Modélisation et statistique bayésienne computationnelle

Corrigé des exercices et démonstrations des résultats
(document évolutif)

nicolas.bousquet@sorbonne-universite.fr

30 mars 2023

Master 2, Sorbonne Université, 2023

Table des matières

1 Corrigés et preuves "Introduction et Théorie de la décision"

1.1 Exercices

Exercice 1 (Adapté de [1]) *Soient (x_1, x_2) deux réalisations aléatoires. Nous disposons de deux candidats pour la loi jointe de ces observations : $x_i \sim \mathcal{N}(\theta, 1)$ ou encore*

$$g(x_1, x_2|\theta) \quad = \quad \pi^{-3/2} \frac{\exp\left\{-(x_1 + x_2 - 2\theta)^2/4\right\}}{1 + (x_1 - x_2)^2}.$$

Quel est l'estimateur du maximum de vraisemblance de θ dans chacun des cas ? Que constate-on ?

Réponse. La vraisemblance dans les deux cas est

$$\ell(\theta|x_1, x_2) \quad \propto \quad \exp\left\{-(\bar{x} - \theta)^2\right\}$$

et qui devrait donc conduire à la même inférence sur θ. Mais $g(x_1, x_2|\theta)$ est très différente de la première distribution (par exemple, l'espérance de $x_1 - x_2$ n'est pas définie). Les estimateurs de θ auront donc des propriétés fréquentistes différentes s'ils ne dépendent pas que de $\bar{x}$ (**ex** : estimateur des moments). En particulier, les régions de confiance pour θ peuvent différer fortement car g possède des queues plus épaisses.

Exercice 2 (Bayes (1763)) *Une boule de billard Y_1 roule sur une ligne de longueur 1, avec une probabilité uniforme de s'arrêter n'importe où. Supposons qu'elle s'arrête à la position θ. Une seconde boule Y_2 roule alors n fois dans les mêmes conditions, et on note X le nombre de fois où Y_2 s'arrête à gauche de Y_1. Connaissant X, quelle inférence peut-on mener sur θ ?*

Réponse. Notons $\mathbb{P}(Y_2|\theta)$ la mesure de probabilité associée à Y_2. On définit la gauche par l'événement $0 \leq Y_2 < \theta \leq 1$. Alors

$$\mathbb{P}(Y_2 < \theta|\theta) \quad = \quad \theta$$

et l'indicatrice $\mathbb{1}_{Y_2<\theta}$ suit alors une loi de Bernoulli $\mathcal{B}(\theta)$. Alors après n répétitions iid, on a

$$X \quad = \quad \sum_{i=1}^{n} \mathbb{1}_{Y_{2,i}<\theta} \quad \sim \quad \mathcal{B}(n, \theta) \quad \text{(loi binomiale)}.$$

Si on connaît une réalisation x de X, la vraisemblance statistique associée est donc

$$\mathbb{P}(X = x|\theta) \quad = \quad \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1 - \theta)^{n-x}.$$

Comme $\theta \in [0, 1]$ et qu'on ne dispose pas d'information *a priori* sur θ, on peut simplement supposer *a priori*

$$\pi(\theta) \quad = \mathbb{1}_{\{0 \leq \theta \leq 1\}}(\theta).$$

On en déduit la loi *a posteriori* (en utilisant le symbole de proportionalité $\propto$)

$$\pi(\theta|X = x) \quad \propto \quad \theta^x (1 - \theta)^{n-x} \mathbb{1}_{\{0 \leq \theta \leq 1\}}(\theta)$$

qui correspond au terme général de la loi bêta $\mathcal{B}_e(1 + x, 1 + n - x)$ d'espérance $(1 + x)/(2 + n) \sim x/n$ si $(x, n) \gg 1$.

Exercice 3 (Loi gaussienne / loi exponentielle) *Soit une observation $x \sim \mathcal{N}(\theta, \sigma^2)$ où σ^2 est connu. On choisit a priori*

$$\theta \quad \sim \quad \mathcal{N}(m, \rho\sigma^2)$$

Quelle est la loi a posteriori de θ sachant x ? Même question en supposant que $X \sim \mathcal{E}(\lambda)$ et

$$\lambda \quad \sim \quad \mathcal{G}(a, b).$$

Exercice 4 (Loi uniforme généralisée) *Soit* $X \sim \mathcal{N}(\mu, \sigma^2)$ *et* $d\pi(\mu) = d\mu$ *(mesure de Lebesgue). Que vaut* $m_\pi(x)$ *?*

Réponse. On a

$$m_\pi(x) \quad = \quad \sigma\sqrt{2\pi} \; < \; \infty$$

donc la mesure de Lebesgue sur $\mathbb{R}$ peut être utilisée.

Exercice 5 (Loi d'échelle) *Soit* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ *et* $\pi(\mu, \sigma) = 1/\sigma$ *avec* $\Theta = \mathbb{R} \times \mathbb{R}_*^+$. *Que vaut* $m_\pi(x_1, \ldots, x_n)$ *? La mesure* $\pi(\mu, \sigma)$ *peut-elle être utilisable ?*

Réponse. L'intégrale de la loi *posteriori* s'écrit

$$\begin{aligned} m_\pi(x_1, \ldots, x_n) &= \int_{\mathbb{R}} \int_0^\infty \exp\left(-\frac{\bar{x_n} - \mu)^2}{2\sigma^2}\right) \exp\left(\frac{\sum_{i=1}^n (x_i - \bar{x_n})^2}{2\sigma^2}\right) \frac{d\mu d\sigma}{\sigma^{n+1}} \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\sum_{i=1}^n (x_i - \bar{x_n})^2}{2\sigma^2}\right) \frac{d\sigma}{\sigma^n}. \end{aligned}$$

Pour que l'expression converge, il faut avoir $n > 1$ et la propriété suivante vérifiée :

$$\exists (i,j) \in \mathbb{N}^*, \; i \neq j \;\text{ tel que }\; X_i \neq X_j.$$

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 ci-dessus : pour estimer la dispersion (variance), au moins deux observations non égales sont nécessaires.

Exercice 6 *Soient* x_1 *et* x_2 *deux observations de la loi définie par*

$$P_\theta(x = \theta - 1) \quad = \quad P_\theta(x = \theta + 1) \; = \; 1/2 \quad \textit{avec } \theta \in \mathbb{R}$$

Le paramètre d'intérêt est θ *(donc* $\mathcal{D} = \Theta$*) et il est estimé par* δ *sous le coût*

$$L(\theta, \delta) \quad = \quad 1 - \mathbb{1}_\theta(\delta)$$

appelé coût 0-1*, qui pénalise par 1 toutes les erreurs d'estimation quelle que soit leur magnitude (grandeur). Soit les estimateurs*

$$\begin{aligned} \delta_1(x_1, x_2) &= \frac{x_1 + x_2}{2}, \\ \delta_2(x_1, x_2) &= x_1 + 1, \\ \delta_3(x_1, x_2) &= x_2 - 1. \end{aligned}$$

Calculez les risques $R(\theta, \delta_1)$*,* $R(\theta, \delta_2)$ *et* $R(\theta, \delta_3)$*. Quelle conclusion en tirez-vous ?*

Réponse. Calculons le risque fréquentiste pour δ_1. On obtient :

$$\begin{aligned} R(\theta, \delta_1) &= \mathbb{E}_\theta \left[L(\theta, \delta_1(X)) \right], \\ &= \int_\Omega L(\theta, \delta_1(x)) f(x|\theta) \, dx. \end{aligned}$$

La fonction $L(\theta, \delta_1(x))$ vaut 0 si $\delta = \theta$ (bonne décision) et 1 sinon. On en déduit que

$$\begin{aligned} R(\theta, \delta_1) &= \int_{\{\theta-1, \theta+1\}} (1 - \mathbb{1}_\theta((x_1 + x_2)/2) \, f(x_1, x_2|\theta) \, dx_1 dx_2, \\ &= \frac{1}{4} \left\{ (1 - \mathbb{1}_\theta((\theta - 1 + \theta - 1)/2)) + (1 - \mathbb{1}_\theta((\theta + 1 + \theta - 1)/2)) + \right. \\ &\quad \left. (1 - \mathbb{1}_\theta((\theta + 1 + \theta + 1)/2)) + (1 - \mathbb{1}_\theta((\theta - 1 + \theta + 1)/2)) \right\}, \\ &= \frac{1}{4}(1 + 1) \; = \; 1/2. \end{aligned}$$

(l'estimateur est correct la moitié du temps). On trouve alors, similairement,

$$R(\theta,\delta_1) \;=\; R(\theta,\delta_2) \;=\; R(\theta,\delta_3) \;=\; 1/2$$

ce qui signifie qu'on ne peut pas classer les estimateurs sous le coût fréquentiste 0-1.

Exercice 7 *Lorsqu'on fait un choix de fonction de coût $L(\theta,\delta)$ dans un ensemble $U : \Theta \times \mathcal{D} \to \Lambda \in I\!R^+$, on commet une erreur par rapport à la meilleure fonction de coût possible pour le problème. On peut donc proposer un estimateur bayésien de cette fonction de coût en introduisant une fonction de coût sur les fonctions de coût $L(\theta,\delta)$:*

$$\begin{aligned}\tilde{L} : \Theta \times U \times \mathcal{D} &\to I\!R^+\\ (\theta,\ell,\delta) &\mapsto \tilde{L}(\theta,\ell,\delta).\end{aligned}$$

Quel est l'estimateur bayésien de $\tilde{L}(\theta,\ell,\delta)$ sous un coût quadratique, lorsque $L(\theta,\delta)$ est elle-même quadratique ?

Réponse. Si $\tilde{L}(\theta,\ell,\delta) = (\ell - L(\theta,\delta))^2$ et $L(\theta,\delta) = (\theta-\delta)^2$, alors l'estimateur bayésien est

$$\begin{aligned}\ell^\pi &= \mathbb{E}_\pi\left[L(\theta,\delta)|X\right]\\ &= \mathbb{E}_\pi\left[(\theta-\delta)^2|X\right].\end{aligned}$$

En choisissant en toute logique $\delta = \delta^\pi = \mathbb{E}[\theta|X]$ il vient donc

$$\ell^\pi = \mathrm{Var}_\pi[\theta|X].$$

Exercice 8 *Soit $X \sim \mathcal{B}(\theta)$ (loi de Bernoulli) avec $\Theta = [0,1]$. Soit M_0 un modèle défini par $\{\theta = 1/2\}$ et M_1 un modèle défini par un θ inconnu dans $[0,1]$, avec $\pi_1(\theta) = \mathcal{U}[0,1]$. Un échantillon de 200 tirages fournit 115 succès et 85 échecs. Au vu de ces données, quel modèle choisir ? Ce résultat diffère-t-il significativement d'un test fréquentiste ?*

Réponse. La vraisemblance des données X est binomiale :

$$f(x|\theta) = \left(\begin{array}{c}200\\115\end{array}\right)\theta^{115}(1-\theta)^{85}$$

ce qui permet de calculer

$$P(X|M_0) \;=\; \left(\begin{array}{c}200\\115\end{array}\right)(1/2)^{200} \;\simeq\; 0.006$$

alors que pour M_1 on a

$$P(X|M_1) \;=\; \int_0^1 \left(\begin{array}{c}200\\115\end{array}\right)\theta^{115}(1-\theta)^{85}\,d\theta \;=\; 1/201 \;\simeq\; 0.005.$$

Le facteur de Bayes vaut alors 1.2, ce qui indique uniquement que la certitude que H_0 est vraie est faible (on pointe très légèrement vers le modèle M_0).

Un test d'hypothèse fréquentiste de M_0 indiquerait que M_0 doit être rejeté par exemple au niveau de signification 5%, car la probabilité d'obtenir 115 succès ou plus à partir d'un échantillon de 200 si $\theta = 1/2$ est de 0.02. On en conclut qu'un test classique donnerait des résultats significatifs permettant de rejeter H_0 tandis qu'un test bayésien ne pourrait considérer le résultat comme extrême.

Exercice 9 Sélection de modèle discret avec des priors impropres.

1. *Pour des données discrètes $x_1,\ldots,x_n$, on considère un modèle de Poisson $\mathcal{P}(\lambda)$ ou une loi binomiale négative $\mathcal{NB}(m,p)$ avec les a priori*

$$\begin{aligned}\pi_1(\lambda) &\propto 1/\lambda\\ \pi_2(m,p) &= \frac{1}{M}\mathbb{1}_{\{1,\ldots,M\}}(m)\mathbb{1}_{[0,1]}(p)\end{aligned}$$

Peut-on sélectionner l'un des deux modèles ?

2. *Si on remplace $\pi_1(\lambda)$ par un a priori* **vague**

$$\pi_1(\lambda) \equiv \mathcal{G}(\alpha, \beta)$$

avec $\alpha(\beta)$ ou/et $\beta(\alpha) \to 0$, peut-on de nouveau résoudre le problème ?

Réponse.

1. Il existe une constante inconnue $\gamma > 0$ telle que

$$\begin{aligned} B_{12}(\mathbf{x_n}) &= \gamma \frac{\int_0^\infty \frac{\lambda^{\sum_i (x_i-1)}}{\prod_i x_i!} \exp(-n\lambda)\ d\lambda}{\frac{1}{M} \sum_{m=1}^{M} \int_0 \infty \left(\prod_i \begin{pmatrix} m \\ x_i - 1 \end{pmatrix} \right) p^{\sum_i x_i} (1-p)^{m \cdot n - \sum_i x_i}\ dp}, \\ &= \gamma M \left(\sum_{m=1}^{M} \begin{pmatrix} m \\ x-1 \end{pmatrix} \frac{x!(m-x)!}{m!} \right)^{-1} \quad \text{si } n = 1 \text{ et } x_i = x \\ &= \gamma M \left(\sum_{m=1}^{M} x/(m-x+1) \right)^{-1} \end{aligned}$$

Impossible de faire un choix car γ n'est pas connu !

2. Si on remplace $\pi_1(\lambda)$ par un *a priori* **vague**

$$\pi_1(\lambda) \equiv \mathcal{G}(\alpha, \beta)$$

avec $\alpha(\beta)$ ou/et $\beta(\alpha) \to 0$, on obtient après quelques calculs (pour $n = 1$ et $x_i = x$)

$$\begin{aligned} B_{12} &= \frac{\Gamma(\alpha + x)}{x!\Gamma(\alpha)} \beta^{-x} \left[\frac{1}{M} \sum_{m=1}^{M} \frac{x}{m-x+1} \right]^{-1} \\ &= \frac{(x+\alpha-1)\ldots\alpha}{x(x-1)\ldots 1} \beta^{-x} \left[\frac{1}{M} \sum_{m=1}^{M} \frac{x}{m-x+1} \right]^{-1} \end{aligned}$$

qui dépend fortement du choix de $\alpha(\beta)$ ou/et $\beta(\alpha) \to 0$. On ne résout donc pas le problème...

1.2 Démonstrations

Théorème 1 *(Th.2 dans le poly de cours) Pour chaque $x \in \Omega$,*

$$\delta^\pi(x) = \arg\min_{d \in \mathcal{D}} R_P(d|\pi, x). \tag{1}$$

Un corollaire est le suivant : s'il existe $\delta \in \mathcal{D}$ tel que $R_B(\delta|\pi) < \infty$, et si $\forall x \in \Omega$ l'équation (1) est vérifiée, alors $\delta^\pi(x)$ est un estimateur de Bayes.

Preuve. Nous prouvons ici qu'un estimateur minimisant le risque intégré R_B est obtenu par sélection, pour chaque valeur $x \in \Omega$, de la valeur $\delta(x)$ qui minimise le coût moyen *a posteriori*. Il s'agit en pratique d'une méthode de calcul d'un estimateur bayésien. En effet, L'application du théorème de Fubini est possible par la

finitude des intégrales impliquées ci-dessous : avec $L(\theta, \delta(x)) \geq 0$, il vient

$$\begin{aligned} R_B(\delta|\pi) &= \int_\Theta \int_\Omega L(\theta, \delta(x)) f(x|\theta)\, dx \pi(\theta)\, d\theta, \\ &= \int_\Omega \int_\Theta L(\theta, \delta(x)) f(x|\theta) \pi(\theta)\, d\theta dx, \\ &= \int_\Omega \int_\Theta L(\theta, \delta(x)) \pi(\theta|x)\, d\theta m_\pi(x)\, dx, \\ &= \int_\Omega R_P(\delta|\pi, x) m_\pi(x)\, dx. \end{aligned}$$

On peut en déduire que pour tout $\delta \in \mathcal{D}$, $R_P\left(\delta^\pi(x)|\pi, x\right) \leq R_P\left(\delta|\pi, x\right)$implique $R_B\left(\delta^\pi|\pi\right) \leq R_B\left(\delta|\pi\right)$ce qui permet de conclure la démonstration du corollaire.

Théorème 2 *(Th.4 dans le poly de cours) Si un estimateur de Bayes δ^π associé à une mesure a priori π (probabiliste ou non) est tel que le risque $R(\theta, \delta^\pi) < \infty$ et si la fonction $\theta \mapsto R(\theta, \delta)$ est continue sur Θ, alors δ^π est admissible.*

Preuve. Soit un estimateur bayésien δ^π de risque R fini. Pour $\delta_0 \in \mathcal{D}$ tel que $\forall \theta \in \Theta$, $R(\theta, \delta) \leq R(\theta, \delta_0)$, on note

$$\mathcal{A}_0 = \left\{\theta \in \Theta,\ R(\theta, \delta) \leq R(\theta, \delta_0)\right\}.$$

On a alors

$$\int_\Theta R(\theta, \delta_0) d\Pi(\theta) - \int_\Theta R(\theta, \delta^\pi) d\Pi(\theta) = \int_{\mathcal{A}_0} \left(R(\theta, \delta) - R(\theta, \delta_0)\right) d\Pi(\theta) \leq 0$$

avec égalité si et seulement si $\pi(\mathcal{A}_0) = 0$. Or, comme δ^π est bayésien et le risque fini, $R(\theta, \delta_0) \geq R(\theta, \delta^\pi)$. Donc l'intégrale ci-dessus est négative et positive, donc nulle, ce qui sous-entend qu'en effet $\pi(\mathcal{A}_0) = 0$ (on dit alors que δ^π est $\pi-$admissible). Supposons cependant que δ^π n'est pas admissible. On déduit de la démarche précédente que $\exists \delta_0$ tel que $\forall \theta$ tel que $R(\theta, \delta_0) \leq R(\theta, \delta^\pi)$ et $\theta_0 \in \Theta$ tel que $R(\theta, \delta_0) < R(\theta, \delta^\pi)$. La fonction définie sur Θ par $\theta \to R(\theta, \delta_0) - R(\theta, \delta^\pi)$ est continue par hypothèse. Donc il existe un voisinage ouvert $V_0 \subset \Theta$ de θ_0 tel que $\forall \theta \in V_0$, $R(\theta, \delta_0) < R(\theta, \delta^\pi)$. On a $\pi(\mathcal{A}_0) \geq \pi(V_0)$. Or π est supposé strictement positive sur Θ, donc $\pi(V_0) > 0$. L'ensemble $\mathcal{A}_0$ est donc de mesure non nulle, ce qui contredit la première partie de la démonstration. En conclusion, δ^π est admissible.

Théorème 3 *(Th.5 dans le poly de cours) Si un estimateur de Bayes δ^π associé à une mesure a priori π (probabiliste ou non) et une fonction de coût L est unique, alors il est admissible.*

Preuve.

Supposons que δ^π est non admissible. Alors, d'après la Définition 12 du poly de cours, $\exists \delta_0 \in \mathcal{D}$ tel que $\forall \theta \in \Theta$, $R(\theta, \delta^\pi) \geq R(\theta, \delta_0)$, et $\exists \theta_0 \in \Theta$ tel que $R(\theta_0, \delta^\pi) > R(\theta_0, \delta_0)$. En intégrant la première inégalité, il vient :

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

donc δ_0 est aussi un estimateur bayésien associé à L et π, et $\delta_0 \neq \delta^\pi$ d'après la seconde inégalité. Ce qui contredit à l'hypothèse du théorème. Par contraposée, on en déduit le résultat de ce théorème. Remarquons que l'unicité de l'estimateur impliqué la finitude du risque :

$$\int_\Theta R(\theta, \delta^\pi)\, d\Pi(\theta) < \infty$$

sinon tout estimateur minimise le risque.

Proposition 1 *L'estimateur de Bayes associé à toute loi a priori π et au coût*

$$L(\theta,\delta) \quad = \quad \|\theta-\delta\|^2. \tag{2}$$

est l'espérance (moyenne) de la loi a posteriori $\pi(\theta|\mathbf{x_n})$

Preuve. On a $\mathcal{D}=\Theta\in I\!R^d$ (ou plus généralement un espace de Hilbert) et $L(\theta,\delta)=\|\theta-\delta\|^2$ (norme euclidienne au carré). Par simplificité travaillons sur $\Theta=I\!R$ (d=1). Alors

$$\begin{aligned} R_P\left(\delta|\pi\right) &= \int_\Theta (\theta-\delta)^2\pi(\theta|x)\,d\theta, \\ &= \mathbb{E}[\theta^2|x]-2\delta\mathbb{E}[\theta|x]+\delta^2. \end{aligned}$$

En dérivant en δ, on obtient

$$R'_P\left(\delta|\pi\right) \quad = \quad 2\delta-2\mathbb{E}[\theta|x]$$

valant 0 en $\delta=\mathbb{E}[\theta|x]$. Or

$$R''_P\left(\delta|\pi\right) \quad = \quad 2 \;>\; 0$$

donc $\delta\to R_B\left(\delta|\pi\right)$ est convexe, ce qui signifie que la solution de $R'_P\left(\delta|\pi\right)=0$ est bien un minimisateur du risque.

Proposition 2 *L'estimateur de Bayes associé à toute loi a priori π et au coût*

$$L_{c_1,c_2}(\theta,\delta) \quad = \quad \left\{\begin{array}{ll} c_2(\theta-\delta) & \textit{si } \theta>\delta \\ c_1(\delta-\theta) & \textit{sinon} \end{array}\right. \tag{3}$$

est le fractile $c_1/(c_1+c_2)$ de la loi a posteriori $\pi(\theta|\mathbf{x_n})$. En particulier, la médiane de la loi a posteriori est l'estimateur de Bayes lorsque $c_1=c_2$ (qui sont donc des coûts associés à la sous-estimation et la surestimation de θ).

Preuve. Considérons la situation générique où

$$L_{c_1,c_2}(\theta,\delta) \quad = \quad \left\{\begin{array}{ll} c_2(\theta-\delta) & \text{si } \theta>\delta \\ c_1(\delta-\theta) & \text{sinon} \end{array}\right. \tag{4}$$

Le risque *a posteriori* s'écrit

$$R_P(\delta|\pi) \quad = \quad \int_{-\infty}^{\delta} c_1(\delta-\theta)\pi(\theta|x)\,d\theta+\int_{\delta}^{\infty} c_2(\theta-\delta)\pi(\theta|x)\,d\theta.$$

En raisonnant par intégration par parties (IPP), il vient :

$$\begin{aligned} R_P(\delta|\pi) &= \left[c_1(\delta-\theta)\Pi(\theta|x)\right]_{-\infty}^{\delta}+c_1\Pi(\theta<\delta|x)+\left[c_2(\theta-\delta)\Pi(\theta\right]_{\delta}^{\infty}-c_2(1-\Pi(\theta>\delta|x)), \\ &= c_1\Pi(\theta<\delta|x)-c_2\left(1-\Pi(\theta<\delta|x)\right) \end{aligned}$$

car $\lim_{\theta\to-\infty}\theta\Pi(\theta)=\lim_{\theta\to\infty}\theta\Pi(\theta)=0$. Ce risque est minimum lorsque $R_P(\delta|\pi)=0$ soit lorsque

$$\Pi(\theta<\delta^\pi|x) \quad = \quad \frac{c_2}{c_1+c_2}$$

ce qui confère donc à l'estimateur δ^π le sens du quantile de seuil $\frac{c_1}{c_1+c_2}$.

Proposition 3 *L'estimateur de Bayes associé à toute loi a priori π et au coût 0-1 est*

$$\delta^\pi \quad = \quad \left\{\begin{array}{ll} 1 & \textit{si } \Pi(\theta\in\Theta_0|\mathbf{x_n})>\Pi(\theta\notin\Theta_0|\mathbf{x_n}) \\ 0 & \textit{sinon} \end{array}\right.$$

Preuve. Cette fonction de perte est utilisée dans le contexte des tests statistiques. On suppose partitionner Θ en Θ_0 et Θ_1. La fonction de perte correspondante est alors

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

Le risque *a posteriori* est alors

$$R_P(\delta|\pi) \quad = \quad \mathbb{1}_{\delta=1}\Pi(\theta\in\Theta_0|x)+\mathbb{1}_{\delta=0}\Pi(\theta\in\Theta_0|x).$$

Ainsi $\delta^\pi=1$ équivaut à $\Pi(\theta\in\Theta_0|X)\leq\Pi(\theta\in\Theta_1|X)$.

1.3 TP : Création d'un système d'alerte pour la circulation routière

On s'intéresse à un évènement routier $X = x$ relevé par un système de détection vivant dans l'espace χ de dimension finie. Ce système de détection peut prédire des évènements répétés du type "un animal sur la voie", "accrochage", "accident", "bouchon"... La question est de déterminer si, à chaque fois qu'un événement routier x est collecté, il est utile qu'une intervention de secours soit menée.

Nommons θ une variable indiquant la gravité de l'évènement. Cette variable a des valeurs dans les ensembles disjoints Θ_0 (incidents sans gravité) et Θ_1 (accidents nécessitant possiblement une intervention). On suppose disposer d'un échantillon labélisé $\mathbf{e_n} = (\mathbf{x_n}, \theta_\mathbf{n})$.

Questions.

1. Lorsqu'une observation x apparaît, comment prévoir θ ?
2. Comment peut-on en déduire une alarme efficace ?

Réponses. **1.** Dans un cadre bayésien, probabilisons l'espace $\Theta = \Theta_0 \oplus \Theta_1$ et nommons Π la mesure de probabilité associée. On note également P la mesure de probabilité associée à l'espace χ supposé probabilisé. On souhaite prévoir la probabilité que $\theta \in \Theta_0$ sachant x et e_n. *Via* une règle de Bayes, on produit le classifieur classique (dit *classifieur de Bayes*)

$$\begin{aligned}\Pi(\theta \in \Theta_0 | x, e_n) &= \frac{P(X = x|\theta \in \Theta_0, e_n)\Pi(\theta \in \Theta_0|e_n)}{P(X = x|e_n)}, \\ &= \frac{P(X = x|\theta \in \Theta_0, e_n)\Pi(\theta \in \Theta_0|e_n)}{P(X = x|\theta \in \Theta_0)\Pi(\theta \in \Theta_0|e_n) + P(X = x|\theta \notin \Theta_0)\left[1 - \Pi(\theta \in \Theta_0|e_n)\right]}\end{aligned}$$

Le terme de vraisemblance $P(X = x|\theta \in \Theta_0, e_n)$ peut être estimé de nombreuses manières (ex : par la fréquence d'observation de l' évènement $X = x$ dans les situations recensées pour lesquelles $\theta \in \Theta_0$). Le classifieur de Bayes *naïf* repose ainsi sur une simplification de cette vraisemblance, etc. (voir cours d'apprentissage statistique).

2. Le calcul de la probabilité $\Pi(\theta \in \Theta_0|x, e_n)$ ne suffit cependant pas pour prendre une décision opérationnelle. Intuitivement, on souhaiterait pourtant choisir de mener une intervention si

$$\begin{aligned}\Pi(\theta \in \Theta_0|X = x, e_n) \geq \Pi(\theta \notin \Theta_0|X = x, e_n) &= \Pi(\theta \in \Theta_1|X = x, e_n), \\ &= 1 - \Pi(\theta \in \Theta_0|X = x, e_n)\end{aligned}$$

et donc à choisir (ou recommander) d'intervenir si

$$\Pi(\theta \in \Theta_0|X = x, e_n) \geq 1/2. \tag{5}$$

Mais cette règle est en fait simpliste, car elle n'intègre pas les risques d'erreur liés au fait qu'on utilise un échantillon de taille finie n pour mener le calcul de cette probabilité. Une bonne façon de faire est de placer le problème de classification dans un problème de décision plus vaste.

La décision qu'un possible intervenant sur le réseau routier souhaite prendre est binaire : sachant $X = x$, on intervient ou non. À partir des données e_n, il tente donc de définir un estimateur statistique $\hat{\delta}_n(x)$ d'une décision *idéale* $\delta(x)$ vivant dans un espace de décision $\mathcal{D} = \{0, 1\}$ où :

- $\delta(x) = 0 \Leftrightarrow$ pas d'intervention,

- $\delta(x) = 1 \Leftrightarrow$ intervention.

Pourquoi parle-t-on d'estimateur statistique ? Parce que la décision idéale $\delta(x)$ est inaccessible par nature – elle sous-entend que le possible intervenant est omniscient, que θ est parfaitement connu et que nécessairement $n = \infty$.

On construit tout estimateur statistique comme le minimiseur d'une *fonction de coût*

$$\delta(x) \in \mathcal{D} \mapsto L(\theta, \delta(x))$$

que l'on cherche à définir si la vérité sur θ pouvait être connue. Dans le cas qui nous intéresse, on aurait :

- $L(\theta, \delta(x)) = C_1 =$ le coût prévisionnel d'une intervention à raison, donc si $\theta \in \Theta_1$ (ou $\theta \notin \Theta_0$) et $\delta(x) = 1$;
- $L(\theta, \delta(x)) = C_2 =$ le coût prévisionnel d'une non-intervention à tort (*erreur de 1ère espèce*), si $\theta \in \Theta_1$ et $\delta(x) = 0$;
- $L(\theta, \delta(x)) = C_3 =$ le coût prévisionnel d'une intervention à tort (*erreur de 2ème espèce*), si $\theta \in \Theta_0$ et $\delta(x) = 1$;
- $L(\theta, \delta(x)) = C_4 = 0$ le coût (nul) d'une non-intervention à raison, si $\theta \in \Theta_0$ et $\delta(x) = 0$.

On peut alors écrire, de façon plus condensée :

$$L(\theta, \delta(x)) \quad = \quad C_1\delta(x)\mathbb{1}_{\{\theta\in\Theta_1\}} + C_2(1-\delta(x))\mathbb{1}_{\{\theta\in\Theta_1\}} + C_3\delta(x)\mathbb{1}_{\{\theta\in\Theta_0\}}. \tag{6}$$

Or la vérité sur θ ne peut être parfaitement connue. On dispose simplement de la connaissance *a posteriori* $\Pi(\theta \in \Theta | X = x, e_n)$. Si l'on souhaite prendre une décision qui prenne en compte l'incertitude épistémique sur θ, il faut définir un *risque* $R(\delta(x), \Pi, e_n)$ qui puisse intégrer la connaissance de $\Pi(\theta \in \Theta | X = x, e_n)$, la construction de $L(\theta, \delta(x))$ et qui soit minimisable en un choix unique d'estimateur de $\delta(x)$. Pour obtenir un *ordre total* sur l'espace des applications $\delta(x) \mapsto R(\delta(x), \Pi, e_n)$, il faut nécessairement définir ce risque comme le *risque de Bayes*

$$R(\delta(x), \Pi, e_n) \quad = \quad \int_\Theta L(\theta, \delta(x)) d\Pi(\theta \in \Theta | X = x, e_n)$$

et d'en déduire donc la *décision optimale* (et non pas *idéale*)

$$\hat{\delta}_n(x) \quad = \quad \arg\min_{\delta(x)\in\mathcal{D}} R(\delta(x), \Pi, e_n).$$

Après intégration,

$$\begin{aligned} R(\delta(x), \Pi, e_n) \quad = \quad & C_1\delta(x)\Pi(\theta \in \Theta_1 | X = x, e_n) + C_2(1-\delta(x))\Pi(\theta \in \Theta_1 | X = x, e_n) \\ & + C_3\delta(x)\left[1 - \Pi(\theta \in \Theta_1 | X = x, e_n)\right]. \end{aligned}$$

On en déduit la règle de décision (ou de recommandation) suivante : ayant observé l'évènement $X = x$, on décide d'intervenir ($\hat{\delta}_n(x) = 1$) si le risque associé à cette décision est moins élevé que le risque associé à la décision contraire ($\hat{\delta}_n(x) = 0$), soit si

$$\begin{aligned} R(0, \Pi, e_n) \; = \; C_2\Pi(\theta \in \Theta_1 | X = x, e_n) \quad \geq \quad R(1, \Pi, e_n) \; = \; & C_1\Pi(\theta \in \Theta_1 | X = x, e_n) \\ & + C_3\left[1 - \Pi(\theta \in \Theta_1 | X = x, e_n)\right], \end{aligned}$$

c'est-à-dire quand

$$\Pi(\theta \in \Theta_1 | X = x, e_n) \quad \geq \quad \frac{C_3}{C_2 - C_1 + C_3} \tag{7}$$

Finissons par quelques remarques importantes :

- il n'est pas utile de disposer des coûts absolus C_i pour prendre une décision, car des rapports de coûts suffisent, ce qui est en général plus simple à estimer dans une vraie démarche opérationnelle ;
- on peut légitimement supposer que $C_1 \leq C_3 < C_2$, car le coût C_1 d'une intervention utile peut être réduit par l'effet d'assurances, tandis que le coût C_3 d'une intervention inutile ne l'est pas. Enfin, le coût (prévisionnel) C_2 d'une non-intervention qui aurait pu être utile peut éventuellement intégrer celui de vies humaines ; remarquons qu'en toute rigueur, $C_2 = C_2(t)$ où t représente le temps depuis l'occurence de l'événement, et que cette fonction est très certainement croissante.
- Il faut que $C_3 = C_2$ et $C_1 = 0$ pour obtenir l'équivalent décisionnel de la règle (5). On perçoit bien qu'une décision purement intuitive est à rejeter, car elle pré-suppose une contradiction forte avec la deuxième remarque.
- Cette règle de décision prend intégralement en compte les incertitudes sur la véritable nature de l'événement θ, conditionnellement à la validité des hypothèses.
- Le choix de la fonction de coût (6) est arbitraire ; mais retenons qu'en l'absence d'arguments permettant de rationaliser ce choix, les coûts sont généralement assemblés de façon additive

2 Corrigés et preuves "Propriétés"

2.1 Exercices

Exercice 10 *Soit $x_1, \ldots, x_n$ des réalisations iid de loi $\mathcal{N}(\mu, \sigma^2)$. On choisit la mesure a priori (non probabiliste) jointe*

$$\pi(\mu, \sigma^2) \quad \propto \quad 1/\sigma^2.$$

1. *Déterminez la loi a posteriori jointe $\pi(\mu, \sigma^2|x_1, \ldots, x_n)$*
2. *Déterminez la loi a posteriori marginale $\pi(\mu|x_1, \ldots, x_n)$*
3. *Calculez la région HPD de seuil α pour μ et comparez-la à la région de confiance fréquentiste, de même seuil, qu'on pourrait calculer par l'emploi du maximum de vraisemblance.*
4. *Déterminez la loi a posteriori marginale $\pi(\sigma^2|x_1, \ldots, x_n)$; le calcul de la région HPD est-il simple ?*

Remarque. *"Déterminez" signifie indiquer si la loi appartient à une famille connue, par exemple largement implémentée sur machines. La connaissance des lois gamma, inverse gamma et Student est peut-être nécessaire pour répondre aux questions.*

Réponse.

1. Loi jointe *a posteriori***.** On a

$$\begin{aligned}
\pi(\mu, \sigma^2|x_1, \ldots, x_n) &\propto \sigma^{-(n+2)} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right), \\
&= \sigma^{2(-(n/2+1)} \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{2\sigma^2} + \frac{2\mu \sum_{i=1}^{n} x_i}{2\sigma^2} - n\mu^2/2\sigma^2\right), \\
&= \sigma^{2(-(n/2+1)} \exp\left(-\frac{n\,(\mu - \bar{x}_n)^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2} S_n^2\right)
\end{aligned}$$

avec

$$S_n^2 \quad = \quad -\sum_{i=1}^{n} \left(\bar{x}_n^2 - x_i^2\right) \; = \; \sum_{i=1}^{n} (\bar{x}_n - x_i)^2 \; \geq \; 0.$$

2. Loi marginale *a posteriori* **de μ.** On obtient alors, par intégration,

$$\begin{aligned}
\pi(\mu|x_1, \ldots, x_n) &= \int_{\mathbb{R}^+} \pi(\mu, \sigma^2|x_1, \ldots, x_n)\, d\sigma^2, \\
&\propto \int_{\mathbb{R}^+} \sigma^{2(-(n/2+1)} \exp\left(-\frac{n\,(\mu - \bar{x}_n)^2}{2\sigma^2} - \frac{1}{2\sigma^2} S_n^2\right)\, d\sigma^2.
\end{aligned}$$

En reconnaissant dans l'intégrande le terme général d'un loi inverse gamma $\mathcal{IG}(n/2, (S_n^2 + n(\mu - \bar{x}_n)^2)/2)$, on a alors

$$\begin{aligned}
\pi(\mu|x_1, \ldots, x_n) &\propto \Gamma(1) \left(S_n^2 + n(\mu - \bar{x}_n)^2\right)^{-n/2}, \\
&\propto \left(1 + k_n u^2\right)^{-(k-1)/2}
\end{aligned}$$

avec $k = n + 1$ et $u = \sqrt{k k_n}(\mu - \bar{x}_n)$ et $k_n = n/S_n^2$. On peut alors réécrire plus simplement

$$\pi(\mu|x_1, \ldots, x_n) \quad \propto \quad \left(1 + u^2/k\right)^{-(k-1)/2}$$

qui définit le terme général d'une loi de Student en u à k degrés de liberté. Posons alors formellement la variable

$$u = \frac{\sqrt{n(n+1)}}{S_n}(\mu - \bar{x}_n) \;=\; g(\mu)$$

et procédons à changement de variable (cf. Proposition 4) pour connaître la loi de u. On a

$$\begin{aligned} g^{-1}(u) &= u\frac{S_n}{\sqrt{n(n+1)}} + \bar{x}_n, \\ g'(u) &= \frac{\sqrt{n(n+1)}}{S_n}. \end{aligned}$$

Donc

$$\begin{aligned} \pi_U(u) &\propto \pi_\mu(g^{-1}(u)), \\ &\propto (1 + u^2/k)^{-\frac{k+1}{/}2}. \end{aligned}$$

Donc la loi *posteriori* de la variable u est bien une Student à $k = n + 1$ degrés de liberté :

$$u|x_1, \ldots, x_n \quad = \quad \frac{\sqrt{n(n+1)}}{S_n}(\mu - \bar{x}_n) \;\sim\; \mathcal{S}_t(n+1).$$

On dit alors que la loi de μ est une *Student décentrée* à k degrés de liberté.

3. Région HPD pour μ. On veut alors déterminer

$$\mathcal{A}_{\alpha,\pi} \quad = \quad \{\mu,\; \pi(\mu|x_1, \ldots, x_n) \geq 1 - \alpha\}\,.$$

Donc en notant $t_{n+1}(\alpha)$ le quantile de seuil α de la loi $\mathcal{S}_t(n+1)$, on a (par symmétrie de cette loi)

$$\Pi\left(-t_{n+1}(1-\alpha/2) \leq u \leq (t_{n+1}(1-\alpha/2)|x_1, \ldots, x_n\right) \quad = \quad 1 - \alpha$$

soit, de fa con équivalente,

$$\Pi\left(\bar{x}_n - \frac{S_n}{\sqrt{n(n+1}}t_{n+1}(1-\alpha/2) \leq \mu \leq \bar{x}_n + \frac{S_n}{\sqrt{n(n+1}}t_{n+1}(1-\alpha/2)|x_1, \ldots, x_n\right) \quad = \quad 1 - \alpha.$$

Rappelons que la région fréquentiste de seuil $1 - \alpha$ connue pour l'estimation du maximum de vraisemblance (EMV), dès qu'on a pris conscience que l'EMV de σ^2 est

$$\hat{\sigma}_n^2 \quad = \quad S_n^2/n,$$

est :

$$\left[\bar{x}_n - \frac{S_n}{n}t_{n+1}(1-\alpha/2)\;;\; \bar{x}_n + \frac{S_n}{n}t_{n+1}(1-\alpha/2)\right].$$

Avec $n < \sqrt{n(n+)}$ on voit que la région bayésienne est légèrement plus étroite que la région fréquentiste, mais qu'il y a équivalence asymptotique (ce qui est logique).

4. Loi marginale *a posteriori* de σ^2. De la même façon que précédemment, on a

$$\begin{aligned} \pi(\sigma^2|x_1, \ldots, x_n) &\propto \frac{1}{\sigma^{n+2}} \exp\left(-S_n^2/2\sigma^2\right) \int_{I\!R} \exp\left(-\frac{(\mu - \bar{x}_n)^2}{2\sigma^2}\right) d\mu, \\ &\propto \frac{1}{\sigma^{n+2}} \exp\left(-S_n^2/2\sigma^2\right) \frac{\sqrt{2\pi}}{\sqrt{n}}\sigma, \\ &\propto \frac{1}{(\sigma^2)^{n/2+1/2}} \exp\left(-S_n^2/2\sigma^2\right) \end{aligned}$$

et on reconnaît ici le terme général d'une loi inverse gamma :

$$\sigma^2|x_1,\ldots,x_n \quad \sim \quad \mathcal{IG}(n/2-1/2, S_n^2/2)$$

qui n'est définie que si $n > 1$ (ce qui semble logique : il faut au moins avoir deux données pour inférer sur la variance σ^2). De plus, pour avoir $S_n^2 \neq 0$ si $n = 2$, il faut que $x_1 \neq x_2$. Comme cette loi est explicite, déterminer ses régions HPD peut être fait formellement.

Proposition 4 *Changement de variable (rappel). Soit $X \sim f_X$ sur $\mathbb{R}$ et $Y = g(X) \sim f_Y$ avec g bijectif. Alors*

$$f_Y(y) \quad = \quad \left|g'(g^{-1}(y))\right|^{-1} f_X(g^{-1}(y)).$$

2.2 Démonstrations

Théorème 4 *Si Θ est fini et discret et $\Pi(\theta = \theta_0) > 0$, alors pour tout échantillon iid $X_1,\ldots,X_n|\theta \sim f(X|\theta)$,*

$$\Pi\left(\theta = \theta_0|X_1,\ldots,X_n\right) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

Preuve. On va montrer que $\Pi(\theta|X_1,\ldots,X_n) \to 0 \; \forall \theta \neq \theta_0$. On a

$$\log \frac{\Pi(\theta|X_1,\ldots,X_n)}{\Pi(\theta_0|X_1,\ldots,X_n)} \quad = \quad \log \frac{\Pi(\theta)}{\Pi(\theta_0)} + \sum_{i=1} Y_i \tag{8}$$

où les

$$Y_i = \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)}$$

sont des v.a. iid, telles ques

$$\mathbb{E}\left[Y_i\right] \quad = \quad KL\left(\theta_0\right) - KL\left(\theta\right)$$

(rappelons que l'intégration se fait par rapport à la vraie loi inconnue des données) qui vaut 0 si $\theta = \theta_0$, et qui est négatif sinon car θ_0 est l'unique minimiseur de $KL(\theta)$. Ainsi, si $\theta \neq \theta_0$, le second terme de (8) est une somme de termes iid avec une espérance négative. Par la loi des grands nombres, on obtient que $\lim_{n\to\infty} \sum_{i=1} Y_i = -\infty$ si $\theta \neq \theta_0$. Tant que le premier terme de (8) est fini (soit tant que $\Pi(\theta = \theta_0) > 0$), l'expression totale pour (8) tend également vers $-\infty$. Nécessairement,

$$\frac{\Pi(\theta|X_1,\ldots,X_n)}{\Pi(\theta_0|X_1,\ldots,X_n)} \quad \xrightarrow[\mathbb{P}]{n\to\infty} 0$$

et donc $\Pi(\theta|X_1,\ldots,X_n) \to 0 \; \forall \theta \neq \theta_0$. Comme toutes les probabilités somment à 1, nécessairement

$$\Pi\left(\theta = \theta_0|X_1,\ldots,X_n\right) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

Théorème 5 *Si Θ est un ensemble compact et si V_{θ_0} est tel que $\Pi(\theta \in V_{\theta_0}) > 0$ avec*

$$\theta_0 \quad = \quad \arg\min_{\theta\in\Theta} KL(\theta)$$

alors

$$\Pi(\theta \in V_{\theta_0}|x_1,\ldots,x_n) \quad \xrightarrow[\mathbb{P}]{n\to\infty} \quad 1.$$

Preuve. On admet ce résultat.

Théorème 6 *Consistance Si $f(.|\theta)$ est suffisamment régulière et identifiable, soit si $\theta_1 \neq \theta_2 \Rightarrow f(x|\theta_1) \neq f(x|\theta_2)$ $\forall x \in \Omega$, alors pour tout échantillon $\mathbf{x_n}$ iid*

$$\pi(\theta|\mathbf{x_n}) \quad \xrightarrow{p.s.} \quad \delta_{\theta_0}.$$

Par ailleurs, si $g : \Theta \to I\!R$ est mesurable et telle que $\mathbb{E}[g(\theta)] < \infty$, alors sous les mêmes hypothèses

$$\lim_{n\to\infty} \mathbb{E}\left[g(\theta)|X_1, \ldots, X_n\right] \quad = \quad g(\theta) \quad p.s.$$

Preuve. On admet ce résultat.

Théorème 7 *Normalité asymptotique (**Bernstein-von Mises**) Soit I_θ la matrice d'information de Fisher du modèle $f(.|\theta)$ et soit $g(\theta)$ la densité de la gaussienne $\mathcal{N}(0, I_{\theta_0}^{-1})$. Soit $\hat{\theta}_n$ le maximum de vraisemblance. Alors, dans les conditions précédentes,*

$$\int_\Theta \left|\pi\left(\sqrt{n}\left\{\theta - \hat{\theta}_n\right\}|\mathbf{x_n}\right) - g(\theta)\right| \, d\theta \quad \to \quad 0.$$

Preuve. Le théorème 6 montre qu'on peut concentrer l'étude sur un voisinage de θ_0. Obtenir la loi limite requiert deux étapes :

- montrer que le mode *a posteriori* est consistant, c'est-à-dire qu'il se situe dans le voisinage de θ_0 où se situe presque toute la masse ;
- montrer l'approximation gaussienne centrée en le mode *a posteriori*.

Pour simplifier, le schéma de preuve donné ici considère que θ est un scalaire. Notons $\tilde{\theta}_n$ le mode *a posteriori*

$$\tilde{\theta}_n = \arg\max_{\theta\in\Theta} \left\{\log f(x_1, \ldots, x_n|\theta) + \log \pi(\theta)\right\}.$$

La preuve de consistance du MLE peut être adaptée pour montrer que $\tilde{\theta}_n \to \theta_0$ quand $n \to \infty$, presque sûrement. On peut alors approximer la log-densité *a posteriori* par un développement de Taylor centré autour de $\tilde{\theta}_n$ (approximation quadratique de $\log \pi(\theta|x_1, \ldots, x_n)$) :

$$\log \pi(\theta|x_1, \ldots, x_n) \quad = \quad \log \pi(\tilde{\theta}_n|x_1, \ldots, x_n) + \frac{1}{3}(\theta - \tilde{\theta}_n)^2 \frac{\partial^2}{\partial\theta^2}\left[\log \pi(\theta|x_1, \ldots, x_n)\right]_{\theta=\tilde{\theta}_n} \tag{9}$$

$$+ \frac{1}{6}(\theta - \tilde{\theta}_n)^3 \frac{\partial^3}{\partial\theta^3}\left[\log \pi(\theta|x_1, \ldots, x_n)\right]_{\theta=\tilde{\theta}_n} + \ldots \tag{10}$$

Le terme linéaire est nul car par définition du mode :

$$\frac{\partial}{\partial\theta}\left[\log \pi(\theta|x_1, \ldots, x_n)\right]_{\theta=\tilde{\theta}_n} \quad = \quad 0.$$

Le premier terme de (10) est constant. Le coefficient du second terme (sous l'hypothèse iid) est

$$\frac{\partial^2}{\partial\theta^2}\left[\log \pi(\theta|x_1, \ldots, x_n)\right]_{\theta=\tilde{\theta}_n} \quad = \quad \frac{\partial^2}{\partial\theta^2}\left[\log \pi(\theta)\right]_{\theta=\tilde{\theta}_n} + \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\left[\log f(x_i|\theta)\right]_{\theta=\tilde{\theta}_n},$$

$$= \quad cte + \sum_{i=1}^n Y_i$$

où les Y_i sont des variables aléatoires iid d'espérance négative sous l'hypothèse $X \sim \tilde{f}(x)$. En effet, si $\tilde{\theta}_n = \theta_0$, on a

$$\mathbb{E}\left[Y_i\right] \quad = \quad \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta_0)\right] \quad = \quad -I_{\theta_0}$$

et sinon

$$\mathbb{E}\left[Y_i\right] \quad = \quad -\,-\,\frac{\partial^2}{\partial\theta^2} KL\left(\theta\right)_{\theta=\tilde{\theta}_n} \;<\; 0$$

par convexité. Ainsi, le coefficient du second terme converge vers $-\infty$ à la vitesse n. Quand $n \to \infty$, $|\tilde{\theta}_n - \theta_0| \to 0$, et les termes suivants du développement de Taylor tendent vers 0. On a donc

$$\log \pi(\theta|x_1, \ldots, x_n) \quad \sim \quad -\alpha(\theta - \tilde{\theta}_n)^2$$

et la loi limite de $\pi(\theta|x_1, \ldots, x_n)$ est donc une gaussienne.

3 Corrigés et preuves "Modélisation *a priori*"

3.1 Exercices

Exercice 11 Prior de Jeffreys pour une loi binomiale. *Soit x un nombre de boules tirées dans une urne en contenant n avec probabilité θ. Alors $x \sim \mathcal{B}(n, \theta)$. Calculer la mesure a priori de Jeffreys sur θ. Mener un calcul bayésien sur un échantillon simulé. Peut-on dire que cette mesure est défavorable a priori ?*

Réponse. Soit x un nombre de boules tirées dans une urne en contenant n avec probabilité θ. Alors $x \sim \mathcal{B}(n, \theta)$ et

$$\begin{aligned} f(x|\theta) &= \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}, \\ \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} &= \frac{x}{\theta^2} + \frac{n-\theta}{(1-\theta)^2} \\ \text{et } I(\theta) &= n \left[\frac{1}{\theta} + \frac{1}{1-\theta} \right] = \frac{n}{\theta(1-\theta)} \end{aligned}$$

Donc la loi de Jeffreys est

$$\pi(\theta) \propto [\theta(1-\theta)]^{-1/2}$$

et elle est propre, il s'agit de la distribution $\mathcal{B}_e(1/2, 1/2)$ de densité générale

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

avec $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t)\, dt$. La mesure de Jeffreys est en effet défavorable dans le sens où elle privilégie fortement des valeurs de θ très proches de 0 ou de 1, peu probables.

Exercice 12 Problème de Neyman-Scott. *On considère le problème suivant*

$$x_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \textit{pour } i = 1, \ldots, n$$

où les x_i sont indépendants. Soit $\theta = (\mu_1, \ldots, \mu_n, \sigma)$. Calculez le prior de Jeffreys $\pi^J(\theta)$ puis l'espérance a posteriori de σ^2. Est-ce un estimateur consistant ?

Réponse. La loi jointe des données, qui est aussi la vraisemblance, s'écrit

$$f(x_1, \ldots, x_n|\theta) = \frac{\sigma^{-n}}{(2\pi)^{n/2}} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu_i)^2}{2\sigma^2} \right). \tag{11}$$

La matrice d'information de Fisher s'écrit, dans ce cas régulier, comme

$$I = -\mathbb{E}\left[\begin{array}{cccc} \frac{\partial^2}{\partial \theta_{i_1}^2} \log f(x|\theta) & \frac{\partial^2}{\partial \theta_{i_1} \theta_{i_2}} \log f(x|\theta) & \ldots & \frac{\partial^2}{\partial \theta_{i_1} \theta_{i_d}} \log f(x|\theta) \\ \frac{\partial^2}{\partial \theta_{i_1} \theta_{i_2}} \log f(x|\theta) & \frac{\partial^2}{\partial \theta_{i_2}^2} \log f(x|\theta) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{array} \right]$$

où $x = (x_1, \ldots, x_n)$ et $d = n$. Or

$$\begin{aligned} \frac{\partial^2}{\partial \sigma^2} \log f(x|\theta) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu_i)^2, \\ \frac{\partial^2}{\partial \mu_i \partial \mu_j} \log f(x|\theta) &= 0 \quad \text{si } i \neq j, \\ \frac{\partial^2}{\partial \mu_i^2} \log f(x|\theta) &= -\frac{1}{2\sigma^2} \end{aligned}$$

et

$$\frac{\partial^2}{\partial\sigma\partial\mu_i}\log f(x|\theta) \quad = \quad -\frac{1}{\sigma^4}(x_i-\mu_i).$$

Avec $\mathbb{E}[X_i-\mu_i]=0$ et $\mathbb{E}[(X_i-\mu_i)^2]=\sigma^2$, il vient donc

$$I=-\mathbb{E}\begin{bmatrix} \frac{1}{2\sigma^2} & & & \\ & \frac{1}{2\sigma^2} & & \\ & & \dots & (\mathbf{0}) \\ & (\mathbf{0}) & & \dots \\ & & & \frac{n}{2\sigma^2} \end{bmatrix}$$

et donc

$$\pi^J(\theta) \quad \propto \quad \sigma^{-n-1}.$$

Réponse. **(suite)** En utilisant (11), la loi *a posteriori* s'écrit sous une forme condensée comme

$$\pi^J(\theta|x_1,\dots,x_n) \quad \propto \quad \sigma^{-2n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_i)^2\right).$$

En opérant le changement de variable $\sigma\to\sigma^2$, on obtient alors

$$\begin{aligned}\pi^J(\sigma^2|x_1,\dots,x_n,\mu_1,\dots,\mu_n) &\propto \sigma^{-1}\pi^J(\theta|x_1,\dots,x_n),\\ &\propto \sigma^{-2(n+1)}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_i)^2\right)\end{aligned}$$

et on reconnaît le terme général d'une loi inverse gamma $\mathcal{IG}\left(n,\frac{1}{2}\sum_{i=1}^n(x_i-\mu_i)^2\right)$ pour la variable aléatoire σ^2. On déduit du résultat précédent que

$$\mathbb{E}[\sigma^2|x_1,\dots,x_n] \quad = \quad \frac{1}{2(n-1)}\sum_{i=1}^n(x_i-\mu_i)^2.$$

Or, avec $X_i-\mu_i|\sigma^2\sim\mathcal{N}(0,\sigma^2)$, par indépendance des X_i il vient que $\sigma^{-2}\sum_{i=1}^n(x_i-\mu_i)^2|\sigma^2\sim\chi^2_n$ et donc que

$$\mathbb{E}_X\left[\frac{1}{2(n-1)}\sum_{i=1}^n(x_i-\mu_i)^2|\sigma^2\right] \quad = \quad \frac{n\sigma^2}{2(n-1)} \quad \xrightarrow{n\to\infty} \quad \sigma^2/2.$$

Ainsi, dans un cadre asymptotique, l'estimateur bayésien $\mathbb{E}[\sigma^2|x_1,\dots,x_n]$ est assimilable à un estimateur fréquentiste non consistant.

Exercice 13 Exemple industriel. *On cherche la distribution de la profondeur X d'un défaut de fabrication dans une enceinte d'acier difficile d'accès. À partir d'anciennes données mesurées sur des aciers différents, on suppose connaître un modèle exponentiel pour $X|\theta\sim\mathcal{E}(\theta)$ de densité $f(x|\theta)=\theta^{-1}\exp(-x/\theta)\mathbb{1}_{X\geq 0}$. Les mesures ultrasonores pour estimer la distribution de X sont cependant coûteuses et ont besoin d'être calibrées a priori. D'où les questions posées successivement à un expert en métallurgie :*

1. *Pouvez-vous préciser la profondeur moyenne θ_e d'un défaut de fabrication dans cette coulée ?*
2. *Pouvez-vous préciser un écart-type σ_e pour cette profondeur en général ?*

(Remarque : la question à l'expert est ici quelque peu idéalisée. Il faut en pratique plutôt passer par des questions sur X et les connecter à $\pi(\theta)$ via la loi prédictive a priori.

Réponse. Il est pertinent au vu de l'expertise disponible de : (a) modéliser l'incertitude sur X par la loi *a priori* prédictive dans un premier calcul de tenue de cuve ; cela implique de calibrer le prior ; (b) se servir de l'amplitude de cette loi pour préciser un domaine ultrasonone où chercher des défauts (constitution d'un échantillon de *retour d'expérience*) ; (c) mélanger les deux informations pour améliorer la connaissance de X *a posteriori*.

1. La solution au problème du maximum d'entropie sous la contrainte linéaire (et en utilisant $\pi_0(\theta) \propto 1$

$$\int_\theta \theta\pi(\theta)\, d\theta \quad = \quad \theta_e$$

mène directement à la solution

$$\pi(\theta) \quad \propto \quad \exp(-\lambda_1\theta)$$

et l'on reconnaît le terme général d'une loi exponentielle d'espérance $1/\lambda_1 = \theta_e$.

2. On suppose ici que σ_e est un estimateur de la variance prédictive *a priori*

$$\begin{aligned}\sqrt{\mathbb{V}[X]} &= \sqrt{\mathbb{E}[\mathbb{V}[X|\theta]] + \mathbb{V}[\mathbb{E}[X|\theta]]},\\ &= \sqrt{\mathbb{E}[\theta^2] + \mathbb{V}[\theta]},\\ &= \sqrt{2\mathbb{E}[\theta^2] - \mathbb{E}^2[\theta]} \;=\; \sqrt{2\mathbb{E}[\theta^2] - \theta_e^2}.\end{aligned}$$

On peut alors maximiser l'entropie sous les contraintes linéaires

$$\int_\theta \theta\pi(\theta)\, d\theta \;=\; \theta_e \quad \text{et} \quad \int_\theta \theta^2\pi(\theta)\, d\theta \;=\; \frac{1}{2}(\sigma_e^2 + \theta_e^2).$$

La loi *a priori* obtenue est gaussienne :

$$\theta \quad \sim \quad \mathcal{N}\left(\theta_e, \frac{1}{2}(\sigma_e^2 - \theta_e^2)\right)$$

et une condition de cohérence de l'expertise vis-à-vis de la modélisation paramétrique $X|\theta \sim f(x|\theta)$ est d'avoir $\sigma_e > \theta_e$.

Exercice 14 *On considère les contraintes suivantes sur une mesure a priori sur $\theta \geq 0$: pour $\beta > 0$,*

$$\mathbb{E}[\theta^\beta] \quad = \quad 1 \tag{12}$$

$$\mathbb{E}[\log\theta] \quad = \quad -\frac{\gamma}{\beta}. \tag{13}$$

où γ est la constante d'Euler ($\simeq 0,577$). Calculer la mesure de maximum d'entropie $\pi(\theta)$ relativement à une mesure $\pi^J(\theta) \propto \theta^{-1}$. Quelles sont les conditions pour que cette mesure soit une vraie mesure de probabilité ?

Indication : si $X \sim \mathcal{G}(a,b)$, alors $\mathbb{E}[\log X] = \psi(a) - \log b$ où ψ est la fonction digamma, dérivée logarithmique de la fonction gamma :

$$\psi(x) \quad = \quad \frac{\Gamma'(x)}{\Gamma(x)},$$

et $\psi(1) = -\gamma$.

Réponse. La mesure de maximum d'entropie s'écrit

$$\pi(\theta) \quad \propto \quad \theta^{-1} \exp\left(-\lambda_1 \theta^\beta + \lambda_2 \log \theta\right) \mathbb{1}_{\{\theta \geq 0\}}$$

où $(\lambda_1, \lambda_2) \in I\!R^2$ sont deux multiplicateurs de Lagrange. Soit :

$$\pi(\theta) \quad \propto \quad \theta^{\lambda_2 - 1} \exp\left(-\lambda_1 \theta^\beta\right) \mathbb{1}_{\{\theta \geq 0\}}.$$

Si l'on souhaite définir une mesure $\sigma-$finie sur $I\!R^+$, il nous faut avoir

$$\lim_{\theta \to \infty} \pi(\theta) \quad = \quad 0$$

ce qui implique $\lambda_1 > 0$. Dans ce cas,

$$\pi(\theta) \quad = \quad \Delta^{-1} \theta^{\lambda_2 - 1} \exp\left(-\lambda_1 \theta^\beta\right) \mathbb{1}_{\{\theta \geq 0\}}.$$

où la constante d'intégration Δ est définie par

$$\Delta \quad = \quad \int_{I\!R^+} \theta^{\lambda_2 - 1} \exp\left(-\lambda_1 \theta^\beta\right) \, d\theta.$$

En opérant le changement de variable $u = \lambda_1 \theta^\beta$, il vient

$$du \quad = \quad \beta u (u/\lambda_1)^{-1/\beta} d\theta$$

et

$$\Delta \quad = \quad \frac{1}{\lambda_1^{\frac{\lambda_2}{\beta}} \beta} \int_{I\!R^+} u^{\lambda_2/\beta - 1} \exp\left(-u\right) \, du.$$

On reconnaît dans l'intégrale le terme général d'une loi gamma $\mathcal{G}(\lambda_2/\beta, 1)$, qui est bien définie (intégrable) si et seulement si $\lambda_2 > 0$. On en déduit alors que

$$\Delta \quad = \quad \frac{\Gamma\left(\frac{\lambda_2}{\beta}\right)}{\lambda_1^{\frac{\lambda_2}{\beta}} \beta}.$$

Réponse. (suite). On peut alors réécrire la contrainte (12) :

$$\mathbb{E}[\theta^\beta] \quad = \quad \Delta^{-1} \int_{I\!R^+} \theta^{\beta + \lambda_2 - 1} \exp\left(-\lambda_1 \theta^\beta\right) \, d\theta.$$

En réutilisant le changement de variable $\theta \to u$, on peut réécrire

$$\begin{aligned} \mathbb{E}[\theta^\beta] \quad &= \quad \frac{\Delta^{-1}}{\beta \lambda_1} \int_{I\!R^+} u^{\lambda_2} \exp\left(-u\right) \, du, \\ &= \quad \frac{\Gamma(1 + \lambda_2/\beta)}{\Gamma(\lambda_2/\beta)} \lambda_1^{\frac{\lambda_2}{\beta}} \quad = \quad 1. \end{aligned}$$

Enfin, on s'intéresse à la contrainte (13) :

$$\mathbb{E}[\log \theta] \quad = \quad \Delta^{-1} \int_{I\!R^+} \theta^{\lambda_2 - 1} \log \theta \exp\left(-\lambda_1 \theta^\beta\right) \, d\theta.$$

En réutilisant une nouvelle fois le changement de variable $\theta \to u$, on peut réécrire

$$\begin{aligned} \mathbb{E}[\log \theta] \quad &= \quad \frac{\Delta^{-1}}{\beta^2 \lambda_1^{\frac{\lambda_2}{\beta}}} \int_{I\!R^+} u^{\lambda_2/\beta - 1} \exp\left(-u\right) \left[\log u - \log \lambda_1\right] \, du, \\ &= \quad \frac{\Delta^{-1} \Gamma\left(\frac{\lambda_2}{\beta}\right)}{\beta^2 \lambda_1^{\frac{\lambda_2}{\beta}}} \mathbb{E}_g \left[\log u - \log \lambda_1\right] \end{aligned}$$

où l'espérance $\mathbb{E}_g$ est définie par rapport à la loi $\mathcal{G}(\lambda_2/\beta, 1)$. Donc

$$\begin{aligned}\mathbb{E}[\log\theta] &= \frac{\Delta^{-1}\Gamma\left(\frac{\lambda_2}{\beta}\right)}{\beta^2\lambda_1^{\frac{\lambda_2}{\beta}}}\left[\psi\left(\frac{\lambda_2}{\beta}\right)-\log\lambda_1\right],\\ &= \frac{1}{\beta}\left[\psi\left(\frac{\lambda_2}{\beta}\right)-\log\lambda_1\right] = -\frac{\gamma}{\beta}.\end{aligned}$$

Une solution triviale est $\lambda_1 = 1$ et $\lambda_2 = \beta$. Dans ce cas, on reconnaît pour $\pi(\theta)$ une loi de Weibull de paramètre de forme β et d'échelle 1.

Exercice 15 *Loi inverse Wishart. Soient des observations $x_1, \ldots, x_n \sim \mathcal{N}_p(\mu, \Sigma)$, de loi jointe*

$$f(x_1,\ldots,x_n|\theta=(\mu,\Sigma)) \propto (\det\Sigma)^{-n/2}\exp\left(-\frac{1}{2}\left[n(\bar{x}_n-\mu)^T\Sigma^{-1}(\bar{x}_n-\mu)+tr(\Sigma^{-1}S_n)\right]\right)$$

avec $S_n = \sum_{i=1}^n (x_i-\bar{x}_n)(x_i-\bar{x}_n)^T$. On suppose prendre a priori

$$\begin{aligned}\mu|\Sigma &\sim \mathcal{N}_p\left(\mu_0, \frac{1}{n_0}\Sigma\right)\\ \Sigma &\sim \mathcal{IW}(\alpha, V)\end{aligned}$$

la loi de Wishart inverse $\mathcal{IW}$ étant définie (sur l'espace des matrices symétriques non indépendantes de rang d) par la densité

$$f(x) = \frac{|V|^{\alpha/2}}{2^{\alpha d/2}\Gamma_d(\alpha/2)}|x|^{-\frac{\alpha+d+1}{2}}\exp\left\{-\frac{1}{2}tr\left(Vx^{-1}\right)\right\}$$

où Γ_d est la fonction gamma multivariée. Le prior est-il conjugué ? En dimension 1, à quelle loi se réduit-il ?

Réponse. Les lois *a posteriori* peuvent s'écrire directement sous la forme conditionnelle

$$\begin{aligned}\mu|\Sigma,x_1,\ldots,x_n &\sim \mathcal{N}_p\left(\frac{n_0\mu_0+n\bar{x}_n}{n_0+n}, \frac{1}{n_0+n}\Sigma\right)\\ \Sigma|x_1,\ldots,x_n &\sim \mathcal{IW}\left(\alpha+n, V^{-1}+S_n+\frac{nn_0}{n+n_0}(\bar{x}_n-x_0)(\bar{x}_n-x_0)^T\right).\end{aligned}$$

Elle est donc bien conjuguée et se réduit à un mélange gaussien - inverse gamma en dimension 1.

Exercice 16 *Soit $X \sim \mathcal{N}(\theta, \theta)$ avec $\theta > 0$.*

1. *Déterminer la loi a priori de Jeffreys $\pi^J(\theta)$*
2. *Établir si la loi de X appartient à la famille exponentielle et construire les lois a priori conjuguées sur θ.*
3. *Utiliser la propriété de linéarité des espérances des familles exponentielles pour relier les hyperparamètres des lois conjuguées à l'espérance de θ.*

Réponse.

1. Avec $f(x|\theta) \propto \theta^{-1/2}\exp\left(-\frac{(x-\theta)^2}{2\theta}\right)$, on a donc

$$\begin{aligned}\log f(x|\theta) &= -\frac{1}{2}\log\theta-(x-\theta)^2/2\theta,\\ &= -\frac{1}{2}\log\theta-\frac{x^2}{2\theta}+x-\theta/2\end{aligned}$$

puis

$$\begin{aligned}\frac{\partial \log f(x|\theta)}{\partial \theta} &= -\frac{1}{2\theta}+\frac{x^2}{2\theta^2}-1/2,\\ \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} &= -\frac{1}{2\theta^2}-\frac{x^2}{\theta^3}.\end{aligned}$$

Avec $\mathbb{E}[X]=\theta$ et $\mathbb{E}[X^2]=\mathbb{V}[X]+\mathbb{E}^2[X]=\theta+\theta^2$, l'information de Fisher vaut, puisque la densité est absolument continue par rapport à $\theta\in I\!R_*^+$,

$$-\mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}(\theta)\right] \propto \frac{1}{\theta}+\frac{3}{2\theta^2},$$

et donc

$$\pi^J(\theta) \propto \sqrt{\frac{1}{\theta}+\frac{3}{2\theta^2}}.$$

2. On peut écrire

$$f(x|\theta) \propto \theta^{-1/2}\exp\left(-\frac{(x-\theta)^2}{2\theta}\right) \propto C(\theta)h(x)\exp\left(R(\theta)\cdot T(X)\right)$$

avec

$$R(\theta) = -1/\theta, \quad T(x) = x^2/2,$$

$$h(x) = \exp(x), \quad C(\theta) = \exp(-\log(\theta)/2-\theta/2)$$

ce qui correspond à l'écriture canonique des lois exponentielles :

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

Nous souhaiterions arriver à l'écriture suivante : pour une reparamétrisation z de x et une reparamétrisation η de θ,

$$Z|\eta \sim f(z|\theta) = h(z)\exp\left(\eta\cdot z-\psi(\eta)\right) \tag{14}$$

afin de proposer la famille de priors conjugués :

$$\pi(\eta) \propto \exp\left(\eta\cdot z_0-\psi(\eta)\right).$$

Essayons avec la reparamétrisation $\eta=1/\theta$ et $z=x^2/2$ (on se restreint donc à $I\!R^+$) et on peut alors écrire

$$\psi(\eta) = \log\int_0^\infty h(z)\exp\left(\eta\cdot z\right)\,dz$$

Il faut ici écrire $h(z)$ proprement, puis définir $\pi(\eta)$ puis enfin $\pi(\theta)$.

Exercice 17 *Soit $\Theta=I\!N$ et $\Pi_n=\mathcal{U}(\{0,1,\ldots,n\})$ la distribution uniforme discrète sur le compact discret $\{0,\ldots,n\}$. Prouver que $\{\Pi_n\}_n$ converge $q-$vaguement vers la mesure de comptage.*

Réponse. On a

$$\pi_n(\theta) = \frac{1}{n+1}\mathbb{1}_{\{0,1,\ldots,n\}}(\theta).$$

Choisissons $a_n=n+1$. Alors pour tout $\theta\in I\!N$, $\lim_{n\to\infty}a_n\pi_n(\theta)=\lim_{n\to\infty}\mathbb{1}_{\{0,1,\ldots,n\}}(\theta)=1$.

Exercice 18 *Soit $\Theta = \mathbb{R}$ et $\Pi_n = \mathcal{N}(0, n)$. Prouver que $\{\Pi_n\}_n$ converge $q-$vaguement vers la mesure de Lebesgue.*

Réponse. On a

$$\pi_n(\theta) \quad = \quad \frac{1}{\sqrt{2\pi n}} \exp\left(-\theta^2/2n\right).$$

et $\pi(\theta) = 1$. Choisissons $a_n = \sqrt{2\pi n}$ pour $n > 0$. Alors pour tout $\theta \in \mathbb{N}$, $\lim_{n\to\infty} a_n \pi_n(\theta) = 1$. De plus, $\forall(n, \theta)$, $a_n \pi_n(\theta) < 2$. On utilise ensuite la proposition 5.

Proposition 5 *Soient μ et μ_n des mesures a priori sur Θ. Supposons que :*

1. *il existe un suite de réels positifs $\{a_n\}_n$ tel que la suite $\{a_n \mu_n\}_n$ converge ponctuellement vers μ ;*
2. *pour tout ensemble compact K, il existe un scalaire M et $N \in \mathbb{N}$ tels que, pour tout $n > N$,*

$$\sup_{\theta \in K} a_n \mu_n(\theta) < M.$$

Alors $\{\mu_n\}_n$ converge $q-$vaguement vers μ.

3.2 Démonstrations

Théorème 8 Propriété d'invariance du prior de Jeffreys. *Soit $\pi_\theta(\theta)$ le prior de Jeffreys pour la paramétrisation θ, et soit $\eta = g(\theta)$ n'importe quelle reparamétrisation bijective de θ. Alors*

$$\pi_\eta(\eta) \quad \propto \quad \sqrt{\det I(\eta)}.$$

Le prior de Jeffreys vérifie donc le principe d'invariance (intrinsèque) proposé par la Définition 1 pour n'importe quelle reparamétrisation.

Définition 1 Principe d'invariance par reparamétrisation. *Si on passe de θ à $\eta = g(\theta)$ par une bijection g, l'information a priori reste inexistante et ne devrait pas être modifiée.*

Preuve. On propose une preuve pour $\dim \Theta = 1$. On pose $\eta = g(\theta)$. On rappelle que

$$\begin{aligned}
I(g(\theta)) &= \mathbb{E}_{g(\theta)}\left[\left(\frac{\partial}{\partial g(\theta)} \log f(x|g(\theta)\right)^2\right], \\
&= \mathbb{E}_{g(\theta)}\left[\left(\frac{\partial}{\partial \theta} \log f(x|g(\theta)\right)^2\right] \left(\frac{\partial \theta}{\partial g(\theta)}\right)^2, \\
&= \mathbb{E}_{g(\theta)}\left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta\right)^2\right] \frac{1}{(g'(\theta))^2} \quad \text{car } g \text{ est bijective}, \\
&= I(\theta)\frac{1}{(g'(\theta))^2} \;=\; I(\theta)\frac{1}{\left(\frac{\partial \eta}{\partial \theta}\right)^2}.
\end{aligned}$$

On en déduit, en utilisant la règle du changement de variable,

$$\begin{aligned}
\pi_\eta(\eta) &= \pi_\theta\left(g^{-1}(\eta)\right)\left|\det \frac{\partial \theta}{\partial \eta}\right|, \\
&\propto \sqrt{I(g^{-1}(\eta)}\left|\frac{\partial \theta}{\partial \eta}\right|, \\
&\propto \sqrt{I(\eta)}\left|\frac{\partial \eta}{\partial \theta}\frac{\partial \theta}{\partial \eta}\right|, \\
&\propto \sqrt{I(\eta)}
\end{aligned}$$

(dans ce cas unidimensionnel $\det I(\theta) = I(\theta)$).

Théorème 9 Prior par maximum d'entropie. *Le problème*

$$\pi^*(\theta) \quad = \quad \arg\max_{\pi\in\mathcal{P}} -\int_\Theta \pi(\theta)\log\frac{\pi(\theta)}{\pi_0(\theta)}\,d\theta$$

dans l'ensemble $\mathcal{P}$ des mesures positives, sous M contraintes linéaires

$$\int_\Theta g_i(\theta)\pi(\theta)\,d\theta \quad = \quad c_i,\ ,\quad i=1,\ldots,M,$$

a pour solution unique presque partout

$$\pi(\theta) \quad \propto \quad \pi_0(\theta)\exp\left(\sum_{i=1}^M \lambda_i g_i(\theta)\right)$$

où les λ_i sont des réels.

Preuve. Dans ce cadre fonctionnel, on définit le problème global d'optimisation

$$\pi^*(\theta) \quad = \quad \arg\max_{\pi\in\mathcal{P}} \mathcal{L}_\pi$$

où $\mathcal{L}(\pi)$ est le Lagrangien défini par

$$\mathcal{L}(\pi) \quad = \quad -\int_\Theta \pi(\theta)\log\frac{\pi(\theta)}{\pi_0(\theta)}\,d\theta - \sum_{i=1}^M \lambda_i\int_\Theta g_i(\theta)\pi(\theta)\,d\theta + C$$

et où les λ_i sont des réels, nommés *multiplicateurs de Lagrange*, et C est une constante. L'optimum π^*, s'il existe dans $\mathcal{P}$, est défini comme une solution du *problème variationnel*

$$\frac{\partial\mathcal{L}(\pi^*)}{\partial\pi} \quad = \quad 0$$

où la différentielle fonctionnelle $\frac{\partial\mathcal{L}(\pi^*)}{\partial\pi}$ (ici grossièrement écrite) doit être redéfinie. On remarque que l'application $\pi\mapsto\mathcal{L}(\pi)$ est convexe, ce qui implique l'existence d'un optimum unique. On introduit une *direction* h (une densité) sur Θ suffisamment intégrable (L^2) et $\epsilon\in\mathbb{R}$; alors la différentielle de $\mathcal{L}(\pi)$ dans la direction h est définie par

$$\begin{aligned}
\nabla_h\mathcal{L}(\pi) &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\{\mathcal{L}(\pi+\epsilon h)-\mathcal{L}(\pi)\right\},\\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\{\int_\Theta (\pi(\theta)+\epsilon h(\theta))\left(\log\left[\pi(\theta)+\epsilon h(\theta)\right]-\log\pi_0(\theta)\right)\,d\theta\right.\\
&\qquad \left.-\int_\Theta \pi(\theta)\log\frac{\pi(\theta)}{\pi_0(\theta)}\right\} - \sum_{i=1}^M\lambda_i\int_\Theta g_i(\theta)h(\theta)\,d\theta,\\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\{\int_\Theta (\pi(\theta)+\epsilon h(\theta))\left(\log\left[1+\epsilon\frac{h(\theta)}{\pi(\theta)}\right]-\log\pi_0(\theta)\right)\,d\theta\right.\\
&\qquad \left.+\int_\Theta \pi(\theta)\log\pi_0(\theta)\,d\theta\right\} + \int_\Theta h(\theta)\log\pi(\theta)\,d\theta - \sum_{i=1}^M\lambda_i\int_\Theta g_i(\theta)h(\theta)\,d\theta,\\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\{\int_\Theta (\pi(\theta)+\epsilon h(\theta))\left[\frac{\epsilon h(\theta)}{\pi(\theta)}(1+k(\theta))-\log\pi_0(\theta)\right]\,d\theta\right.\\
&\qquad \left.+\int_\Theta \pi(\theta)\log\pi_0(\theta)\,d\theta\right\} + \int_\Theta h(\theta)\log\pi(\theta)\,d\theta\\
&\qquad -\sum_{i=1}^M\lambda_i\int_\Theta g_i(\theta)h(\theta)\,d\theta,
\end{aligned}$$

en faisant un développement limité à l'ordre 1, où $k(\theta)$ est une fonction dominée par $h(\theta)$, telle que $h(\theta)k(\theta)$ et$(h^2(\theta)/\pi(\theta))(1+k(\theta))$ soient intégrables sur Θ. Alors

$$\begin{aligned}\nabla_h \mathcal{L}(\pi) &= \lim_{\epsilon \to 0} \left[\int_\Theta h(\theta)(1+k(\theta))\, d\theta + \epsilon \int_\Theta \frac{h^2(\theta)}{\pi(\theta)}(1+k(\theta))\, d\theta + \int_\Theta h(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)}\, d\theta \right] \\ &\quad - \sum_{i=1}^M \lambda_i \int_\Theta g_i(\theta) h(\theta)\, d\theta.\end{aligned}$$

Preuve. **(suite)** Avec $\int_\Theta h(\theta)\, d\theta = 1$, on en déduit donc que π^* est tel que (au premier ordre variationnel)

$$\int_\Theta h(\theta) \log \frac{\pi^*(\theta)}{\pi_0(\theta)}\, d\theta = \sum_{i=1}^M \lambda_i \int_\Theta g_i(\theta) h(\theta) + D$$

où D est une constante indépendante de h, pour toute direction h respectant les conditions d'intégrabilité.

On en déduit que

$$\int_\Theta h(\theta) \left[\log \frac{\pi^*(\theta)}{\pi_0(\theta)} - \sum_{i=1}^M \lambda_i g_i(\theta) - D \right] d\theta = 0$$

et en reconnaissant un produit scalaire $< h, g >= 0$ pour tout h (en supposant que Θ est un borélien sur $\mathbb{R}^d$ avec $d < \infty$ et des conditions de bornitude), on a que

$$\log \frac{\pi^*(\theta)}{\pi_0(\theta)} = \sum_{i=1}^M \lambda_i g_i(\theta) + D \quad \text{presque partout.}$$

On en déduit l'expression

$$\pi^*(\theta) \propto \pi_0(\theta) \exp\left(\sum_{i=1}^M \lambda_i g_i(\theta) \right).$$

Proposition 6 *Le minimiseur de la perte pondérée*

$$\pi^*(\theta) = \arg\min_\pi \sum_{i=1}^M \omega_i KL(\pi, \pi_i)$$

est l'a priori opérant la fusion logarithmique.

Preuve. Soit $\{\pi_i\}_{i=1,\ldots,M}$ un ensemble de lois *a priori* sur Θ. On pose

$$\pi^*(\theta) = \arg\min_\pi \sum_{i=1}^M \omega_i KL(\pi, \pi_i)$$

avec $\sum_{i=1}^M \omega_i = 1$ et $0 \leq \omega_i \leq 1\ \forall i = 1, \ldots, M$. Alors, sous réserve que

$$A = \int_\Theta \prod_{i=1}^M \pi_i^{\omega_i}(\theta)\, d\theta < \infty,$$

on a

$$\begin{aligned}\pi^*(\theta) &= \arg\min_\pi \sum_{i=1}^M \omega_i \int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{\pi_i(\theta)}\, d\theta, \\ &= \arg\min_\pi \int_\Theta \pi(\theta) \left(\log \pi(\theta) - \log A^{-1} \prod_{i=1}^M \pi_i^{\omega_i}(\theta) - A \right) d\theta, \\ &= \arg\min_\pi \left\{ KL\left(\pi || A^{-1} \prod_{i=1}^M \pi_i^{\omega_i}(.) \right) - A \right\}, \\ &= \arg\min_\pi KL\left(\pi || A^{-1} \prod_{i=1}^M \pi_i^{\omega_i}(.) \right).\end{aligned}$$

et on en déduit que

$$\pi^*(\theta) \quad = \quad A^{-1} \prod_{i=1}^{M} \pi_i^{\omega_i}(\theta).$$

3.3 TP : Un exemple complet dans un cadre de fiabilité industrielle

Soit X la durée de vie d'un composant Σ, supposé tomber uniquement en panne par hasard. Le taux de défaillance λ de Σ est donc constant, ce qui implique $X \sim \mathcal{E}(\lambda)$. Il est courant de disposer d'un expert industriel familier de λ, avec qui le dialogue suivant peut être engagé. "Considérons une décision de management (remplacement) établie sur une valeur donnée $\bar{\lambda}$ *(différente de la vraie valeur inconnue λ)*

Pour un *coût* similaire $|\bar{\lambda} - \lambda|$, il y a 2 conséquences possibles au remplacement :

- soit C_1 le coût positif moyen d'être trop optimiste (d'avoir $\bar{\lambda} \leq \lambda$) ;
- soit C_2 le coût positif moyen d'être trop pessimiste (d'avoir $\bar{\lambda} > \lambda$).
- Pouvez-vous donner un estimé $\hat{\delta}$ du rapport des coûts moyens $\delta = C_2/C_1$" ?

L'axiome de rationalité dit que si l'expert n'est pas *averse au risque*, alors

$$\bar{\lambda} \quad = \quad \arg\min_x \underbrace{\int_0^\infty |x - \lambda| \left(C_1 \mathbb{1}_{\{x \leq \lambda\}} + C_2 \mathbb{1}_{\{x > \lambda\}} \right) \pi(\lambda) \, d\lambda}_{\text{fonction de coût intégrée sur toutes les valeurs possibles } a \text{ priori} \text{ du vrai } \lambda}$$

$$\text{Il s'ensuit que} \quad \int_0^{\bar{\lambda}} d\Pi(\lambda) \quad = \quad \Pi(\lambda < \bar{\lambda}) \; = \; \frac{C_1}{C_1 + C_2}.$$

L'interprétation de la réponse de l'expert est que $1/(1 + \hat{\delta})$ est un estimé du quantile *a priori* d'ordre $\alpha = C_1/(C_1 + C_2)$. Avec $P(\lambda < \bar{\lambda}) = \frac{C_1}{C_1 + C_2} = \alpha$, on a :

- tant que les coûts sont équilibrés, un expert de plus en plus optimiste fournira un $\bar{\lambda}$ de plus en plus petit, car la durée moyenne avant la prochaine défaillance est

$$\mathbb{E}[X|\lambda] \quad = \quad \frac{1}{\lambda}.$$

- cependant l'expert s'exprime plutôt sur les coûts lorsqu'on lui fournit une valeur représentative de $\bar{\lambda}$:
  - plus l'expert est optimiste, plus le coût C_2 d'être optimiste (selon lui) est petit, donc α grandit vers 1 et

$$\mathbb{E}[X] \quad = \quad \mathbb{E}[\mathbb{E}[X|\lambda]] \; = \; \mathbb{E}[1/\lambda] \text{ augmente.}$$

  - plus l'expert est pessimiste, plus le coût C_2 d'être optimiste augmente, donc α tombe vers 0 et

$$\mathbb{E}[X] \quad = \quad \mathbb{E}[1/\lambda] \text{ diminue.}$$

Quelle choix de loi *a priori* pouvons-nous proposer au décideur ?

4 Corrigés et preuves "Calcul bayésien"

4.1 Exercices

Exercice 19 *On suppose $X \sim \mathcal{N}(\theta, 1)$ et on suppose connaître un échantillon $\mathbf{x_n}$ composé de :*

- *quelques observations $x_1, \ldots, x_{n-1}$ supposées iid.*
- *une pseudo-observation y qui est un cas-limite masquant (censurant) une observation x_n qui aurait dû être faite : $y < x_n$*

A priori, on suppose $\theta \sim \mathcal{N}(\mu, 1)$. Pouvez-vous produire un algorithme d'AR qui génère des réalisations de la loi a posteriori de θ ?

Réponse. La vraisemblance s'écrit

$$f(\mathbf{x_n}|\theta) \quad \propto \quad \underbrace{\exp\left(-\frac{1}{2}\sum_{k=1}^{n-1}(x_k-\theta)^2\right)}_{\text{terme régulier}} \quad \left(\underbrace{1-\Phi(y-\theta)}_{\substack{\text{terme dû à la censure} \\ = P(X>y)}}\right).$$

L'*a posteriori* sur θ s'écrit alors

$$\pi(\theta|\mathbf{x_n}) \quad \propto \quad \tilde{\pi}(\theta|\mathbf{x_n}) \;=\; \exp\left\{-\frac{n}{2}\left[\theta - \frac{1}{n}\left(\mu + \sum_{k=1}^{n-1} x_k\right)\right]^2\right\}\{1-\Phi(y-\theta)\}\,.$$

On remarque que si $y = x_n$, on aurait un modèle conjugué et $\pi(\theta|\mathbf{x_n})$ serait une loi normale. Il semble donc pertinent de proposer, comme choix de loi instrumentale,

$$\rho(\theta) \equiv \mathcal{N}\left(\frac{1}{n}\left(\mu + \sum_{k=1}^{n-1} x_k\right), 1/n\right).$$

Puisque $1-\Phi(y-\theta) \leq 1$, on a

$$\tilde{\pi}(\theta|\mathbf{x_n}) \quad \leq \quad \underbrace{\sqrt{\frac{2\pi}{n}}}_{K} \cdot \{1-\Phi(y-\theta)\} \cdot \rho(\theta).$$

On peut donc mettre en oeuvre l'algorithme comme suit : on accepte θ_i si $U_i \leq 1-\Phi(y-\theta_i)$. Le nombre moyen d'appels nécessaires à $\rho(\theta)$ varie proportionnellement à $1/\sqrt{n}$, donc plus l'échantillon de données grandit, plus l'algorithme est efficace. Si cependant, on fait le choix $\rho(\theta) = \pi(\theta)$, alors

$$K \quad = \quad \sqrt{2\pi}\exp\left(\frac{1}{2}\left[\frac{1}{n-1}\sum_{k=1}^{n} x_k - \mu\right](1-\sqrt{n})\right)$$

Voir la figure 1 pour une illustration de la mise en oeuvre de cet algorithme.

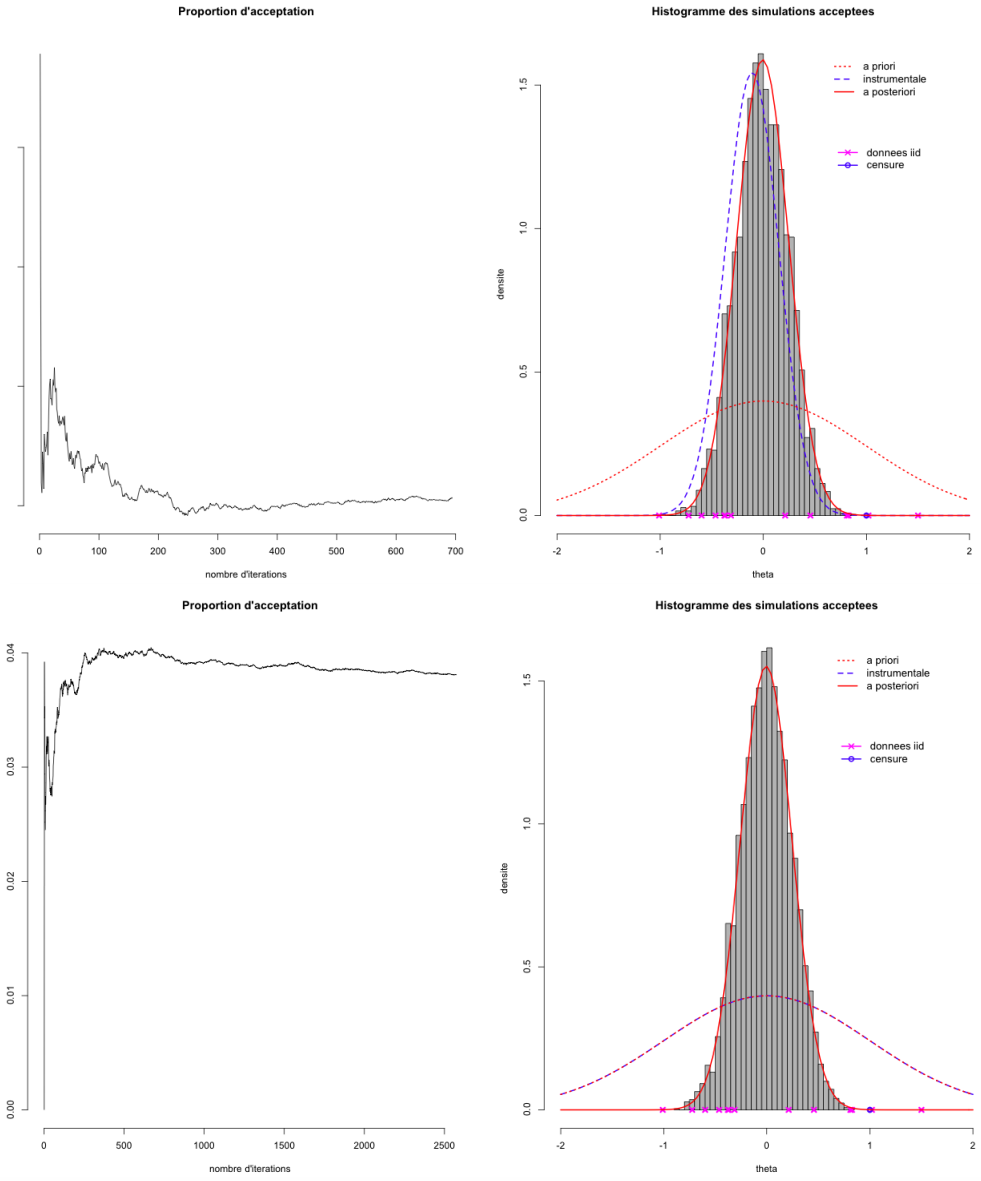


FIGURE 1 – Essai de simulation par AR en utilisant (en haut) une loi instrumentale “proche" du vrai posterior ; (en bas) le prior comme loi instrumentale.

Exercice 20 *Soit un échantillon de loi gamma* $x_1, \ldots, x_n \overset{iid}{\sim} \mathcal{G}(a, \theta)$ *où* a *est connu. On suppose* $\pi(\theta) \equiv \mathcal{G}(c, d)$*. Produisez une méthode par AR pour simuler la loi a posteriori* $\pi(\theta|x_1, , x_n)$ *et vérifiez que les tirages obtenus sont bien issus de cette loi, par ailleurs explicite.*

Réponse. Il est aisé de vérifier qu'on connaît parfaitement la loi *a posteriori* de θ :

$$\theta|x_1, \ldots, x_n \quad \sim \quad \mathcal{G}\left(c + an, d + \sum_{i=1}^{n} x_i\right).$$

On peut donc vérifier l'accord entre un échantillon iid simulé par AR et cette loi, via des tests statistiques classiques comme Kolmogorov-Smirnov, Cramer-von Mises ou Anderson-Darling. Pour construire cet algorithme d'AR, faisons par exemple le choix d'une loi instrumentale lognormale (qui est bien à support positif, car $\theta > 0$) :

$$\rho(\theta|\mu, \sigma) \quad = \quad \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \log\theta\right)^2\right).$$

Il vient alors

$$\kappa(\theta|\mu, \sigma) \; = \; \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{\rho(\theta)} \quad = \quad \frac{\sqrt{2\pi}\sigma\theta^{c+an} \exp\left(-\theta(d + \sum_{i=1}^{n} x_i)\right)}{\exp\left(-\frac{1}{2\sigma^2}\left(\mu - \log\theta\right)^2\right)}$$

et

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log\kappa(\theta|\mu, \sigma) &= \frac{c + an}{\theta} - \left(d + \sum_{i=1}^{n} x_i\right) - \frac{1}{\sigma^2\theta}\left(\mu - \log(\theta)\right), \\
\frac{\partial^2}{\partial\theta^2} \log\kappa(\theta|\mu, \sigma) &= \frac{1}{\theta^2}\left(-(c + an) + \frac{1}{\sigma^2}(1 + \mu - \log(\theta)\right).
\end{aligned}$$

Notons $\theta_0(\mu, \sigma) = \exp(-(c + an) + (1 + \mu)/\sigma^2)$. Si $0 \leq \theta \leq \theta_0$, alors $\frac{\partial}{\partial\theta} \log\kappa(\theta|\mu, \sigma)$ est croissante. Si $\theta > \theta_0$, elle est décroissante vers $-(d + \sum_{i=1}^{n} x_i) < 0$. Pour permettre à $\kappa(\theta|\mu, \sigma)$ d'être maximisable sur $\theta > 0$, il faut donc que

$$\frac{\partial}{\partial\theta} \log\kappa(\theta_0(\mu, \sigma)|\mu, \sigma) = -\left(d + \sum_{i=1}^{n} x_i\right) + 1/\sigma^2\theta_0 \; > \; 0.$$

Sous cette contrainte, on peut résoudre numériquement en $\theta = \theta_1(\mu, \sigma) > \theta_0(\mu, \sigma)$ l'équation $\frac{\partial}{\partial\theta} \log\kappa(\theta|\mu, \sigma) = 0$, et $\theta_1(\mu, \sigma)$ maximise alors $\kappa(\theta|\mu, \sigma)$. Dans ce cas, on peut définir

$$K \quad = \quad \arg\min_{\mu,\sigma} \kappa(\theta_1(\mu, \sigma)|\mu, \sigma).$$

et mettre en place l'algorithme AR.

Exercice 21 *Considérons une fonction d'intérêt* $h(\theta)$ *que l'on cherche à résumer par un estimateur calculé sous un coût quadratique ; il s'agit donc de l'espérance a posteriori*

$$h \; = \; \mathbb{E}_\pi[h(\theta)|x_1, \ldots, x_n] \quad = \quad \int_\Theta h(\theta)\pi(\theta|x_1, \ldots, x_n)\, d\theta \tag{15}$$

que l'on suppose pouvoir estimer simplement, de façon consistante, par Monte Carlo. Supposons vouloir modifier le prior : $\pi(\theta) \to \pi'(\theta)$*, sans modifier le support, mais de façon à ce que la nouvelle loi a posteriori ne soit plus directement simulable. Peut-on (et sous quelles conditions) ne pas faire de calcul supplémentaire pour simuler le nouveau posterior* $\pi'(\theta_1, \ldots, x_n)$ *?*

Réponse. On considére donc une fonction d'intérêt $h(\theta)$ que l'on cherche à résumer par son espérance *a posteriori*

$$h \;=\; \mathbb{E}_\pi[h(\theta)|x_1,\ldots,x_n] \;=\; \int_\Theta h(\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta \tag{16}$$

que l'on suppose pouvoir estimer simplement, de façon consistante, par Monte Carlo :

$$\hat{h}_M \;=\; \frac{1}{M}\sum_{k=1}^{M} h(\theta_i) \quad \text{avec } \theta_i \overset{iid}{\sim} \pi(\theta|x_1,\ldots,x_n)$$

Supposons vouloir modifier le prior : $\pi(\theta) \to \pi'(\theta)$, sans modifier le support, mais de façon à ce que la nouvelle loi *a posteriori* ne soit plus directement simulable. En supposant que $\pi(\theta|x_1,\ldots,x_n) > 0$ pour tout $\theta \in \Theta$, on peut néanmoins recalculer facilement le nouvel estimateur de Bayes :

$$\begin{aligned} h' \;=\; \mathbb{E}_{\pi'}[h(\theta)|x_1,\ldots,x_n] \;&=\; \int_\Theta h(\theta)\pi'(\theta|x_1,\ldots,x_n)\,d\theta, \\ &=\; \int_\Theta \omega(\theta_i)h(\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta \end{aligned}$$

avec

$$\omega(\theta_i) \;=\; \frac{\pi'(\theta|x_1,\ldots,x_n)}{\pi(\theta|x_1,\ldots,x_n)} \;=\; C\tilde{\omega}_i^*$$

où

$$\begin{aligned} \tilde{\omega}_i^* \;&=\; \left(\frac{\pi'(\theta_i)}{\pi(\theta_i)}\right), \\ C \;&=\; \frac{m_\pi(x_1,\ldots,x_n)}{m_{\pi'}(x_1,\ldots,x_n)}. \end{aligned}$$

Le calcul de la constante de proportionnalité C nécessiterait usuellement de disposer de deux échantillons *a posteriori*. On peut cependant s'en passer en remarquant que d'après la loi forte des grands nombres,

$$\frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_i^* \;\xrightarrow[n\to\infty]{p.s.}\; \frac{1}{C}\int_\Theta \omega(\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta \;=\; 1/C$$

et on en déduit donc qu'un estimateur IS consistant de h', qui réutilise les calculs faits pour l'estimateur $\hat{h}_M$ sans coût calculatoire additionnel, est

$$\hat{h''}_M \;=\; \frac{1}{M}\sum_{k=1}^{M}\hat{\omega}_i^* h(\theta_i)$$

avec

$$\hat{\omega}_i^* \;=\; \frac{\tilde{\omega}_i^*}{\sum_{j=1}^{M}\tilde{\omega}_j^*}.$$

Notons que ce faisant, on crée de la corrélation entre les deux estimateurs de h.

Exercice 22 *Soit X la variable "débit maximal de rivière". Elle est supposée suivre une loi des extrêmes (Gumbel) de densité*

$$f(x|\theta) \;=\; \lambda\mu\exp(-\lambda x)\exp(-\mu\exp(-\lambda x)).$$

avec $\theta = (\mu, \lambda)$.

Considérons n observations $\mathbf{x_n} = (x_1, \ldots, x_n)$ supposées iid selon cette distribution.

1. *Comment s'écrit la vraisemblance ?*
2. *On considère l'a priori $\pi(\mu, \lambda) = \pi(\mu|\lambda)\pi(\lambda)$ avec*

$$\begin{aligned} \mu|\lambda &\sim \mathcal{G}\left(m, b_m(\lambda)\right), \\ \lambda &\sim \mathcal{G}\left(m, m/\lambda_e\right) \end{aligned}$$

et $b_m(\lambda) = \left[\alpha^{-1/m} - 1\right]^{-1} \exp(-\lambda x_{e,\alpha})$.. Ces hyperparamètres ont le sens suivant :

- *$x_{e,\alpha}$ = quantile prédictif a priori d'ordre α :*

$$P\left(X < x_{e,\alpha}\right) = \int P\left(X < x_{e,\alpha}|\mu, \lambda\right) \pi(\mu, \lambda)\, d\mu d\lambda = \alpha;$$

- *m = taille d'échantillon fictif, associée à la "force" de la connaissance a priori $x_{e,\alpha}$;*
- *$1/\lambda_e$ = moyenne de cet échantillon fictif.*

Pouvez-vous produire un algorithme de type MCMC qui permette de générer une loi jointe a posteriori pour (μ, λ) ?

Réponse. En posant $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{b}_{\mathbf{x_n}}(\lambda) = \sum_{i=1}^{n} \exp\left(-\lambda x_i\right)$, la vraisemblance s'écrit alors

$$f\left(\mathbf{x_n}\right) = \lambda^n \mu^n \exp(-\lambda n \bar{x}_n) \exp\{-\mu \bar{b}_{\mathbf{x_n}}(\lambda)\}.$$

En conséquence, la loi *a posteriori* s'obtient sous la forme *hiérarchisée* suivante :

$$\pi\left(\mu, \lambda|\mathbf{x_n}\right) = \pi\left(\mu|\lambda, \mathbf{x_n}\right) \pi\left(\lambda|\mathbf{x_n}\right)$$

où

$$\mu|\lambda, \mathbf{x_n} \sim \mathcal{G}\left(m + n, b_m(\lambda) + \bar{b}_{\mathbf{x_n}}(\lambda)\right)$$

et

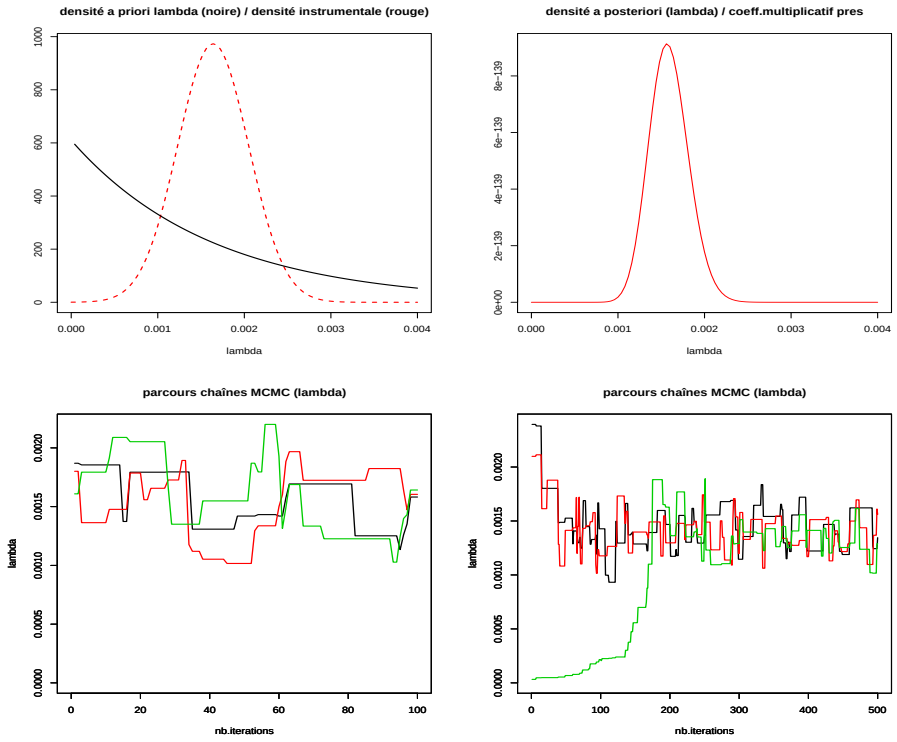
$$\pi\left(\lambda|\mathbf{x_n}\right) = \gamma(\lambda) \cdot \mathcal{G}\left(m + n, m/\lambda_e + n\bar{x}_n\right)$$

avec

$$\gamma(\lambda) \quad \propto \quad \frac{b_m^m(\lambda)}{\left(b_m(\lambda) + \bar{b}_{\mathbf{x_n}}(\lambda)\right)^{m+n}}$$

La loi *a priori* est donc *semi-conjuguée*, et il suffit de simuler λ *a posteriori* pour obtenir un tirage joint *a posteriori* de (μ, λ). On trace ci-dessous quelques graphiques typiquement obtenus. Plusieurs choix de lois instrumentales pour $\rho(\lambda|\lambda^{(i-1)})$ peuvent être faits. On peut notamment proposer :

- la loi *a priori* $\pi(\lambda)$;
- une loi qui "semble proche" : $\mathcal{G}\left(m+n, m/\lambda_e + n\bar{x}_n\right)$;
- une loi normale de moyenne $\lambda^{(i-1)}$ et de coefficient de variation petit (5%) ou grand (25 ou 50%).



Exercice 23 (Retour à l'exercice 19).. *On suppose de nouveau connaître un échantillon* $\mathbf{x_n} \sim \mathcal{N}(\theta, 1)$ *composé de quelques observations* $x_1, \ldots, x_{n-1}$ *supposées iid de loi* $\mathcal{N}(\theta, 1)$*, et d'une pseudo-observation* y *qui est un cas-limite masquant (censurant) une observation* x_n *qui aurait dû être faite :* $y < x_n$*. On considère toujours* $\theta \sim \mathcal{N}(\mu, 1)$ *a priori. Pouvez-vous produire un algorithme d'échantillonnage par Gibbs qui génère des réalisations de la loi a posteriori de* θ *?*

Réponse. Si on connaissait x_n, le modèle bayésien serait conjugué et

$$\theta|\mathbf{x_n} \quad \sim \quad \mathcal{N}\left(\frac{1}{n}\left(\mu + \sum_{i=1}^{n} x_i\right), (n+1)^{-1}\right)$$

On considère alors la donnée manquante x_n comme un paramètre inconnu et aléatoire. Sachant θ et $\mathbf{x_n}$, on peut montrer par la règle de Bayes que la loi de la variable aléatoire manquante X_n est la normale tronquée

$$\mathcal{N}(\theta, 1) \cdot \mathbb{1}_{\{x_n \geq y\}}.$$

En effet, la fonction de répartition de X_n est conditionnelle : $P(X_n < x | X_n > y)$. Par la règle de Bayes

$$P(X_n < x | X_n > y) \quad = \quad \frac{P(X_n < x \cap X_n > y)}{P(X_n > y)} \quad = \quad \frac{P(y < X_n < x)}{P(X_n > y)}.$$

Le dénominateur est une constante (indépendante de x). Donc

$$P(X_n < x | X_n > y) \quad \propto \quad \int_y^x f_X(u)\, du \quad = \quad \int_{-\infty}^x f_X(u) \mathbb{1}_{\{y \leq u\}}\, du$$

où f_X est la densité d'un X non-contraint (ici gaussienne). On en déduit que la densité de X_n est

$$f_{X_n}(x) \quad = \quad \frac{f(x)\mathbb{1}_{\{y \leq x\}}}{\int_{-\infty}^{\infty} f(u)\mathbb{1}_{\{y \leq u\}}\, du}.$$

L'algorithme de Gibbs à mettre en oeuvre est donc le suivant :

- `On part d'une valeur` $\theta^{(0)}$
- `Itération` $i \geq 1$ `:`
  1. `on simule` $x_n^{(i)} \sim \mathcal{N}\left(\theta^{(i-1)}, 1\right) \cdot \mathbb{1}_{\{x_n \geq y\}}$
  2. `on simule` $\theta^{(i)} \sim \mathcal{N}\left(\frac{1}{n}\left(\mu + \sum_{i=1}^{n-1} x_i + x_n^{(i)}\right), (n+1)^{-1}\right)$

Exercice 24 Modèle à effets aléatoires autour d'une constante (Hobert-Casella). *Pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$, on considère*

$$x_{ij} \quad = \quad \beta + u_i + \epsilon_{ij}$$

où $u_i \sim \mathcal{N}(0, \sigma^2)$ et $\epsilon_{ij} \sim \mathcal{N}(0, \tau^2)$. Ce type de modèle permet de représenter la distribution d'une caractéristique au sein d'une population, où β est une tendance moyenne, u_i correspond à une variation d'un groupe et ϵ_{ij} à une variation au sein d'un sous-groupe. On suppose choisir

$$\pi(\beta, \sigma^2, \tau^2) \quad \propto \quad \frac{1}{\sigma^2 \tau^2} \quad \textit{(prior de Jeffreys).}$$

On note $\mathbf{x_{IJ}}$ l'échantillon des données observées, $\bar{x}_i$ la moyenne sur les j. On note $\mathbf{u_I}$ l'échantillon manquant des $u_1, \ldots, u_I$ (reconstitué dans l'inférence).

1. *Calculer les lois conditionnelles a posteriori de*

$$\begin{aligned} U_i | \mathbf{x_{IJ}}, \beta, \sigma^2, \tau^2 \\ \beta | \mathbf{x_{IJ}}, \sigma^2, \tau^2, \mathbf{u_I} \\ \sigma^2 | \mathbf{x_{IJ}}, \beta, \tau^2, \mathbf{u_I} \\ \tau^2 | \mathbf{x_{IJ}}, \beta, \sigma^2, \mathbf{u_I} \end{aligned}$$

   Ces lois sont-elles bien définies ?

2. *Donner une formule (à un coefficient proportionnel près) pour la loi a posteriori jointe $\pi(\sigma^2, \tau^2 | \mathbf{x_{IJ}})$. Comment se comporte-t-elle au voisinage de $\sigma = 0$, pour $\tau \neq 0$? Que pouvez-vous en déduire ?*

3. *Mettre en place un algorithme de Gibbs permettant d'inférer sur $(\beta, \sigma^2, \tau^2)$. Que pouvez-vous dire sur la convergence des chaînes MCMC ?*

Réponse.

1. Des calculs algébriques montrent que les lois conditionnelles sont

$$\begin{aligned}
U_i|\mathbf{x_{IJ}},\beta,\sigma^2,\tau^2 &\sim \mathcal{N}\left(\frac{J(\bar{x}_i-\beta)}{J+\tau^2\sigma^{-2}},(J\tau^{-2}+\sigma^{-2})^{-1}\right)\\
\beta|\mathbf{x_{IJ}},\sigma^2,\tau^2,\mathbf{u_I} &\sim \mathcal{N}\left(\bar{x}-\bar{u},\tau^2/IJ\right)\\
\sigma^2|\mathbf{x_{IJ}},\beta,\tau^2,\mathbf{u_I} &\sim \mathcal{IG}\left(I/2,(1/2)\sum_{i=1}^{I}u_i^2\right) \quad \textit{(loi inverse gamma)}\\
\tau^2|\mathbf{x_{IJ}},\beta,\sigma^2,\mathbf{u_I} &\sim \mathcal{IG}\left(IJ/2,(1/2)\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-u_i-\beta)^2\right)
\end{aligned}$$

   qui sont donc bien définies.

2. La loi *a posteriori* jointe

$$\begin{aligned}
\pi(\sigma^2,\tau^2|\mathbf{x_{IJ}}) &= \int \pi(\beta,\sigma^2,\tau^2|\mathbf{x_{IJ}})\,d\beta\\
&= \int\left[\int_1\ldots\int_i\ldots\int_I \pi(\beta,\sigma^2,\tau^2|\mathbf{x_{IJ}})\,du_i\right]d\beta
\end{aligned}$$

   est proportionnelle à

$$\frac{\sigma^{-2-I}\tau^{-2-IJ}}{(J\tau^{-2}+\sigma^{-2})^{I/2}}\sqrt{\tau^2+J\sigma^2}\exp\left\{-\frac{1}{2\tau^2}\sum_{i,j}(y_{ij}-\bar{y}_i)^2-\frac{J}{2'\tau^2+J\sigma^2)}\sum_i(\bar{y}_i-\bar{y})^2\right\}$$

   qui se comporte comme σ^{-2} au voisinage de $\sigma=0$, pour $\tau\neq 0$. Cette loi jointe n'est donc pas intégrable (*propre*).

3. On constate normalement une absence de convergence claire des chaînes, ou une stationnarité (momentanée) trompeuse.

4.2 Démonstrations

Algorithme d'acceptation-rejet (AR).

1. simulation indirecte : soit $\theta_i \sim \rho(\cdot)$
2. test :
   - soit $U_i \sim \mathcal{U}[0,1]$
   - si $U_i \leq \dfrac{f(\mathbf{x_n}|\theta_i)\pi(\theta_i)}{K\rho(\theta_i)}$ alors θ_i suit la loi $\pi(\theta|\mathbf{x_n})$

Preuve. Soit θ la variable aléatoire dont les tirages sont acceptées par le test. Alors, en définissant P la mesure de probabilité usuelle sur $[0,1]$, et $\tilde{\Pi}$ la mesure de probabilité produit sur $\Theta \times [0,1]$, d'après la formule de Bayes,

$$\begin{aligned}
\Pi\left(\tilde{\theta} \leq y | U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right) &= \frac{\tilde{\Pi}\left(\tilde{\theta} \leq y\ ,\ U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right)}{\tilde{\Pi}\left(U \leq \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\right)}, \\
&= \frac{\int_{-\infty}^{y}\int_0^1 \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}\, du\rho(\theta)\, d\theta}{\int_{-\infty}^{\infty}\int_0^{\frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{K\rho(\theta)}} du\rho(\theta)\, d\theta}, \\
&= \frac{\int_{-\infty}^{y} f(\mathbf{x_n}|\theta)\pi(\theta)\, d\theta}{K\int_{-\infty}^{\infty}\frac{1}{K} f(\mathbf{x_n}|\theta)\pi(\theta)\, d\theta}, \\
&= \Pi(\theta \leq y|\mathbf{x_n}).
\end{aligned}$$

Théorème 10 *Importance sampling optimal. Soit l'estimateur de la fonction d'intérêt $h(\theta) \in \mathbb{R}$ par IS :*

$$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M}\frac{\pi(\theta_i|x)}{\rho(\theta_i)}h(\theta_i) \rightarrow \mathbb{E}_\pi[h(\theta|X] \quad p.s.$$

où les $\theta_i \overset{iid}{\sim} \rho(\theta)$. Alors le choix de ρ qui minimise la variance de l'estimateur $\hat{h}_M$ est

$$\rho^*(\theta) = \frac{|h(\theta)|\pi(\theta|X)}{\int_\Theta |h(\theta)|\pi(\theta|X)\, d\theta}.$$

Preuve. On note d'abord que

$$\mathbb{V}_\rho\left[\frac{h(\theta)\pi(\theta|x)}{\rho(\theta)}\right] = \mathbb{E}_\rho\left[\frac{h^2(\theta)\pi^2(\theta|x)}{\rho^2(\theta)}\right] - \left(\mathbb{E}_\rho\left[\frac{h(\theta)\pi(\theta|x)}{\rho(\theta)}\right]\right)^2$$

et que le second terme ne dépend pas de ρ. Il suffit donc de minimiser le premier terme. D'après l'inégalité de Jensen, on a donc,

$$\begin{aligned}
\mathbb{E}_\rho\left[\frac{h^2(\theta)\pi^2(\theta|x)}{\rho^2(\theta)}\right] &\geq \left(\mathbb{E}_\rho\left[\frac{|h(\theta)|\pi(\theta|x)}{\rho(\theta)}\right]\right)^2, \\
&\geq \left(\int_\Theta |h(\theta)|\pi(\theta|x)\, d\theta\right)^2
\end{aligned}$$

qui est une borne inférieure indépendante de ρ, qui est atteinte en ρ^*.

Références

[1] C.P. Robert. *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation (2nd edition)*. Springer, 2007.