ENSAE ParisTech

Troisième année
2009 – 2010

Statistique Bayésienne

Notes de cours

Judith Rousseau [1]
rousseau@ceremade.dauphine.fr

1. Ce support de cours a été rédigé par Mathias André et Alexis Eidelman, élèves en 2008 – 2009 puis relu et corrigé par Julyan Arbel, correspondant de statistiques. Il s'appuie notamment sur l'ouvrage de référence de la statistique bayésienne, disponible dans toute bonne bibliothèque : *Le Choix Bayésien - Principes et pratique* de Christian P. Robert [2].

Table des matières

Chapitre 1

Introduction : Les principes bayésiens

1.1 L'inférence bayésienne

Définition 1.1.1 *Modèle classique*
On se place dans un espace probabilisé paramétrique classique :

$$X \in (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Theta\})$$

$\mathcal{X}$ désigne l'espace des données, Θ celui des paramètres θ. Le but de l'analyse statistique est de faire de l'inférence sur θ, c'est-à-dire décrire un phénomène passé ou à venir dans un cadre probabiliste.

L'idée centrale de l'analyse bayésienne est de considérer le paramètre inconnu θ comme aléatoire : l'espace des paramètres Θ est muni d'une probabilité π tel que $(\Theta, \mathcal{A}, \pi)$ est un espace probabilisé. Nous noterons $\theta \sim \pi$. π est appelée *loi a priori*. Intuitivement et en termes informationnels, elle détermine ce qu'on sait **et** ce qu'on ne sait pas avant d'observer X.

Définition 1.1.2 *Modèle dominé*
Le modèle est dit dominé s'il existe une mesure commune dominante μ, c'est-à-dire pour tout θ, P_θ admet une densité par rapport à μ :[1]

$$f(X|\theta) = \frac{dP_\theta}{d\mu}$$

Cette fonction $\ell(\theta) = f(X|\theta)$, vue comme une fonction de θ une fois qu'on a observé un tirage de X, est appelée vraisemblance du modèle. C'est la loi de X conditionnellement à θ.

1. Pour des mesures σ-finies et en vertu du théorème de Radon-Nikodym, ceci est équivalent à être absolument continue par rapport à μ.

Définition 1.1.3 *Loi jointe et loi a posteriori*
Dans le cas d'un modèle dominé, la loi jointe de (X, θ) *s'écrit* $\lambda_\pi(X, \theta) = f(X|\theta)d\pi(\theta) = f(X|\theta)\pi(\theta)d\nu(\theta)$*, la dernière égalité étant valable dans le cas absolument continu par rapport à* ν*, la mesure de Lebesgue*[2]*. La loi a posteriori est définie par sa densité :*

$$d\pi(\theta|X) = \frac{f(X|\theta)d\pi(\theta)}{\int_\Theta f(X|\theta)d\pi(\theta)} \tag{1.1}$$

La quantité $m_\pi(X) = \int_\Theta f(X|\theta)d\pi(\theta)$ est la loi marginale de X et est une constante de normalisation de la loi *a posteriori*, indépendante de θ. Nous travaillerons donc très régulièrement à une constante multiplicative près : $\pi(X|\theta) \propto f(X|\theta)\pi(\theta)$. Nous ajoutons que par construction la loi *a posteriori* est absolument continue par rapport à la loi *a priori* π.

Exemple 1.1 *Dans le cas gaussien, à variance connue :* $X \sim N(\mu, \sigma^2)$ *et* $\theta = \mu$ *(*σ^2 *connnu) :*

$$\pi(\mu) = \frac{e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}}}{\tau\sqrt{2\pi}}$$

$$\begin{aligned}
\pi(\mu|X) &\propto e^{-\frac{(X-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}} \\
\pi(\mu|X) &\propto \exp\left(-\frac{1}{2}(\frac{1}{\sigma^2}+\frac{1}{\tau^2})\left(\mu - (\frac{X}{\sigma^2}+\frac{\mu_0}{\tau^2})(\frac{1}{\sigma^2}+\frac{1}{\tau^2})^{-1}\right)\right)
\end{aligned}$$

Ainsi

$$\pi(\mu|X) \sim N(X\frac{\tau^2}{\sigma^2+\tau^2} + \mu_0\frac{\sigma^2}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$$

On remarque sur cet exemple que la loi a posteriori est plus resserrée (pointée) que la loi a priori. Cela s'avère être intuitif : la loi a posteriori est la loi de θ *en ayant une information supplémentaire à savoir la donnée de* X*, l'incertitude sur* θ *ne peut donc que diminuer, en d'autres termes la variance diminue. En considérant* $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ *dans le cas indépendant et identiquement distribué, la loi a posteriori se centre sur* $\overline{X}_n$ *avec un nombre d'observations qui augmente. Dans ce cas, elle se rapproche du maximum de vraisemblance.*

1.2 Extension aux lois impropres

Nous généralisons l'approche précédente aux lois *a priori* impropres, ce qui est notamment utile dans les modèles non-informatifs.

2. $\mathcal{B}$ est la tribu borélienne dans $\mathbb{R}^p$.

Définition 1.2.1 *Loi impropre*
Une loi impropre est une mesure σ-finie[3] *qui vérifie* $\int_\Theta \pi(\theta)d\nu(\theta) = +\infty$.

L'utilisation d'une loi *a priori* impropre peut sembler saugrenue mais cela peut s'avérer particulièrement intéressant. Il est ainsi envisageable de travailler avec une loi normale centrée à grande variance pour approcher une « loi uniforme sur $\mathbb{R}$ » ; avec de bonnes propriétés, il est judicieux (et préconisé !) de travailler alors avec une loi impropre, la mesure de Lebesgue par exemple. Cependant une telle loi est utile du moins tant que la loi *a posteriori* existe. Aussi, on se limite aux lois impropres telles que :

$$m_\pi(X) = \int_\Theta f(X|\theta)d\pi(\theta) < \infty \quad \mu\text{-pp}$$

Exemple 1.2 *Loi uniforme généralisée*
– *Loi a priori uniforme*[4]

$$\begin{aligned} d\pi(\mu) &= d\mu \\ X &\sim N(\mu, \sigma^2) \\ \forall X\ m_\pi(X) &= \int e^{-\frac{(X-\mu)^2}{2\sigma^2}} d\mu = \sigma\sqrt{2\pi} < \infty \end{aligned}$$

La mesure de Lebesgue sur $\mathbb{R}$ est donc une loi impropre qui peut être utilisée.
– *Dans le cas suivant :*

$$\begin{aligned} X_1, \ldots, X_n &\sim N(\mu, \sigma^2) \\ \pi(\mu, \sigma) = \frac{1}{\sigma} \quad et \quad \Theta &= \mathbb{R} \times \mathbb{R}_+^* \end{aligned}$$

L'intégrale de la loi a priori s'écrit :

$$\int_{\mathbb{R}} \int_0^{+\infty} e^{-\frac{(\bar{X}_n-\mu)^2}{2\sigma^2}} e^{\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{2\sigma^2}} \frac{d\sigma d\mu}{\sigma^{n+1}} = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{2\sigma^2}} \frac{d\sigma}{\sigma^n}$$

Pour que l'expression converge, il faut avoir $n > 1$ et la propriété suivante vérifiée : $(\exists i, j; i \neq j, tqX_i \neq X_j)$. Lorsque $n > 1$ l'ensemble des vecteurs qui ne vérifient pas cette propriété est de mesure nulle et donc n'affecte pas la finitude de l'intégrale définie ci-dessus. Il est possible de donner une interprétation intuitive du résultat précédent. Pour estimer la dispersion (variance), au moins deux observations sont nécessaires. L'interprétation est peut-être un peu moins évidente en ce qui concerne la deuxième remarque mais elle reste cohérente si toutes les observations sont égales. L'inférence sur σ conduit alors à considérer ce paramètre comme nul ; dans un tel cas, la distribution a

3. Cela peut être défini dans un cas général sans σ-finitude mais, comme nous allons le voir, sans grand intérêt...
4. Dans ce cas, π est la mesure de Lebesgue.

priori $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$ *n'est pas définie, il n'est donc pas classique de la choisir de cette forme.*

Dans le cas où π est une mesure impropre σ-finie, on considère $\pi^* = c \cdot \pi$, cette constante arbitraire doit être sans influence pour être cohérent. C'est bien le cas car c se simplifie haut et bas dans l'équation (1.1) définissant la loi *a posteriori* de telle sorte que $d\pi^*(\theta|X) = d\pi(\theta|X)$.

En conclusion, l'usage de lois impropres *a priori* est justifié si la loi *a posteriori* est propre (non impropre [5]) car elle ne dépend pas de constante multiplicative de la loi *a priori*. Ceci est à rapprocher du principe de vraisemblance indiqué par Christian Robert. D'une manière plus générale, l'inférence bayésienne se base sur $\pi(\theta|X)$.

1.3 Extension aux modèles non dominés

Le paramètre a jusqu'à présent été choisi parmi les éléments de $\mathbb{R}^d$ avec d fini dans un modèle dominé. Il est envisageable de choisir $\Theta = [0,1]^{\mathbb{R}}$, l'ensemble des distributions sur $[0,1]$ ou encore l'ensemble des probabilités sur $\mathbb{R}$ (qui est non dominé).

Dans ce dernier cas, si $P \sim \pi$ une probabilité sur Θ, par exemple un processus de Dirichlet, on définit la loi *a posteriori* comme une version de la loi conditionnelle. Sur $(\mathbb{R}, \mathcal{B})$ et $(\Theta, \mathcal{A})$, pour $(A, B) \in \mathcal{A} \times \mathcal{B}$:

$$\int_A P(X \in B) d\pi(P) = \int_B \int_A m_\pi(X) d\pi(P|X)$$

La loi de Dirichlet, notée $\mathcal{D}(\alpha_1, \cdots, \alpha_p)$, est définie sur le simplexe $S_p = \{(x_1, \cdots, x_p), \sum_{i=1}^p x_i = 1, x_i \geq 0 \text{ pour } 0 \leq i \leq p\}$ par :

$$d\pi_\alpha(x) = x_1^{\alpha_1} \dots x_p^{\alpha_p} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}$$

Pour une mesure α sur $\mathbb{R}$, le processus de Dirichlet se note $\mathcal{D}(\alpha(.))$. En utilisant la topologie faible, on peut montrer pour toute partition de $\mathbb{R}$ $(B_1, \cdots, B_p)$, si $P \sim \mathcal{D}(\alpha(.))$ alors

$$(P(B_1), \cdots, P(B_p)) \sim \mathcal{D}(\alpha(B_1), \cdots, \alpha(B_p))$$

.

5. C'est-à-dire une loi de probabilité qui mesure les informations une fois les données observées.

Chapitre 2

Une introduction à la théorie de la décision

2.1 Fonction de perte et risque

Pour le modèle $X \in (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Theta\})$, on définit $\mathcal{D}$ l'ensemble des décisions possibles. C'est-à-dire l'ensemble des fonctions de Θ dans $g(\Theta)$ où g dépend du contexte :

- si le but est d'estimer θ alors $\mathcal{D} = \Theta$
- pour un test, $\mathcal{D} = \{0, 1\}$

La fonction de perte est une fonction mesurable de $(\Theta \times \mathcal{D})$ à valeurs réelles positives : $L : \Theta \times \mathcal{D} \to \mathbb{R}_+$. Elle est définie selon le problème étudié et constitue l'armature du problème statistique [1].

Définition 2.1.1 *Risque fréquentiste*
Pour $(\theta, \delta) \in \Theta \times \mathcal{D}$, le risque fréquentiste [2] est défini par :

$$\begin{aligned} R(\theta, \delta) &= E_\theta \left[L(\theta, \delta(X))) \right] \\ &= \int_X L(\theta, \delta(X)))f(X|\theta)d\mu(X) \end{aligned}$$

C'est une fonction de θ et ne définit donc par un ordre total sur $\mathcal{D}$ et ne permet donc pas de comparer toutes décisions et estimateurs. Il n'existe donc pas de meilleur estimateur dans un sens absolu. Ainsi, l'approche fréquentiste restreint l'espace d'estimation en préférant la classe des estimateurs sans biais dans laquelle il existe des estimateurs de risque uniformément minimal ; l'école bayésienne ne perd pas en généralité en définissant un risque *a posteriori*. L'idée est d'intégrer sur l'espace des paramètres pour pallier cette difficulté.

1. cf 2.2 pour des exemples.
2. Dans le cas d'une fonction de perte quadratique, il est appelé risque quadratique.

Définition 2.1.2 *Risque a posteriori*
Une fois données la loi a priori sur le paramètre et la fonction de perte, le risque a posteriori est défini par :

$$\begin{aligned}\rho(\pi,\delta|X) &= E^{\pi}(L(\theta,\delta(X)))|X) \\ &= \int_{\Theta} L(\theta,\delta(X)))d\pi(\theta|X)\end{aligned}$$

Ainsi, le problème change selon les données ; ceci est dû à la non existence d'un ordre total sur les estimateurs.

Définition 2.1.3 *Risque intégré*
A fonction de perte donnée, le risque intégré est défini par :

$$r(\pi,\delta) = \int_{\Theta} R(\theta,\delta)d\pi(\theta)$$

Définition 2.1.4 *Estimateur bayésien*
Un estimateur bayésien est un estimateur vérifiant :

$$r(\pi,\delta^{\pi}) = \inf_{\delta\in\mathcal{D}} r(\pi,\delta) < \infty$$

Pour obtenir la valeur de l'infimum du risque intégré il faut donc en théorie minimiser une intégrale double δ. L'introduction du risque intégré se justifie par le théorème suivant. Il suffira de minimiser une grandeur qui ne dépend plus que des données, ceci permet donc d'arriver à des estimateur satisfaisants.

Théorème 2.1 *Méthode de calcul*
Si $\exists\delta\in\mathcal{D}, r(\pi,\delta) < \infty$ et $\forall X\in\mathcal{X}\delta^{\pi}(X) = \text{Argmin}_{\delta}\rho(\pi,\delta|X)$,
alors $\delta^{\pi}(X)$ est un estimateur bayésien.
Démonstration :

$$\begin{aligned}r(\pi,\delta) &= \int_{\Theta} R(\theta,\delta)d\pi(\theta) \\ &= \int_{\Theta}\int_{X} L(\theta,\delta(X)))f(X|\theta)d\mu(X)d\pi(\theta) \\ &= \int_{X}\int_{\Theta} L(\theta,\delta(X)))\frac{f(X|\theta))d\pi(\theta)}{m_{\pi}(X)}m_{\pi}(X)d\mu(X) \quad \text{(Fubini)} \\ &= \int_{X}\int_{\Theta} L(\theta,\delta(X)))d\pi(\theta|X)m_{\pi}(X)d\mu(X) \\ &= \int_{X} \rho(\pi,\delta|X)m_{\pi}(X)d\mu(X)\end{aligned}$$

Ainsi, pour $\delta\in\mathcal{D}$, $\rho(\pi,\delta^{\pi}|X) \leq \rho(\pi,\delta|X) \Rightarrow r(\pi,\delta^{\pi}) \leq r(\pi,\delta)$. Ce qui permet de conclure. ∎

2.2 Exemples

Exemple 2.1 *Perte quadratique*
$\mathcal{D} = \Theta \subset \mathbb{R}^d$ [3] et $L(\theta, \delta) = \|\theta - \delta\|^2$

Comme la norme au carré est une fonction convexe deux fois dérivable sur Θ, pour trouver δ^π estimateur bayésien, il suffit de déterminer les points critiques du risque *a posteriori* [4].

$$\begin{aligned} \rho(\pi, \delta|X) &= E^\pi(\|\theta - \delta\|^2 |X) \\ \frac{\partial \rho(\pi, \delta|X)}{\partial \delta} &= -2\int_\Theta (\pi - \delta(X)) d\pi(\theta, X) \\ \text{Donc} \quad \frac{\partial \rho(\pi, \delta|X)}{\partial \delta} = 0 &\iff \delta(X) = E^\pi(\theta|X) \end{aligned}$$

D'après l'inégalité de Jensen, $\delta \longmapsto \rho(\pi, \delta|X)$ est aussi convexe. L'estimateur bayésien vaut donc $\delta^\pi(X) = E^\pi(\theta|X) \quad \mu$-pp. Dans le cas gaussien de l'exemple 1.1 $X \sim N(\mu, \sigma^2)$ et $\mu \sim N(\mu_0, \tau)$, l'estimateur bayésien s'exprime :

$$\delta^\pi(X) = \frac{\tau^2}{\tau^2 + \sigma^2} X + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu_0$$

Exemple 2.2 *Perte L^1*
$\mathcal{D} = \Theta \subset R$ et $L(\theta, \delta) = \sum_{i=1}^d |\theta_i - \delta_i|$

Dans le cas simple où $d = 1$:

$$\begin{aligned} \rho(\pi, \delta|X) &= \int_\Theta |\theta - \delta| \, d\pi(\theta|X) \\ &= \int_{-\infty}^{\delta} (\delta - \theta)\pi(\theta|X)d\theta + \int_\delta^\infty (\theta - \delta)\pi(\theta|X)d\theta \end{aligned}$$

Comme $\delta \longmapsto \rho(\pi, \delta|X)$ est convexe et dérivable μ-presque partout, il suffit là encore de déterminer les points critiques :

$$\begin{aligned} \frac{\partial \rho(\pi, \delta|X)}{\partial \delta} &= \int_{-\infty}^{\delta} \pi(\theta|X)d\theta - \int_\delta^\infty \pi(\theta|X)d\theta \\ \text{Donc} \quad \frac{\partial \rho(\pi, \delta|X)}{\partial \delta} = 0 &\iff P^\pi(\theta \leq \delta|X) = P^\pi(\theta \geq \delta|X) \end{aligned}$$

C'est-à-dire $\delta^\pi(X)$ est la médiane de $\pi(\theta|X)$ [5].

3. Ceci peut facilement être généralisé au cas d'un Hilbert.
4. En s'appuyant sur le théorème 2.1 et pour des solutions intérieures évidemment...
5. Nous soulignons la robustesse vis à vis des valeurs extrêmes en comparaison de la moyenne, ceci s'explique par les fonctions de perte correspondantes qui accordent plus ou moins d'importance aux valeurs élevées.

Exemple 2.3 *Perte 0–1*
Cette fonction de perte est utilisée dans le contexte des tests. Un test est la donnée d'une partition de Θ en Θ_0 et Θ_1 [6]. $\theta \in \Theta_i$ correspond à l'hypothèse H_i, H_0 est appelée l'*l'hypothèse nulle*. Le principe du test (décisions) δ est défini comme suit :

$$\delta = \begin{cases} 0 & \text{si} \quad \theta \in \Theta_1 \\ 1 & \text{si} \quad \theta \in \Theta_0 \end{cases}$$

La fonction de perte correspondant au test est définie par :

$$L(\theta, \delta) = \mathbb{1}_{\theta \in \Theta_1} \times \mathbb{1}_{\delta=0} + \mathbb{1}_{\theta \in \Theta_0} \times \mathbb{1}_{\delta=1}$$

Le risque *a posteriori* est alors le suivant :

$$\rho(\pi, \delta|X) = \mathbb{1}_{\delta=0} P^\pi(\Theta_1 \,|X) + \mathbb{1}_{\delta=1} P^\pi(\Theta_0 \,|X)$$

Ainsi :

$$\delta^\pi(X) = 1 \Leftrightarrow P^\pi(\Theta_0 \,|X) \leq P^\pi(\Theta_1 \,|X)$$

C'est-à-dire que l'estimation permet d'accepter H_0 si c'est l'hypothèse la plus probable *a posteriori*, ce qui est une réponse naturelle.

Une variante du test 0–1 est le test de Neymann Person qui permet de distinguer risques de première et de deuxième espèce :

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

Le risque *a posteriori* est donné par :

$$\rho(\pi, \delta|X) = a_0 \delta P^\pi(\Theta_1 \,|X) + a_1(1-\delta) P^\pi(\Theta_0 \,|X)$$

Ainsi :

$$\delta^\pi(X) = 1 \Leftrightarrow a_0 P^\pi(\Theta_0 \,|X) \leq a_1 P^\pi(\Theta_1 \,|X)$$

Ceci permet de modéliser des raisonnements de la forme « j'ai plus ou moins tort » en jouant sur le rapport $\frac{a_1}{a_0}$ [7] et accorder plus ou moins d'importance relative à Θ_0 selon des arguments *a priori*.

2.3 Fonction de perte intrinsèque

Nous nous plaçons dans le cas où $X|\theta \sim f(X|\theta)$ et $L(\theta, \delta) = d(f_\theta, f_\delta)$ avec d une distance entre densités.

Avant de voir la définition de fonction de perte intrinsèque, nous commençons par des exemples de distances classiques entre densités.

6. $\Theta_0 \cup \Theta_1 = \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$.
7. $a_1 = a_0$ correspond au cas précédent de test 0–1.

Exemple 2.4 *Distances usuelles entre densités $(f_\theta, f_{\theta'})$ de fonctions de répartition $(F_\theta, F_{\theta'})$:*

1. *Distance de Kolmogoroff-Smirnoff :*
$$d_{KS}(f_\theta, f_{\theta'}) = \sup_x |F_\theta(x) - F_{\theta'}(x)|$$
2. *Distance L^1 :*
$$\begin{aligned} d_1(f_\theta, f_{\theta'}) &= \int |f_\theta(x) - f_{\theta'}(x)|\, dx \qquad (2.1) \\ &= 2 \sup_A |P_\theta(A) - P_{\theta'}(A)| \qquad (2.2) \end{aligned}$$
3. *Distance de Hellinger :*
$$d_H(f_\theta, f_{\theta'}) = \left(\int (\sqrt{f_\theta(x)} - \sqrt{f_{\theta'}(x)})^2 dx \right)^{\frac{1}{2}}$$
4. *Pseudo-distance*[8] *de Kullback-Liebler :*
$$K(f_\theta, f_{\theta'}) = \int f_\theta(x) \log \frac{f_\theta(x)}{f_{\theta'}(x)} dx$$
Avec l'inégalité de Jensen, on prouve l'inégalité $K(f_\theta, f_\delta) \geq 0$. De plus, $K(f_\theta, f_\delta) = 0$ si et seulement si $f_\theta = f_{\theta'}$ μ-presque sûrement.
5. *Distance L^2 :*
$$d_2(f_\theta, f_{\theta'}) = \int (f_\theta(x) - f_{\theta'}(x))^2 dx$$
Ceci peut s'utiliser si les densités sont de carré intégrable.

La notion de perte intrinsèque provient de l'exigence d'invariance par transformation monotone inversible sur les données : les distances et donc les fonctions de perte, utilisées doivent donc être inchangées par l'action d'un $\mathcal{C}^1$-difféomorphisme sur $\mathcal{X}$. Ainsi, pour $y = g(x)$ et $X \sim f_\theta$ où g est un $\mathcal{C}^1$-difféomorphisme, on note $y \sim g_\theta(y) = f_\theta(g^{-1}(y)) \left| \frac{dx}{dy} \right|$. Pour une perte intrinsèque, la distance entre f_θ et $f_{\theta'}$ est la même que celle entre g_θ et $g_{\theta'}$ pour un g donné ; ceci correspond à une distance entre distributions (et non plus entre densités).

L'invariance par $\mathcal{C}^1$-difféomorphisme est vérifiée pour les distances précédentes à l'exception de la norme L^2. Nous donnons à titre d'exemple la démonstration pour la norme L^1. Les autres cas sont laissés aux lecteurs et lectrices :

$$\begin{aligned} d_1(g_\theta, g_{\theta'}) &= \int \left| f_\theta(g^{-1}(y)) - f_{\theta'}(g^{-1}(y)) \right| \left| \frac{dx}{dy} \right| dy \\ &= \int |f_\theta(x) - f_{\theta'}(x)|\, dx \\ &= d_1(f_\theta, f_{\theta'}) \end{aligned}$$

8. Car non symétrique.

Définition 2.3.1 *Fonction de perte intrinsèque*
En s'appuyant sur ce qui précède, une fonction de perte intrinsèque est une fonction de perte définie à partir d'une distance entre distributions, c'est-à-dire invariante par transformation monotone inversible.

Nous terminons cette section par un exemple où nous calculons la distance de Hellinger entre deux gaussiennes.

Exemple 2.5 *Distance d'Hellinger de lois normales*
En notant Φ_{μ,σ^2} la densité de la loi $\mathcal{N}(\mu,\sigma^2)$ et $\theta=(\mu,\sigma^2)$, nous avons :

$$\begin{aligned} d_H^2(\Phi_\theta,\Phi_{\theta'}) &= \int_{\mathbb{R}}(\sqrt{\Phi_\theta}-\sqrt{\Phi_{\theta'}})^2dx \\ &= 2-2\int_{\mathbb{R}}\sqrt{\Phi_\theta-\Phi_{\theta'}}dx \\ &= 2-\frac{2}{\sqrt{2\pi\sigma\sigma'}}\underbrace{\int_{\mathbb{R}}exp(-\frac{(x-\mu)^2}{4\sigma^2}-\frac{(x-\mu')^2}{4\sigma'^2})dx}_{=I} \end{aligned}$$

$$I=\int_{\mathbb{R}}exp\left(-\frac{1}{4\Sigma^2}\{x-(\frac{\mu}{\sigma^2}+\frac{\mu}{\sigma'^2})\Sigma^2\}^2-\frac{\Sigma^2}{4}(\frac{\mu}{\sigma^2}+\frac{\mu}{\sigma'^2})^2+\frac{\mu^2}{4\sigma^2}+\frac{\mu'^2}{4\sigma'^2}\right)dx$$

en notant $\Sigma^2=(\frac{1}{\sigma^2}+\frac{1}{\sigma'^2})^{-1}$. D'où, en reconnaissant une intégrale gaussienne de variance $2\Sigma^2$ et en factorisant le facteur exponentiel :

$$I=\Sigma\sqrt{2\pi}e^{-\frac{(\mu-\mu')^2}{4(\sigma^2+\sigma'^2)}}$$

Ainsi, nous arrivons à :

$$\begin{aligned} \frac{1}{2}d_H^2(\Phi_\theta,\Phi_{\theta'}) &= 1-\sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}}\exp\left(-\frac{(\mu-\mu')^2}{4(\sigma^2+\sigma'^2)}\right) \\ &= 1-\sqrt{\frac{2(u+1)}{(u+1)^2+1}}\left(1-\frac{h^2}{4(\sigma^2+\sigma'^2)}\right)+o(h^2) \end{aligned}$$

en notant $h=\mu-\mu'$ *et* $u=\dfrac{\sigma}{\sigma'}-1$ *(développement limité en (h,u))*

C'est la somme de deux termes du second ordre :

$$1-\sqrt{\frac{2(u+1)}{(u+1)^2+1}}=\frac{u^2}{((u+1)^2+1)+\sqrt{2(u+1)((u+1)^2+1)}}\sim\frac{u^2}{4}$$

$$et\qquad\sqrt{\frac{2(u+1)}{(u+1)^2+1}}\frac{h^2}{\sigma'^2(1+(u+1)^2)}\sim\frac{h^2}{8\sigma'^2}$$

d'où

$$d_H^2(\Phi_\theta,\Phi_{\theta'})=\frac{2(\sigma-\sigma')^2+(\mu-\mu')^2}{4\sigma'^2}+o(\|h,u\|^2)$$

La partie régulière de ce développement limité s'écrit $(\theta - \theta')' \frac{I(\theta)}{2}(\theta - \theta')$ *où* $I(\theta)$ *est l'information de Fisher du paramètre* $\theta = (\mu, \sigma)$ *; ceci est l'expression d'une forme quadratique au voisinage de* θ*. L'information de Fisher s'interprète donc comme la courbure (locale) du modèle*[9]*. Intuitivement, cela s'explique par le fait que si l'information est grande, elle permet de distinguer deux paramètres distincts (et réciproquement), c'est une quantité intrinsèque au modèle et explique en partie la récurrence avec laquelle elle apparaît dans les cours de statistique.*

2.4 Admissibilité et minimaxité

2.4.1 Admissibilité

Définition 2.4.1 *Estimateur admissible*
Soit $X \in (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Theta\})$ un modèle paramétrique et L une fonction de perte sur $\Theta \times \mathcal{D}$ où $\mathcal{D}$ est l'ensemble des décisions (fonctions des données vers une transformation de l'ensemble des paramètres). On dit que $\delta \in \Theta$ est inadmissible si et seulement si $(\exists \delta_0 \in \mathcal{D}, \forall \theta \in \Theta R(\theta, \delta) \geq R(\theta, \delta_0)$ et $\exists \theta_0 \in \Theta, R(\theta_0, \delta) > R(\theta_0, \delta_0))$. Un estimateur est dit admissible si et seulement si il n'est pas inadmissible.

Même si les estimateurs admissibles apparaissent comme ceux qu'il faut rechercher en priorité, le gain de l'admissibilité peut s'avérer réduit et coûteux. Ainsi, dans le cas gaussien $X \sim \mathcal{N}(\theta, Id)$ et $(X, \theta) \in (\mathbb{R}^p)^2$, on peut montrer que le maximum de vraisemblance est admissible si et seulement si $p \leq 2$ pour la fonction de perte quadratique. En outre, le gain à être admissible augmente avec p, la dimension de l'espace.

Théorème 2.2 *Estimateurs bayésiens admissibles*
Si l'estimateur bayésien δ^π associé à une fonction de perte L et une loi *a priori* π est unique, alors il est admissible.
Démonstration : Supposons δ^π estimateur bayésien non admissible : $\exists \delta_0 \in \mathcal{D}, \forall \theta \in \Theta,\ R(\theta, \delta) \geq R(\theta, \delta_0)$ et $\exists \theta_0 \in \Theta, R(\theta_0, \delta) > R(\theta_0, \delta_0)$. En intégrant la première inégalité :

$$\int_\Theta R(\theta, \delta_0) d\pi(\theta) \leq \int_\Theta R(\theta, \delta^\pi) d\pi(\theta) = r(\pi)$$

donc δ_0 est aussi un estimateur bayésien associé à L et π et $\delta_0 \neq \delta^\pi$ d'après la seconde inégalité. Le théorème se déduit par contraposée. ■ Ce théorème s'applique notamment dans le cas d'un risque fini et d'une fonction de coût convexe. En outre, l'unicité de l'estimateur bayésien implique la finitude du risque : $r(\pi) = \int R(\theta, \delta^\pi) d\pi(\theta) < \infty$ (sinon, tout estimateur minimise le risque).

9. Il est nécessaire de se placer dans le cadre de la géométrie différentielle et d'étudier la variété de Riemann de géométrie l'information de Fisher.

Définition 2.4.2 *π-admissibilité*
Un estimateur δ_0 est π-admissible si et seulement si

$$\forall(\delta,\theta), R(\theta,\delta) \leq R(\theta,\delta_0) \Rightarrow \pi(\{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0)\}) = 0$$

Cette notion est utile pour démontrer le théorème important de cette section par le biais de la proposition suivante.

Proposition 2.3 *Tout estimateur bayésien tel que $r(\pi) < \infty$ est π-admissible*
Démonstration : Soit δ^π un estimateur bayésien à risque fini. Pour δ_0 tel que $\forall\theta R(\theta,\delta) \leq R(\theta,\delta_0)$, on note $A = \{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0)\}$. Nous avons alors :

$$\int_\Theta R(\theta,\delta_0)d\pi(\theta) - \int_\Theta R(\theta,\delta^\pi)d\pi(\theta) = \int_A (R(\theta,\delta) - R(\theta,\delta_0))d\pi(\theta) \leq 0$$

avec égalité si et seulement si $\pi(A) = 0$. Or, comme δ^π est bayésien et le risque fini, $r(\theta,\delta_0) \geq r(\theta,\delta^\pi)$ donc l'intégrale est nulle (positive et négative !), d'où $\pi(A) = 0$: δ^π est π-admissible. ∎

Nous pouvons maintenant énoncer une condition suffisante d'admissibilité des estimateurs bayésiens.

Théorème 2.4 *Continuité et π-admissibilité*
Si $\pi > 0$ sur Θ, $r(\pi) < \infty$ pour une fonction de perte L donnée, si δ^π estimateur bayésien correspondant existe et si $\theta \longmapsto R(\theta,\delta)$ est continu, alors δ^π est admissible.
Démonstration : Supposons que δ^π est non admissible. D'après la proposition précédent, δ^π est π-admissible. Ainsi, il existe δ_0 tel que pour tout θ, $R(\theta,\delta_0) \leq R(\theta,\delta^\pi)$ et $\theta_0 \in \Theta$, $R(\theta_0,\delta_0) < R(\theta_0,\delta^\pi)$. La fonction définie sur Θ par $\theta \longmapsto R(\theta,\delta_0) - R(\theta,\delta^\pi)$ est continue. Donc il existe un voisinage (ouvert) de θ_0, $\mathcal{V}_0 \subset \Theta$ tel que $\forall\theta \in \mathcal{V}_0, R(\theta,\delta_0) < R(\theta,\delta^\pi)$. En considérant A la même région que dans la proposition précédente, $A = \{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0)\}$, $\pi(A) \geq \pi(\mathcal{V}_0)$. Or π est supposée strictement positive que Θ donc en prenant un modèle dominé par une mesure qui charge positivement les ouverts (la mesure de Lebesgue par exemple) $\pi(\mathcal{V}_0) > 0$, A est donc non négligeable (de mesure non nulle), ce qui n'est pas conforme avec la π-admissibilité. En conclusion, δ^π est admissible. ∎

Exemple 2.6 *Discussion sur l'hypothèse de risque fini*
L'hypothèse $r(\pi) < \infty$ est cruciale, comme le montre l'exemple suivant. Dans le cas gaussien, $X \sim \mathcal{N}(\theta, I_d)$ et $(X,\theta) \in \mathbb{R}^{p2}$, pour une fonction de perte $L(\theta,\delta) = (\|\theta\|^2 - \delta)^2$ et pour la mesure de Lebesgue ($\pi(\theta) = 1$), on montre que $\delta^\pi(X) = \mathbb{E}^\pi[\|\theta\|^2|X] = \|X^2\| + p$. En prenant une loi a posteriori gaussienne $\pi(\theta|X) \sim \mathcal{N}(X, I_d)$, $\pi(\theta|X) \propto exp(-\frac{\|X-\theta\|^2}{2})$ et on peut montrer que pour $\delta_0(X) = \|X\|^2 - p$, $\forall\theta \in \Theta, R(\theta,\delta_0) \leq R(\theta,\delta^\pi)$ avec inégalité stricte pour $\theta = 0$. Ici, $r(\pi) = \infty$ met en défaut le théorème précédent.

2.4.2 Minimaxité

Définition 2.4.3 *Estimateur randomisé*
Pour le modèle $X \in (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Theta\})$, un ensemble de décisions $\mathcal{D}$, on définit $\mathcal{D}^*$ comme l'ensemble des probabilités sur $\mathcal{D}$. $\delta^* \in \mathcal{D}^*$ est appelé estimateur randomisé.

L'idée à l'origine de cette notion est de rendre $\mathcal{D}$ convexe pour pouvoir maximiser facilement. L'estimateur de Neyman-Pearson introduit dans la section 2.2 s'appuie sur ce principe.

On définit de même des fonctions de perte et risque randomisés par :

$$L^*(\delta^*, \theta) = \int_{\mathcal{D}} L(\theta, a) d\delta^*(a) \quad \text{et} \quad R(\theta, \delta^*) = \mathbb{E}_\theta[L^*(\delta^*, \theta)]$$

Définition 2.4.4 *Risque minimax et estimateur minimax*
Le risque minimax est défini par $\overline{R} = \inf_{\delta^* \in \mathcal{D}^*} \sup_{\theta \in \Theta} R(\theta, \delta^*)$. On dit que δ_0 est un estimateur minimax si et seulement si $\overline{R} = \sup_{\theta \in \Theta} R(\theta, \delta_0)$.

L'estimateur minimax correspond au point de vue (conservateur !) de faire le mieux dans le pire des cas, c'est-à-dire à s'assurer contre le pire. Il est utile dans des cadres complexes [10] mais trop conservateur dans certains cas où le pire est très peu probable. Il peut être judicieux de voir l'estimation comme un jeu entre le statisticien (choix de δ) et la Nature (choix de θ), l'estimation minimax rejoint alors celle de la Théorie des Jeux.

Théorème 2.5 *Valeur et risque minimax*
En se plaçant dans le cadre précédent pour $\Theta \subset \mathbb{R}$ et Θ compact, on note $\underline{R} = \sup_{\pi \ proba} r(\pi)$; si pour π_0 tel que $r(\pi_0) = \overline{R}$, δ^{π_0} est un estimateur bayésien, alors, en notant $f = \theta \longmapsto R(\theta, \delta^{\pi_0})$:

$$(f \text{ analytique}) \Rightarrow (f \text{ constante ou } supp(\pi) \text{ [11] fini})$$

Démonstration : Vérifions dans un premier temps le résultat général : $\underline{R} \leq \overline{R}$.

$$r(\pi) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq r(\pi, \delta) \leq \int_\Theta R(\theta, \delta) d\pi(\theta) \leq \max_{\theta \in \Theta} R(\theta, \delta) \text{ si } \delta \in \mathcal{D}$$

Donc pour tout $\delta \in \mathcal{D}$, $r(\pi) \leq \max_{\theta \in \Theta} R(\theta, \delta)$. Par combinaison convexe [12] sur $\mathcal{D}$, on passe à l'infimum sur $\mathcal{D}^*$. D'où $r(\pi) \leq \overline{R}$, et pour conclure on passe au sup sur π une probabilité.

10. Par exemple, dans le cas non paramétrique ; il est alors utile de regarder la vitesse de convergence de cet estimateur avec le nombre d'observations.

11. support de la loi π, i.e. l'ensemble des estimateurs θ tels que $\pi(\theta) > 0$.

12. argument un peu rapide mais clair dans le cas fini et s'étend par densité dans le cas continu.

Plaçons nous maintenant dans le cadre du théorème et considérons δ_0^π l'estimateur minimax bayésien : $r(\pi) = r(\pi, \delta_0^\pi)$. Supposons le risque non constant : il existe (θ_1, θ_2) tel que $R(\theta_1, \delta_0) < R(\theta_2, \delta_0)$. Comme Θ compact, le maximum est atteint : il existe θ_0 tel que $R(\theta_0, \delta_0) = \max_{\theta \in \Theta} R(\theta, \delta_0)$, on définit $A = \{\theta \in \Theta, R(\theta, \delta_0) < R(\theta_0, \delta_0)\}$. En sachant que $r(\pi_0, \delta^{\pi_0}) \leq \overline{R}$ et δ_0 estimateur minimax, on écrit :

$$r(\pi_0, \delta^{\pi_0}) = \int_A R(\theta, \delta^{\pi_0}) \pi_0(\theta) d\theta + \int_{A^c} R(\theta, \delta^{\pi_0}) \pi_0(\theta) d\theta$$

Si $\pi_0(A) > 0$, $r(\pi_0, \delta^{\pi_0}) < \overline{R}$; ce qui est impossible donc $\pi_0(A) = \pi_0(\{\theta \in \Theta, R(\theta, \delta_0) < R(\theta_0, \delta_0)\}) = 0$. En utilisant une conséquence du théorème de Picard en analyse complexe, f analytique implique que le support de π_0 est fini. ∎

Chapitre 3

Estimation ponctuelle

Contrairement à l'estimation ensembliste, le but de ce chapitre est d'estimer une fonction du paramètre $g(\theta)$, par exemple $g(\theta) = \theta$ ou $g(\theta) = \|\theta\|^2$. Nous nous plaçons à L et π donnés de telle sorte que nous disposons d'un estimateur bayésien δ^π correspondant.

3.1 Estimateur du maximum *a posteriori*

Définition 3.1.1 *estimateur du maximum a posteriori (*MAP*)*
On appelle estimateur MAP *tout estimateur* $\delta^\pi(X) \in Argmax_\theta \pi(\theta|X)$.

Cette notion est le pendant bayésien du maximum de vraisemblance fréquentiste. Il a le grand avantage de ne pas dépendre d'une fonction de perte et est utile pour les approches théoriques. Ses inconvénients sont les mêmes que l'estimateur du maximum de vraisemblance : non unicité, instabilité (dus aux calculs d'optimisation) et dépendance vis-à-vis de la mesure de référence (dominant Θ). En outre, il ne vérifie pas la non invariance par reparamétrisation qui peut apparaître importante intuitivement. Le dernier point se formalise ainsi : pour g un $\mathcal{C}^1$-difféomorphisme et $\eta = g(\theta), \theta \in \Theta$, $\text{Argmax}\pi'(\eta|X) \neq \text{Argmax}\pi(\theta|X)$ avec $\pi'(\eta|X) \propto \pi(\theta(\eta), X) \cdot \left|\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right|$

Exemple 3.1 *Modèles de mélange*
Ceci a pour but de modéliser des populations hétérogènes non distinguées. Dans ce cas, la loi conditionnelle s'écrit comme suit :

$$f(X|\theta_k) = \sum_{j=1}^{k} p_j g(X|\gamma_j) \quad où \quad \sum_{j=1}^{k} p_j = 1 \text{ et pour tout } j \; p_j \geq 0$$

En outre, le nombre de composantes k n'est pas forcément connu.

Les paramètres sont donc $(k, (p_j, \gamma_j)_{0 \leq j \leq k})$ et la loi a priori se met sous la forme $d\pi(\theta) = p(k)\pi_k(p_1, \cdots, p_k, \gamma_1, \cdots, \gamma_k)$. Un estimateur naturel maxi-

mise la loi a posteriori :

$$\hat{k}^{\pi} = Arg\max_{k} \pi(k|X) = \frac{p(k)\int_{\Theta_k} f(X|\theta_k)d\pi_k(\theta_k)}{\sum_{k=1}^{k_{max}} p(k)\int_{\Theta_k} f(X|\theta_k)d\pi_k(\theta_k)} \ .$$

3.2 Importance de la statistique exhaustive

Définition 3.2.1 *Statistique exhaustive*
On appelle statistique exhaustive une statistique[1] $S(X)$ *telle que la loi conditionnelle se décompose sous la forme :*

$$f(X|\theta) = h(X|S(X)) \cdot \tilde{f}(S(X)|\theta)$$

Ceci se récrit $\pi(\theta|X) = \pi(\theta|S(X))$. En termes informationnels, ceci signifie que S résume l'information a priori.

Lorsqu'on utilise des lois impropres, le comportement d'une statistique peut être capricieux. Par exemple, si on considère un paramètre de la forme suivante $\theta = (\theta_1, \theta_2)$ avec $\pi(\theta)$ une loi impropre et pour S une statistique exhaustive pour θ_1, nous pouvons écrire :

$$f(X|\theta_1, \theta_2) = g(X|\theta_2) \cdot f(S|X)$$

Dans ce cas, il peut arriver que $\pi(\theta_1|X) = \pi(\theta_1|\delta)$ mais sans qu'il n'existe de $\pi_1(\theta_1)$ tel que $\pi(\theta_2, S) = \frac{f(S|X)\pi_1(\theta_1)}{\int_{\Theta_1} f(S|\theta_1)\pi_1(\theta_1)d\theta_1}$. Ceci est dû au fait que la loi *a priori* est impropre et s'appelle le *paradoxe de marginalisation* qui nous illustrons dans l'exemple qui suit.

Exemple 3.2 *Paradoxe de marginalisation*

- *Si* $X_1, \cdots, X_n$ *indépendants tels que* $X_i \sim \exp(\eta)$ *et* $X_j \sim \exp(c\eta)$ *pour* $0 \leq i \leq \xi < j \leq n$ *avec* c *connu.*
  Le paramètre est $\theta = (\eta, \xi)$ *où* $\xi \in \{1, 2, \ldots, n-1\}$ *et nous supposons que la loi a priori vérifie* $\pi(\eta, \xi) = \pi(\xi)$*, cela implique notamment que la fonction a priori est impropre en* η*.*
  On peut montrer que $\pi(\xi|X) = \pi(\xi|Z)$ *en posant* $Z = (\frac{X_2}{X_1}, \ldots, \frac{X_n}{X_1})$ *et* $f(Z|\xi, \eta) = f(Z|\eta)$ *et pourtant il n'existe pas de* $\pi(\xi)$ *telle que* $\pi(\xi|Z) \propto f(Z|\xi)\pi(\xi)$*.*

- *Dans le cas où* $U_1 \sim N(\mu_1, \sigma^2)$*,* $U_2 \sim N(\mu_2, \sigma^2)$ *et* $\frac{S^2}{\sigma^2} \sim \frac{\chi^2(p)}{p}$ *avec* U_1, U_2, S^2 *indépendants, en prenant comme paramètre une transformation de* (μ_1, μ_2, σ^2)*,* $(\xi = \frac{\mu_1 - \mu_2}{\sigma}, \mu_1, \sigma^2)$*, pour* $\pi(\mu_1, \mu_2, \sigma^2) = \frac{1}{\sigma^2}$ *et en posant* $Z = \frac{U_1 - U_2}{S}$ *on montre que* $f(Z|\mu_1, \mu_2, \sigma^2) = f(Z, \xi)$ *et*

1. C'est-à-dire une fonction des données.

$\pi(\xi|U_1, U_2, S) = \pi(\xi|Z)$ mais $\pi(\xi|Z) \neq f(Z|\xi)\pi(\xi)$. Pour poursuivre cet exemple, les lecteurs-trices peuvent se référer à l'exercice 3.47 (David et Ziteck) du Choix Bayésien.

3.3 Prédiction

Le contexte du problème de la prédiction est le suivant : les observations X sont identiquement distribuées selon P_θ, absolument continue par rapport à une mesure dominante μ et donc qu'il existe une fonction de densité conditionnelle $f(\cdot|\theta)$. Par ailleurs on suppose que θ suit une loi *a priori* π.

Il s'agit alors à partir de n tirages $X_1, \cdots, X_n$ de déterminer le plus précisément possible ce que pourrait être le tirage suivant X_{n+1}.
Dans l'approche fréquentiste, on calcule $f(X_{n+1}|X_1, \cdots, X_n, \hat{\theta})$. Comme le révèle la notation $\hat{\theta}$, on ne connaît pas exactement θ. On doit donc l'estimer dans un premier temps et de ce fait, on utilise deux fois les données : une fois pour l'estimation du paramètre et une nouvelle fois pour la prédiction dans la fonction f. En général, ceci amène à sous-estimer les intervalles de confiance.

La stratégie du paradigme bayésien, désormais bien comprise par la lectrice et peut-être un peu assimilé par le lecteur, consiste à intégrer la prévision suivant une loi *a priori* sur θ et ce, afin d'avoir la meilleure prédiction compte tenu à la fois de notre savoir et de notre ignorance sur le paramètre. La loi prédictive s'écrit ainsi :

$$f^\pi(X_{n+1}|X_1, \cdots, X_n) = \int_\Theta f(X_{n+1}|X_1, \cdots, X_n, \theta)\pi(\theta|X_1, \cdots, X_n)d\theta$$

Dans le cas des tirages indépendants et identiquement distribués, ceci devient :

$$f^\pi(X_{n+1}|X_1, \cdots, X_n) = \int_\Theta f(X_{n+1}|\theta)\pi(\theta|X_1, \cdots, X_n)d\theta$$

En considérant le coût quadratique $L(\theta, \delta) = \|\theta - \delta\|^2$, on peut proposer le prédicteur :

$$\hat{X}^\pi_{n+1} = E^\pi(X_{n+1}|X_1, \cdots, X_n) = \int X_{n+1} f(X_{n+1}|X_1, \cdots, X_n)dX_{n+1}$$

3.4 Modèle Gaussien

Nous supposons $Y_1, \cdots, Y_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma)$, $\pi(\mu, \sigma) = \frac{1}{\sigma}$ (la mesure de Haar) et $\pi(\mu, \sigma^2) = \mathcal{N}(\mu_0, (\tau\sigma)^2) \otimes IG(a, b)$ (loi Inverse Gamma). On considère le modèle linéaire $Y = X\beta + \varepsilon$, où $\varepsilon \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma = \sigma^2 I_d$. Le paramètre est $\theta = (\beta, \sigma^2)$. Si σ^2 est connu, alors on peut écrire $\pi(\beta) = \mathcal{N}(\mu, A)$ où A est un hyperparamètre et $\pi(\beta|Y, X) = \mathcal{N}((\hat{\Sigma}^{-1} + A^{-1})^{-1}(\hat{\Sigma}^{-1}\hat{\beta} +$

$A^{-1}\mu), (\hat{\Sigma}^{-1} + A^{-1})^{-1})$. La regression linéaire bayésienne retrouve les formules $\hat{\beta} = (X'X)^{-1}X'Y$ et $\hat{\Sigma} = \sigma^2(X'X)^{-1}$. Les modèles hétéroscédastiques peuvent aussi être étudiés.

Liens entre les modèles à effets aléatoires et le paradigme bayésien

Considérons le modèle suivant : Y une variable d'intérêt quelconque, $X = 1$ si un individu est un hommme $X = 0$ si c'est une femme, avec $(Y|X) \sim N(\beta X, \sigma^2)$. L'hétérogénéité de gendres se traduit par :

$$\begin{aligned} Y &= u_1 + \varepsilon \quad si X = 1 \\ Y &= u_2 + \varepsilon \quad si X = 0 \\ \Rightarrow Y &= X u_1 + (1 - X)u_2 + \varepsilon \end{aligned}$$

On suppose que $u_1 \sim \mathcal{N}(0, \tau_1^2)$ et que $u_2 \sim \mathcal{N}(0, \tau_2^2)$. Les effets aléatoires signifient que β est aléatoire : $\beta \sim \mathcal{N}(\beta_0, A)$

La variabilité est plus grande que ce que le modèle aléatoire propose, les effets aléatoires augmentent la variance sans changer le modèle. La limite entre le modèle bayésien hiérarchique et modèle à effet aléatoire est fine.

3.5 Mesure d'erreur

Pour mesurer une erreur, il est nécessaire d'introduire une fonction de perte, nous l'appellerons ici traditionnellement $L = L(\theta, \delta)$. On définit ensuite le risque fréquentiste $R(\theta, \delta)$ et le risque *a posteriori* $\rho(\pi, \delta|X)$.

Le but est d'estimer la fonction de perte pour δ donné mais l'utilisation double des données tend à sous-estimer les erreurs, ce qui limite cette démarche. On a évidement $L(\delta, \delta) = 0$ donc estimer une erreur en un point n'a pas de sens. L'idée est d'introduire une fonction de perte sur les fonctions de perte et d'estimer $\ell = L(\theta, \delta) \in \Lambda \subset \mathbb{R}_+$:

$$\begin{aligned} \tilde{L} : \Theta \times \Lambda \times \mathcal{D} &\to \mathbb{R} \\ (\theta, \ell, \delta) &\longmapsto \tilde{L}(\theta, \ell, \delta) \end{aligned}$$

En appliquant dans ce contexte, le principe bayésien, maintenant maîtrisé par les lecteurs-trices et qui consiste à intégrer sur les paramètres, on définit le risque : $\rho(\pi, \delta, \ell|X) = E^\pi \left[\hat{L}(\theta, \ell, \delta|X)\right]$. On cherche alors à calculer $\hat{\ell} = \arg\min_\ell \rho(\pi, \delta, \ell|X)$.

Exemple 3.3 *Perte quadratique*
Si $\hat{L}(\theta, \ell, \delta|X) = (\ell - L(\theta, \delta))^2$, *alors* $\hat{\rho}^\pi = E^\pi [L(\theta, \delta|X)]$. *Si de plus,* $L(\theta, \delta) = \|\theta - \delta\|^2$, *alors* $\hat{\rho}^\pi = E^\pi \left[(\theta - \delta)^2|X\right]$ *et puisqu'en toute logique, on choisit* $\delta = \delta^* = E^\pi [\theta|X]$, *on a :*

$$\hat{\rho}^\pi = V^\pi [\theta|X]$$

Chapitre 4

Tests et régions de confiance

4.1 Région de confiance

4.1.1 Définitions

Définition 4.1.1 *Région α-crédible*
Une région C de Θ est dite α-crédible si et seulement si $P^\pi(\theta \in C|X) > 1-\alpha$.

Notons que le paradigme bayésien permet une nouvelle fois de s'affranchir d'un inconvénient de l'approche fréquentiste. En effet, au sens fréquentiste, une région de confiance C est définie par $\forall\theta,\ P_\theta(\theta \in C) \geq 1-\alpha$ et correspond à l'interprétation suivante. En refaisant l'expérience un grand nombre de fois, la probabilité que θ soit dans C est plus grand que $1-\alpha$. Une région de confiance n'a donc de sens que pour un très grand nombre d'expériences tandis que la définition bayésienne exprime que la probabilité que θ soit dans C au vue des celles déjà réalisées est plus grande que $1-\alpha$. Il n'y a donc pas besoin ici d'avoir recours à un nombre infini d'expériences pour définir une région α-crédible, seule compte l'expérience effectivement réalisée.

Il y a une infinité de régions α-crédibles, il est donc logique de s'intéresser à la région qui a le *volume* minimal. Le volume étant défini par $vol(C) = \int_C d\nu(\theta)$, si $\pi(\theta|X)$ est absolument continue par rapport à une mesure de référence ν.

Définition 4.1.2 *Région* HPD *(highest posteriori density)*
C_α^π est une région HPD *si et seulement si $C_\alpha^\pi = \{\theta, \pi(\theta|X) \geq h_\alpha\}$ où h_α est défini par $h_\alpha = \sup\{h, P^\pi(\{\theta, \pi(\theta|X) \geq h\}\,|X) \geq 1-\alpha\}$.*

C_α^π est parmi les régions qui ont une probabilité supérieure à $1-\alpha$ de contenir θ (et qui sont donc α-crédibles) et sur lesquelles la densité *a posteriori* ne descend pas sous un certain niveau (restant au dessus de la valeur la plus élevée possible).

Théorème 4.1 *Régions* HPD
C_α^π est parmi les régions α-crédibles celle de volume minimal si et seulement

si elle est HPD.

Démonstration : On considère le problème dual.

$$\min_{C,\ P^\pi(C|X)\geq 1-\alpha} vol(C) \Leftrightarrow \max_{C,\ vol(C)=\overline{V}} P^\pi(C|X) \quad \text{où } \overline{V} \text{ constante}$$

Il suffit alors de montrer que pour tout C telle que $\pi(C\Delta C_\alpha^\pi) > 0$ (en notant $C\Delta C_\alpha^\pi$ l'ensemble tel que $C\Delta C_\alpha^\pi = C_\alpha^\pi \cap C^c \bigcup C_\alpha^{\pi c} \cap C$), si $vol(C) = vol(C_\alpha^\pi)$ alors $P^\pi(C|X) < P^\pi(C_\alpha^\pi|X)$.

$$\int_C \pi(\theta|X)d\nu(\theta) - \int_{C_\alpha^\pi} \pi(\theta|X)d\nu(\theta) = \int_{C\cap C_\alpha^\pi} \pi(\theta|X)d\nu(\theta) - \int_{C_\alpha^\pi \cap C^c} \pi(\theta|X)d\nu(\theta)$$

Or, comme C et C_α^π ont même volume, $\int_{C\cap C_\alpha^\pi} d\nu(\theta) = \int_{C_\alpha^\pi \cap C^c} d\nu(\theta)$. Comme π est dominé par ν, ces volumes sont strictement positifs. En outre :

- $\int_{C\cap C_\alpha^\pi} \pi(\theta|X)d\nu(\theta) < h_\alpha\ vol(C \cap C_\alpha^\pi)$
- $\int_{C_\alpha^\pi \cap C^c} \pi(\theta|X)d\nu(\theta) \geq h_\alpha\ vol(C_\alpha^\pi \cap C^c)$

d'où finalement $P^\pi(C|X) < P^\pi(C_\alpha^\pi|X)$. ∎

Exemple 4.1 *Par les calculs*
Si $X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ *et* $\pi(\mu, \sigma) = \frac{1}{\sigma}$, *alors*

$$\pi(\mu, \sigma^2|X_1, \cdots, X_n) \propto e^{-\frac{n(\overline{X}_n-\mu)^2}{2\sigma^2}} e^{\frac{-S_n^2}{2\sigma^2}}$$

où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *et* $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2$. *On a alors :*

$$\pi(\mu|\sigma^2, X_1, \cdots, X_n) = \mathcal{N}(\overline{X}_n, \frac{\sigma^2}{n}) \quad \textit{de densité :}\ \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{\frac{-n(\overline{X}_n-\mu)^2}{2\sigma^2}}$$

$$\pi(\sigma^2|X_1, \cdots, X_n) \propto \frac{e^{\frac{-S_n}{2\sigma^2}}}{\sigma^{n+1}}$$

Il est important de remarquer qu'on a ici une formule explicite en μ, *c'est ce qui va nous permettre de poursuivre le calcul dans ce cas exceptionnel. Notons que la calculabilité repose parfois sur la notion de loi conjuguée. Ici on est dans un cas limite, la famille conjuguée de la loi normale est la famille des lois normales ; la loi a priori* $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$ *est en fait une loi normale dégénérée (la variance est infinie).*

Des lois précédentes, on déduit :

$$\begin{aligned}\pi(\mu|X_1, \cdots, X_n) &\propto \int_0^\infty \frac{1}{\sigma} e^{\frac{n(\overline{X}_n-\mu)^2}{2\sigma^2}} \frac{e^{\frac{-S_n}{2\sigma^2}}}{\sigma^{n+1}} d\sigma^2 \\ &\propto \int_0^\infty \frac{1}{\sigma^{n+2}} e^{\frac{-1}{2\sigma^2}(n(\overline{X}_n-\mu)^2+S_n)} d\sigma^2\end{aligned}$$

On reconnaît une loi inverse gamma[1] *:*

$$\begin{aligned}\sigma^2 &\sim IG\left(\frac{(n(\overline{X}_n-\mu)^2+S_n)}{2},\frac{n}{2}\right)\\ &\propto \left(1+\frac{n(\overline{X}_n-\mu)^2}{{S_n}^2}\right)^{-\frac{n}{2}}\end{aligned}$$

Ainsi $\frac{\mu-\overline{X}_n}{\sqrt{\frac{S_n}{n}}} \sim St(n)$[2]*.*
On peut désormais calculer $C^{\pi}_{\alpha,\mu}=\{\mu,\pi(\mu|X_1,\cdots,X_n)\geq 1-\alpha\}$ *et* $C^{\pi}_{\alpha,\sigma}=\{\sigma,\pi(\sigma|X_1,\cdots,X_n)\geq 1-\alpha\}$*.*

$$\begin{aligned}\pi(\mu|X_1,\cdots,X_n)\geq k_\alpha \quad &= \quad (1+\frac{n(\overline{X}_n-\mu)^2}{{S_n}^2})^{-\frac{n}{2}}\geq \widetilde{k}_\alpha\\ &\iff \frac{n(\overline{X}_n-\mu)^2}{{S_n}^2}\leq \widetilde{\widetilde{k}}_\alpha\\ &\iff -\sqrt{\widetilde{\widetilde{k}}_\alpha}\leq \frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n}\leq \sqrt{\widetilde{\widetilde{k}}_\alpha}\end{aligned}$$

Comme $\frac{\sqrt{n}(\overline{X}_n-\mu)}{S_n}$ *suit un Student à* $\frac{n}{2}$ *degré de liberté, on connaît la valeur de* $\widetilde{\widetilde{k}}_\alpha$ *qui est le carré du quantile d'ordre* $\frac{\alpha}{2}$*. On peut ensuite remonter à* k_α *puisque* $\widetilde{\widetilde{k}}_\alpha$ *est une transformation bijective de* k_α*. Résumons :* k_α *est tel que*

$$P(St(\frac{n}{2})\leq\sqrt{\widetilde{\widetilde{k}}_\alpha})=1-\alpha$$

Ainsi :

$$C^{\pi}_{\alpha,\mu}=\left[\bar{X}_n-t_{\frac{n}{2}}(1-\frac{\alpha}{2})\sqrt{\frac{S_n^2}{n}};\overline{X}_n+t_{\frac{n}{2}}(1-\frac{\alpha}{2})\sqrt{\frac{S_n^2}{n}}\right]$$

Pour comparer, nous rappelons la région de confiance fréquentiste bien connue du lecteur une fois que l'on a pris conscience que $\sqrt{\frac{S_n^2}{n}}$ *n'est autre que* $\hat{\sigma_n^2}$ *(au maximum de vraisemblance)*[3] *:*

$$C^{\pi}_{\alpha,\mu}=\left[\overline{X}_n-t_{(n-1)}(1-\frac{\alpha}{2})\sqrt{\frac{\hat{\sigma_n^2}}{n}};\overline{X}_n+t_{(n-1)}(1-\frac{\alpha}{2})\sqrt{\frac{\hat{\sigma_n^2}}{n}}\right]$$

On s'interesse maintenant au calcul de la région de confiance relative à σ^2 *:*

$$C^{\pi}_{\alpha,\sigma^2}=\left\{\sigma^2;\pi(\sigma^2|X_1,\cdots,X_n)\geq g_\alpha\right\}$$

1. La densité de $IG(a,b)$ est $g_{a,b}(x)=x^{-(a+1)}e^{-\frac{b}{x}}\frac{b^a}{\Gamma(a)}$.
2. On rappelle que u suit une loi de Student à p degré de liberté si $g_p(u)\propto(1+u^2)^{-p}$.
3. La différence est dans l'ordre du quantile de Student, il y a donc équivalence asymptotique.

Comme on a montré que $\pi(\sigma^2|X_1,\cdots,X_n) \propto \frac{e^{\frac{-S_n}{2\sigma^2}}}{\sigma^{n+1}}$ *:*

$$\pi(\sigma^2|X_1,\cdots,X_n) \geq g_\alpha \Longleftrightarrow \frac{e^{\frac{-S_n}{2\sigma^2}}}{\sigma^{n+1}} \geq \widetilde{g}_\alpha$$

On ne reconnaît pas là de loi connue. Il faut donc établir une table de valeur pour déterminer g_α*. Remarquons que le mode de la loi a posteriori est unique et vaut* $\frac{S_n^2}{n+1}$*, il est alors assez simple de déterminer une procédure numérique calculant* g_α*. Pour un* k *donné, les deux solutions de* $\frac{e^{\frac{-S_n}{2u}}}{u^{\frac{n+1}{2}}} = k$ [4] *permettent de calculer l'aire sous la courbe et délimitée par ces deux solutions : si cette aire est plus grande que* $1-\alpha$ *on baisse* k *et on l'augmente sinon ; ensuite cette démarche est renouvelée en fonction de la précision souhaitée.*

Pour prolonger cet exemple, il est envisageable d'étudier les régions de confiance liées au paramètre $\theta = (\mu, \sigma^2)$*.*

4.1.2 Calcul de région HPD

Pour calculer les régions HPD, il y a plusieurs méthodes :

1. Méthode analytique et numérique : c'est ce qui a été fait lors de l'exemple précédent. Précisons une nouvelle fois que cette méthode ne peut s'appliquer que dans des cas assez rares.
2. Méthode par approximation : cette méthode peut être appliquée si le modèle est régulier : $\log(\pi)$ est $\mathcal{C}^2$ et $\hat{\theta}$ tend vers θ pour P_θ.
3. Méthode de simulation que nous développons ci-après.

- **Méthode par approximation**

Considérons l'estimateur MAP $\widetilde{\theta}^\pi = \text{Arg}\max_\theta \log(\theta) + \log\pi(\theta)$ avec $\log(\theta) = \log f(X^n|\theta)$ pour un échantillon $X^n = (X_1,\cdots,X_n)$. On peut montrer que : $\pi(\theta|X_1,\cdots,X_n) = \frac{e^{-n(\theta-\tilde{\theta})'\frac{\widetilde{I}_n}{2}(\theta-\tilde{\theta})}}{|\widetilde{I_n}|^{-\frac{1}{2}}(2\pi)^{\frac{k}{2}}}$ où $\widetilde{I}_n = I(\widetilde{\theta}^\pi)$ est l'information de Fisher pour une observation prise en la valeur du MAP et k est tel que $\Theta \subset \mathbb{R}^k$. Ainsi, les propriétés asymptotiques bayésiennes ressemblent à celles de l'estimateur du maximum de vraisemblance. L'idée est d'appliquer ce résultat aux régions HPD : $\pi(\theta|X_1,\cdots,X_n) \geq k_\alpha \Longleftrightarrow n(\theta-\widetilde{\theta}')\widetilde{I}_n(\theta-\widetilde{\theta})) \leq g_\alpha$. On trouve alors des ellipses de taille $\frac{1}{n}$ telles que g_α proche du quantile d'ordre $1-\alpha$ d'un chi-deux χ^2_k.

- **Méthode de simulation**

Le principe repose sur la méthode MCMC qui seront développées plus loin [5]. Si $\theta^1,\cdots,\theta^T \overset{iid}{\sim} \pi(\theta|X_1,\cdots,X_n)$ alors lorsque $\theta \in \mathbb{R}$, on s'intéresse

4. Il peut être judicieux de faire un dessin.
5. Méthodes de Monte-Carlo par Chaînes de Markov ; voir notamment **?**].

aux intervalles (quantiles empiriques) de la forme $\left[\theta^{(\frac{\alpha}{2})}, \theta^{(\frac{1-\alpha}{2})}\right]$ tels que

$$P^{\pi}(\theta \in \left[\theta^{(\frac{\alpha}{2})}, \theta^{(\frac{1-\alpha}{2})}\right] | X_1, \cdots, X_n) \underset{T\to+\infty}{\longrightarrow} 1-\alpha$$

Pour T grand, $\theta^{\frac{\alpha}{2}}$ s'approche du quantile d'ordre $\frac{\alpha}{2}$ de la loi *a posteriori* $\pi(\theta|X_1, \cdots, X_n)$. Cette région n'est pas nécessairement HPD mais reste α-crédible. Cette méthode est particulièrement adaptée lorsque la loi *a priori* est unimodale. Il est toujours utile de représenter graphiquement les sorties pour fixer les idées. Enfin, il est aussi envisageable d'avoir recours à une estimation non paramétrique par noyaux.

- **Remarques sur les régions HPD**

Non-invariance par reparamétrisation

Considérons $C_\alpha^\pi = \{\theta; \pi(\theta|X) \geq h_\alpha\}$ une région HPD et un C^1-difféomorphisme $\eta = g(\theta)$ alors on peut définir : $\widetilde{C}_\alpha^\pi = \left\{\theta; \widetilde{\pi}(\eta|X) \geq \widetilde{k}_\alpha\right\}$.

En général, $\widetilde{C}_\alpha^\pi \neq g(C_\alpha^\pi)$ [6]. En effet, $\widetilde{\pi}(\eta|X) = \pi(\theta(\eta)|X) \times \left|\frac{d\theta}{d\eta}\right|$ donc $\left\{\theta; \widetilde{\pi}(\eta|X) \geq \widetilde{k}_\alpha\right\} = \left\{\theta; \pi(\theta(\eta)|X) \left|\frac{d\theta}{d\eta}\right| \geq \widetilde{k}_\alpha\right\}$.

Exemple 4.2 $\mathcal{N}(\theta, 1)$
Supposons $\pi(\theta) = 1$ *et* $\pi(\theta|X) = \mathcal{N}(X, 1)$, *posons* $\eta = e^\theta$.
Comme $\pi(\theta|X) \geq h \iff (\theta - X)^2 \leq (\phi^{-1}(1-\alpha))^2$,
$C_{\alpha,\theta}^\pi = \left[X - \phi^{-1}(1-\alpha); X + \phi^{-1}(1-\alpha)\right]$.
D'un autre côté, $\pi(\eta|X) \propto \frac{1}{\eta} e^{-\frac{(X-\log(\eta))^2}{2}}$. *Ainsi :*

$$\begin{aligned}
\pi(\eta|X) \geq \widetilde{g}_\alpha &\iff \frac{(X-\log(\eta))^2}{2} + \log(\eta) \leq g_\alpha \\
&\iff \frac{(2X-\log(\eta))^2}{2} - \frac{X^2}{2} \leq g_\alpha \\
&\iff (2X-\log(\eta))^2 \leq g_\alpha \\
&\iff \log(\eta) \in \left[2X - \phi^{-1}(\frac{\alpha}{2}); 2X + \phi^{-1}(\frac{\alpha}{2})\right]
\end{aligned}$$

On vérifie sur cet exemple que le lien entre $C_{\alpha,\theta}^\pi$ *et* $C_{\alpha,\eta}^\pi$ *ne correspond pas à la transformation initiale (exponentielle).*

Nous pouvons comprendre pourquoi une région de confiance n'est pas invariante par reparamétrisation. En effet, cette région se définit comme une solution du problème de minimisation suivant :

$$C_\alpha^\pi = \underset{C,\ P^\pi(C|X)\geq 1-\alpha}{\operatorname{argmin}} vol(C)$$

6. Toujours à cause du jacobien...

où $Vol(C) = \int_C d\theta$. C'est là que le bât blesse : la mesure de Lebesgue n'est pas invariante par reparamétrisation. Une idée pour lever cette difficulté est donc logiquement d'abandonner la mesure de Lebesgue et de considérer pour une mesure s :

$$C^{\pi}_{\alpha,s} = \underset{C,\ P^{\pi}(C|X)\geq 1-\alpha}{\operatorname{argmin}} \int ds(\theta)$$

Lien avec la théorie de la décision

En revenant à la notion de région de confiance, il est judicieux de remarquer qu'il s'agit d'un problème de minimisation sous contrainte. L'idée est alors que la région de confiance minimise un fonction de perte. En particulier, on peut penser à la forme suivante :

$$L(C, \theta) = Vol(C) + k(1 - \mathbb{1}_{\theta \in C})$$

qui correspond au risque *a posteriori* :

$$\rho^{\pi}(C|X) = vol(C) + k(1 - P^{\pi}(C|X))$$

Nous pouvons observer un problème dans l'expression suivante, le second terme varie entre 0 et k, tandis que le volume peut prendre des valeurs dans R_+. Pour résoudre ce problème, une idée est de définir une probabilité sur C (avec évidemment l'épineuse question du choix de cette probabilité), une autre idée étudiée est de prendre dans la fonction de perte $\frac{vol(C)}{1+vol(C)}$ à la place de $vol(C)$ mais le comportement qui s'en déduit ne suit pas le maximum de vraisemblance, qualité pourtant recherchée.

4.2 Test

Comme nous l'avons vu précédemment, un test se formalise de la manière suivante : on définit une hypothèse nulle $H_0 : \theta \in \Theta_0$ et une hypothèse alternative $H_1 : \theta \in \Theta_1$ de telle sorte que $\Theta_0 \cap \Theta_1 = \emptyset$ et $\Theta_0 \cup \Theta_1 \subset \Theta$.

Pour le modèle paramétrique $X \in (\mathcal{X}, \mathcal{B}, \{P_\theta, \theta \in \Theta\})$, $\Theta_0 \cup \Theta_1 \subset \Theta$. Le cas propre correspond au cas où la loi *a priori* est telle que si θ suit la loi π alors $\pi(\Theta_0) + \pi(\Theta_1) = 1 = \pi(\Theta)$ [7]. Cette dernière condition peut s'exprimer aussi sous la forme suivante : $\pi(\Theta \setminus (\Theta_0 \cup \Theta_1)) = 0$.

4.2.1 Approche par la fonction de perte de type 0–1

Dans cette partie, l'espace des décisions est $\mathcal{D} = \{0, 1\}$, la décision i correspondant à accepter H_i, $i = 0, 1$. La fonction de perte est alors la suivante :

$$L(\theta, \delta) = \begin{cases} 0 & si \quad \delta = \mathbb{1}_{\theta=\theta_1} \\ 1 & si \quad \delta = \mathbb{1}_{\theta=\theta_0} \end{cases}$$

7. c'est-à-dire $\pi(\Theta \setminus \Theta_1 \cup \Theta_0) = 0$.

Comme cela a été montré précédemment, l'estimateur bayésien associé à cette fonction de perte est :

$$\delta^\pi(X) = \begin{cases} 1 & si \quad P^\pi(\Theta_1|X) \geq P^\pi(\Theta_0|X) \\ 0 & sinon \end{cases}$$

Exemple 4.3 *Loi binomiale*

Exemple 4.4 *Loi normale*
$X \sim \mathcal{N}(\theta, \sigma^2)$ *avec* σ^2 *connu,* $\theta \sim \mathcal{N}(\mu, \tau^2)$ *et pour :*

$$L(\theta, \delta) = \begin{cases} 0 & si \quad \delta = \mathbb{1}_{\theta \in \Theta_1} \\ a_0 & si \quad \delta = \mathbb{1}_{\theta \in \Theta_0} \\ a_1 & si \quad \delta = 0, \theta \in \Theta_1 \end{cases}$$

Alors :

$$\delta^\pi(X) \quad = 1 \iff a_0 P^\pi(\Theta_0) < a_1 P^\pi(\Theta_1|X)$$

Ainsi, $\delta^\pi(X) = 1 \iff P^\pi(\Theta_1|X) > \frac{a_0}{a_0+a_1}$.

Pour tester l'hypothèse $H_0 : \theta < 0$ *contre* $H_1 : \theta \leq 0$*, on calcule*

$$\begin{aligned} \pi(\theta|X) &\propto e^{-\frac{(X-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \\ &\propto e^{-\frac{(\theta-\mu(X))^2}{2V^2}} = \mathcal{N}(\mu(X), V) \\ \text{où} \qquad & \mu(X) = \frac{X\tau^2}{\sigma^2+\tau^2} + \frac{\mu\sigma^2}{\sigma^2+\tau^2} \quad et \quad V = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2} \end{aligned}$$

Notons une dualité entre la fonction de perte μ *et la loi a priori* (a_0, a_1)*. On peut prendre à l'une pour donner à l'autre comme l'illustre l'exemple précédent. La détermination d'une fonction de risque ne peut se dissocier de celle de la loi a priori puisqu'elle correspond à une façon de sanctionner l'erreur. Ceci s'explique par le fait que le statisticien a une idée a priori sur la répartition de l'erreur et donc sur la loi a priori du paramètre. Prendre* $\mu >> 0$ *équivaut à* $a_0 << a_1$.

Ce dernier point se retrouve dans le cas des fonctions de perte quadratiques de la forme $L(\theta, \delta) = \omega(\theta)(\theta - \delta)^2$ où seul le produit $\omega \cdot \pi$ compte dans l'estimateur bayésien associé $\delta^\pi(X) = \frac{\int_\Theta \theta\omega(\theta) f(X|\theta) d\pi(\theta)}{\int_\Theta \omega(\theta) f(X|\theta) d\pi(\theta)}$.
Afin de rendre *absolue* la fonction de perte et qu'elle ne dépende plus de la loi *a priori*, nous introduisons dans la partie suivante le facteur de Bayes.

4.2.2 Facteur de Bayes

L'idée à l'origine de cette notion est de limiter l'importance du choix *a priori* de a_0 et a_0 introduits ci-dessus.

Définition 4.2.1 *Facteur de Bayes*
Le facteur de Bayes est défini par $B_{0/1} = \frac{P^\pi(\Theta_0|X)}{P^\pi(\Theta_1|X)} \times \frac{\pi(\Theta_1)}{\pi(\Theta_0)}$ *dès que* $\pi(\Theta_0) > 0$ *et* $\pi(\Theta_1) > 0$.

Le problème de définition relatif au Θ_i n'est en réalité que peu gênant puisque si $\pi(\Theta_i) > 0,\ i = 0, 1$ alors cela signifie que l'hypothèse i ne peut être acceptée a priori et donc, le risque correspondant ne peut pas être correctement défini. Cependant, même s'il peut être naturel de ne pas prendre en compte le cas de probabilité nulle pour Θ_0, le cas où Θ_0 est de mesure nulle sans être vide peut s'avérer être problématique. A ce moment là, toute loi *a priori* est nulle sur Θ_0 sans qu'il soit absurde de considérer des tests.

C'est le cas notamment pour l'estimation ponctuelle : $\Theta_0 = \{\theta_0\}$ ou les tests de significativité dans les modèles de régression. Pour une loi *a priori* absolument continue par rapport à la mesure de Lebesgue, $\pi(\Theta_0) = 0$. Il faut donc veiller à ne pas se placer dans un tel cas.

En écrivant un modèle linéaire sous la forme :

$$y = x_1\beta_1 + \cdots + x_p\beta_p + \varepsilon$$

, le test de significativité de x_p s'écrit : $H_0 : \beta_p = 0$ et $H_1 : \beta_p \neq 0$. A seuil de significativité (et p-value correspondante) fixé, l'augmentation du nombre d'observations mène à refuser plus souvent H_0, hypothèse ponctuelle. En réalité, le problème peut se reformuler en précisant que le statisticien [8] veut en réalité tester $|\beta_p| < \varepsilon$, mais reste la difficulté majeure que ε est inconnu et dépend du contexte de l'étude.

L'idée est d'introduire une masse de Dirac en 0 et de considérer à ε fixé, $\rho_\varepsilon = \pi(\theta,\ |\theta - \theta_0| < \varepsilon)$ et de choisir comme loi :

$$\pi(\theta) = \rho_\varepsilon \delta_{\theta_0} + (1 - \rho_\varepsilon)\pi_1(\theta)$$

où δ_{θ_0} mesure de Dirac en θ_0 et $\pi_1(\theta)$ absolument continue par rapport à la mesure de Lebesgue.

Le facteur de Bayes permet de s'affranchir du choix de ρ et de limiter l'influence du choix de ε [9]. En effet, dans ce cas $\Theta_1 = \Theta \backslash \Theta_0$ et le facteur de Bayes s'écrit (les termes en ρ se simplifient) :

$$B_{0/1} = \frac{f(X|\theta_0)}{\int_\Theta f(X|\theta)\pi_1(\theta)d\theta}$$

Ainsi, le facteur de Bayes peut s'interpréter comme le rapport de la vraisemblance sous H_0 par celle sous H_1. Il peut s'utiliser comme une p-value : pour

8. ou économètre...
9. Noter la correspondance entre choisir ρ et ε.

$B_{0/1}$ grand par rapport à 1, on accepte H_0 [10], et H_1 pour des petites valeurs, autour de 1 il n'est pas possible de conclure, c'est une *zone floue.*

Plus généralement, pour

$$d\pi(\theta) = \rho d\pi_0(\theta)\mathbb{1}_{\theta\in\Theta_0} + (1-\rho)d\pi_1(\theta)\mathbb{1}_{\theta\in\Theta_1}$$

avec

$$\int_{\Theta_0} d\pi_0(\theta)d\theta = \int_{\Theta_1} d\pi_1(\theta)d\theta = 1$$

le facteur de Bayes s'exprime :

$$B_{0/1} = \frac{\rho \int_{\Theta_0} f(X|\theta)d\pi_0(\theta)}{(1-\rho)\int_{\Theta_1} f(X|\theta)d\pi_1(\theta)} \times \frac{1-\rho}{\rho} = \frac{m_0(X)}{m_1(X)}$$

où $m_i(X) = \int_{\Theta_i} f(X|\theta)d\pi_i(\theta)$ est la vraisemblance intégrée sous H_i.

Une remarque importante concerne le cas où $d\pi_0$ ou $d\pi_1$ n'est pas propre : $B_{0/1}$ n'est alors pas défini de manière unique [11]. En effet, si une loi est impropre (par exemple $d\pi_0$) alors elle est définie à une constante multiplicative près. Pour $d\pi_0^*(\theta) = c\, d\pi_0(\theta)$ alors le facteur de Bayes est lui aussi multiplié par c : $B_{0/1}^* = B_{0/1}$ et par conséquent les ordres de grandeur de $B_{0/1}$ n'ont plus de sens : il n'est plus possible de comparer ses valeurs à 1 comme le précise l'exemple qui suit.

Exemple 4.5 *Cas impropre*
Pour $X \sim \mathcal{N}(\theta, 1)$ et $\pi(\theta) = c$, on considère le test suivant : $H_0 : \theta = 0$ et $H_1 : \theta \neq 0$. Dans ce cas, le facteur de Bayes est :

$$B_{0/1} = \frac{1}{\int_{\Theta_1} e^{-\frac{(x-\theta)^2}{2}} d\theta} = \frac{1}{C}$$

Le facteur de Bayes est donc une constante et n'a pas d'interprétation.

Pour remédier à ce problème, on envisage de réduire l'intervalle des paramètres de $\mathbb{R}$ à $[-M; M]$, mais de la même façon que le choix de ε plus haut était influent, celui de M n'est pas sans conséquence sur l'expression du facteur de Bayes.

On peut aussi jouer sur la variance ; en adaptant l'exemple précédent avec $X \sim \mathcal{N}(\theta, 1)$ et $\theta \sim \mathcal{N}(0, V)$, le choix d'une grande variance V permet de jouer sur le facteur de Bayes. Ainsi, cette stratégie n'est pas satisfaisante.

En conclusion, retenons que les tests requièrent des lois *a priori* propres et bien justifiées. C'est un problème parce qu'une loi non informative est en général impropre. Notons que ces défauts proviennent en partie du choix de la fonction de perte. La section suivante a pour but de généraliser le facteur de Bayes aux lois impropres.

10. par exemple, pour une valeur proche de 12, H_0 est 12 fois plus vraisemblable que H_1.

11. Et n'a donc pas beaucoup de sens...

4.2.3 Variations autour du facteur de Bayes

Nous nous plaçons dans le cas des lois *a priori* impropres pour définir le facteur de Bayes intrinsèque. L'idée est d'utiliser une partie de l'échantillon de départ pour travailler avec des lois propres.

Définition 4.2.2 *Echantillon d'apprentissage (Training sample)*
Si $X^n = (X_1, \cdots, X_n)$ *sont les données, un sous-échantillon de* X^n *de taille* ℓ *s'écrit* $X^{(\ell)} = (X_{i_1}, \cdots), X_{i_l}$, $\ell = i_1, \cdots, i_l \subset 1, \cdots, n$. *Cet échantillon est dit d'apprentissage si et seulement si* $\int_\Theta f(X^{(\ell)}(\theta)\pi(\theta)d\theta < +\infty$.

Il est ainsi possible de construire une loi *a priori* à partir de $X^{(\ell)}$.

Définition 4.2.3 *Echantillon d'apprentissage minimal*
$X^{(\ell)}$ *est un échantillon d'apprentissage minimal si pour tout sous-échantillon strict de* $X^{(\ell)}$ *l'intégrale précédente sur ce sous-échantillon est impropre.*

Le facteur de Bayes intrinsèque est alors défini pour $H_i : \theta \in \Theta_i$, muni de la loi $d\pi_i$ $(i = 0, 1)$.

Définition 4.2.4 *Facteur de Bayes intrinsèque*
Si $X^{(\ell)}$ *échantillon d'apprentissage pour* Θ_0 *et* Θ_1 *et minimal pour la réunion des deux*[12]*, le facteur de Bayes intrinsèque est défini par :*

$$B_{0/1}^{(\ell)} = \frac{\int_{\Theta_0} f(X^{(-\ell)}|\theta)\pi_0(\theta|X^{(\ell)})d\nu(\theta)}{\int_{\Theta_1} f(X^{(-\ell)}|\theta)\pi_1(\theta|X^{(\ell)})d\nu(\theta)}$$

où $X^{(-\ell)} = X^i, \notin \ell$.

En faisant apparaître la loi jointe de $X^{(\ell)}$ et $X^{(-\ell)}$ [13], le facteur de Bayes intrinsèque s'écrit :

$$B_{0/1}^{(\ell)} = \frac{B_{0/1}^{\pi_0,\pi_1}(X^n)}{B_{0/1}^{\pi_0,\pi_1}(X^{(\ell)})}$$

avec des notations « claires », étant donné que le facteur de Bayes dépend de l'échantillon considéré. Moyenner sur les différents sous-échantillons $X^{(\ell)}$ permet de s'affranchir de cette propriété.

Pour éviter de moyenner sur les sous-échantillons, O'Hagan a introduit le facteur de Bayes fractionnaire.

Définition 4.2.5 *Facteur de Bayes fractionnaire*
Il est défini par :

$$FBF_{0/1}^{(b)} = B_{0/1} \times \frac{\int_{\Theta_1} f^b(X|\theta)d\pi_1(\theta)}{\int_{\Theta_0} f^b(X|\theta)d\pi_0(\theta)}$$

12. Nous supposons ainsi qu'il existe ; c'est presque toujours le cas, sinon c'est que le problème doit être repensé à la base.
13. Celle de X^n !

où

$$b = \inf\left\{t > 0; \int_{\Theta_1} f^{(t)}(X|\theta)d\pi_1(\theta) < +\infty, \int_{\Theta_0} f^{(t)}(X|\theta)d\pi_0(\theta) < +\infty\right\}$$

Notons que cette méthode présente l'inconvénient d'utiliser deux fois les données. Cependant, tout ce qui a été présenté dans cette section a une justification asymptotique ; en particulier, b défini ci-dessus est très vraisemblablement de la forme $\frac{1}{n}$.

Exemple 4.6 *Interprétation du facteur de Bayes fractionnaire comme vraisemblance pénalisée*
Considérons le modèle linéaire $M_{[1,n]}$ *:*

$$y = X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Le problème consiste à sélectionner les X_j *significatifs. A un sous-ensemble* $[\ell]$ *donné d'éléments, on associe le modèle* $M_{[\ell]}$ *et l'on s'intéresse à* $\beta_{[\ell]} = \{\beta_j, j \in [\ell]\}$.

Le maximum de vraisemblance est donné par $f(X|\hat{\beta}_\ell, \sigma^2, M_{[\ell]})$ *et on regarde* $Arg\max_\ell f(X|\hat{\beta}_\ell, \sigma^2, M_{[\ell]})$. *Le maximum est donné par le modèle le plus gros qui inclut toutes les variables proposées dans le modèle de départ. Cependant, il faut garder en tête que* β *et* σ *sont considérés comme connus. En fait, c'est là la justification de la pénalisation, il faut tenir compte de l'incertitude sur les paramètres qui s'ajoute en gardant toutes les variables. C'est pourquoi, pénaliser permet de ne prendre en compte qu'un nombre limité et raisonnable de variables.*

Pour une loi a priori dépendant des variables retenues $\pi_{(l)}$*, le critère retenu est de la forme :*

$$\ell_n(\pi_{(l)}) = \ell_n(\hat{\theta}_{\pi_{(l)}}) - \underbrace{pen(n, \ell)}_{\text{critère de pénalité}}$$

Le rapport suivant permet ainsi de comparer les deux modèles et de prendre en compte l'incertitude sur les paramètres :

$$BF_{\ell/\ell'} = \frac{\int_{\Theta_\ell} f_\ell(X^n|\theta_\ell)d\pi_\ell(\theta_\ell)}{\int_{\Theta_{\ell'}} f_{\ell'}(X^n|\theta_{\ell'})d\pi_{\ell'}(\theta_{\ell'})}$$

4.2.4 Propriétés asymptotiques des facteurs de Bayes

Cette section cherche en outre à établir un lien entre le facteur de Bayes et la vraisemblance pénalisée. Le contexte est désormais usuel : $X^n = (X_1, \cdots, X_n)$

et $X^n \overset{iid}{\sim} f(X,\theta),\ \theta \sim \pi$, le test s'écrit toujours $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ dans le cadre un peu plus restreint $\Theta \subset \mathbb{R}^d$.

Si le modèle est régulier et si $\theta_0 \in \Theta_0$ tel que $\pi_1(\theta_0) > 0$ et de surcroît $d\pi_j(\theta) = \pi_j(\theta)d(\theta)$, alors dans le cas où :

$$d\pi(\theta) = \alpha\pi_0(\theta)\mathbb{1}_{\Theta_0}(\theta) + (1-\alpha)\pi_1(\theta)\mathbb{1}_{\Theta_1}(\theta),$$

on peut définir une sorte de rapport de vraisemblance par :

$$B_{0/1} = \frac{\int_{\Theta_0} e^{\log(\theta_n)} d\pi_0(\theta)d\theta}{\int_{\Theta_1} e^{\log(\theta_n)} d\pi_1(\theta)d\theta}$$

où $\log(\theta_n)$ désigne $\log(f(X^n|\theta))$ la valeur de la log-vraisemblance.

Les propriétés asymptotiques de $B_{0/1}$ découlent du maximum du rapport de vraisemblance et sont les suivantes :

- si $\theta_0 \in \Theta_0$, $B_{0/1} \xrightarrow[n\to+\infty]{P_{\theta_0}} +\infty$
- si $\theta_0 \in \Theta_1$, $B_{0/1} \xrightarrow[n\to+\infty]{P_{\theta_0}} 0$

Ainsi le facteur de Bayes distingue la bonne hypothèse, avec une probabilité qui tend exponentiellement rapidement avec n vers 1. Le facteur de Bayes est dit **consistant**.

Remarquons que l'hypothèse $\pi_1(\theta_0) > 0$ peut paraître illusoire étant donnée la propriété $\Theta_0 \cap \Theta_1 = \emptyset$. En réalité, cela peut être le cas si par exemple Θ_0 est dans l'adhérence de Θ_1 ; en particulier, si Θ_1 est de dimension k, on peut trouver de tels Θ_0 de dimension $k-1$. Même si cela semble contraignant, il faut se rendre compte que ce cas recouvre un grand nombre de problèmes comme nous le verrons plus loin.

Dans le cadre d'hypothèses « pseudo ponctuelles » de la forme $\Theta_0 = \{\theta_0\}$ ou $\Theta = \{\theta_1, \theta_2\}$ et $\Theta_0 = \{\theta_1 = \theta_{1,0}\}$ et pour $\Theta_1 = \Theta \setminus \Theta_0$, si π_1 est absolument continu par rapport à la mesure de Lebesgue alors $\pi_1(\Theta_0) = 0$ et $\pi(\Theta_1) = \pi(\Theta) = 1$. Comme l'illustre l'exemple suivant, l'hypothèse ainsi formulée est « trop petite ». L'idée est que θ_0 est autant dans Θ_1 que dans Θ_0 au sens où θ_0 appartient à l'adhérence de Θ_1.

Exemple 4.7 *Hypothèse pseudo-ponctuelle*
On regarde un modèle que l'on suppose mélangé au sens où il se pourrait que certains éléments soient tirés suivant une loi d'un certain paramètre et d'autres suivant la loi avec un autre paramètre. La vraisemblance est alors : $p \cdot f(X|\theta_1) + (1-p) \cdot f(X|\theta_2)$.

Ainsi, $\Theta = (p, \theta_1, \theta_2)$ *et on définit :* $H_0 : p = 1$ *ou* $\theta_1 = \theta_2$. *Cela correspond à n'avoir qu'un paramètre ou deux :* Θ_1 *est de dimension 2,* Θ_0 *est de dimension 1. Le facteur de Bayes converge exponentiellement vers* 0 *sous* H_O *et de façon polynômiale vers l'infini sous* H_1. *Le problème est que l'on teste une hypothèse « trop petite » et qu'il est difficile d'accepter* H_0.

Considérons le cadre particulier d'un choix de modèle dans une famille croissante de modèles : $\forall j \Theta_j \subset \Theta_{j+1}$ tels que $\pi_j(\Theta_j) = 1$. L'inclusion doit se comprendre comme le fait que lorsqu'on est dans un modèle on est aussi dans un autre « plus grand ». L'exemple du nombre de variables explicatives dans une regression linéaire est en ce sens parlant. Le but du choix de modèle est d'établir une procédure qui permet de sélectionner le modèle « le plus petit » possible, c'est-à-dire le plus précis [14]. Si θ_0, la vraie valeur, est contenue dans $\Theta_j \setminus \Theta_{j-1}$ alors on veut sélectionner Θ_j comme étant le plus petit modèle contenant θ_0 et ce avec un probabilité tendant vers 1 lorsque $n \to +\infty$.

Théorème 4.2 *Hypothèses emboîtées*
Si $\forall j \ \theta \in \Theta_j \longmapsto f_j(X, \theta)$ est régulière sur Θ_j, et si $\forall j$ la loi *a priori*, continue sur Θ_j, est telle que $\pi_j(\Theta_j) = 1$ et $\pi_j(\Theta_{j-1}) > 0$ alors en notant $\widehat{\theta_k}$ le maximum de vraisemblance sur Θ_k et d_k la dimension de Θ_k :

$$log\left(\frac{m_{n,h}(X^n)}{m_{n,h-1}(X^n)}\right) = log(\widehat{\theta_k}) - log(\widehat{\theta_{k-1}}) - \left(\frac{d_k - d_{k-1}}{2}\right) log\left(\frac{n}{2\pi}\right) + O(1)$$

On voit apparaître ici un rapport de vraisemblance pénalisé [15]. Cela correspond au critère BIC. Il est utile de formuler les hypothèses sous la forme $d_1 < d_2 < \cdots$ où $\forall j \ \Theta_j \subset \mathbb{R}^{d_j}$. L'idée à la base des critères BIC et AIC est que l'augmentation du nombre de paramètres à estimer augmente l'incertitude et qu'il faut pénaliser le fait qu'on fait comme si les valeurs estimées sont les vraies valeurs. On pénalise donc pour prendre en compte cette incertitude.

Comme le montre l'exemple qui suit, le choix de modèle peut être considéré comme un cas particulier de test et donc on peut utiliser le facteur de Bayes.

Exemple 4.8 *Choix de modèle*
Considérons $H0 : X \sim \mathcal{N}(\theta, \sigma^2)\theta \in \mathbb{R}, \sigma > 0$ *sous* $\pi_0(\theta, \sigma^2)$ *et* $H1 : X \sim \mathfrak{C}(\theta, \sigma)$ [16] *sous* $\pi_0(\theta, \sigma)$.

Le facteur de Bayes s'écrit :

$$B_{0/1} = \frac{\int_{\mathbb{R}\times\mathbb{R}^+} e^{-\frac{(X-\theta)^2}{2\sigma^2}} \frac{\pi_0(\theta,\sigma^2)}{\sqrt{2\pi\sigma^2}} d\theta d\sigma}{\int_{\mathbb{R}\times\mathbb{R}^+} \frac{\pi}{\sigma} \frac{1}{1+\left(\frac{X-\theta}{\sigma}\right)^2} \pi_1(\theta, \sigma) d\theta d\sigma}$$

14. En effet, on rappelle au lecteur que dans une régression linéaire la qualité de la représentation ne peut qu'augmenter avec le nombre de variables explicatives. C'est un moyen parfois artificiel de faire augmenter le sacro-saint R^2 mais en augmentant les procédures de calcul. S'il veut plus d'explications, le lecteur peut s'adresser à la lectrice qui a bien suivi l'enseignement d'économétrie dispensé à l'Ensae.

15. Notons que $O(1)$ masque le logarithme du déterminant de la matrice d'information de Fisher.

16. Une loi de Cauchy de paramètre (θ, σ) admet comme densité par rapport à la mesure de Lebesgue $f(X|\theta, \sigma) = \frac{\pi}{\sigma\left[1+\left(\frac{X-\theta}{\sigma}\right)^2\right]}$

Remarquons que le rapport de Bayes est une réponse à la fonction de perte 0–1 correspondant à $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f$, $\{f_\theta\ \theta \in \Theta\} = \mathcal{F}$, $H_0 : f \in \mathcal{F}$.

Lors d'une simplification de modèle, si on cherche à savoir dans quelle mesure un modèle en approche bien un autre [17], il faut considérer une distance entre la fonction f à approcher et la famille de modèle $\mathcal{F}$ par laquelle on veut approcher f. On considère une fonction de perte au lieu de la perte 0–1 simple. Pour $\delta \in 0, 1$:

$$L(\theta, \delta) = \begin{cases} (d(f, \mathcal{F}) - \varepsilon)\mathbb{1}_{d(f,\mathcal{F})>\varepsilon} & \text{si} \quad \delta = 0 \\ (\varepsilon - d(f, \mathcal{F}))\mathbb{1}_{d(f,\mathcal{F})<\varepsilon} & \text{si} \quad \delta = 1 \end{cases}$$

Il faut que les lois qu'on regarde (de mesure nulle dans l'espace des lois on le rappelle) ait un critère particulier pour qu'on puisse les différencier des autres lois. Ces lois particulières définissent un ensemble $\mathcal{F}_0$.

Pour résoudre notre problème il faut définir une loi *a priori* sur l'espace de dimension infinie contenant les lois de distribution [18]. Cette loi π_1 doit être centrée autour de $\mathcal{F}_0$. Si l'hypothèse H_0 était vraie alors il existerait θ tel que $F_\theta(X) \sim \mathcal{U}[0, 1]$, $\forall \theta \in \Theta$, $F_\theta(X)$ est une variable aléatoire sur $[0, 1]$. Et puisque si X ne suit une loi F_θ pour aucun θ alors $F_\theta(X)$ n'est pas une variable aléatoire à répartition uniforme sur $[0, 1]$, on a un élément qui nous permet d'établir une stratégie de test.

Exemple 4.9 *On suppose ici* $\Theta = \left\{f, \log(f) \in \mathcal{L}^2[0, 1]\right\}$ *et on considère une base orthonormale* $(\phi_j)_{j=0}^{+\infty}$ *quelconque de* $\mathrm{L}^2[0, 1]$. *Sur cette base, on décompose* $\log f$ *par* $\psi = (\psi_j)_{j\geq 0}$ *en écrivant* $\log f = \sum_{j=0}^{\infty} \psi_j \phi_j(X)$. *On définit alors* $g_p si = e^{\sum_{j=0}^{\infty} \psi_j \phi_j(X) - c(\psi)}$ *: s'il existe* ψ_0 *tel que* $\forall u \in [0, 1]$, $g_{\psi_0}(u) = u$, *ce sera centré.*

4.2.5 Calcul du facteur de Bayes

On se place dans le cadre habituel : pour $j = 0, 1$, $H_j : \theta \in \Theta_j$ muni de $d\pi_j$ (loi *a priori*) et $f_j(X|\theta)$ (loi conditionnelle). Rappelons l'expression du facteur de Bayes :

$$B_{0/1} = \frac{\int_{\Theta_0} f_0(X|\theta) d\pi_0(\theta)}{\int_{\Theta_1} f_1(X|\theta) d\pi_1(\theta)}$$

Les différentes méthodes de calcul sont les suivantes :

- Formules analytiques de $B_{0/1}$ (rare, correspond au forme conjuguée)
- Approximation par BIC si : les modèles sont réguliers, le max de vraisemblance se calcule sous H_0 et H_1 et de nombreuses observations sont disponibles
- Méthodes par simulations (ou de Monte-Carlo) à partir de tirages indépendants des lois $\pi_j, j = 0, 1$, on estime les intégrales grâce aux contreparties empiriques (très grande nombre d'observations nécessaires)

17. c'est-à-dire avoir une idée de qualité (distance) au lieu d'une simple réponse binaire
18. d'où une approche non paramétrique.

Chapitre 5

Propriétés asymptotiques des approches bayésiennes

5.1 Théorie générale

Cette partie s'appuie en particulier sur les travaux de Ibragimov & Khas'minskii (1985) [1].

On considère le cadre désormais classique : $X^n \sim f(X^n|\theta)$ où f est une densité absolument continue par rapport à une mesure μ et $\theta \in \Theta \subset \mathbb{R}^d$. On définit la forme suivante :

$$Z_{n,\theta}(u) = \frac{f(X^n|\theta + \phi_n u)}{f(X^n|\theta)}$$

où $(\phi_n)_{n\in\mathbb{N}*} \searrow 0$, défini à une constante multiplicative près et $u_n \in U_n = \phi_n^{-1}(\Theta \setminus \theta)$. L'idée est de faire varier le paramètre avec n autour d'une valeur donnée.

La fonction de perte est alors donnée par $L : \Theta \times \mathcal{D} \to \mathbb{R}^+$ avec δ un estimateur potentiel de θ dans $\mathcal{D}$ et π une loi de probabilité (*a priori*) sur Θ. On fait maintenant les hypothèses suivantes :

1. Θ est un ensemble convexe [1]
2. $\exists\ (\phi_n)_{n\in\mathbb{N}^*} \searrow 0$ tel que $\forall u \in U_n\ Z_{n,\theta}(u) \xrightarrow{\mathcal{L}} Z_\theta(u)$ uniformément en θ
3. $\exists m, M_1$ tels que :
   (a) $\exists \alpha > 0$ tel que $\forall \theta \in \Theta \forall R > 0$

$$\sup_{|u_1|\leq R|u_2|\leq R} |u_1 - u_2|^{-\alpha}\, \mathbb{E}_\theta^n \left[Z_{n,\theta}(u_1) - Z_{n,\theta}(u_2)\right] \leq M_1(1 + R^n)$$

   (b) $\exists \beta > 0, \forall u \in U_n, \forall \theta \in \Theta$

$$\mathbb{E}_\theta^n(Z_{n,\theta}^\beta(u)) \leq h_n(u) = r_n(u)^\gamma u^{-K}$$

1. Il semble que compact peut aussi convenir.

où K est une constante dépendant de γ, β, m, α et où

$$\lim_{H\to+\infty} \lim_{n\to+\infty} \sum_{k>H} r_n(k) = 0$$

Ce qui est important ici est de majorer $\mathbb{E}_\theta^n(Z_{n,\theta}^\beta(u))$ par une fonction décroissante de u. Il est conseillé de se référer à l'article correspondant pour d'avantage de précisions.

4. π est continue sur Θ et Θ est l'enveloppe convexe du support de π ; de plus $\forall\theta \in \Theta,\ \pi(\theta) > 0$
5. $L : \delta \longmapsto L(\theta, \delta) \in C^3(\Theta \times \mathcal{D})$; $\forall x \left| x \frac{{}^t\partial^3}{\partial\delta_i^3} L(\theta, \delta) x \right| \leq M \left|x\right|^2$
6. $\ell(\theta) = \frac{\partial^2}{\partial\delta^2}\ L(\theta, \delta)|_{\delta=\theta}$ définie positive. L'idée est d'effectuer un développement limité au second ordre pour se ramener à une fonction de perte quadratique [2].

Théorème 5.1 *Propriétés asymptotiques*
Sous ces hypothèses, en posant $\tau(\theta) = \frac{\int u Z_\theta(u)du}{\int Z_\theta(u)du}$, les propriétés suivantes sont vraies :

1. $\phi_n^{-2}\mathbb{E}_\theta^n\left[L(\theta, \delta^\pi(X^n))\right] \underset{n\to\infty}{\longrightarrow} \frac{1}{2}\mathbb{E}\left[\tau(\theta)'\ell(\theta)\tau(\theta)\right]$
2. $\phi_n^{-1}(\theta - \delta^\pi(X^n)) \xrightarrow{\mathcal{L}} \frac{\int u Z_\theta(u)du}{\int Z_\theta(u)du} = \tau(\theta)$

On précise que l'opérateur $\mathbb{E}_\theta^n$ fait à chaque fois référence à l'espérance selon $\pi(\theta)f(X^n|\theta)$. Avant de donner des pistes de démonstration du théorème précédent, étudions un exemple.

Exemple 5.1 *Modèles réguliers en dimension 1*

$$\begin{aligned}\log Z_{n,\theta}(u) &= \log f(X^n|\theta + \phi_n u) - \log f(X^n|\theta) \\ &= \phi_n u \frac{\partial}{\partial\theta}\log f(X^n|\theta) + \frac{(\phi_n u)^2}{2}\frac{\partial^2}{\partial\theta^2}\log f(X^n|\theta) + o((\phi_n u)^3)\end{aligned}$$

Comme $S_n(\theta) = \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log f(X^x|\theta) \to \mathcal{N}(0, I_0(\theta))$ *($I_0(\theta)$ est l'information de Fisher), en choisissant* $(\phi_n)_{n\in\mathbb{N}^*} = (\frac{1}{\sqrt{n}})_{n\in\mathbb{N}^*}$ *et si on note* ξ *une variable aléatoire suivant la loi normale, on a :*

$$\log Z_{n,\theta}(u) = u\xi - \frac{u^2}{2}I(\theta) + \mathcal{O}_p(1)$$

et $\quad \log Z_{n,\theta}(u) \xrightarrow{\mathcal{L}} \mathcal{Z}$ *, loi de densité* $\log Z_\theta(u) = e^{-\frac{u^2}{2}I(\theta)+u\xi}$

2. qui est donc locale...

Si $L(\theta,\delta)=(\theta-\delta)^2$ *et avec* $\ell(\theta)=2$ *on a alors :*

$$\begin{aligned}\sqrt{n}(\delta^\pi(X^n)-\theta) &\xrightarrow{\mathcal{L}} \frac{\int uZ_\theta(u)du}{\int Z_\theta(u)du}\\ &= \frac{\int ue^{-\frac{u^2}{2}I(\theta)+u\xi}du}{\int e^{-\frac{u^2}{2}I(\theta)+u\xi}du}\\ &= \frac{\int e^{-\frac{I(\theta)}{2}(u-I^{-1}(\theta)\xi)^2}udu}{\sqrt{I^{-1}(\theta)}\sqrt{2\pi}}\\ &= I^{-1}(\theta)\xi\\ &\sim \mathcal{N}(0,I^{-1}(\theta))\end{aligned}$$

A titre d'exercice, le lecteur peut appliquer le théorème ci-dessus à la loi uniforme sur un intervalle inconnu : $\mathcal{U}(\theta_1,\theta_2)$ avec une fonction de perte de la forme $L(\theta,\delta)=\int(\sqrt{f_\theta}-\sqrt{f_\delta})^2dx$, ou encore à $\theta_1e^{-\theta_1(X-\theta_2)}\mathbb{1}_{X\leq\theta 2}$ avec la fonction de perte $L(\theta,\delta)=(\theta_1-\delta_1)^2+(\theta_2-\delta_2)^2$ où l'on prendra ϕ_n bidimensionnel.

Démonstration : (idée rapide, consulter l'article cité pour la démonstration détaillée et rigoureuse)

Considérons l'estimateur bayésien $\delta^\pi=\text{Argmin}_\delta\mathbb{E}[L(\theta,\delta)]$. En effectuant un développement limité au second ordre de la vraisemblance, il s'agit de montrer en intégrant que $\mathbb{E}[L(\theta,\delta)]\phi_n^{-2}\leq\frac{1}{2}\mathbb{E}[\tau(\theta)^t l(\theta)\tau(\theta)]$ et de passer ensuite au minimum. ∎

5.2 Normalité asymptotique de la loi *a posteriori*

Si le modèle est régulier, si $\pi\in C^0$ et $\pi>0$, en notant $z_n=z+\sqrt{n}\hat{\theta}$, où $\hat{\theta}$ est le maximum de vraisemblance :

$$P^\pi(\sqrt{n}\theta\leq z|X^n)=P^\pi((\theta-\hat{\theta})\sqrt{n}\leq z_n|X^n)\to\phi(z\sqrt{I_0(\theta)})$$

Or $\pi(\theta|X^n)=\frac{e^{\ell_n(\theta)+\log\pi(\theta)}}{\int_\Theta e^{\ell_n(\theta)+\log\pi(\theta)}d\theta}$ et $\ell_n(\theta)=\sum_{1\leq i\leq n}\log f(X_i|\theta)$, donc pour $|\theta-\hat{\theta}|<\eta$ (petit intuitivement), $\ell_n(\theta)-\ell_n(\hat{\theta})\simeq -nI(\theta)\frac{(\theta-\hat{\theta})^2}{2}+\mathcal{O}_p(n|\theta-\hat{\theta}|^3)$. Le comportement gaussien asymptotique est obtenu en repassant aux exponentielles et en encadrant les termes non quadratiques des ordres supérieurs.

Chapitre 6

Détermination de lois *a priori*

Un grand intérêt de la théorie bayésienne est sa grande cohérence et sa méthodologie unifiée. Ainsi, donner les lois *a priori* et *a posteriori* ainsi que la fonction de perte suffit pour déterminer, entre autres, un estimateur optimal, des régions α-crédibles. Le choix de la loi *a priori* π est donc crucial. Avec beaucoup d'observations, le comportement asymptotique peut guider ce choix mais sinon il est nécessaire de le justifier avec précision. Il existe principalement deux méthodes : subjectives ou informatives, présentées successivement dans ce chapitre.

6.1 Lois subjectives

Précisons tout d'abord que cette démarche n'est pas forcément facile dans la pratique. L'idée est d'utiliser les données antérieures. Par exemple, dans un cadre paramétrique, cela revient à choisir une valeur particulière du paramètre.

Dans un cas concret, il peut être judicieux de baser son raisonnement sur les dires d'experts, notamment à l'aide de questionnaires. Il est alors nécessaire de veiller à ce que les questions soient compréhensibles [1], par exemple en prenant comme base les quantiles plutôt que les moments. Pour plusieurs experts, il peut être utile de pondérer leurs réponses et d'utiliser des modèles hiérarchiques.

Ainsi, la difficulté ici n'est pas mathématique mais plus psychométrique pour réduire les biais sur les réponses fournies. Nous allons nous concentrer sur le second aspect de la détermination.

1. On évitera : « Quelle est la loi a priori ? ».

6.2 Approche partiellement informative

6.2.1 Maximum d'entropie

Si l'on possède des informations partielles du type $\mathbb{E}^{\pi}[g_k(\theta)] = \mu_k$ où pour chaque $k = 1, \cdots, n$, g_k est une fonction donnée, on cherche la loi la moins informative sous ces contraintes, seules informations dont on dispose. Pour comparer le caractère informatif, il est nécessaire d'avoir recours à un critère d'information. L'entropie de Shannon permet de définir ce niveau d'informativité, nous présentons dans un premier temps cette entropie dans le cas fini et discret.

Pour $\theta \in \{1, \cdots, n\}$ et $\pi(\theta) = (\pi_1, \cdots, \pi_n)$ tel que $\pi_i \geq 0$ et $\sum_{i=1}^{m} \pi_i = 1$, l'entropie de la loi est définie par :

$$\mathrm{Ent}(\pi) = -\sum_{i=1}^{m} \pi_i \log(\pi_i) \leq -\sum_{i=1}^{m} \frac{1}{m} \log(\frac{1}{m}) = \log m$$

Ce dernier terme correspond à une répartition uniforme, la loi la plus « plate », la plus désordonnée. Pour la masse de Dirac [2] $\delta(j)$, $\mathrm{Ent}(\delta(j)) = 0$ ce qui correspond à l'intuition puisqu'alors il n'y a plus d'incertitude et l'information est totale. Une entropie petite s'interprète comme une loi concentrée et informative. La maximisation de l'entropie sous les contraintes permet de chercher la loi qui apporte le moins d'information. Le principe à la base de cette méthode est donc de chercher à calculer :

$$\mathrm{Arg}\max_{\pi} \ \mathrm{Ent}(\pi) \text{ sous la contrainte } \mathbb{E}^{\pi}[g_k(\theta)] = \mu_k$$

La solution de ce problème est alors donnée par :

$$\pi^{\star} \propto e^{\sum_{k=1}^{n} \lambda_k g_k(\theta)}$$

où les λ_k sont les multiplicateurs de Lagrange associés. Dans la pratique, on détermine ces valeurs λ à partir des contraintes (systèmes d'équations) comme l'indique l'exemple à suivre.

Exemple 6.1 *Un cas dénombrable*
Ici, $\Theta = \mathbb{N}$ et $\mathbb{E}^{\pi}[\theta] = x > 1$, c'est-à-dire qu'ici $g(\theta) = \theta$ et $\mu = x$. On sait que $\pi^{\star} \propto e^{\lambda\theta}$ et que λ est déterminé par :

$$\frac{\sum_{\theta \in \mathbb{N}} \theta e^{\lambda\theta}}{\sum_{\theta \in \mathbb{N}} e^{\lambda\theta}} = x$$

Cela conduit à résoudre :

$$\begin{aligned} \frac{x}{1 - e^{\lambda}} &= \frac{1}{e^{\lambda}} \frac{e^{\lambda}}{(1 - e^{\lambda})^2} \\ \text{d'où } e^{\lambda} &= \frac{x - 1}{x} \end{aligned}$$

2. telle que $\pi_j = 1$ et $\forall i \neq j, \pi_i = 0$.

Par exemple si $x = \frac{12}{11}$ *alors* $\lambda = -\log(12)$.

En continu, il n'est pas possible de définir l'entropie comme ci-dessus puisqu'on ne peut dénombrer les états (pas de mesure de comptage) en l'absence de mesure de référence. Dans le cas continu, on définit alors l'équivalent de l'entropie par rapport à une mesure π_0 :

$$\mathrm{Ent}(\pi/\pi_0) = \int_\Theta \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$$

C'est en fait la divergence de Kullback. Dans l'idée π_0 est la plus plate possible, la plus proche de la répartition uniforme, c'est en fait l'équivalent de la répartition en $\frac{1}{m}$ de l'information discrète. L'objectif est donc de maximiser $\mathrm{Ent}(\pi/\pi_0)$ sous les contraintes $\mathbb{E}^\pi[g_k(\theta)] = \mu_k$. Là encore, la solution générale est connue :

$$\pi^\star(\theta) \propto e^{\sum_{k=1}^n \lambda_k g_k(\theta)} \pi_0(\theta)$$

Lorsque la structure est bonne, des choix raisonnables de π_0 sont la mesure de Haar (pour les groupes) ou bien la loi de Jeffreys.

Exemple 6.2 *Cas continu*
Si le modèle est de la forme $f(X-\theta)$ *et si l'on choisit* $\pi_0(\theta) = 1$ *alors :*

- *si* $\mathbb{E}^\pi[\theta] = \mu$ *et* $\mathbb{V}^\pi(\theta) = \sigma^2$ *sont connus alors la théorie prédit :*

$$\pi(\theta) \propto e^{\lambda_1\theta + \lambda_2\theta^2}$$

  C'est donc la loi normale $\mathcal{N}(\theta, \sigma^2)$
- *si* $\mathbb{E}^\pi[\theta] = \mu$ *alors la théorie donne :*

$$\pi(\theta) = e^{\lambda\theta}$$

  On n'a donc pas de solution sur $\mathbb{R}$ *puisque dans ce cas ou bien* $\theta < 0$ *ou bien* $\theta > 0$

Ce dernier résultat est paradoxal : avec une information supplémentaire, la variance de θ, l'intervalle dans lequel évolue θ est agrandi, et une région exclue dans un cas plus large (le second) devient accessible ; cela n'est pas loin de signifier que la conclusion antérieure qu'une région doit être exclue n'est pas si évidente. Suivant les contraintes, il est donc possible de ne pas trouver de solutions.

De plus, le problème repose sur le choix de π_0 et non du modèle (ou de sa géométrie), cela constitue une limite de cette approche ou tout du moins, un point important à souligner.

6.2.2 Familles conjuguées

On considère une variable X suivant une fonction de densité paramétrique absolument continue par rapport à la mesure de Lebesgue : $X \sim f(X|\theta)$.

Définition 6.2.1 *Famille conjuguée*
On dit que la famille de lois *a priori* $\{\pi_\gamma, \gamma \in \Gamma\}$ est conjuguée si et seulement si :

$$\forall X, \forall \gamma \in \Gamma, \pi_\gamma(\theta|X) \in \{\pi_\gamma, \gamma \in \Gamma\}$$
$$\Leftrightarrow \forall \gamma \in \Gamma, \forall X, \exists \gamma'(X) \in \Gamma \text{ tel que } \pi_\gamma(\theta|X) = \pi_{\gamma'(X)}(\theta)$$

L'avantage des familles conjuguées est avant tout de simplifier les calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs.

Exemple 6.3 *Lois normales et inverse-gamma*
Pour $X \sim \mathcal{N}(\theta, \sigma^2)$, $\pi(\theta, \sigma^2) = \pi(\theta|X, \sigma^2)\pi(\sigma^2)$ *et* $\theta|\sigma^2 \sim \mathcal{N}(\mu, \tau\sigma^2)$, *et* $\sigma^2 \sim IG(a, b)$ *:*

$$\begin{aligned}\pi(\theta, \sigma^2, |X) &\propto \frac{\exp(-\frac{(X-\theta)^2}{2\sigma^2} - \frac{-(\theta-\mu)^2}{2\tau\sigma^2})}{\sigma^2} \times (\sigma^2)^{-(a+b)} e^{\frac{-b}{\sigma^2}} \\ &\propto \frac{1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(1+\frac{1}{\tau})\left[\theta - (X+\frac{\mu}{\tau})(1+\frac{1}{\tau})^{-1}\right]\right) \times \\ &\qquad \frac{e^{-\frac{X^2}{2\sigma^2} - \frac{\mu}{2\tau\sigma^2} - \frac{b}{\sigma^2}}}{(\sigma^2)^{a+\frac{3}{2}}} \times e^{\frac{1}{2\sigma^2}(X+\frac{\mu}{\tau})^2 \frac{\tau}{1+\tau}}\end{aligned}$$

Donc la loi a posteriori est : $(\theta|\sigma^2, X) \sim \mathcal{N}\left((X+\frac{\mu}{\tau})\frac{\tau}{1+\tau}, \sigma^2\frac{\tau}{1+\tau}\right)$
et $(\sigma^2|X) \sim IG\left(a+\frac{1}{2}, b+\frac{X^2}{2}+\frac{\mu^2}{2\tau} - \frac{(X+\mu)^2}{2}\frac{\tau}{1+\tau}\right)$.

Un autre exemple est celui des lois binomiales dont les lois $\pi(p) = Beta(a, b)$ constituent une famille conjuguée.

Définition 6.2.2 *Familles exponentielles*
La famille exponentielle regroupe les lois de probabilité qui admettent une densité de la forme : $f(X|\theta) = e^{\alpha(\theta)'T(X) - \phi(\theta)}h(X); \theta \in \Theta$. T *est alors une statistique exhaustive.*

Une telle famille est dite *régulière* si Θ est un ouvert tel que $\Theta = \left\{\theta / \int e^{\alpha(\theta)'T(X)}h(X)d\mu(X) < \infty\right\}$. En outre, on appelle *paramétrisation canonique*, l'écriture : $f(X|\theta) = e^{\theta'T(X) - \phi(\theta)}h(X)$ et *famille naturelle*, l'expression $f(X|\theta) = e^{\theta'T(X)}k(X)$. Nous rappelons qu'il est possible de montrer que :

$$\mathbb{E}_\theta(T(X)) = \nabla\phi(\theta)$$
$$\mathbb{V}_\theta(T(X)) = \nabla^2\phi(\theta)$$

Exemple 6.4 *Loi de Poisson*
Pour $X \sim \mathcal{P}(\lambda)$, $\mathbb{P}(X|\lambda) = e^{-\lambda}\frac{\lambda^X}{\lambda!} = \frac{e^{X\log\lambda - \lambda}}{\lambda!}$. *Le paramètre canonique est dans ce cas* $\theta = \log\lambda$.

Théorème 6.1 *Familles exponentielles*
Si $X \sim f(X|\theta) = e^{\theta' T(X) - \phi(\theta)} h(X)$, alors la famille de loi *a priori*

$$\left\{ \pi_{\lambda,\mu}(\theta) \propto e^{\theta'\mu - \lambda\phi(\theta)} h(X); \lambda, \mu \right\}$$

est conjuguée. On note que $\pi_{\lambda,\mu}$ est une densité de probabilité si et seulement si $\lambda > 0$ et $\frac{\mu}{\lambda} \in \Theta$.
Démonstration :

$$\begin{aligned} \pi_{\lambda,\mu}(\theta|X) &\propto e^{\theta' T(X) - \phi(\theta)} h(X) \cdot e^{\theta'\mu - \lambda\phi(\theta)} \\ &\propto e^{\theta'(T(X)+\mu) - (\lambda+1)\phi(\theta)} h(X) \\ &= \pi_{\lambda+1,\ \mu+T(X)}(\theta) \end{aligned}$$

■

Dans un cadre partiellement informatif, on utilise $\mathrm{E}^{\pi}[g_k(\theta)] = \mu_k$ pour calculer γ dans $\pi_\gamma, \gamma \in \Gamma$. Une fois déterminé $\hat{\gamma}_{\text{él}}$ on utilise $\pi_{\hat{\gamma}_{\text{él}}}$ [3]. Il faut noter qu'avec un nombre d'expériences réduit, une modification de Γ ou le choix d'un $\hat{\gamma}_{\text{él}}$ différent peuvent modifier sensiblement les résultats. Ce n'est plus le cas avec un grand nombre d'expériences comme on pouvait s'y attendre.

Exemple 6.5 *Si $X \sim \mathcal{P}(\lambda)$; la famille conjuguée classique est l'ensemble des lois Gamma : $\{\Gamma(a,b); a,b > 0\}$. Pour les données élicitées suivantes : $E^{\pi}(\lambda) = 2 : V^{\pi}(\lambda) = 4$*

Alors $E^{\pi}(\Gamma(a,b)) = \frac{a}{b} = 2$ et $V^{\pi}(\Gamma(a,b)) = \frac{a}{b^2} = 4$. D'où l'on tire $a = 1$ et $b = \frac{1}{2}$, c'est-à-dire : $\hat{\gamma}_{él} = (1, \frac{1}{2})$.

En l'absence de connaissance *a priori* sur γ, il peut être judicieux d'avoir recours à un modèle hiérarchique, c'est-à-dire à plusieurs étages. Si $(X|\theta) \sim f(X|\theta)$ et $(\theta|\gamma) \sim \pi_\gamma$, on munit le paramètre d'une loi $\gamma \sim q$, qui peut être non informative. Les calculs sont simplifiés dans le cas des lois conjuguées.

6.3 Approche non informative

Précisons dans un premier temps que la terminologie du titre est assez discutée dans la littérature. C'est l'approche utilisée pour construire une loi de référence, on parle aussi de loi par défaut ou encore de loi objective car c'est une démarche à mettre en place en l'absence d'information *apriori*, quand on ne veut pas influencer l'étude.

Nous nous plaçons dans un contexte différent : $X \sim f(X|\theta)$ et $\theta \in \Theta = \mathbb{R}^d$. Une première idée naturelle pour ne pas influencer *a priori* les résultats est de considérer chaque cas comme équiprobable et donc de prendre une mesure uniforme par rapport à la mesure de Lebesgue ou la mesure de comptage dans le cas discret. Le problème est que ce choix correspond à

3. L'indice « él » se réfère à *élicité* signifiant mis en valeur.

la sélection d'un paramètre particulier ; en effet, il n'est pas invariant par reparamétrisation comme nous l'avons déjà montré précédemment. Cette méthode peut aussi présenter l'inconvénient présenté dans l'exemple qui suit.

Exemple 6.6 *Problèmes pathogènes*
Dans le cas où $X \sim \mathcal{B}(n,p)$, $\theta = p$ *et pour* $p(\lambda) = 1 - e^{-\theta d}$, $\theta > 0$ *avec* $d \geq 10^2$, *cette démarche s'avère être très (trop) informative (proche d'une Dirac).*

6.3.1 Lois de Jeffreys et Bernardo

Une seconde idée est d'adopter une approche invariante par reparamétrisation. Jeffreys introduit la mesure suivante $\pi_J(\theta) \propto \sqrt{|I(\theta)|}$ en utilisant l'information de Fisher I. Vérifions que c'est effectivement une loi de probabilité invariante par reparamétrisation : si $\eta = g(\theta)$ avec $g \in C^1$, alors $\pi_J(\eta) = \sqrt{|I(\eta)|}$. En outre, $\tilde{I}(\eta) = \nabla g(\theta)' I(\theta) \nabla g(\theta)$; il s'en suit que dans le cas de la dimension 1, on a :

$$\pi_J(\eta) = \sqrt{|I(\eta)|} = \sqrt{|I(\theta)|} \left|\frac{d\theta}{d\eta}\right| = \pi_J(\theta) \left|\frac{d\theta}{d\eta}\right|$$

Ainsi la loi *a priori* de Jeffreys est invariante par reparamétrisation. Malgré l'immense intérêt d'une telle propriété, il faut savoir que la loi *a priori* de Jeffreys n'a de bonnes propriétés que dans le cas des petites dimensions et en particulier de la dimension 1. La limite de l'*a priori* de Jeffreys en grande dimension peut se comprendre sur l'exemple suivant.

Si $X \sim \mathcal{N}(\mu, \sigma^2 I_p)$ et si l'on veut estimer $\theta = \|\mu\|^2$ alors comme $\ell(\mu, \sigma^2) = \frac{-\|X-\mu\|^2}{2\sigma^2} - \frac{p}{2}\log\sigma^2$ on a :

$$\frac{\partial \ell}{\partial \mu_j} = \frac{X_j - \mu_j}{2\sigma^2} \quad ; \quad \frac{\partial^2 \ell}{\partial \mu^2} = \frac{1}{\sigma^2}$$
$$\frac{\partial \ell}{\partial \sigma} = \frac{\|X-\mu\|^2}{\sigma^3} - \frac{p}{2} \quad ; \quad \frac{\partial^2 \ell}{\partial \sigma^2} = -3\frac{\|X-\mu\|^2}{\sigma^4} + \frac{p}{\sigma^2}$$

La matrice d'information de Fisher est alors :

$$I = \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 1 & 1 \\ 0 & 0 & p - 3\frac{\|X-\mu\|^2}{\sigma^2} \end{pmatrix}$$

On a donc $|I| \propto \frac{1}{\sigma^{2p}}$ et par conséquent $\pi_J \propto \frac{1}{\sigma^p}$. Avec l'augmentation de la dimension, trop d'importance est accordée à σ et c'est sous optimal. Car, comme $\mathrm{E}^\pi\left[\|\mu\|^2 | X\right] = \|X\|^2 + p$, on sait que le meilleur estimateur se trouve dans la classe $\left\{\|X\|^2 + c; c \in \mathbb{R}\right\}$ et c'est $\|X\|^2 - p$.

Il existe une variante de l'approche précédente : la loi de référence de Bernardo dont l'idée est de reposer sur un critère d'objectivité. On regarde ici la loi a priori qui amène le moins d'information par rapport à ce que pouvait fournir les données. On s'intéresse donc à la fonction de π, la loi *a priori*, suivante :

$$I_n(\pi) = \int K\left(\pi(\theta|X^n); \pi(\theta)\right) m_\pi(X^n) dX^n$$

En outre, on peut vérifier que :

$$I_n(\pi) \underset{n\to+\infty}{\sim} \frac{d}{2}\log\frac{n}{2\pi} - \int \pi(\theta)\log\frac{\pi}{\sqrt{I(\theta)}}\, d\theta + \mathcal{O}_p(1)$$

Dans le cas d'un modèle régulier $I_n(\pi)$ est donc maximal lorsque n tend vers $+\infty$ si $\pi(\theta) \propto \sqrt{I(\theta)}$.

Bernardo introduit la stratégie suivante avec un paramètre θ de la forme $\theta = (\theta_1, \theta_2)$ où θ_1 est un paramètre d'intérêt et θ_2 est un paramètre de nuisance. On raisonne alors séquentiellement en écrivant : $\pi(\theta_1, .) = \pi(.|\theta_1)\pi(\theta_1)$ et en prenant pour θ_1 la loi de Jeffreys. Cette manière de faire peut se généraliser si $\theta = (\theta_1, \cdots, \theta_n)$ où l'on a ordonné sans perte de généralité les θ_i par intérêt croissant. Un algorithme consiste à prendre pour θ_1 la loi de Jeffreys, puis de calculer $\pi(\theta_2|\theta_1)$ la plus objective puis $\pi(\theta_3|\theta_1, \theta_2)$ la plus objective, etc. Cependant, il est clair ce raisonnement n'est pas purement objectif parce que donner plus d'importance à un paramètre qu'à un autre relève une fois encore d'un choix.

6.3.2 Loi a priori de concordance – (*matching priors*)

Le but est de trouver une loi a priori cernant le paramètre θ qui se rapproche le plus possible de la méthode de choix fréquentiste, cela revient à faire en sorte que le tirage X n'influence pas le résultat.

On rappelle dans un premier temps des exemples d'une région de confiance ou $\alpha-$crédible. Elle peut être par exemple un intervalle unilatéral $\left\{\theta \leq \theta_d^{(X)}\right\}$ ou bien bilatéral $\{\theta_{\alpha,1} \leq \theta \leq \theta_{\alpha,2}\}$. Il peut s'agir aussi de région HPD, $\theta \in C_\alpha^\pi$ avec par exemple $\left\{\log(\hat{(}\theta) - \log(\theta) \leq h_\alpha\right\}$ tel que $P^\pi(\theta \in C|X) = 1 - \alpha$.

On cherche π tel que $\forall\theta; P_\theta(\theta \in C) = 1 - \alpha$, appelé la *parfaite concordance*. C'est en général impossible. On va alors chercher r_n le plus petit possible tel que :

$$\forall\theta \in \Theta, \forall\alpha \in]0,1[\,,\ P_\theta(\theta \in C) = 1 - \alpha + \mathcal{O}(r_m)$$

La loi a priori est alors dite *concordante à l'ordre* r_m.

Lorsque le modèle est régulier et si π est continu alors :

$$P^\pi\left[\sqrt{n}(\hat{\theta} - \theta) \leq t|X\right] = \Phi(tI^{\frac{1}{2}}) + \mathcal{O}_p(\frac{1}{n})$$

Si $C = \{\theta \leq \theta_\alpha^\pi\}$ où θ_α^π est tel que $P^\pi(\theta \leq \theta_\alpha^\pi|X) = 1-\alpha$. Alors pour évaluer $P_\theta(\theta \leq \theta_\alpha^\pi)$, il faut une approximation de θ_α^π :

$$\begin{aligned} P^\pi(\theta \in [\hat{\theta} + \frac{I^{\frac{-1}{2}}}{\sqrt{n}}\Phi^{-1}(1-\alpha)]|X) &= 1-\alpha+\mathcal{O}(\frac{1}{\sqrt{n}}) \\ \theta_\alpha^\pi &= \hat{\theta} + \frac{I^{\frac{-1}{2}}}{\sqrt{n}}\Phi^{-1}(1-\alpha) \\ P_\theta(\theta \leq \hat{\theta} + \frac{I^{-\frac{1}{2}}}{\sqrt{n}}\Phi^{-1}(1-\alpha)) &= P_\theta(\sqrt{nI}(\hat{\theta}-\theta) \geq -\Phi^{-1}(1-\alpha)) \\ &= 1-\alpha+\mathcal{O}(\frac{1}{\sqrt{n}}) \end{aligned}$$

C'est dans le terme $\mathcal{O}(\frac{1}{\sqrt{n}})$ qu'intervient l'influence de la loi *a priori* π. Eliminer le terme dominant permet d'obtenir une loi concordante à un ordre plus poussé et donc plus précise.

Pour tout π continu, $P^\pi(\theta \leq \theta_\alpha^\pi|X) = P_\theta(\theta \leq \theta_\alpha^\pi) + \mathcal{O}(\frac{1}{\sqrt{n}})$, or nous cherchons π tel que $\forall\alpha;\forall\theta; P^\pi(\theta \leq \theta_\alpha^\pi|X) = P_\theta(\theta \leq \theta_\alpha) + \mathcal{O}(\frac{1}{n})$. Il faut alors développer avec la formule de Bernstein Von Mises.

Les développements de $\pi(\theta|X)$ à des ordres supérieurs sont basés sur des expressions de type Laplace :

$$\theta_\alpha^\pi = \hat{\theta} + \frac{I^{-\frac{1}{2}}}{\sqrt{n}}\Phi^{-1}(1-\alpha) + \frac{H(X^n)\phi(\Phi^{-1}(1-\alpha))}{n} + \mathcal{O}(\frac{1}{n^{\frac{3}{2}}})$$

où $H(X^n)$ est une fonction des paramètres $\hat{\theta}$ et $\frac{\partial^2 \log(\theta)}{\partial\theta^2}$.

A partir de cette expression, on calcule

$$P_\theta\left(\theta \leq \hat{\theta} + \frac{I^{-\frac{1}{2}}}{\sqrt{n}}\Phi^{-1}(1-\alpha) + \frac{H(X^n)\phi(\Phi^{-1}(1-\alpha))}{n} + \mathcal{O}(n^{-\frac{3}{2}})\right)$$

et on obtient le développement d'Edgeworth, à savoir que cette probabilité vaut : $1-\alpha+\frac{P_1(\pi,\theta)(1-\alpha)}{\sqrt{n}}+\mathcal{O}(\frac{1}{n})$.

On en tire un résultat important, π sera de concordance à l'ordre 1 pour les intervalles unilatéraux si et seulement si $\forall\theta \in \Theta, P_1(\pi,\theta) = 0$. Il est ensuite possible de montrer que $\forall\theta P_1(\pi,\theta) = 0 \Longleftrightarrow \forall\theta\pi(\theta) \propto \sqrt{I(\theta)}$. Nous retrouvons alors la loi *a priori* de Jeffreys.

Chapitre 7

Méthodes numériques

Ce chapitre développe succinctement les deux principales approches des méthodes numériques bayésiennes.

7.1 Approches indépendantes

Le problème généralement rencontré est celui du calcul d'une intégrale de la forme :

$$I_\pi(h) = \int h(\theta)\pi(\theta)d\theta$$

Par exemple, le calcul de la loi *a posteriori* : $\pi(\theta|X) = \frac{f(X|\theta\pi(\theta)}{m(X)}$ exige le calcul de $m(X)$. De même, $\hat{\theta} = \int \theta\pi(\theta)d\theta$ peut aussi s'écrire $I_\pi(id)$.

Dès que la dimension de l'espace d'intégration est supérieure ou égale à 3 le calcul numérique est délicat, c'est pourquoi on a recours à des méthodes de simulation.

La première idée se base sur le principe de Monte Carlo, à l'origine développé pour les sciences physique. Pour $\theta^t \overset{iid}{\sim} \pi$ où $t = 1, \cdots T$ avec un T grand devant 1, $I_\pi(h)$ est approché par :

$$\widehat{I}_\pi(h) = \frac{1}{T}\sum_{t=1}^{T} h(\theta^t)$$

Notons qu'il peut être difficile de simuler sur π.

Une deuxième idée est de baser l'estimation sur l'échantillonnage d'importance. Il s'agit dans un premier temps de définir une densité instrumentale q puis par un jeu d'écriture de noter :

$$I_\pi(h) = \int h(\theta)\frac{\pi(\theta)}{q(\theta)}q(\theta)d\theta$$

Le principe repose sur $\theta^t \sim q$ et l'estimation est alors :

$$\widehat{I}_\pi(h) = \frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta^t)}{q(\theta)}\pi(\theta^t)$$

En se qui concerne la variance :

$$\mathbb{V}(\widehat{I}_\pi(h)) = \int \frac{h^2(\theta)\pi^2(\theta)}{q(\theta)} q(\theta)d\theta - \widehat{I}_\pi(h)^2$$

D'après cette formule, si $\frac{\pi^2}{q}$ est grand alors la variance explose et le calcul devient périlleux, il est nécessaire que, dans une certaine mesure, q domine π. Ainsi le choix de q est critique alors qu'il est arbitraire. Le problème se résout si $q \approx \pi$; seulement trouver un tel q n'est pas évident. Et de toute façon ceci marche très mal en grande dimension [1].

La section suivante présente à grands traits les principes des méthodes de Monte Carlo par Chaînes de Markov MCMC. Lectrices et lecteurs sont incités à consulter l'ouvrage de référence sur le sujet de Christian P. Robert [3].

7.2 Méthodes MCMC

Le but des méthodes MCMC est de simuler selon π et l'idée de base est de construire une chaîne de Markov ergodique de loi stationnaire π.

7.2.1 Algorithme Hasting-Metropolis

Pour $\theta^{(0)}$ est une valeur initiale, on définit par récurrence les valeurs de $\theta^{(t)}$.

A l'étape t, à partir de $\theta^{(t-1)}$, $\theta^{(t)}$ est construit en tirant un θ' à l'aide d'une distribution de probabilité instrumentale : $\theta' \sim q(\big|\theta^{(t-1)})$. $\theta^{(t)}$ est alors donné par :

$$\theta^{(t)} = \begin{cases} \theta' & \text{avec une probabilité } \alpha(\theta', \theta^{(t-1)}) \\ \theta^{(t-1)} & \text{avec une probabilité } 1 - \alpha(\theta', \theta^{(t-1)}) \end{cases}$$

où $\alpha(\theta', \theta^{(t-1)}) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{t-1})}\frac{q(\theta^{(t-1)}|\theta')}{q(\theta'|\theta^{(t-1)})}, 1\right)$. Notons qu'il est possible suivant cette construction de rester au même endroit après une itération. On peut alors montrer en écrivant la condition de balance, que pour ce choix de α, on obtient une chaîne de Markov de loi stationnaire π.

Cette chaîne de Markov est ergodique si et seulement si $(\theta^{(t)})_t$ est irréductible et apériodique.

1. l'erreur est exponentielle en la dimension, c'est le problème bien connu de *curse of dimensionality*.

Exemple 7.1 *Proposition indépendante*
$\theta \sim q(\theta')$ *ne dépend pas de* θ^{t+1}. $\alpha(\theta'|\theta^{t-1}) = \frac{\pi(\theta')}{\pi(\theta^{t-1})}\frac{q(\theta^{t-1})}{q(\theta')}$ *marche bien si* $q > \pi$ *ou* $\frac{\pi}{q}$ *borné.*

Exemple 7.2 *Marche aléatoire (ou symétrique)*
$q(\theta|\theta') = q(\theta'|\theta)$ *et* $\theta' = \theta^{(t-1)} + \epsilon\sigma$ *où* ϵ *a une loi symétrique par rapport à 0.*

Le choix de σ *est ici primordial. Comme* $\alpha(\theta', \theta^{t-1}) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{t-1})}, 1\right)$, *un* σ *trop faible ne permet pas d'explorer tout le support puisque les* $\theta^{(t)}$ *sont proches les uns des autres ; si* σ *est grand, alors on refuse souvent* θ' *et donc on ne profite pas bien du nombre de simulations. Empiriquement, le taux d'acceptation optimal est entre 0,1 et 0,6, il faut donc choisir* σ *en fonction.*

7.2.2 Algorithme de type Gibbs

Pour $\theta = (\theta_1, \cdots, \theta_p)$, on veut simuler $\pi(\theta)$ à partir de $\pi_i(\theta_i|\theta_{(-i)}) = \pi_i(\theta_i|\theta_j, j \neq i)$ pour tout i. On initialise avec $\theta^{(0)}$ et à l'instant t, on écrit :

$$\begin{aligned}
(\theta_1^{(t)}|\theta^{(t-1)}) &\sim \pi_1(\theta_1^{(t)}|\theta_{(-1)}^{(t-1)}) \\
(\theta_2^{(t)}|\theta^{(t-1)}, \theta_1^{(t)}) &\sim \pi_2(\theta_2^{(t)}|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}) \\
\vdots &\sim \vdots \\
(\theta_p^{(t)}|\theta^{(t-1)}, \theta_{(-p)}^{(t)}) &\sim \pi_p(\theta_p^{(t)}|\theta_{(-p)}^{(t)})
\end{aligned}$$

Dans le cas où une telle loi π existe, $\theta^{(t)}$ issu de cet algorithme est une chaîne de Markov ergodique de loi stationnaire π.

En pratique, on combine souvent les deux grandes approches précédentes comme le montre l'exemple qui suit.

Exemple 7.3 *Lois normales*
$(X_{ij}|\mu_j, \sigma_j^2) \sim \mathcal{N}(\mu_j, \sigma_j^2)$ *avec* $\mu_j \sim \mathcal{N}(\mu_0, \tau_0^2)$ *et* $\sigma_j^2 \sim IG(a, b)$ *on peut récupérer* τ_0, μ_0, a, b *à partir des lois conjuguées mais les calculs ne sont pas si simples. On considère plutôt* $\pi(\mu_0|X_i, \tau_0, a, b) \propto \tilde{\pi}(\tau_0, \mu_0, a, b)$ *et* $(\mu_0^t|\tau_0^t, a^t, b^t) \sim q(\cdot|\mu_0^{t-1}, \tau_0^t, a^t, b^t)$ *; c'est alors une chaîne de Markov de loi stationnaire* π.

Conclusion générale

Comme conclusion, nous suggérons :

Les fréquentistes sont-ils fréquentables ? La question est posée... [2]

Bibliographie

[1] I. A. Ibragimov and R. Z. Khas'minskii. On nonparametric estimation of the value of a linear functional in gaussian white noise. *Theory of Probability and its Applications*, 29(1), 1985.

[2] Christian P. Robert. *Le Choix Bayésien - Principes et pratique*. Springer, 2006.

[3] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

2. Bien entendu, ceci n'est qu'un point de vue.