(\theta\mid x)\propto\frac{e^{-\|x-\theta\mathbf{1}_n\|^2/2}}{\theta^2+1}$$

et là par contre, la loi a posteriori n'est plus gaussienne. Le MAP est alors donné par la résolution de l'équation

$$-\frac{n}{2\sigma^2}(\theta-\bar{x})-\frac{2\theta}{\theta^2+1}=0$$

et la moyenne a posteriori par

$$\frac{\int\frac{\theta}{\theta^2+1}e^{-\|x-\theta\mathbf{1}_n\|^2/2}\mathrm{d}\theta}{\int\frac{1}{(\theta^2+1)}e^{-\|x-\theta\mathbf{1}_n\|^2/2}\mathrm{d}\theta}.$$

■

6 — Les tests d'hypothèses bayésiens

Le cadre théorique des tests d'hypothèse est le suivant : on se donne deux sous-ensembles disjoints $\Theta_0, \Theta_1 \subset \Theta$ de l'ensemble des paramètres, et l'on souhaite décider si le vrai paramètre appartient à Θ_0 ou à Θ_1. Formellement, il s'agit donc essentiellement de faire de l'inférence sur la fonction $\mathbb{1}(\theta \in \Theta_0)$, ou bien, en terme de théorie de la décision, de prendre une décision à valeurs dans $\{0, 1\}$, mais ce cadre inférentiel présente certaines particularités propres qui motivent son étude séparée.

6.1 Cadre fréquentiste

Dans le cadre fréquentiste, **tous les raisonnements et calculs sont conditionnés par une hypothèse**. On se pose donc des questions du genre : *Si l'hypothèse nulle ou alternative est vraie, quelle est la probabilité de tel ou tel évènement ?* Par exemple, le risque de première espèce est la probabilité de rejeter l'hypothèse nulle en supposant qu'elle est vraie. Ainsi, un risque de première espèce faible garantit qu'on ne rejettera pas l'hypothèse nulle à tort. Le risque de deuxième espèce est quant à lui la probabilité de rejeter l'hypothèse alternative en supposant qu'elle est vraie. Ces deux risques sont antagonistes, i.e., avoir un risque de première espèce faible se traduit en général par un risque de deuxième espèce élevé. L'approche classique veut que l'on fixe le risque de première espèce, puis que l'on choisisse le test qui, à ce risque de première espèce fixé, présente le plus faible risque de deuxième espèce. Cette approche induit donc une asymétrie très forte entre les hypothèses nulle et alternative, qui ne jouent pas des rôles interchangeables. Ainsi, intervertir les deux rôles peut mener à des conclusions différentes. Avec cette approche, où le risque de première espèce est privilégié, il est cohérent de choisir une hypothèse nulle "conservative", i.e., telle que les conséquences de l'accepter à tort sont limitées (puisque le risque de la rejeter est faible).

Par ailleurs, les tests d'hypothèse fréquentistes sont souvent utilisés, ou tout du moins interprétés, à tort, et l'on est souvent enclins de parler de probabilité d'une hypothèse, par exemple d'interpréter un risque comme la probabilité de l'hypothèse : dans le cadre fréquentiste, cette interprétation n'a aucun sens. Néanmoins, dans le cadre bayésien cela devient tout à fait normal.

6.2 Cadre bayésien

6.2.1 Estimateurs bayésiens

On retrouve cela à l'aide de la théorie de la décision. Neyman et Pearson ont proposé le coût 0-1 :

$$L(\theta,d) = \begin{cases} 0 & \text{si } d = \mathbb{1}(\theta \in \Theta_0), \\ 1 & \text{sinon.} \end{cases}$$

Une approche plus générale consiste à pénaliser différemment les erreurs de type I et II, ce qui correspond à la fonction de coût suivante :

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

L'estimateur de Bayes associé est alors donné par le résultat suivant.

Proposition 6.1. *Sous le coût a_0–a_1 ci-dessus, l'estimateur de Bayes associé à la loi a priori π est*

$$\delta_\pi(x) = \begin{cases} 1 & \textit{si } \mathbb{P}(\theta \in \Theta_0 \mid x) > \frac{a_1}{a_0+a_1}, \\ 0 & \textit{sinon.} \end{cases}$$

Démonstration. On utilise le Théorème 5.15 : on a

$$\begin{aligned} \mathbb{E}[L(\theta,d) \mid x] &= a_0 \mathbb{1}(d=0)\mathbb{P}(\theta \in \Theta_0 \mid x) + a_1 \mathbb{1}(d=1)\mathbb{P}(\theta \in \Theta_1 \mid x) \\ &= a_0(1-d)\mathbb{P}(\theta \in \Theta_0 \mid x) + a_1 d \mathbb{P}(\theta \in \Theta_1 \mid x) \end{aligned}$$

et donc

$$\arg\min_{d\in\{0,1\}} \mathbb{E}[L(\theta,d) \mid x] = \begin{cases} 1 & \text{si } a_0\mathbb{P}(\theta \in \Theta_0 \mid x) \geq a_1 \mathbb{P}(\theta \in \Theta_1 \mid x) \\ 0 & \text{sinon} \end{cases}$$

□

On voit donc que la décision est uniquement basée sur la probabilité a posteriori que l'hypothèse soit vraie, ce qui est naturel dans le cadre bayésien. Par ailleurs, la décision ne dépend que de la fonction de coût via le ratio a_0/a_1, contrairement au cas fréquentiste qui, en plus de la fonction de coût, nécessite de fixer un seuil α. On remarque que plus a_0/a_1 est grand, i.e., plus une réponse incorrecte est pénalisée sous H_0 relativement à H_1, plus la probabilité a posteriori de H_0 doit être petite pour être rejetée.

EXEMPLE 6.2 (FIL ROUGE) Pour tester $H_0 : \theta < 0$ dans l'exemple du fil rouge, on calcule à l'aide du Lemme 1.5

$$\mathbb{P}(\theta < 0 \mid x) = \mathbb{P}(\tau N + \mu(\bar{x}) < 0 \mid x) = \phi(-\mu(\bar{x})/\tau)$$

avec $\mu(s) = ps + q\mu_0$ et $\tau^2 = 1/\sigma_0^2 + n/\sigma^2$, et H_0 est donc acceptée lorsque $-\mu(\bar{x}) > z_{a_0,a_1}\tau$ avec $\phi(z_{a_0,a_1}) = a_1/(a_0+a_1)$, i.e.,

$$\bar{x} < -\frac{\sigma^2}{n\sigma_0^2}\mu_0 - \left(1 + \frac{\sigma^2}{n\sigma_0^2}\right)\tau z.$$

■

6.2.2 Le facteur de Bayes

Définition 6.3. Le facteur de Bayes est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport des probabilités a priori de ces mêmes hypothèses, soit

$$B_{01}^{\pi}(x) = \frac{\mathbb{P}(\theta \in \Theta_0 \mid x)}{\mathbb{P}(\theta \in \Theta_1 \mid x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}$$

Ce rapport évalue la modification de la vraisemblance de l'ensemble Θ_0 par rapport à celle de l'ensemble Θ_1 due à l'observation et peut se comparer naturellement à 1, bien qu'une échelle de comparaison exacte doive être fondée sur une fonction de coût. Dans le cas particulier d'hypothèses simples où $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, le facteur de Bayes se simplifie et devient le rapport de vraisemblance classique

$$B_{01}^{\pi}(x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)}.$$

De manière plus générale, ce rapport peut être perçu comme un rapport de vraisemblance bayésien, car, si $\pi_i(\theta) \propto \mathbb{1}(\theta \in \Theta_i)\,\pi(\theta)$ est la loi a priori sous H_i, alors $B_{01}^{\pi}(x)$ peut s'écrire

$$\begin{aligned} B_{01}^{\pi}(x) &= \frac{\int_{\Theta_0} f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta_1} f(x \mid \theta)\pi(\theta)\mathrm{d}\theta} \Big/ \frac{\int_{\Theta_0} \pi(\theta)\mathrm{d}\theta}{\int_{\Theta_1} \pi(\theta)\mathrm{d}\theta} \\ &= \frac{\int_{\Theta_0} f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta_0} \pi(\theta)\mathrm{d}\theta} \Big/ \frac{\int_{\Theta_1} f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta_1} \pi(\theta)\mathrm{d}\theta} \\ &= \frac{\int_{\Theta_0} f(x \mid \theta)\pi_0(\theta)\mathrm{d}\theta}{\int_{\Theta_1} f(x \mid \theta)\pi_1(\theta)\mathrm{d}\theta} = \frac{m_0(x)}{m_1(x)} = \frac{f(x \mid \theta \in \Theta_0)}{f(x \mid \theta \in \Theta_1)} \end{aligned}$$

ce qui revient donc à remplacer les vraisemblances par des marginales sous les deux hypothèses.

Définition 6.4. L'échelle de Jeffreys est la suivante :

1. si $\log_{10}(B_{01}^{\pi})$ varie entre 0 et 0,5, la certitude que H_0 est vraie est **faible**, on opte pour H_1 ;
2. si elle varie entre 0,5 et 1, cette certitude est **substantielle** ;
3. si elle est entre 1 et 2, elle est **forte** ;
4. si elle est au-dessus de 2, elle est **décisive**,

avec la même échelle en faveur de H_0 pour les valeurs négatives.

6.2.3 Hypothèses nulles simples

Une hypothèse nulle simple du genre $\Theta_0 = \{\theta_0\}$ peut ne pas avoir de sens, par exemple il ne semble pas y avoir de sens à se demander si la probabilité qu'il pleuve demain vaut 0,7163891256… Dans certains cas néanmoins, cela peut avoir un sens, par exemple dans le cas discret, ou dans le cas de sélection de modèle si l'on teste la nullité d'un paramètre, ou encore, en astrophysique, tester si l'univers est en expansion, s'il se contracte ou s'il est stable

revient à tester si la constante de Hubble est plus grande, plus petite ou égale à une valeur spécifique $\mathfrak{h}_0$.

Dans le cas à densité, la probabilité a posteriori d'une telle hypothèse est toujours nulle et il faut donc modifier la démarche. Pour cela, on procède à une modification de la loi a priori, qui est une démarche qui s'applique de manière plus générale dès qu'on teste un ensemble de mesure nulle. Dans le cas présent, on se donne ρ_0 la probabilité a priori que $\theta = \theta_0$ et π_1 la densité alternative, si bien que la loi a priori s'écrit

$$\pi(\theta) = \rho_0 \mathbb{1}(\theta = \theta_0) + (1-\rho_0)\,\pi_1(\theta)$$

et la probabilité a posteriori de H_0 est donnée par

$$\pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)\rho_0}{\int f(x \mid \theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(x \mid \theta_0)\rho_0}{f(x \mid \theta_0)\rho_0 + (1-\rho_0)\int_{\Theta_1} f(x \mid \theta)\pi_1(\theta)\mathrm{d}\theta}$$

et le facteur de Bayes par

$$B^{\pi}_{01}(x) = \frac{f(x \mid \theta_0)}{m_1(x)}.$$

C'est ce que l'on obtient lorsque l'on considère $\Theta_0 = \{\theta : |\theta - \theta_0| \leq \varepsilon\}$ et que l'on fait tendre ε vers 0.

EXEMPLE 6.5 (FIL ROUGE) Si dans l'exemple fil rouge on cherche maintenant à tester $H_0 : \theta = 0$, il semble raisonnable de prendre $\pi_1 = \mathcal{N}(0, \sigma_0^2)$. Alors, puisque $\Theta_1 = \mathbb{R} \setminus \{0\}$,

$$m_1(x) = \int_{\mathbb{R}} f_\theta(x)\pi_1(\theta)\mathrm{d}\theta \propto \int_{\mathbb{R}} e^{-(x-\theta)^2/(2\sigma^2)} e^{-\theta^2/(2\sigma_0^2)}\mathrm{d}\theta$$

et donc

$$m_1(x) \propto e^{-x^2/(2\sigma^2)} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\theta^2 + \frac{1}{\sigma^2}x\theta\right)\mathrm{d}\theta.$$

Si on note $z = x/\sigma$ et $\alpha = \sigma^2/\sigma_0^2$ et qu'on fait le changement de variables $\theta \to \theta/\sigma$, on obtient

$$\begin{aligned}
m_1(x) &\propto e^{-z^2/2} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(1+\alpha)\,\theta^2 + z\theta\right)\mathrm{d}\theta \\
&= e^{-z^2/2} \int_{\mathbb{R}} \exp\left(-\frac{1+\alpha}{2}\left(\theta^2 - \frac{2z\theta}{1+\alpha}\right)\right)\mathrm{d}\theta \\
&= e^{-z^2/2} \int_{\mathbb{R}} \exp\left(-\frac{1+\alpha}{2}\left(\left(\theta - \frac{z}{1+\alpha}\right)^2 - \frac{z^2}{(1+\alpha)^2}\right)\right)\mathrm{d}\theta \\
&\propto e^{-z^2/2} e^{z^2/(2(1+\alpha))} = e^{-z^2/2\times\alpha/(1+\alpha)}
\end{aligned}$$

ce qui donne

$$\frac{m_1(x)}{f(x \mid 0)} = \frac{\sigma}{\sqrt{\sigma^2+\sigma_0^2}} \frac{e^{-x^2/(2(\sigma^2+\sigma_0^2))}}{e^{-x^2/(2\sigma^2)}} = \frac{1}{\sqrt{1+1/\alpha}} \exp\left(\frac{z^2}{2(1+\alpha)}\right)$$

ce qui donne pour la probabilité

$$\pi(\theta = 0 \mid x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{\sqrt{1+\alpha}} \exp\left(\frac{\alpha z^2}{2(1+\alpha)}\right)\right]^{-1}.$$

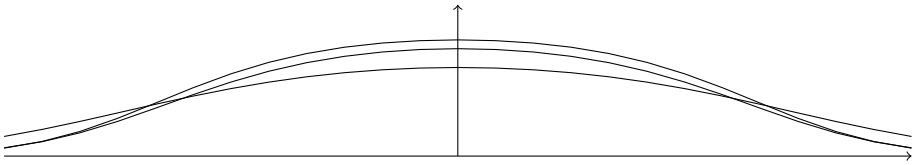

FIGURE 6.1 – Evolution de la probabilité a posteriori de H_0 en fonction de $z = x/\rho$ pour $\rho_0 = 1/2$ et différentes valeurs de $\alpha = \sigma^2/\sigma_0^2$.

La Figure 6.1 montre l'évolution de la probabilité de H_0 en fonction de $z = x/\rho$ pour $\rho_0 = 1/2$ et différentes valeurs de $\alpha = \sigma_0^2/\sigma^2$. Paradoxalement, pour certaines valeurs de z, augmenter la variance a priori augmente la probabilité d'accepter H_0. ■

6.2.4 Lois a priori impropres

Si l'on a ardemment défendu l'utilisation des lois a priori impropres pour l'inférence, ça n'est plus aussi clair dans le cadre des tests d'hypothèses. D'une part, le simple fait de tester des hypothèses, et pas n'importe lesquelles, est en contradiction avec l'absence d'information supposée à la base des lois a priori impropres. D'autre part, l'utilisation de lois impropres dans le cadre des tests d'hypothèses mènent à des contradictions apparemment superficielles : par exemple, les résultats ne sont plus invariants par multiplication de la loi a priori par une constante.

Nous illustrons ce point sur l'exemple du fil rouge. On souhaite tester $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$. Si nous utilisons la loi a priori impropre $\pi_1(\theta) = 1$, la loi a priori est alors

$$\pi(\theta) = \frac{1}{2}\mathbb{1}(\theta = 0) + \frac{1}{2}$$

et la probabilité a posteriori de H_0 est

$$\pi(\theta = 0 \mid x) = \frac{e^{-x^2/2}}{e^{-x^2/2} + \int e^{-(x-\theta)^2/2} \mathrm{d}\theta} = \frac{1}{1 + \sqrt{2\pi} e^{x^2/2}}$$

Ainsi, la probabilité a posteriori de H_0 est bornée par $1/(1+\sqrt{2\pi}) \approx 0{,}285$. Ceci implique que la loi a posteriori est plutôt biaisée contre H_0, même dans le cas le plus favorable.

En outre, une difficulté conceptuelle est que dans le cadre inférentiel précédent, les lois a priori impropres avaient été justifiées comme limite de lois propres, par exemple la mesure de Lebesgue comme limite de la densité normale avec une variance infinie. Néanmoins, dans le cadre des tests d'hypothèse cela n'est plus le cas. Par exemple, on a calculé dans l'Exemple 6.5 une probabilité a posteriori égale à

$$\pi(\theta = 0 \mid x) = \left[1 + \frac{1-\rho_0}{\rho_0} \frac{1}{\sqrt{1+\alpha}} \exp\left(\frac{\alpha z^2}{2(1+\alpha)}\right)\right]^{-1}.$$

Lorsque la variance a priori σ_0^2 tend vers $+\infty$, i.e., que $\alpha \to \infty$, cette probabilité tend vers 1 qui est à la fois inutile et différent de la réponse obtenue en considérant directement la loi a priori impropre. Ce comportement est une manifestation du paradoxe de Jeffreys-Lindley.

6.2.5 Régions de crédibilité

Définition 6.6. Une région de crédibilité de niveau $\alpha \in]0,1[$ est un sous-ensemble $R_\alpha = R_\alpha(x)$ de Θ tel que $\mathbb{P}(\theta \in R_\alpha \mid x) \geq 1-\alpha$. Une région Highest Posterior Density (HPD) de niveau α est une région de crédibilité de niveau α de la forme

$$\{\theta : \pi(\theta \mid x) \geq h\}$$

où h est le plus grand seuil tel que $\mathbb{P}(\theta \in R \mid x) \geq 1-\alpha$.

Une fois de plus, le fait que, dans la formulation bayésienne, θ ait une probabilité donnée d'appartenir à une région fixée R est plus attrayant que l'interprétation fréquentiste d'une région aléatoire ayant une probabilité donnée de contenir le paramètre inconnu θ.

Il est assez clair d'après sa définition qu'une région HPD est de volume minimal parmi les régions α-crédible. Une région HPD n'est en général pas connexe, mais on a le résultat suivant.

Proposition 6.7. *Si $\pi(\cdot \mid x)$ est continue et n'a qu'un seul maximum local, alors toute région HPD est connexe.*

Si la densité a posteriori admet un unique maximum local dans R, les région HPD sont des intervalles dont les bornes sont des quantiles de la loi a posteriori. Même si ces quantiles ont des formes explicites, comme pour certaines familles de lois conjuguées, calculer analytiquement des quantiles optimaux n'est pas forcément simple. On peut néanmoins obtenir une approximation des régions HPFD en utilisant le comportement asymptotiquement gaussien de la densité a posteriori.

Proposition 6.8. *Sous les hypothèses du théorème de Bernstein-von Mises (Théorème 4.15), l'intervalle*

$$I_\alpha = \left[\hat{\theta}^{\mathrm{MAP}} - q_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta}^{\mathrm{MAP}})n}}, \hat{\theta}^{\mathrm{MAP}} + q_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta}^{\mathrm{MAP}})n}}\right]$$

est un intervalle de crédibilité de niveau α asymptotiquement en probabilités : $\mathbb{P}(\theta \in I_\alpha \mid x) \xrightarrow{\mathbb{P}} 1-\alpha$.

On ne fournit qu'une ébauche de preuve : si $\Pi(\cdot \mid x)$ est la fonction de répartition associée à $\pi(\cdot \mid x)$, on a

$$\begin{aligned}\mathbb{P}(\theta \in I_\alpha \mid x) &= \mathbb{P}\left(\theta \leq \hat{\theta}^{\mathrm{MAP}} + \frac{q}{\sqrt{In}} \mid x\right) - \mathbb{P}\left(\theta \leq \hat{\theta}^{\mathrm{MAP}} - \frac{q}{\sqrt{In}} \mid x\right) \\ &= \mathbb{P}\left(\sqrt{n}(\theta - \hat{\theta}^{\mathrm{MAP}}) \leq \frac{q}{\sqrt{I}} \mid x\right) - \mathbb{P}\left(\sqrt{n}(\theta - \hat{\theta}^{\mathrm{MAP}}) \leq -\frac{q}{\sqrt{I}} \mid x\right)\end{aligned}$$

et donc le théorème de Bernstein von Mises (Théorème 4.15) donne

$$\mathbb{P}(\theta \in I_\alpha \mid x) \to \mathbb{P}(N \leq q \mid x) - \mathbb{P}(N \leq -q \mid x) = 1-\alpha.$$

A — Tableau de quelques lois absolument continues

Loi	Paramètres	Densité	Moyenne
Loi gaussienne	$\sigma > 0, \mu \in \mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ
Loi exponentielle	$\lambda > 0$	$\lambda e^{-\lambda x} \mathbb{1}(x > 0)$	$\frac{1}{\lambda}$
Loi exponentielle symétrique	$\lambda > 0$	$\frac{\lambda}{2} e^{-\lambda \lvert x \rvert}$	0
Loi de Cauchy	—	$\frac{1}{\pi(x^2+1)}$	Non définie
Loi Gamma	$\alpha, \beta > 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}(x > 0)$	$\frac{\alpha}{\beta}$
Loi Beta	$\alpha, \beta > 0$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}(0 < x < 1)$	$\frac{\alpha}{\alpha+\beta}$

TABLE A.1 – Densités des lois continues usuelles

Bibliography

[1] Keith Conrad. Probability distributions and maximum entropy. Available at https://kconrad.math.uconn.edu/blurbs/.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014.

[4] Richard D. Gill and Boris Y. Levit. Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.

[5] I. A. Ibragimov and R. Z. Has′minskiĭ. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin, 1981. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[6] Erwin Lutwak, Deane Yang, and Gaoyong Zhang. Moment-entropy inequalities. *The Annals of Probability*, 32(1B):757–774, 2004.

[7] Lawrence R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984.

[8] Christian P. Robert. *Le choix bayésien*. Statistique et probabilités appliquées. Springer, Paris, second edition, 2006.

[9] R. J. Samworth. Steins's paradox. *Eureka*, 62:38–41, 2012.

[10] Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995.

[11] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[12] E. Weinstein and A. J. Weiss. Lower bounds on the mean square estimation error. *Proceedings of the IEEE*, 73(9):1433–1434, Sept 1985.

[13] Ehud Weinstein and Anthony J. Weiss. A general class of lower bounds in parameter estimation. *IEEE Trans. Inform. Theory*, 34(2):338–342, 1988.

[14] Anthony J. Weiss and Ehud Weinstein. A lower bound on the mean-square error in random parameter estimation. *IEEE Trans. Inform. Theory*, 31(5):680–682, 1985.