

# Table des matières

<b>Table des matières</b>	<b>1</b>
<b>Avant-propos</b>	<b>3</b>
<b>Introduction générale</b>	<b>4</b>
<b>1 Principe Bayésien</b>	<b>5</b>
1.1 Introduction	5
1.2 Rappels sur les notions de probabilité	5
1.3 Inférence bayésienne	7
1.3.1 Information a priori	7
1.3.2 Loi des observations	7
1.3.3 Loi a posteriori	7
1.3.4 Modèle statistique bayésien	8
1.4 Raisonnement proportionnel	9
1.5 Approche décisionnelle	10
1.5.1 La fonction Coût (ou de Perte)	10
1.5.2 Risque fréquentiste (coût moyen)	10
1.5.3 Risque a posteriori	11
1.5.4 Risque Intégré	11
1.5.5 Risque de Bayes	11
1.5.6 Risque minimax	12
1.5.7 Admissibilité	13
<b>2 Estimation Bayésienne</b>	<b>14</b>
2.1 Estimateurs de Bayes : cas unidimensionnel	14
2.1.1 Estimateur MMSE	14
2.1.2 Estimateur MAP	15
2.1.3 Propriétés des estimateurs de bayes	16
2.2 Estimateurs de Bayes : cas multidimensionnel	17
2.3 Probabilité prédictive	17
2.4 Extension aux lois impropres	17
2.4.1 Loi a priori impropre	17
2.4.2 Estimateur de bayes généralisé	18

2.5	Intervalles de confiance bayésiens . . . . .	18
2.5.1	Intervalle de confiance a priori . . . . .	19
2.5.2	Intervalle de confiance a posteriori . . . . .	19
2.6	Approche bayésienne des tests . . . . .	20
2.6.1	Critère de décision . . . . .	20
2.6.2	Facteur de Bayes . . . . .	22
<b>3</b>	<b>Modélisation de l'information a priori</b>	<b>25</b>
3.1	Lois conjuguées naturelles . . . . .	25
3.2	Cas du modèle Exponentiel . . . . .	26
3.3	Lois non informatives . . . . .	27
3.3.1	A priori Uniforme . . . . .	27
3.3.2	A priori invariant . . . . .	28
3.3.3	Règle de Jeffreys (cas unidimensionnel) . . . . .	28
3.3.4	Généralisation : cas multidimensionnel . . . . .	29
3.3.5	Loi a priori de Référence . . . . .	30
3.4	Approche hiérarchique . . . . .	31
3.5	Incorporation de l'information a priori en pratique . . . . .	32
3.6	Le poids de l'a priori dans la réponse bayésienne . . . . .	33
3.7	Sensibilité de la réponse bayésienne à la loi a priori . . . . .	35
<b>4</b>	<b>Méthodes de simulation de Monte-Carlo</b>	<b>36</b>
4.1	Méthode de Monte Carlo . . . . .	36
4.2	Méthode de Monte Carlo par Chaînes de Markov (méthode MCMC) . . . . .	37
4.2.1	Algorithme de Metropolis-Hastings . . . . .	38
4.2.2	Echantionnage de Gibbs . . . . .	38
4.3	Mise en oeuvre des méthodes de simulation de Monte-Carlo . . . . .	39
<b>5</b>	<b>Exercices de TD</b>	<b>41</b>
5.1	Énoncé des exercices . . . . .	41
5.2	Éléments de réponses . . . . .	43
	<b>Bibliographie</b>	<b>50</b>

# Avant-Propos

Ce document est adressé à tous les étudiants ayant déjà acquis les connaissances nécessaires sur la statistique descriptive, inférentielle et en probabilité. Il correspond particulièrement aux enseignements en Master 2 de la Statistique et l'Analyse Décisionnelle (SAD). Il constitue un support du cours consacré au traitement des données statistiques par l'approche bayésienne.

Nous donnons dans ce cours les principes de l'inférence bayésienne permettant de définir les estimateurs bayésiens. Les deux difficultés majeures de l'approche bayésienne sont le choix de la loi a priori et le calcul des procédures bayésiennes qui peuvent être surmontées en suivant des règles simples.

Les exercices avec des éléments de réponses qui figurent dans le chapitre 5 sont des exercices des séries de TD effectuées au sein du département de Mathématiques pour les étudiants de Master 2 SAD. La bibliographie donne la liste des documents que nous avons consulté pour la réalisation de ce cours.

# Introduction

L'observation d'un phénomène aléatoire peut mener à une inférence sur la distribution de probabilité à l'origine de ce phénomène pour une analyse ou une prévision.

La statistique bayésienne est une approche cohérente et pratique pour la résolution des problèmes d'inférence statistique, adaptée aux outils informatiques de simulation et apte à répondre aux problèmes de modélisation les plus avancés dans toutes les disciplines. Cette approche, appelée aussi "Théorie moderne de la décision", tient compte d'informations a priori basées souvent sur des probabilités subjectives. Cette information exogène aux données, fournie par l'expérimentateur, est susceptible d'améliorer l'estimation des paramètres. Elle permet une analyse complète des incertitudes et aussi la modélisation (choix de la loi a priori).

L'estimation des paramètres est basée sur une distribution dite "loi a posteriori" qui permet de fusionner l'information donnée par les observations avec l'information a priori.

La densité a priori étant la densité qui décrit l'état de connaissance ou l'ignorance concernant les paramètres avant la prise en compte des observations, le choix de cette densité est le problème le plus délicat et le plus critique de l'analyse bayésienne. En effet, il est très rare que l'information disponible soit suffisamment précise pour pouvoir déterminer exactement cette densité. Il faut donc faire des approximations dont le choix aura des répercussions sur les inférences statistiques bayésiennes. Une revue complète concernant l'assignement des densités a priori est faite dans la référence [14]. Nous présentons dans ce document deux types de densités a priori les plus courants : les densités conjuguées et les densités a priori non informative.

Les algorithmes de simulation des méthodes de Monte Carlo par Chaînes de Markov (MCMC) sont très puissants et suffisamment généraux pour permettre de simuler des densités de probabilité complexes et de dimensions élevées et, donc, pour résoudre les problèmes numériques rencontrés en statistique paramétrique bayésienne. Une revue complète sur l'analyse bayésienne peut être trouvée dans les références (voir [16], [17] et [3]).

La statistique bayésienne connaît une croissance exponentielle par ses applications dans différents domaines à savoir : Science de l'ingénieur (Fiabilité), Science expérimentale, Sciences sociales, Médecine, Economie, ...etc. Elle est devenue un outil incontournable pour la modélisation des différents phénomènes.

# Principe Bayésien

## 1.1 Introduction

Soit  $X_1, X_2, \dots, X_n$  un  $n$ -échantillon issu d'une variable aléatoire (v.a.)  $X$  dont la loi dépend d'un paramètre inconnu  $\theta \in \mathbb{R}$ .

Le but de l'analyse statistique est de faire de l'inférence sur  $\theta$  (décrire un phénomène passé ou à venir dans un cadre probabiliste). Si toute l'information à propos de  $\theta$  est contenue dans les données  $\underline{x} = (x_1, x_2, \dots, x_n)$  on applique l'approche classique donnée par le Maximum de Vraisemblance "MV" (Maximum Likelihood Estimator "MLE").

On considère le cas où nous avons des connaissances a priori sur  $\theta$ . L'approche bayésienne consiste à utiliser cette information pour l'estimation du paramètre  $\theta$ .

## 1.2 Rappels sur les notions de probabilité

### Mesure

Soit  $\Omega$  un ensemble. Une mesure  $\mu$  sur  $\Omega$  est une application  $\mu : \mathcal{P}(\Omega) \rightarrow \mathbb{R}^+$  définie sur l'ensemble de toutes les parties de  $\Omega$ , additive ( $\mu(A \cup B) = \mu(A) + \mu(B)$  si  $A$  et  $B$  sont deux parties disjointes avec  $\mu(\emptyset) = 0$ ).

**Exemple 1.2.1.** Sur  $\mathbb{R}$  on peut définir, pour toute partie  $A \subset \mathbb{R}$ , la mesure  $\mu_n(A) = \frac{1}{n} \#\{0, \dots, n-1\} \cap A$ . En passant à la limite sur  $n$ , on obtient ce qu'on appelle une "moyenne"  $\mu$  qui est additive, positive et vérifie  $\mu(\mathbb{R}) = 1$ .

### Tribu

Soit  $\Omega$  un ensemble. Une tribu, ou  $\sigma$ -Algèbre est une collection  $\mathcal{A} \subset \mathcal{P}(\Omega)$  de parties de  $\Omega$  qui contient  $\emptyset$  et  $\Omega$ , stable par union dénombrable et par passage au complémentaire.

Les exemples classiques et les plus utilisés en pratique sont les suivants :

Si  $\Omega$  est dénombrable  $\mathcal{P}(\Omega)$  est une tribu. Sur  $\mathbb{R}$  (ou plus généralement sur  $\mathbb{R}^d$ ) on considère la tribu borélienne engendrée par les ouverts.  $\mathcal{P}(\mathbb{N})$  et  $\mathcal{P}(\mathbb{R})$  sont les tribus les plus utilisées.

Si  $\mathcal{A}$  est une tribu sur  $\Omega$ , un ensemble mesurable est simplement un ensemble  $A \in \mathcal{P}(\Omega)$  qui est dans la tribu  $\mathcal{A}$ . Autrement dit,  $A \in \mathcal{A} \subset \mathcal{P}(\Omega)$ .

## Mesure de probabilité

Soit  $(\Omega, \mathcal{A})$  un espace mesurable. Une mesure de probabilité est une application  $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$  vérifiant  $\mathbb{P}(\Omega) = 1$  et si  $(A_n)_{n \in \mathbb{N}}$  est une suite de sous ensembles mesurables de  $\Omega$  deux à deux disjoints, alors  $\mathbb{P}(\bigcup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$ .

## Espace de probabilité

Un espace de probabilité, ou espace probabilisé  $(\Omega, \mathcal{A}, \mathbb{P})$  est la donnée d'un ensemble  $\Omega$  muni d'une tribu  $\mathcal{A}$  et d'une mesure de probabilité  $\mathbb{P}$ .

En pratique, les espaces les plus utilisés sont  $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ ,  $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$  et plus généralement  $(\mathbb{R}^d, \mathcal{P}(\mathbb{R}^d))$ ,  $d > 1$ , munis des probabilités associées.

La théorie de probabilité permet de fournir des modèles mathématiques pertinents à des phénomènes faisant intervenir du hasard et de l'aléa : lancer de dé, cours d'une action en bourse, durée de vie d'une ampoule, temps d'attente d'un bus, fluctuations de certaines quantités (variation des différentes mesures d'une grandeur physique autour de la valeur théorique attendue, variation de la température autour de la valeur moyenne attendue, etc).

On modélise un ensemble de situations possibles par un espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ . L'ensemble  $\Omega$  est l'espace des états, ou espace des résultats possibles et on appelle événement un ensemble  $A \in \mathcal{A}$ .  $\mathbb{P}(A)$  est la probabilité que l'événement  $A$  se réalise.

## Probabilité conditionnelle

Soit  $(\Omega, \mathcal{A}, \mathbb{P})$  un espace de probabilité.

Soit  $B \in \mathcal{A}$  un ensemble mesurable, tel que  $\mathbb{P}(B) > 0$ . La fonction  $\mathbb{P}(\cdot|B)$  définie par  $A \in \mathcal{A} \rightarrow \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$  est une probabilité sur  $(\Omega, \mathcal{A})$ , et la quantité  $\mathbb{P}(A|B)$  est appelé la probabilité conditionnelle de  $A$  sachant  $B$ .

$\mathbb{P}(A|B)$  représente la probabilité que  $A$  se réalise sachant que  $B$  se réalise.

Si  $A$  et  $B$  sont indépendants, i.e.,  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ , alors  $\mathbb{P}(A|B) = \mathbb{P}(A)$ .

## Probabilités composées

Soit  $(A_1, \dots, A_n) \in \mathcal{A}^n$  tel que  $\mathbb{P}(A_1 \cup \dots \cup A_n) > 0$ . Alors,

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cup A_2) \dots \mathbb{P}(A_n|A_1 \cup \dots \cup A_{n-1}).$$

## Probabilités totales

Une partition mesurable de  $\Omega$  est une famille d'ensembles  $(E_i)_{i \in I}$ , mesurables deux à deux disjoints tels que  $\bigcup_{i \in I} E_i = \Omega$ . On suppose que  $\mathbb{P}(E_i) > 0$  pour tout  $i \in I$ .

Soit  $(E_i)_{i \in I} \subset \mathcal{A}^I$  une partition mesurable finie ou dénombrable de  $\Omega$ . Alors, pour tout  $A \in \mathcal{A}$ , on a  $\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A \cap E_i) = \sum_{i \in I} \mathbb{P}(A|E_i)\mathbb{P}(E_i)$ .

## Formule de Bayes

Soit  $(E_i)_{i \in I} \subset \mathcal{A}^I$  une partition mesurable finie ou dénombrable de  $\Omega$ . Si  $A \in \mathcal{A}$  vérifie  $\mathbb{P}(A) > 0$ , alors

$$\mathbb{P}(E_i|A) = \frac{\mathbb{P}(A|E_i)\mathbb{P}(E_i)}{\sum_{j \in I} \mathbb{P}(A|E_j)\mathbb{P}(E_j)}, \quad i \in I.$$

## 1.3 Inférence bayésienne

Considérons un modèle paramétrique statistique pour lequel l'observation d'une v.a.  $X$  est distribuée selon  $f(x|\theta)$ , où seulement le paramètre  $\theta$  est inconnu et appartient à un espace de dimension finie  $\Theta$ .

### 1.3.1 Information a priori

L'information a priori est toute l'information disponible sur le paramètre  $\theta$ , en dehors de celle apportée par les observations. Cette information est modélisée au travers une loi de probabilité appelée "Loi a priori". On note la densité de probabilité de la loi a priori par  $\pi(\theta)$ .

### 1.3.2 Loi des observations

On interprète la loi des observations (la fonction vraisemblance) comme la loi conditionnelle des observations  $\underline{x}$  sachant  $\theta$ . On note sa densité par  $L(\underline{x}, \theta)$ . Elle est définie par :

$$L(\underline{x}, \theta) = \begin{cases} f_\theta(\underline{x}) = f(\underline{x}|\theta) = \prod_{i=1}^n f(x_i, \theta), & \text{cas continu;} \\ P_\theta(\underline{x}) = P(\underline{X} = \underline{x}|\theta) = \prod_{i=1}^n P(X_i = x_i, \theta), & \text{cas discret.} \end{cases}$$

### 1.3.3 Loi a posteriori

Dans la théorie bayésienne, l'incertitude sur  $\theta$  étant décrite par une distribution a priori  $\pi$  sur  $\Theta$ , l'inférence est alors fondée sur la distribution a posteriori donnée par la loi conditionnelle de  $\theta$  sachant  $\underline{x}$ . On note sa densité par  $\pi(\theta|\underline{x})$ , elle est définie par :

$$\pi(\theta|\underline{x}) = \frac{\varphi(\underline{x}, \theta)}{m(\underline{x})}.$$

Où,  $\varphi(\underline{x}, \theta)$  est la distribution conjointe de  $\underline{x}$  et  $\theta$

$$\Rightarrow \varphi(\underline{x}, \theta) = L(\underline{x}, \theta)\pi(\theta)$$

et  $m(\underline{x})$  la distribution marginale de  $\underline{X}$ ,

$$m(\underline{x}) = \int_{\Theta} \varphi(\underline{x}, \theta).$$

La densité de la loi a posteriori est alors donnée par

$$\pi(\theta|\underline{x}) = \frac{L(\underline{x}, \theta)\pi(\theta)}{\int_{\Theta} L(\underline{x}, \theta)\pi(\theta) d\theta}.$$

$\pi(\theta|\underline{x})$  est proportionnelle à la distribution de  $\underline{x}$  conditionnelle à  $\theta$ .

### 1.3.4 Modèle statistique bayésien

Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  avec  $f(x|\theta)$  densité de  $P_\theta$  et d'une distribution a priori  $\pi(\theta)$  sur le paramètre  $\theta$ . Le modèle est alors noté  $(\mathcal{X}, \mathcal{A}, P_\theta, \pi(\theta), \theta \in \Theta)$ .

L'analyse statistique bayésienne vise à exploiter le plus efficacement possible l'information apportée par  $X$  sur le paramètre  $\theta$ , pour ensuite construire des procédures d'inférence sur  $\theta$ . Le théorème de Bayes actualise l'information sur  $\theta$  en extrayant l'information contenue dans l'observation  $\underline{x}$  en calculant la loi a posteriori  $\pi(\theta)$ .

**Exemple 1.3.1.** Soit un  $n$ -échantillon  $X_1, X_2, \dots, X_n$  issu d'une v.a.  $X$  de loi de Poisson  $\mathcal{P}(\theta)$  de paramètre inconnu  $\theta, \theta \geq 0$ . Considérons la loi a priori de  $\theta$  de type Gamma  $\mathcal{G}(a, b)$ ,  $a > 0$  et  $b > 0$ .

La loi de  $X$  :

$$P(X = x|\theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad x \in \mathbb{N}, \theta > 0.$$

La loi a priori de  $\theta$  :

$$\pi(\theta) = \frac{b^a e^{-\theta b}}{\Gamma(a)} \theta^{a-1}, \quad \theta > 0$$

avec  $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx, a > 0$ .

La vraisemblance (loi de l'échantillon) :

$$L(\underline{x}, \theta) = \prod_{i=1}^n P(X_i = x_i, \theta) = \prod_{i=1}^n \left( e^{-\theta} \frac{\theta^{x_i}}{x_i!} \right) = \frac{e^{-n\theta} \theta^{\sum_i x_i}}{\prod_i x_i!}.$$

La loi a posteriori de  $\theta$  :

$$\pi(\theta|\underline{x}) = \frac{\varphi(\underline{x}, \theta)}{m(\underline{x})}.$$

Avec,

$$\begin{aligned} \varphi(\underline{x}, \theta) &= L(\underline{x}, \theta) \pi(\theta) \\ &= \left( \frac{e^{-n\theta} \theta^{\sum_i x_i}}{\prod_i x_i!} \right) \left( \frac{b^a e^{-\theta b}}{\Gamma(a)} \theta^{a-1} \right) \\ &= \frac{b^a \theta^{\sum_i x_i + a - 1}}{\Gamma(a) \prod_i x_i!} e^{-(n+b)\theta} \end{aligned}$$

et

$$m(\underline{x}) = \int_{\Theta} \varphi(\underline{x}, \theta) = \frac{b^a}{\Gamma(a) \prod_i x_i!} \int_0^{+\infty} \theta^{\sum_i x_i + a - 1} e^{-(n+b)\theta} d\theta.$$

En utilisant le changement de variable  $\alpha = (n+b)\theta$  on obtient,

$$\int_0^{+\infty} \theta^{\sum_i x_i + a - 1} e^{-(n+b)\theta} d\theta = \frac{1}{(n+b)^{\sum_i x_i + a}} \Gamma(\sum_i x_i + a).$$

D'où,

$$m(\underline{x}) = \frac{b^a \Gamma(\sum_i x_i + a)}{\Gamma(a) \prod_i x_i! (n+b)^{\sum_i x_i + a}}.$$

La loi a posteriori est alors donnée par

$$\pi(\theta|\underline{x}) = \frac{(n+b)\theta^{\sum_i x_i + a} e^{-\theta \sum_i x_i + a - 1}}{\Gamma(\sum_i x_i + a)} e^{-(n+b)\theta}.$$

$$\Rightarrow \theta|\underline{x} \sim \mathcal{G}\left(\sum_i x_i + a, n+b\right).$$

*Remarque 1.3.1.* Dans le cas d'une seule observation la loi a posteriori de  $\theta$  s'écrit :

$$\pi(\theta|x) = \frac{\varphi(x, \theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta},$$

où  $x$  est une observation de densité  $f(x|\theta)$ .

## 1.4 Raisonnement proportionnel

Soient  $f$  et  $g$  deux fonctions réelles définies sur le même espace  $\mathcal{X}$ .

On dit que  $f$  et  $g$  sont proportionnelles, et on note  $f \propto g$ , s'il existe une constante  $a > 0$  telle que

$$f(x) = a g(x), \quad \forall x \in \mathcal{X}.$$

### Résultats

1. La relation  $\propto$  est une relation d'équivalence. En particulier,

$$f \propto g \quad \text{et} \quad g \propto h \quad \text{alors} \quad f \propto h.$$

2. Soient  $X$  et  $Y$  deux v.a. et soient  $f$  et  $g$  les densités de  $X$  et  $Y$ , respectivement.

On suppose que la loi de probabilité de  $X$  est inconnue et celle de  $Y$  est connue et est notée par  $\mathbb{P}$ . Alors,

$$f \propto g \quad \Rightarrow \quad X \sim \mathbb{P}.$$

3. La proportionnalité dans le contexte bayésien est donnée par :

Pour  $\pi(\theta|\underline{x}) = \frac{\varphi(\underline{x}, \theta)}{m(\underline{x})} = \frac{L(\underline{x}, \theta)\pi(\theta)}{\int_{\Theta} L(\underline{x}, \theta)\pi(\theta) d\theta}$  et  $a = \frac{1}{m(\underline{x})}$ , indépendante de  $\theta$ , on a :

$$\pi(\theta|\underline{x}) \propto L(\underline{x}, \theta)\pi(\theta).$$

**Exemple 1.4.1.** Considérons le modèle bayésien de l'exemple 1.3.1, donné par

$$X|\theta \sim \mathcal{P}(\theta), \quad \theta > 0 \quad \text{et} \quad \theta \sim \mathcal{G}(a, b).$$

La loi a posteriori est telle que :

$$\begin{aligned} \pi(\theta|\underline{x}) &\propto L(\underline{x}, \theta)\pi(\theta) \\ &\propto \left( \frac{e^{-n\theta} \theta^{\sum_i x_i}}{\prod_i x_i!} \right) \left( \frac{b^a e^{-\theta b}}{\Gamma(a)} \theta^{a-1} \right) \\ &\propto \theta^{\sum_i x_i + a - 1} e^{-(n+b)\theta}. \end{aligned}$$

Par conséquent,

$$\theta|\underline{x} \sim \mathcal{G}\left(\sum_i x_i + a, n+b\right).$$

## 1.5 Approche décisionnelle

L'objectif des études inférentielles est de fournir une décision au statisticien. Les différentes décisions sont comparées au moyen d'un critère d'évaluation donné par la fonction **Coût**.

Soit  $\mathcal{D}$  l'ensemble des règles de décisions  $\delta$ , applications de  $\mathcal{X}$  dans  $\mathcal{A}$  (les estimateurs possibles), en général  $\mathcal{D} = \Theta$ . On note  $a \in \mathcal{A}$  une action, on a  $a = \delta(x)$  est une estimation.

L'inférence consiste à choisir une règle de décision  $\delta \in \mathcal{D}$  concernant  $\theta \in \Theta$  sur la base d'une observation  $x \in \mathcal{X}$ .

### 1.5.1 La fonction Coût (ou de Perte)

On appelle la fonction Coût (ou de Perte), toute fonction  $l$  de  $\Theta \times \mathcal{A}$  dans  $\mathbb{R}$ . C'est à dire,

$$\begin{aligned} l : \Theta \times \mathcal{A} &\rightarrow \mathbb{R} \\ (\theta, \delta(x)) &\mapsto l(\theta, \delta(x)). \end{aligned}$$

$l(\theta, \delta(x))$  évalue le coût (ou la perte ou encore la pénalité) associée à la décision  $a = \delta(x)$  quand le paramètre vaut  $\theta$ . Elle permet donc de quantifier la perte encourue par une mauvaise décision. Un coût négatif correspond alors à un gain.

#### Fonctions de coûts usuelles

- Coût quadratique :

$$l(\theta, \delta(x)) = (\theta - \delta(x))^2.$$

Ce coût pénalise trop fortement les grandes erreurs. C'est une fonction convexe.

- Coût absolu  $L^1$  :

$$l(\theta, \delta(x)) = |\theta - \delta(x)|.$$

Ou plus généralement une fonction linéaire par morceaux

$$l_{k_1, k_2}(\theta, \delta(x)) = \begin{cases} k_2(\theta - \delta(x)), & \theta > \delta(x) \\ k_1(\delta(x) - \theta), & \theta \leq \delta(x), \end{cases} \quad k_1, k_2 \text{ sont des constantes positives.}$$

Ces fonctions croissent plus lentement que le coût quadratique. Le coût absolu est une fonction convexe et ne pénalisent pas les grandes erreurs mais peu vraisemblables.

- Coût 0 – 1 :

C'est un coût non négatif, très utilisé dans la théorie des tests. Il associe à un estimateur  $\delta$  la pénalité 0 si la réponse est correcte et 1 sinon.

$$l(\theta, \delta(x)) = 1_{\delta(x)=0} 1_{\theta \in \Theta_1} + 1_{\delta(x)=1} 1_{\theta \in \Theta_0}.$$

$$\text{Où } \Theta = \Theta_1 \cup \Theta_0 \text{ et } \delta(x) = \begin{cases} 1, & \text{si } \theta \in \Theta_0 \\ 0, & \text{si } \theta \in \Theta_1. \end{cases}$$

Ce coût n'est pas très intéressant de par son caractère non quantitatif.

### 1.5.2 Risque fréquentiste (coût moyen)

On appelle risque fréquentiste, et on note  $R(\theta, \delta)$ , le coût moyen du coût d'une règle de décision, i.e.

$$R(\theta, \delta) = \mathbb{E}_\theta [l(\theta, \delta(x))] = \int_{\mathcal{X}} l(\theta, \delta(\underline{x})) L(\theta, \underline{x}) d\underline{x}.$$

(C'est une fonction de  $\theta$ ).  $\delta$  est appelé estimateur tandis que  $\delta(\underline{x})$  estimation de  $\theta$ .

### Cas particulier :

Dans le cas d'une perte quadratique  $l(\theta, \delta(x)) = (\theta - \delta(x))^2$ , on définit le risque quadratique par :

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[(\theta - \delta(x))^2] \\ &= \theta^2 - 2\theta\mathbb{E}(\delta(x)) + (\mathbb{E}(\delta(x)))^2 \\ &= [(\mathbb{E}(\delta(x)) - \theta)^2] + [\mathbb{E}(\delta(x)^2) - (\mathbb{E}(\delta(x)))^2] \\ &= (\text{biais})^2 + \text{Var}(\delta(x)). \end{aligned}$$

C'est le risque le plus utilisé en pratique.

### 1.5.3 Risque a posteriori

On appelle le risque a posteriori, noté  $\rho(\pi, \delta|\underline{x})$ , la moyenne du coût par rapport à la loi a posteriori.

$$\rho(\pi, \delta|\underline{x}) = \mathbb{E}^{\pi(\cdot|\underline{x})}(l(\theta, \delta(x))|\underline{x}) = \int_{\Theta} l(\theta, \delta(\underline{x})) \pi(\theta|\underline{x}) d\theta.$$

(C'est une fonction de  $\underline{x}$ ).

### 1.5.4 Risque Intégré

On définit le risque intégré ou risque fréquentiste moyenné sur les valeurs de  $\theta$  selon leurs distribution a priori  $\pi$ , noté  $r(\pi, \delta)$ , par :

$$r(\pi, \delta) = \mathbb{E}^{\pi(\cdot)}(R(\theta, \delta)) = \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta.$$

(C'est un nombre réel).

**Théorème 1.5.1.** *Le risque intégré est égal à la moyenne du risque a posteriori suivant la loi marginale de  $\underline{x}$ , i.e.,*

$$r(\pi, \delta) = \mathbb{E}[\rho(\pi, \delta|\underline{x})] = \int_{\mathcal{X}} \rho(\pi, \delta|\underline{x}) m(\underline{x}) d\underline{x}.$$

*Remarque 1.5.1.* Dans l'approche classique la règle de décision optimale  $\delta$  doit minimiser le risque fréquentiste  $R(\theta, \delta)$  pour tout  $\theta \in \Theta$ .

On définit alors les propriétés données ci-dessous.

### 1.5.5 Risque de Bayes

Un estimateur bayésien associé à une distribution a priori  $\pi$  et une fonction de coût  $l$  est un estimateur  $\delta^\pi$  minimisant le risque intégré  $r(\pi, \delta)$ .

Le risque bayésien (ou risque de bayes) est alors la quantité donnée par

$$r(\pi) = r(\pi, \delta^\pi).$$

**Théorème 1.5.2.** *S'il existe  $\delta \in \mathcal{D}$  tel que  $r(\pi, \delta) < \infty$  alors l'estimateur*

$$\delta^\pi(\underline{x}) = \arg \min_{\delta} \rho(\pi, \delta|\underline{x}), \quad \underline{x} \in \mathcal{X}$$

*est un estimateur Bayésien.*

Par conséquent, un estimateur  $\delta$  minimisant le risque intégré est obtenu en minimisant le coût moyen a posteriori.

En effet,

A partir du théorème de Fubini, en utilisant que  $l(\theta, \delta(x)) \geq 0$ ,

$$\begin{aligned}
 r(\pi, \delta) &= \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta \\
 &= \int_{\Theta} \int_{\mathcal{X}} l(\theta, \delta(x)) L(\theta, \underline{x}) \pi(\theta) d\underline{x} d\theta \\
 &= \int_{\mathcal{X}} \left( \int_{\Theta} l(\theta, \delta(x)) L(\theta, \underline{x}) \pi(\theta) d\theta \right) d\underline{x} \\
 &= \int_{\mathcal{X}} \left( \int_{\Theta} l(\theta, \delta(x)) \pi(\theta | \underline{x}) d\theta \right) m(\underline{x}) d\underline{x} \\
 &= \int_{\mathcal{X}} \rho(\pi, \delta | \underline{x}) m(\underline{x}) d\underline{x}.
 \end{aligned}$$

Pour  $\delta \in \mathcal{D}$ ,  $\rho(\pi, \delta^\pi | \underline{x}) \leq \rho(\pi, \delta | \underline{x})$ . D'où,

$$r(\pi, \delta^\pi) \leq r(\pi, \delta).$$

L'estimateur de bayes  $\delta^\pi$  est un estimateur ayant un risque bayésien  $r(\pi, \delta^\pi)$  minimal fini. i.e.,

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

### 1.5.6 Risque minimax

On appelle risque minimax associé à la fonction de coût  $l$ , la valeur

$$\bar{R} = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} \mathbb{E}_\theta [l(\theta, \delta(\underline{x}))].$$

Un estimateur minimax est tout estimateur  $\delta_0$  tel que

$$\sup_{\theta} R(\theta, \delta_0) = \bar{R}.$$

**Théorème 1.5.3.** [19],

*Le risque de bayes est toujours plus petit que le risque minimax*

C'est à dire,

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \bar{R} = \inf_{\delta \in \mathcal{D}} \sup_{\theta} R(\theta, \delta).$$

Les résultats suivants fournissent des conditions suffisantes de minimaxité (voir Robert C.P. 2006).

**Lemme 1.5.1.** *Si  $\delta^\pi$  est un estimateur de bayes pour  $\pi$  et si  $R(\theta, \delta^\pi) \leq r(\pi), \forall \theta$  dans le support de  $\pi$ , l'estimateur  $\delta^\pi$  est minimax et  $\pi$  est la distribution la moins favorable.*

**Exemple 1.5.1.** (Berger J. 1985) Soit le modèle  $X \sim \mathcal{B}(n, \theta)$  et  $\theta \sim \mathcal{B}e(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$ .

L'estimateur de bayes défini par  $\delta^*(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}$  (sous le coût quadratique), a un risque constant par rapport à  $\theta$  donné par  $R(\theta, \delta^*) = \frac{1}{4(1 + \sqrt{n})^2}$ . En intégrant sur  $\theta$ ,  $r(\pi) = R(\theta, \delta^*)$ . Par conséquent,  $\delta^*$  est minimax et  $\pi$  est la distribution la moins défavorable.

*Remarque 1.5.2.* Les estimateurs minimax correspondent généralement à des estimateurs ayant un risque de bayes infini. On doit souvent recourir à un argument limite pour établir la minimaxité.

**Lemme 1.5.2.** *S'il existe une suite  $(\pi_n)$  de lois a priori propres telles que l'estimateur de bayes  $\delta^*$  satisfasse*

$$R(\theta, \delta^*) \leq \lim_{n \rightarrow +\infty} r(\pi_n) < +\infty, \forall \theta \in \Theta,$$

*alors  $\delta^*$  est minimax.*

Le principe minimax n'est pas nécessairement attirant d'un point de vue bayésien : les lois a priori les moins favorables sont souvent irréalistes, il est souvent difficile de la construire et plusieurs estimateurs minimax peuvent exister simultanément. Il est alors nécessaire de présenter un deuxième critère, plus local, pour comparer les estimateurs minimax.

### 1.5.7 Admissibilité

Un estimateur  $\delta_0$  est dit inadmissible s'il existe un estimateur  $\delta_1$  qui domine  $\delta_0$ , i.e.,

$$\forall \theta, R(\theta, \delta_0) \geq R(\theta, \delta_1) \text{ et pour au moins une valeur } \delta_0 \text{ de } \theta, R(\theta_0, \delta_0) > R(\theta_0, \delta_1).$$

Sinon,  $\delta_0$  est dit admissible.

Les résultats suivants fournissent des conditions suffisantes d'admissibilité [19].

**Proposition 1.5.1.** *S'il existe un unique estimateur minimax, cet estimateur est admissible.*

**Proposition 1.5.2.** *Si l'estimateur  $\delta_0$  est admissible et de risque constant alors  $\delta_0$  est l'unique estimateur minimax.*

**Proposition 1.5.3.** *Si la distribution a priori  $\pi$  est strictement positive sur  $\Theta$ , de risque de bayes fini et  $R(\theta, \delta)$  est continu de  $\theta$  pour tout  $\delta$ , alors l'estimateur de bayes  $\delta^\pi$  est admissible.*

**Proposition 1.5.4.** *Si l'estimateur de bayes associé à une loi a priori  $\pi$  est unique, alors il est admissible.*

**Proposition 1.5.5.** *Si un estimateur de bayes  $\delta^\pi$ , associé à une loi a priori  $\pi$  (propre ou impropre), est tel que le risque de bayes soit fini alors il est admissible.*

*Remarque 1.5.3.* Si la fonction de coût est strictement convexe (par exemple, le coût quadratique), alors l'estimateur de bayes est unique et donc admissible.

**Exemple 1.5.2.** Suite de l'exemple 1.5.1.

L'estimateur  $\delta^* = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}$  est un estimateur de bayes ayant un risque de bayes fini. Il est donc admissible. Par conséquent, il est l'unique estimateur minimax (sous le coût quadratique).

## Estimation Bayésienne

Les estimateurs bayésiens du paramètre  $\theta$  sont construits à partir de la loi a posteriori  $\pi(\theta|\underline{x})$  par minimisation d'une fonction de coût appropriée.

### 2.1 Estimateurs de Bayes : cas unidimensionnel

Soit  $\underline{X} = (X_1, X_2, \dots, X_n)$  un  $n$ -échantillon issu d'une v.a.  $X$  de loi dépendant d'un paramètre inconnu  $\theta, \theta \in \mathbb{R}$ .

#### 2.1.1 Estimateur MMSE

L'estimateur MMSE (Minimum Mean Square Error) de  $\theta$ , noté  $\hat{\theta}_{MMSE}(\underline{x})$  est l'estimateur qui minimise le coût quadratique moyen

$$R(\theta, \hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2], \text{ avec } \hat{\theta} = \hat{\theta}(\underline{x}).$$

Il est défini par :

$$\hat{\theta}_{MMSE}(\underline{x}) = \mathbb{E}(\theta|\underline{x}), \forall \underline{x}.$$

L'estimateur  $MMSE$  est donc la moyenne a posteriori de  $\theta$ .

**En effet :**

$$R(\theta, \hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E} \left[ \mathbb{E}[(\hat{\theta} - \theta)^2|\underline{x}] \right] = \int_{\mathcal{X}} \mathbb{E} \left[ (\hat{\theta} - \theta)^2|\underline{x} \right] m(\underline{x}) d\underline{x}.$$

Puisque  $m(\underline{x}) \geq 0$ , l'estimateur qui minimise  $\mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right]$  est tel que  $\hat{\theta}$  minimise  $\mathbb{E} \left[ (\hat{\theta} - \theta)^2|\underline{x} \right]$  pour tout  $\underline{x}$ .

On a

$$\begin{aligned} \mathbb{E} \left[ (\hat{\theta} - \theta)^2|\underline{x} \right] &= \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}(\theta|\underline{x}) + \mathbb{E}(\theta^2|\underline{x}) \\ &= \left( \hat{\theta} - \mathbb{E}(\theta|\underline{x}) \right)^2 + \mathbb{E}(\theta^2|\underline{x}) - \mathbb{E}^2(\theta|\underline{x}) \\ &= \left( \hat{\theta} - \mathbb{E}(\theta|\underline{x}) \right)^2 + \text{Var}(\theta|\underline{x}) \geq \text{Var}(\theta|\underline{x}). \end{aligned}$$

L'égalité est atteinte pour

$$\hat{\theta}(\underline{x}) = \mathbb{E}(\theta|\underline{x}).$$

L'estimateur de bayes du paramètre  $\theta$ , obtenu relativement à la perte quadratique, est alors la moyenne a posteriori  $\theta_{MMSE}$ .

**Exemple 2.1.1.** Considérons le modèle bayésien de l'exemple 1.3.1 donné par :

$$\begin{cases} X|\theta \sim \mathcal{P}(\theta), \theta > 0 \\ \theta \sim \mathcal{G}(a, b), a, b > 0. \end{cases}$$

On a,

$$\theta|\underline{x} \sim \mathcal{G}\left(\sum_i x_i + a; n + b\right).$$

L'estimateur de bayes de  $\theta$ , relativement à la perte quadratique, est donné par :

$$\hat{\theta}_{MMSE}(\underline{x}) = \mathbb{E}(\theta|\underline{x}) = \frac{\sum_i X_i + a}{n + b} = \frac{n\bar{X} + a}{n + b}.$$

*Remarque 2.1.1.*

- Le MMSE est un estimateur admissible. Il n'est minimax que si la loi a priori  $\pi$  est bien choisie.
- L'estimation bayésienne d'une fonction de  $\theta$ ,  $h(\theta) \in \mathbb{R}$ , relativement au coût quadratique, notée  $\hat{h}_B(\theta)$ , est donnée par  $\hat{h}_B(\theta) = \mathbb{E}(h(\theta)|x)$ .
- Les MMSE sont asymptotiquement confondus avec ceux du maximum de vraisemblance. La différence est très petite devant  $1/\sqrt{n}$ , (voir Borovkov A. 1987, Th.36.9)

## 2.1.2 Estimateur MAP

L'estimateur MAP (Maximum A Posteriori) de  $\theta$  se fait par maximisation de la loi a posteriori. C'est à dire qu'on choisit le mode de la probabilité a posteriori  $\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{m(\underline{x})}$ . L'estimateur MAP est alors

$$\hat{\theta}_{MAP}(\underline{x}) = \arg \max_{\theta} \pi(\theta|\underline{x}) = \arg \max_{\theta} \pi(\underline{x}|\theta)\pi(\theta).$$

Cet estimateur est associé au coût 0 – 1. Il peut s'exprimer comme un estimateur du maximum de vraisemblance pénalisé au sens classique.

**Exemple 2.1.2.** [19]

Pour  $X \sim \mathcal{B}(n, \theta)$ ,  $\theta \in (0, 1)$  on considère trois lois a priori  $\pi_0(\theta) = \frac{\theta^{-1/2}(1-\theta)^{-1/2}}{B(1/2, 1/2)}$  (loi béta  $\mathcal{B}e(1/2, 1/2)$ ),  $\pi_1(\theta) = 1$  (loi de Laplace) et  $\pi_2(\theta) = \theta^{-1}(1 - \theta)^{-1}$  (loi de Haldane).

Les estimateurs MAP associés, pour  $n > 2$ , sont

$$\delta_0(x) = \max\left(\frac{x - 1/2}{n - 1}, 0\right), \delta_1(x) = \frac{x}{n} \text{ et } \delta_2(x) = \max\left(\frac{x - 1}{n - 2}, 0\right).$$

On note que  $\delta_1(x) = \frac{x}{n}$  est le MLE de  $\theta$ . Lorsque  $n$  est grand les estimateurs sont équivalents.

*Remarque 2.1.2.*

- Le coût  $0 - 1$  est utilisé dans l'approche classique des tests d'hypothèses, proposée par Neyman-Pearson. C'est un exemple typique d'un coût non quantitatif. Il associe à un estimateur  $\delta$  la pénalité 0 si la réponse est correcte et 1 sinon.

Pour la fonction de coût  $0 - 1$  qui vaut

$$L(\theta, \delta) = \begin{cases} 1 - \delta & \text{si } \theta \in \Theta_0 \\ \delta & \text{sinon,} \end{cases}$$

$$\text{le risque associé est } R(\theta, \delta) = \begin{cases} P(\delta(x) = 0) & \text{si } \theta \in \Theta_0 \\ P(\delta(x) = 1) & \text{sinon,} \end{cases}$$

(on retrouve les erreurs de première et deuxième espèce, voir § 2.6).

L'estimateur de bayes associé à ce coût est alors (voir [19]),

$$\delta^\pi(x) = \begin{cases} 1 & \text{si } P(\theta \in \Theta_0|x) > P(\theta \notin \Theta_0|x) \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent,

$$\delta^\pi(x) = 1 \text{ si et seulement si } P(\theta \in \Theta_0|x) > 1/2.$$

- Les estimateurs MAP sont asymptotiquement équivalents aux estimateurs du maximum de vraisemblance classiques et ont l'avantage d'être disponibles pour des tailles finies d'échantillons.
- Le MAP est utilisé quand le MMSE est difficile à mettre en oeuvre en raison des calculs d'intégrales compliqués qu'il impose. Il n'est pas particulièrement recommandé mais se comporte symptomatiquement comme le maximum de vraisemblance puisque  $\pi(\theta)$  a un poids qui diminue lorsque  $n$  tend vers l'infini.
- Pour les petits échantillons, il est plus habituel de s'intéresser plus précisément à la probabilité a posteriori, le MAP étant considéré comme une information trop succincte [6].

## Autres estimateurs bayésiens

- En minimisant la fonction de coût  $L^1$ , on obtient la médiane de la loi a posteriori.
- Pour une fonction de coût linéaire par morceaux

$$l_{k_1, k_2}(\theta, \delta(x)) = \begin{cases} k_2(\theta - \delta(x)), & \theta > \delta(x) \\ k_1(\delta(x) - \theta), & \text{sinon} \end{cases}$$

l'estimateur de bayes est le fractile  $k_2/(k_1 + k_2)$  de  $\pi(\theta|x)$ .

*Remarque 2.1.3.*

On utilise le mode de la loi a posteriori si on suspecte que la vraisemblance est multimodale (ayant des maxima locaux). On retient la moyenne et le mode de la loi a posteriori comme des estimateurs bayésiens. Le plus souvent on utilise la moyenne a posteriori.

### 2.1.3 Propriétés des estimateurs de bayes

Les estimateurs de bayes sont toujours biaisés. Sous certaines hypothèses de régularité, souvent satisfaites en pratique, ils sont convergents en probabilité (lorsque  $n$  tend vers l'infini). La loi a posteriori peut être asymptotiquement approximée par une loi Normale  $\mathcal{N}(\mathbb{E}(\theta|x), \text{Var}(\theta|x))$  où  $\text{Var}(\theta|x) = \mathbb{E}((\theta - \mathbb{E}(\theta|x))^2|x)$  est la variance a posteriori de  $\theta$ . Cette propriété est utile pour construire les intervalles de confiance a posteriori.

## 2.2 Estimateurs de Bayes : cas multidimensionnel

Soit  $\theta = (\theta_1, \theta_2, \dots, \theta_p) \in \mathbb{R}^p, p > 1$  le paramètre à estimer.

L'estimateur de Bayes de  $\theta$ , relativement à la perte quadratique, est la moyenne a posteriori définie par :

$$\hat{\theta}_{MMSE}(\underline{x}) = \mathbb{E}(\theta|\underline{x}) = (\mathbb{E}(\theta_1|\underline{x}), \dots, \mathbb{E}(\theta_p|\underline{x}))$$

où

$$\mathbb{E}(\theta_j|\underline{x}) = \int_{\Theta_j} \theta_j \pi(\theta_j|\underline{x}) d\theta_j, \quad j = 1, 2, \dots, p,$$

avec,  $\Theta = \Theta_1 \times \Theta_2 \times \dots \times \Theta_p$  et  $\pi(\theta_j|\underline{x}) = \int_{\dots \Theta \setminus \Theta_j} \pi(\theta|\underline{x}) d\theta_1 \dots d\theta_p$  obtenu en intégrant  $\pi(\theta|\underline{x})$  sur toutes les composantes de  $\theta$  autres que  $\theta_j, j = 1, \dots, p$ .

*Remarque 2.2.1.* Le plus souvent les estimateurs de Bayes des  $\theta_j$  ne peuvent pas être calculés de façon explicite. Il faut parfois faire appel aux méthodes de simulation de Monte-carlo.

## 2.3 Probabilité prédictive

Dans le paradigme bayésien, la probabilité d'une observation future  $x^*$  sera estimée par la probabilité prédictive :

$$P(x^*|\underline{x}) = \int_{\Theta} P(x^*|\theta)P(\theta|\underline{x}) d\theta, \text{ où } \underline{x} = (x_1, x_2, \dots, x_n),$$

avec

$$P(\theta|\underline{x}) = \frac{P(x_n|\theta)P(\theta|x_1, x_2, \dots, x_{n-1})}{P(x_n|x_1, x_2, \dots, x_{n-1})} \propto P(x_n|\theta)P(\theta|x_1, x_2, \dots, x_{n-1}) \frac{P(x_1, x_2, \dots, x_{n-1})}{P(x_1, x_2, \dots, x_n)}$$

et

$$P(x_1, x_2, \dots, x_n) = \int_{\Theta} \prod_{i=1}^n P(x_i|\theta)P(\theta) d\theta.$$

Dans ce cas, nous avons

- nouvel a priori :  $P(\theta|x_1, x_2, \dots, x_{n-1})$
- vraisemblance :  $P(x_n|\theta)$
- nouvel a posteriori :  $P(\theta|x_1, x_2, \dots, x_n)$ .

## 2.4 Extension aux lois impropres

L'approche bayésienne peut être généralisée aux lois impropres, ce qui est notamment utile dans les modèles non-informatifs (voir § 3.3.5).

### 2.4.1 Loi a priori impropre

Soit  $\pi(\theta)$  une application de  $\Theta$  dans  $]0, +\infty[$  telle que

$$\int_{\Theta} \pi(\theta) d\theta = +\infty.$$

On appelle  $\pi(\theta)$  la loi a priori impropre.

*Remarque 2.4.1.* Dans le cas d'une loi a priori  $\pi(\theta)$  impropre, la fonction  $\pi(\theta)$  n'est pas une densité de probabilité.

### 2.4.2 Estimateur de bayes généralisé

On se donne une loi a priori impropre caractérisée par la fonction  $\pi(\theta)$ . On suppose que l'intégrale  $\int_{\Theta} L(\theta, \underline{x})\pi(\theta) d\theta$  est convergente. On considère la densité de la loi a posteriori définie par :

$$\pi(\theta|\underline{x}) = \frac{L(\theta, \underline{x})\pi(\theta)}{\int_{\Theta} L(\theta, \underline{x})\pi(\theta) d\theta}.$$

**Définition 2.4.1.** On appelle estimateur de Bayes généralisé de  $\theta$ , relativement au coût quadratique, la moyenne de la loi a posteriori  $\pi(\theta|\underline{x})$ .

**Exemple 2.4.1.** Considérons le modèle bayésien suivant :

$$\begin{cases} X|\lambda \sim \text{Exp}(\lambda), \lambda > 0 \\ \lambda \sim \mathcal{U}(\mathbb{R}^+). \end{cases}$$

La loi a priori est

$$\pi(\lambda) = 1_{\mathbb{R}^+}(\lambda) = 1, \text{ si } \lambda > 0.$$

La vraisemblance de l'échantillon :

$$L(\underline{x}, \lambda) = \prod_{i=1}^n f(x_i|\lambda) = \lambda^n e^{-\lambda \sum_i x_i}.$$

La loi a posteriori est alors donnée par :

$$\begin{aligned} \pi(\lambda|\underline{x}) &= \frac{\lambda^n e^{-\lambda \sum_i x_i}}{\int_0^{+\infty} \lambda^n e^{-\lambda \sum_i x_i} d\lambda} 1_{\mathbb{R}^+}(\lambda) \\ &= \frac{\lambda^n e^{-\lambda \sum_i x_i}}{\frac{1}{(\sum_i x_i)^{n+1}} \int_0^{+\infty} \alpha^n e^{-\alpha} d\alpha} \\ &= \frac{(\sum_i x_i)^{n+1} \lambda^n e^{-\lambda \sum_i x_i}}{\Gamma(n+1)}. \end{aligned}$$

D'où,

$$\pi(\lambda|\underline{x}) \rightarrow \mathcal{G}(n+1; \sum_i x_i)$$

L'estimateur de bayes généralisé de  $\lambda$  est alors donné par :

$$\hat{\lambda}_G = \mathbb{E}(\lambda|\underline{x}) = \frac{n+1}{\sum_i X_i} = \frac{n+1}{n\bar{X}}.$$

*Remarque 2.4.2.* L'inférence bayésienne étant basée sur  $\pi(\theta|\underline{x})$ , l'usage des lois impropres a priori est justifié si la loi a posteriori est propre. Dans ce cas, l'estimateur bayésien est bien défini.

## 2.5 Intervalles de confiance bayésiens

L'approche bayésienne présente l'avantage de permettre une construction directe d'une région de confiance, contrairement à l'approche classique qui nécessite une stratégie particulière comme par exemple l'utilisation d'une fonction pivotale.

**Définition 2.5.1.** On se donne un modèle bayésien  $(\mathcal{X}, \mathcal{A}, P_\theta, \pi(\theta), \theta \in \Theta)$  et on suppose que le paramètre  $\theta$  à estimer est un réel.

On appelle région de confiance de niveau  $\alpha$  ou région  $\alpha$ -crédible tout ensemble  $C_x$  tel que :

$$P^{\pi(\cdot|x)}(\theta \in C_x) \geq 1 - \alpha,$$

où  $P^{\pi(\cdot|x)}$  est la loi de probabilité dont la densité a posteriori est  $\pi(\cdot|x)$ .

On dira qu'une région est  $\alpha$ -crédible de plus forte densité a posteriori, et on note PFDP (ou encore HPD, Highest Posterior Density), si elle s'écrit :

$$C_x^\pi = \{\theta; \pi(\theta|x) \geq k_\alpha\}$$

où  $k_\alpha$  est la plus grande valeur telle que

$$P^{\pi(\cdot|x)}(\theta \in C_x^\pi) \geq 1 - \alpha.$$

On obtient que si  $C_x$  est une région PFDP alors elle est  $\alpha$ -crédible.

### 2.5.1 Intervalle de confiance a priori

On appelle un intervalle de confiance a priori  $J$ , de niveau  $1 - \alpha$ , l'intervalle pour lequel on a

$$P(\theta \in J) = \int_J \pi(\theta) = 1 - \alpha,$$

où  $\alpha$  est une probabilité fixée dans  $]0, 1[$ .

### 2.5.2 Intervalle de confiance a posteriori

On appelle un intervalle de confiance a posteriori  $I$ , de niveau  $1 - \alpha$ , l'intervalle pour lequel on a

$$P(\theta \in I|X) = \int_I \pi(\theta|X) = 1 - \alpha,$$

où  $\alpha$  est une probabilité fixée dans  $]0, 1[$ .

**Exemple 2.5.1.** Considérons dans cet exemple une observation  $x$  de loi normale  $\mathcal{N}(\theta, 1)$  et une loi a priori pour  $\theta$  donnée par la loi normale  $\mathcal{N}(0, 10)$ .

On obtient facilement la loi a posteriori

$$\pi(\theta|x) = \frac{1}{\sqrt{\frac{20\pi}{11}}} \exp\left\{-\frac{(\theta - \frac{10x}{11})^2}{\frac{20}{11}}\right\}, \theta \in \mathbb{R}, x \in \mathbb{R}.$$

La région  $\alpha$ -crédible FPDP est de la forme

$$C_x^\pi = \{\theta, \pi(\theta|x) \geq k_\alpha\} = \{\theta, |\theta - \frac{10}{11}x| \leq K_\alpha\}.$$

En effet,

$$\begin{aligned}
C_x^\pi &= \left\{ \theta, \pi(\theta|x) \geq k_\alpha \right\} = \left\{ \theta, \frac{1}{\sqrt{\frac{20\pi}{11}}} \exp\left\{-\frac{(\theta - \frac{10x}{11})^2}{\frac{20}{11}}\right\} \geq k_\alpha \right\} \\
&= \left\{ \theta, \exp\left\{-\frac{(\theta - \frac{10x}{11})^2}{\frac{20}{11}}\right\} \geq k'_\alpha \right\}, \quad (k'_\alpha = k_\alpha \sqrt{\frac{20\pi}{11}}) \\
&= \left\{ \theta, \left(\theta - \frac{10x}{11}\right)^2 \leq k''_\alpha \right\}, \quad (k''_\alpha = -\frac{20}{11} \log k'_\alpha) \\
&= \left\{ \theta, \left| \theta - \frac{10x}{11} \right| \leq K_\alpha \right\}, \quad (K_\alpha = \sqrt{k''_\alpha}).
\end{aligned}$$

Où  $K_\alpha = q_{1-\alpha/2}$  est le quantile d'ordre  $1 - \alpha/2$  de la loi normale centrée réduite. On obtient donc,

$$C_x^\pi = \left\{ \theta, \frac{10}{11}x - q_{1-\alpha/2} \leq \theta \leq \frac{10}{11}x + q_{1-\alpha/2} \right\}.$$

## 2.6 Approche bayésienne des tests

On suppose que l'ensemble  $\Theta$  des paramètres est partitionné en  $\Theta_0$  et  $\Theta_1$  et que  $P(\theta \in \Theta_0) > 0$  et  $P(\theta \in \Theta_1) > 0$ .

On veut tester :

$$H_0 : \theta \in \Theta_0 \text{ contre } H_1 : \theta \in \Theta_1.$$

En statistique bayésienne, la réponse à un tel test repose sur les probabilités a posteriori de  $H_0$  et de  $H_1$ , i.e.

$$\begin{cases} P(H_0|X) = P(\theta \in \Theta_0|X) = \int_{\Theta_0} \pi(\theta|X) d\theta = \alpha_0 \\ \text{et} \\ P(H_1|X) = 1 - P(\theta \in \Theta_0|X) = P(\theta \in \Theta_1|X) = \alpha_1. \end{cases}$$

### 2.6.1 Critère de décision

Par définition, les décisions bayésiennes sont celles qui minimisent le coût a posteriori

$$\rho(\pi, \delta|x) = \mathbb{E}^{\pi(\cdot|x)}(l(\theta, \delta(x))|x) = \int_{\Theta} l(\theta, \delta(x)) \pi(\theta|x) d\theta.$$

On a deux décisions possibles :  $\delta \in \{d_0, d_1\}$ , avec

$$\begin{aligned}
d_0 &: \text{ On ne rejette pas } H_0 : \theta \in \Theta_0, \\
d_1 &: \text{ On rejette } H_0.
\end{aligned}$$

À ces décisions on associe des coûts  $l(\theta, d_0(x))$  et  $l(\theta, d_1(x))$ .

**1.** En pratique, on accepte l'hypothèse  $H_0$  ou  $H_1$  dès que sa probabilité a posteriori est suffisamment forte ( $\geq 0.9$  ou  $0.95$ ).

Si aucune des probabilités a posteriori n'excède  $0.9$ , deux attitudes sont alors possibles :

- Soit on ne prend pas de décision et on choisit de recueillir davantage d'observations ;

– Soit on choisit l'hypothèse dont la probabilité a posteriori est la plus élevée.

2. En considérant une fonction de coût 0 – 1,

$$l(\theta, d_i(x)) = 1_{\{\theta \notin \Theta_i\}} = \begin{cases} 1, & \text{si } \theta \notin \Theta_i, i = 0, 1 \\ 0, & \text{sinon.} \end{cases}$$

D'où,

$$\begin{aligned} \rho(\pi, d_i|x) &= \mathbb{E}^{\pi(\cdot|x)}(l(\theta, d_i)|x) = \int_{\Theta_j} \pi(\theta|x) d\theta. \\ &= P(\theta \notin \Theta_i|x) = 1 - P(H_i|x) = 1 - \alpha_i, i = 0, 1. \end{aligned}$$

Si  $\alpha_0 < \alpha_1$ , i.e.  $P(\alpha_0|x) < P(\alpha_1|x)$  alors on rejette  $H_0 : \theta \in \Theta_0$ , et on décide de prendre la décision  $\delta = d_1$  (la décision bayésienne est l'hypothèse ayant la plus grande probabilité a posteriori).

3. Considérons une fonction de coût de la forme

$$l(\theta, d_i) = K_i 1_{\{\theta \in \Theta_j\}} = \begin{cases} K_i, & \text{si } \theta \notin \Theta_i \\ 0, & \text{si } \theta \in \Theta_i, i = 0, 1. \end{cases}$$

D'où,  $\rho(\pi, d_i|x) = K_i(1 - \alpha_i)$ ,  $K_i > 0$ ,  $i = 0, 1$ .

La décision bayésienne est alors obtenue en comparant  $K_0\alpha_1$  avec  $K_1\alpha_0$ .

Si  $K_1\alpha_0 < K_0\alpha_1$  alors  $\frac{K_0}{K_1} > \frac{\alpha_0}{\alpha_1}$ . On rejette dans ce cas  $H_0$  et on prend la décision  $\delta = d_1$ . On obtient donc  $\alpha_1 > \frac{K_1}{K_0+K_1}$ , (car  $\alpha_1 + \alpha_2 = 1$ ).

D'où, la région critique du test bayésien

$$C_x = \left\{ x, \alpha_1 > \frac{K_1}{K_0 + K_1} \right\}.$$

**Exemple 2.6.1.** Soit  $X$  une v.a. de loi normale  $\mathcal{N}(\theta, \sigma^2)$  avec  $\sigma$  connu et  $\theta$  le paramètre à estimer.

Il s'agit de tester  $H_0 : \theta \geq \theta_0$  contre  $H_1 : \theta < \theta_0$ .

On considère une loi a priori  $\pi(\theta) = \mathcal{N}(\mu, \tau^2)$  pour  $\theta$ . La loi a posteriori est alors une loi normale  $\mathcal{N}(\mu(x), \rho^{-1})$  avec  $\mu(x) = \frac{\sigma^2\mu + \tau^2x}{\sigma^2 + \tau^2}$  et  $\rho = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$ . On peut écrire  $\mu(x) = \rho\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right)$ .

Dans ce cas, on calcule  $\alpha_1$  :

$$\alpha_1 = P(\Theta_1|x) = \left(\frac{\rho}{\sqrt{2\pi}}\right)^{1/2} \int_{-\infty}^{\theta_0} e^{-\rho(\theta - \mu(x))^2/2} d\theta = \left(\frac{1}{\sqrt{2\pi}}\right) \int_{-\infty}^{\sqrt{\rho}(\theta_0 - \mu(x))} e^{-\eta^2/2} d\eta$$

où,  $\eta = \sqrt{\rho}(\theta - \mu(x))$ . D'où  $\alpha_1 = P(\theta_1|x) = \Phi(\sqrt{\rho}(\theta_0 - \mu(x)))$ .

La région critique est alors donnée par

$$\begin{aligned} C_x &= \left\{ x, \Phi(\sqrt{\rho}(\theta_0 - \mu(x))) > \frac{K_1}{K_0 + K_1} \right\} \\ &= \left\{ x, x < \sigma^2 \left[ \frac{1}{\rho} \left( \theta_0 - \frac{1}{\sqrt{\rho}} \Phi^{-1}\left(\frac{K_1}{K_1 + K_0}\right) \right) - \frac{\mu}{\tau^2} \right] \right\}. \end{aligned}$$

Si  $x \in C_x$ , alors on rejette  $H_0 : \theta \geq \theta_0$ .

## 2.6.2 Facteur de Bayes

Les facteurs de bayes fournissent l'outil bayésien principal pour la sélection de modèles et a un lien avec la pénalité BIC (Bayesian Information Criterion) qui est utilisé par les fréquentistes pour "corriger" le maximum de vraisemblance. On note qu'il existe d'autres pénalité que le BIC et d'autres approches à la sélection de modèles.

Les décisions bayésiennes étant basées sur le calcul des probabilités  $\alpha_0$  et  $\alpha_1$ , elles reposent alors sur la nature du rapport  $\alpha_0/\alpha_1$ .

**Définition 2.6.1.** Le facteur de Bayes est le rapport

$$B_F = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1},$$

où  $\pi_i = P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta) d\theta$ ,  $i = 0, 1$ .

C'est une transformation bijective de la probabilité a posteriori. Il s'agit de comparer le rapport des lois a posteriori et le rapport des lois a priori.

### Cas particuliers

1. Si  $\Theta_0 = \{\theta_0\}$  et  $\Theta_1 = \{\theta_1\}$ , le facteur de bayes n'est que le rapport de vraisemblance classique. En effet,

$$B_F = \frac{f(x|\theta_0)}{f(x|\theta_1)} = \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta_1)\pi_1(\theta) d\theta} = \frac{m_0(x)}{m_1(x)},$$

en utilisant le fait que  $\int_{\Theta_i} f(x|\theta)\pi(\theta) d\theta = f(x|\theta_i)\pi(\theta_i)$ ,  $i = 0, 1$ .

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

2. On veut tester  $H_0 : \theta = \theta_0$ .

Soit  $\pi_0$  la probabilité a priori que  $\theta = \theta_0$ . On note  $g_1$  la densité a priori sur  $\Theta_1 = \{\theta \neq \theta_0\}$ . La loi a priori de  $\theta$  s'écrit :

$$\pi(\theta) = \pi_0 1_{\{\theta=\theta_0\}} + (1 - \pi_0)g_1(\theta)1_{\{\theta \neq \theta_0\}}.$$

On calcule la loi a posteriori de  $\theta$

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

avec

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta) d\theta = \pi_0 f(x|\theta_0) + (1 - \pi_0)m_1(x),$$

où  $m_1(x) = \int_{\{\theta \neq \theta_0\}} f(x|\theta)g_1(\theta) d\theta$  est la loi marginale sous  $H_1$ .

On a

$$\begin{aligned} \alpha_0 &= P(\theta_0|x) = \pi(\theta = \theta_0|x) \\ &= \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + (1 - \pi_0)m_1(x)} \\ &= \left[ 1 + \frac{1 - \pi_0}{\pi_0} \left( \frac{m_1(x)}{f(x|\theta_0)} \right) \right]^{-1} \end{aligned}$$

et

$$\begin{aligned}\alpha_1 &= P(\theta_1|x) = \int_{\Theta_1} P(\theta|x) d\theta \\ &= \int_{\{\theta \neq \theta_0\}} \frac{(1 - \pi_0)g_1(\theta)f(x|\theta)}{\pi_0 f(x|\theta_0) + (1 - \pi_0)m_1(x)} \\ &= \frac{(1 - \pi_0)m_1(x)}{\pi_0 f(x|\theta_0) + (1 - \pi_0)m_1(x)}.\end{aligned}$$

Ainsi, le facteur de Bayes s'écrira

$$B_F = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \left( \frac{\pi_0 f(x|\theta_0)}{(1 - \pi_0)m_1(x)} \right) / \left( \frac{\pi_0}{1 - \pi_0} \right) = \frac{f(x|\theta_0)}{m_1(x)}.$$

C'est une expression ressemblant à un rapport de vraisemblance modifié.

*Remarque 2.6.1.*

$$\pi(\theta = \theta_0|x) = \left[ 1 + \frac{1 - \pi_0}{\pi_0} \frac{1}{B_F} \right]^{-1}.$$

**Exemple 2.6.2.** Considérons l'exemple 2.5.1 et supposons que l'on veut tester les hypothèses :

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

nous avons alors sous  $H_0$ ,  $x \sim \mathcal{N}(0, 1)$  et sous  $H_1$ ,  $x \sim \mathcal{N}(\theta, 1)$ , avec  $\theta \sim \mathcal{N}(0, 10)$ . On note les densités respectives,  $f_0(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$  et  $f_1(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$ . Le facteur de Bayes étant le rapport des densités marginales, on obtient

$$B_F = \sqrt{11} \exp\left(-\frac{5}{11}x^2\right).$$

En effet,

$$\begin{aligned}B_F &= \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{m_0(x)}{m_1(x)} \\ &= \frac{\int_{\theta=0} f_0(x|\theta)\pi_0(\theta) d\theta}{\int_{\theta \neq 0} f_1(x|\theta)\pi_1(\theta) d\theta} = \frac{f_0(x|\theta)}{\int_{\theta \neq 0} f_1(x|\theta)\pi_1(\theta) d\theta} \\ &= \frac{f_0(x|\theta)}{\int_{\theta \in \mathbb{R}} f_1(x|\theta)\pi_1(\theta) d\theta} = \frac{e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{20\pi}} \int_{\theta \in \mathbb{R}} e^{-\left(\frac{\theta^2}{20} + \frac{(x-\theta)^2}{2}\right)} d\theta} \\ &= \frac{\sqrt{20\pi}}{\int_{\theta \in \mathbb{R}} e^{-\left(\frac{11}{20}\theta^2 - \theta x\right)} d\theta} = \frac{\sqrt{20\pi} e^{-\frac{5}{11}x^2}}{\int_{\theta \in \mathbb{R}} e^{-\frac{11}{20}\left(\theta - \frac{10}{11}x\right)^2} d\theta} = \sqrt{11} \exp\left(-\frac{5}{11}x^2\right).\end{aligned}$$

## Echelle de Jeffreys

Jeffreys en (1939) a développé une échelle pour évaluer le degré de certitude de  $H_0$  par les données. L'échelle est la suivante :

- (i) Si  $\ln_{10}(B_F)$  varie entre 0 et 0.5, la certitude que  $H_0$  est fautive est faible.

- (ii) Si elle est entre 0.5 et 1, cette certitude est substantielle (non négligeable ou importante).
- (iii) Si elle est entre 1 et 2, elle est forte.
- (iv) Si elle est au dessus de 2, elle est décisive.

# Modélisation de l'information a priori

Le choix des lois a priori est une étape fondamentale dans l'analyse bayésienne. Les stratégies du choix sont diverses. Elles peuvent se baser sur des expériences du passé ou sur une intuition ou une idée que le praticien a du phénomène aléatoire qu'il étudie.

## 3.1 Lois conjuguées naturelles

Une des difficultés de l'approche bayésienne est le calcul de la loi a posteriori. Ce calcul est facilité lorsque la loi a priori et la loi a posteriori ont la même forme. On parle dans ce cas de la loi a priori conjuguée.

**Définition 3.1.1.** [15]

Une classe  $\mathcal{C}$  de distributions a prioris  $\pi(\theta)$  conjuguées naturelles par rapport à une densité  $f(x|\theta)$  est un ensemble de distributions  $\pi(\theta)$  tel que la loi a posteriori  $\pi(\theta|x)$  appartient à  $\mathcal{C}$ . La classe  $\mathcal{C}$  est dite conjuguée naturelle.

### Exemples de lois conjuguées naturelles

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$	$\mathbb{E}(\theta x)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(x/\sigma^2 + \mu/\tau^2; (1/\sigma^2 + 1/\tau^2)^{-1})$	$x/\sigma^2 + \mu/\tau^2$
$\mathcal{P}(\theta)$	$\mathcal{G}(a, b)$	$\mathcal{G}(a + x, b + 1)$	$(a + x)/(b + 1)$
$\mathcal{B}(n, \theta)$	$\mathcal{Be}(a, b)$	$\mathcal{Be}(a + x, b + 1)$	$(a + x)/(a + b + x + 1)$
$\mathcal{G}(a, b)$	$\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + a, \beta + x)$	$(a + \alpha)/(b + x)$

*Remarque 3.1.1.* Dans le cas de la famille conjuguée naturelle, le praticien induit directement la forme de son estimateur dès qu'il a choisi sa loi a priori.

**Exemple 3.1.1. 1.** Soit  $X$  une v.a. de loi de Paréto de paramètres  $\theta$  et  $a$ ,  $a > 0$  de densité de probabilité définie par :

$$f(x|\theta, a) = \frac{\theta a^\theta}{x^{\theta+1}} 1_{[a, +\infty)}(x), \quad a > 0, \theta > 0.$$

On suppose que  $a$  est connu. On peut écrire,  $f(x|\theta) = \frac{\theta e^{\theta \ln a}}{e^{(\theta+1) \ln x}}$ . D'où,

$$f(x|\theta) \propto \theta e^{\theta \ln(a/x)}.$$

On peut donc prendre une loi a priori de type Gamma.

2. On considère une loi Binomiale Négative  $Neg(n, \theta)$ , définie par :

$$P(X = x|\theta) = C_{n+x-1}^x \theta^x (1 - \theta)^n, \quad 0 < \theta < 1, \quad x \in \mathbb{N}, n \in \mathbb{N}^*.$$

Une loi conjuguée naturelle sera une loi béta car,

$$P(X = x|p) \propto \theta^x (1 - \theta)^n.$$

## 3.2 Cas du modèle Exponentiel

### Définition 3.2.1. [16]

On appelle famille Exponentielle à  $s$ -paramètres, toute famille de lois de distributions  $\{P_\theta\}$  dont la densité a la forme suivante :

$$f(x|\theta) = \exp\left\{\sum_{i=1}^s \eta_i(\theta) T_i(x) - B(\theta)\right\} h(x),$$

où  $\eta_i(\cdot)$  et  $B(\cdot)$  sont des fonctions du paramètre  $\theta$  et les  $T_i(\cdot), i = 1, \dots, s$  sont des statistiques.

### Exemple 3.2.1.

1. Loi Exponentielle  $Exp(1|\theta)$ .

On a,

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} 1_{[0, +\infty[}(x) = \exp\left\{-\frac{1}{\theta} x - \ln \theta\right\} 1_{[0, +\infty[}(x).$$

Dans ce cas :  $s = 1, \eta_1(\theta) = 1/\theta, T_1(x) = x, B(\theta) = \ln \theta$  et  $h(x) = 1_{[0, +\infty[}(x)$ .

2. Loi Binomiale  $\mathcal{B}(n, \theta)$ .

On a pour  $x \in \{0, 1, \dots, n\}$  et  $\theta \in ]0, 1[$ ,

$$P(X = x|\theta) = C_n^x \theta^x (1 - \theta)^{n-x} = C_n^x \exp\left\{x \ln\left(\frac{\theta}{1 - \theta}\right) + n \ln(1 - \theta)\right\}.$$

Dans ce cas :  $s = 1, \eta_1(\theta) = \ln\left(\frac{\theta}{1 - \theta}\right), T_1(x) = x, B(\theta) = n \ln(1 - \theta)$  et  $h(x) = C_n^x$ .

### Forme canonique

**Définition 3.2.2.** Le modèle exponentiel s'écrit sous la forme canonique en reparamétrisant  $\eta_i(\theta) \equiv \theta_i$ , par :

$$f(x|\theta) = \exp\left\{\sum_{i=1}^s \theta_i T_i(x) - B(\theta)\right\} h(x).$$

Le résultat suivant donne la forme des lois naturelles conjuguées dans le cas du modèle exponentiel.

**Proposition 3.2.1.** Soit une densité  $f(x|\theta)$  appartenant à une famille exponentielle, alors une famille de lois a priori conjuguées pour  $f(x|\theta)$  est donnée par

$$\pi(\theta|\lambda, \mu) = K(\lambda, \mu) \exp\{\theta \mu - \lambda B(\theta)\},$$

où  $K(\lambda, \mu)$  est une constante de normalisation,  $\lambda$  et  $\mu$  sont appelés hyper-paramètres.

La loi a posteriori, dans ce cas, est de la forme

$$\pi(\theta|x) \propto \exp\{(\mu + x)\theta - (\lambda + 1)B(\theta)\}.$$

**Exemple 3.2.2.** Soit le modèle, dit logistique, défini par :

$$P(X = x|\theta, \beta) = \frac{e^{(\theta-\beta)x}}{1 + e^{\theta-\beta}}, \quad x \in \{0, 1\}.$$

Ce modèle est très appliqué en économétrie.

- La loi logistique appartient à la famille exponentielle. En effet,

$$P(X = x|\theta, \beta) = \exp\{(\theta - \beta)x - \ln(1 + e^{\theta-\beta})\}.$$

Dans ce cas,  $s = 1$ ,  $T_1(x) = x$ ,  $\eta_1(\theta, \beta) = \theta - \beta$ ,  $B(\theta, \beta) = \ln(1 + e^{\theta-\beta})$  et  $h(x) = 1$ .

- En utilisant la reparamétrisation suivante  $\alpha = {}^t(\theta, \beta)$ , on obtient :  $T(x) = {}^t(x, -x)$ ,  $B(\alpha) = \ln(1 + e^{\theta-\beta})$  et  $h(x) = 1$ . D'où le modèle,

$$P(X = x|\alpha) = \exp\{{}^t\alpha T - B(\alpha)\}.$$

- D'après la proposition précédente on obtient une loi a priori de la forme :

$$\begin{aligned} \pi(\theta, \beta) &\propto \exp\{(\theta, \beta) {}^t(\mu_1, \mu_2) - \lambda B(\theta, \beta)\} \\ &\propto e^{\theta\mu_1 + \beta\mu_2} e^{-\lambda \ln(1 + e^{\theta-\beta})} \\ &\propto \frac{e^{\theta\mu_1 + \beta\mu_2}}{(1 + e^{\theta-\beta})^\lambda}. \end{aligned}$$

- La loi a posteriori aura la forme suivante :

$$\begin{aligned} \pi(\theta, \beta|x) &\propto \exp\{(\mu_1 + x)\theta + (\mu_2 + x)\beta - (\lambda + 1)B(\theta, \beta)\} \\ &\propto \frac{e^{(\mu_1 + x)\theta + (\mu_2 + x)\beta}}{(1 + e^{\theta-\beta})^{\lambda+1}}. \end{aligned}$$

On remarque que la loi a priori obtenue est impropre.

### 3.3 Lois non informatives

**Définition 3.3.1.** Une loi non informative est une loi qui porte une information sur le paramètre à estimer dont le poids dans l'inférence est réduit.

*Remarque 3.3.1.* Certains auteurs définissent la loi non informative comme une loi a priori qui ne contient aucune information sur le paramètre à estimer ou encore comme une loi qui ne donne pas davantage de poids à telle ou telle valeur du paramètre.

#### 3.3.1 A priori Uniforme

La densité a priori non informative la plus simple et la plus communément utilisée est la densité Uniforme. En effet, ce choix repose sur l'équiprobabilité des valeurs possibles du paramètre  $\theta$  sur son domaine de définition, et donc n'apporte aucune informationsupplémentaire sur  $\theta$ . Ainsi, la densité est définie par

$$\pi(\theta) = k, \quad k \text{ est une constante positive.}$$

### 3.3.2 A priori invariant

Si on passe d'un paramètre  $\theta \in \Theta$  au paramètre  $\eta = h(\theta) \in h(\Theta)$  par une transformation bijective  $h$ , l'information a priori n'est pas modifiée, puisqu'elle est toujours inexistante, donc on devrait aussi utiliser une loi a priori non informative pour  $\eta$ . Il s'agit de l'invariance par reparamétrisation.

#### 3.3.2.1 Invariance pour le paramètre de position

Soit  $X$  est une v.a. de densité ne dépendant que de  $x - \theta$ .

Dans ce cas,  $\theta$  est appelé paramètre de position et  $X - \theta$  le pivot.

Un a priori invariant par rapport à un paramètre de position est un a priori invariant par rapport au choix de l'origine (position). L'absence d'information a priori pour  $\theta$  implique nécessairement que le décalage de la position n'a pas d'influence sur l'état de connaissance. On a donc,

$$\theta' = \theta + b \text{ où } b \text{ est une constante.}$$

La loi a priori doit donc vérifier

$$\pi(\theta + b) = \pi(\theta).$$

La seule solution est donnée par la loi Uniforme sur  $\Theta$  (qui est une loi impropre).

Par exemple, on peut établir de l'inférence sur la température en Celsius  $T_C$  ou en Kelvin  $T_K$ , avec  $T_K = T_C + 273,15$ .

#### 3.3.2.2 Invariance pour le paramètre d'échelle

Soit  $X$  est une v.a. de densité ne dépendant que de  $\frac{1}{\theta}f(x|\theta)$ .

Dans ce cas,  $\theta$  est un paramètre d'échelle.

Un a priori invariant par rapport à un paramètre d'échelle est un a priori qui est invariant pour la mesure d'échelle. L'absence d'information a priori pour  $\theta$  implique nécessairement l'invariance par changement d'échelle

$$\theta' = a\theta \text{ où } a \text{ est une constante.}$$

La loi a priori doit donc vérifier

$$\pi(a\theta) = a\pi(\theta).$$

La seule solution est de prendre une densité a priori telle que  $\pi(\theta) = k/\theta$  (impropre).

Par exemple, on peut établir de l'inférence sur les mètres (m) contre le pied (foot), avec 1 pied = 0.3048 m.

*Remarque 3.3.2.* L'estimateur MMSE n'est pas invariant par reparamétrisation alors que le MAP vérifie le principe d'invariance. Si la densité est unimodale et symétrique les deux estimateurs coïncident.

### 3.3.3 Règle de Jeffreys (cas unidimensionnel)

Jeffreys en 1961 [11] propose une méthode de construction de lois a priori non informative, en se basant sur le principe d'invariance par transformation, utilisant l'information de Fisher Euler qui représente une mesure de la quantité d'informations sur  $\theta$  contenue dans l'observation.

**Définition 3.3.2.** Soit  $\theta$  un paramètre réel. On appelle loi a priori non informative de Jeffreys la loi (éventuellement impropre) de densité

$$\pi_J(\theta) \propto \sqrt{I_X(\theta)} 1_{\Theta}(\theta).$$

Ou encore,  $\pi_J(\theta) = C \sqrt{I_X(\theta)}$ ,  $C$  est une constante et  $I_X(\theta)$  l'information de Fisher apportée par  $X$  sur le paramètre  $\theta$ , définie par

$$I_X(\theta) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln f(x|\theta) \right)^2 \right],$$

si le domaine de  $X$  est indépendant de  $\theta$  alors  $I_X(\theta) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \ln f(x|\theta) \right]$ .

*Remarque 3.3.3.*

1. Plus  $I_X(\theta)$  est grande, plus l'observation apporte de l'information. Il est donc naturel de favoriser les valeurs de  $\theta$  pour lesquelles  $I_X(\theta)$  est grande, au sens rendre plus probable suivant  $\pi(\theta)$ , ce qui minimise l'influence de la loi a priori au profit de l'observation.
2. Ce type de construction de lois a priori non informatives conduit très souvent à des lois a priori impropres (appelées aussi quasi a priori).
3. Dans le cas d'un échantillon de taille  $n$ ,  $\pi_J(\theta) \propto (I_n(\theta))^{1/2}$ .
4. L'a priori de Jeffreys offre une méthode automatisée pour obtenir un a priori non informatif pour n'importe quel modèle paramétrique.
5. L'a priori de Jeffreys est invariant par transformation bijective.

**Exemple 3.3.1.** Soit  $X$  une v.a. de densité  $f(x|\lambda) = \lambda e^{-\lambda x}$ ,  $x > 0$ .

On calcule l'information de Fisher,  $I(\lambda) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \lambda^2} \ln f(x|\lambda) \right]$  apportée par l'observation  $X$  sur le paramètre inconnu  $\lambda$ .

On a  $\ln f(x|\lambda) = \ln \lambda - \lambda x$ ,  $\frac{\partial}{\partial \lambda} \ln f(x|\lambda) = 1/\lambda - x$  et  $\frac{\partial^2}{\partial \lambda^2} \ln f(x|\lambda) = -1/\lambda^2$ .

On obtient,  $I(\lambda) = 1/\lambda^2$ .

La loi a priori non informative de Jeffreys est alors donnée par

$$\pi_J(\lambda) \propto 1/\lambda, \quad 1]0, +\infty[ (\lambda).$$

En considérant  $\lambda \in [a, b]$ ,  $a > 0$ , on peut déterminer la constante  $c$ , telle que

$$c \int_a^b \pi_J(\lambda) d\lambda = 1.$$

Dans ce cas,  $c \int_a^b 1/\lambda d\lambda = c \ln \lambda \Big|_a^b = 1 \Rightarrow c = \frac{1}{\ln(b/a)}$ .

La loi a priori propre de Jeffreys est alors donnée par

$$\pi(\lambda) = [\ln(b/a)\lambda]^{-1}, \quad a < \lambda < b.$$

On peut donc toujours normaliser les lois quasi a priori.

### 3.3.4 Généralisation : cas multidimensionnel

Dans le cas multidimensionnel le paramètre  $\theta$  est un vecteur,  $\theta = (\theta_1, \theta_2, \dots, \theta_q)$ . Dans ce cas, on peut considérer des lois de la forme :

$$\pi(\theta) = [\det I(\theta)]^{1/2},$$

où  $I(\theta)$  est la matrice d'information de Fisher, dont les éléments sont donnés par

$$I(\theta)_{ij} = \mathbb{E} \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(x|\theta) \right], \quad 1 \leq i, j \leq q.$$

*Remarque 3.3.4.* Dans le cas de la théorie du Maximum de Vraisemblance, le déterminant de cette quantité représente ce que l'on appelle la Variance Généralisée.

**Exemple 3.3.2.** On considère la loi normale,  $X \sim \mathcal{N}(\mu, \sigma^2)$ , dont le paramètre à estimer est  $\theta = (\mu, \sigma)$ . On a  $f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$  et  $\ln f(x|\theta) = -\ln(\sqrt{2\pi}) - \ln \sigma - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$ .

Les dérivées partielles de premiers et seconds ordres sont données par :

$$\frac{\partial}{\partial \mu} \ln f(x|\theta) = \frac{x-\mu}{\sigma^2}, \quad \frac{\partial^2}{\partial \mu^2} \ln f(x|\theta) = -\frac{1}{\sigma^2}, \quad \frac{\partial^2}{\partial \sigma \partial \mu} \ln f(x|\theta) = -\frac{2(x-\mu)}{\sigma^3} \quad \text{et} \quad \frac{\partial^2}{\partial \sigma^2} \ln f(x|\theta) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}.$$

En utilisant le fait que :  $\mathbb{E}(x - \mu) = 0$  et  $\mathbb{E}(x - \mu)^2 = \sigma^2$ , on obtient :

$$\begin{aligned} I(\theta) &= \mathbb{E} \left[ \begin{pmatrix} 1/\sigma^2 & \frac{2(x-\mu)}{\sigma^3} \\ \frac{2(x-\mu)}{\sigma^3} & -\frac{1}{\sigma^2} + \frac{3(x-\mu)^2}{\sigma^4} \end{pmatrix} \right] \\ &= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & -\frac{1}{\sigma^2} + \frac{3}{\sigma^2} \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \end{aligned}$$

On obtient,  $\det(I(\theta)) = 2/\sigma^4$ .

La loi a priori de Jeffreys est alors  $\pi(\theta) \propto 1/\sigma^2$ ,  $\mu \in \mathbb{R}$  et  $\sigma \in \mathbb{R}^{+*}$ .

### 3.3.5 Loi a priori de Référence

La théorie de l'a priori de référence est introduite par Bernardo (1979) suite à la difficulté rencontrée dans l'utilisation de l'approche de Jeffreys dans le cas multidimensionnel.

Le principe de cette loi est d'assigner comme loi de probabilité a priori une densité  $\pi^*(\theta)$  qui maximise l'information apportée par les données sur le paramètre  $\theta$ , lorsque la loi a priori est  $\pi(\theta)$ .

**Définition 3.3.3.** (Lindley 1956) Considérons l'information apportée par les observations sur le paramètre  $\theta$ , définie par

$$I_{\underline{X}}^{\theta}(\pi) = \mathbb{E} [\mathcal{D}(\pi(\theta|\underline{X}), \pi(\theta))] = \int \mathcal{D}(\pi(\theta|\underline{X}), \pi(\theta)) m(\underline{x}) d\underline{x},$$

où  $m(\underline{x}) = \int_{\Theta} L(\underline{x}|\theta) \pi(\theta) d\theta$  (densité marginale de l'échantillon  $\underline{x}$ ) et  $\mathcal{D}(\pi(\theta|\underline{X}), \pi(\theta))$  est la divergence de Kullback-Leibler entre l'a priori et l'a posteriori, définie par

$$\mathcal{D}(\pi(\theta|\underline{X}), \pi(\theta)) = \int \pi(\theta|\underline{X}) \ln \left( \frac{\pi(\theta|\underline{X})}{\pi(\theta)} \right) d\theta.$$

L'a priori  $\pi^*(\theta)$  qui maximise l'information  $I_{\underline{X}}^{\theta}(\pi)$  vérifie la relation :

$$\pi(\theta) \propto \exp \left\{ \int L(\underline{X}|\theta) \ln \pi(\theta|\underline{X}) d\theta \right\}. \quad (3.1)$$

*Remarque 3.3.5.* En général, la solution exacte vérifiant la relation (3.1) est difficile à obtenir.

### Choix pratique de la loi non informative

Le choix de la loi non informative est motivé par des a priori qui donnent des a posteriori correspondants à des estimations fréquentistes, des a priori qui ont une interprétation attractive ou des a priori permettant une forme analytique pour un a posteriori. Il est difficile de trouver une approche systématique, simple et intéressante sur les petits échantillons. Les choix les plus utilisés en pratique sont :

1. Pour  $\theta \in \mathbb{R}$ , on utilise la distribution normale.
2. Pour  $\theta \in (0, +\infty)$ , on utilise la distribution gamma (pour les paramètres de précision), inverse gamma pour la variance ou lognormale.
3. Pour  $\theta \in (0, 1)$ , on utilise la distribution bêta.

On peut utiliser un mélange de distributions si le choix d'une seule distribution n'est pas suffisant. Il est difficile de spécifier des distributions a priori multidimensionnelles (cas multivarié).

### 3.4 Approche hiérarchique

Considérons une loi a priori  $\pi_1(\theta)$  pour le paramètre inconnu  $\theta$ . Supposant que cette loi s'écrit en fonction d'un paramètre  $\mu$ , appelé hyperparamètre.

L'approche hiérarchique propose de manière systématique un modèle aléatoire  $\pi_2$  pour l'hyperparamètre  $\mu$ .

Le modèle devient :

$$\begin{cases} X|\theta \sim f(x|\theta) \\ \theta|\mu \sim \pi_1(\theta|\mu) \\ \mu \sim \pi_2(\mu). \end{cases}$$

La loi a priori de  $\theta$  est alors,

$$\pi(\theta) = \int_{\mathcal{E}} \pi_1(\theta|\mu)\pi_2(\mu) d\mu,$$

où  $\mathcal{E}$  est l'ensemble de toutes les valeurs de  $\mu$ .

*Remarque 3.4.1.* Dans la plupart des cas l'espace  $\mathcal{E}$  des hyperparamètres est multidimensionnel, donc induit une plus grande complexité que la distribution du modèle  $f(x|\theta)$ .

En général, il n'y a pas d'information a priori portant sur l'hyperparamètre  $\mu$ , on adopte souvent pour  $\pi_2$  une distribution non informative qui ne donne pas une préférence à des valeurs particulières.

L'intérêt essentiel de la hiérarchisation est de lever certaines difficultés calculatoires pour obtenir la loi a posteriori, en utilisant la décomposition en loi conditionnelle.

**Définition 3.4.1.** On appelle modèle bayésien hiérarchique un modèle statistique bayésien où la loi a priori est décomposée en distributions conditionnelles :

$$\pi(\theta|\theta_1), \pi(\theta_1|\theta_2), \dots, \pi(\theta_{k-1}|\theta_k)\pi(\theta_k).$$

Les paramètres  $\theta_i, i = 1, \dots, k$  sont appelés hyperparamètres.

**Exemple 3.4.1.** Considérons le modèle bayésien hiérarchique simple suivant :

$$\begin{cases} X|\theta \sim f(x|\theta) \\ \theta|\lambda \sim \text{Exp}(\lambda), \lambda \text{ est inconnu} \\ \lambda \sim \text{Exp}(\mu), \mu \text{ est connu.} \end{cases}$$

Alors,  $\pi_1(\theta|\lambda) = \lambda e^{-\lambda\theta}$ ,  $\lambda > 0, \theta > 0$  et  $\pi_2(\lambda) = \mu e^{-\mu\lambda}$ ,  $\mu > 0$ .

La loi a priori de  $\theta$  est alors obtenue par :

$$\pi(\theta) = \int_{\mathcal{E}} \pi_1(\theta|\lambda)\pi_2(\lambda) d\lambda = \int_0^{+\infty} \lambda \mu e^{-\lambda(\theta+\mu)} d\lambda = \frac{\mu}{(\theta + \mu)^2}, \theta > 0, \mu > 0.$$

La loi a posteriori de  $\theta$  sachant  $x$  est donnée par

$$\pi(\theta|x) = \int_0^{+\infty} \pi(\theta|\lambda, x)\pi(\lambda|x) d\lambda.$$

Dans ce cas,  $\pi(\theta|\lambda, x) = \frac{f(x|\theta)\pi_1(\theta|\lambda)}{f(x|\lambda)}$ ,  $f(x|\lambda) = \int_0^{+\infty} f(x|\theta)\pi_1(\theta|\lambda) d\theta$  et  $\pi(\lambda|x) = \frac{f(x|\lambda)\pi_2(\lambda)}{f(x)}$  avec  $f(x) = \int_0^{+\infty} f(x|\lambda)\pi_2(\lambda) d\lambda$ .

### 3.5 Incorporation de l'information a priori en pratique

On considère le cas d'un paramètre  $\theta$  réel. On envisage trois situations :

- $\theta \in [0, 1]$ ,
- $\theta \in [0, +\infty[$ ,
- $\theta \in ]-\infty, +\infty[$ .

On suppose que l'information a priori fournie par l'expert consiste en une estimation ponctuelle  $\theta^*$  de  $\theta$  et un intervalle  $I^* = [\alpha^*, \beta^*]$  contenant  $\theta^*$ , tel que  $P(\theta^* \in I^*)$  soit élevée (typiquement 0.95).

#### Situation 1 : $\theta \in [0, 1]$

On choisit dans ce cas une loi a priori de type bêta,

$$\theta \sim \pi(\theta) = \mathcal{B}e(a, b).$$

On procède à la reparamétrisation comme suit :

$$\begin{cases} a = \lambda\mu \\ b = \lambda(1 - \mu), \text{ où } \mu = \mathbb{E}(\theta) \end{cases} \Rightarrow \begin{cases} \mu = \mathbb{E}(\theta) = \frac{a}{a+b} \\ \lambda = a + b. \end{cases}$$

On a

$$\text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = \frac{\mu(1-\mu)}{1+\lambda}.$$

On obtient qu'à  $\mu$  fixé,  $\lambda$  est proportionnel à la variance et donc s'interprète comme la précision de l'information a priori. Autrement dit, plus  $\lambda$  est grand plus  $\text{Var}(\lambda)$  est petit et plus l'information a priori est précise. Inversement, plus  $\lambda$  est petit plus  $\text{Var}(\lambda)$  est grand et moins l'information a priori est précise.

En pratique, on pose  $\mathbb{E}(\theta) = \theta^* = \mu$ . D'où,

$$\theta \sim \pi(\theta) = \mathcal{B}e(\lambda\theta^*, \lambda(1 - \theta^*)).$$

$\lambda$  est déterminé tel que

$$P(\theta^* \in I^*) = \int_{I^*} \pi(\theta) d\theta = 0.95.$$

Pour,  $\mathbb{E}(\theta) = \theta^* = 0.2$  et  $I^* = [0.05, 0.4]$  tel que  $P(\theta \in I^*) = 0.95$ , on obtient sur MATLAB  $\lambda = 17.5$ . D'où la loi a priori,  $\pi(\theta) = \mathcal{B}e(3.5, 14)$ .

**Situation 2 :**  $\theta \in [0, +\infty[$

On choisit, dans ce cas, une loi Gamma  $\mathcal{G}(a, b)$  et on procède à la reparamétrisation comme suit :

$$\begin{cases} a = \lambda\mu \\ b = \lambda, \end{cases} \Rightarrow \begin{cases} \mu = a/b \\ \lambda = b \end{cases} \text{ avec } \mu = \mathbb{E}(\theta).$$

On a  $\mathbb{E}(\theta) = \mu = a/b$  et  $Var(\theta) = a/b^2 = \mu/\lambda$ .

D'où pour  $\mu$  fixé,  $\lambda$  est inversement proportionnel à la variance. Autrement dit, plus  $\lambda$  est grand plus  $Var(\theta)$  est petit et plus l'information a priori est précise. Inversement, plus  $\lambda$  est petit plus  $Var(\theta)$  est grand et moins l'information a priori est précise.

En pratique, on pose  $\mathbb{E}(\theta) = \theta^* = \mu$ . On obtient

$$\theta \sim \mathcal{G}(\lambda\theta^*, \lambda)$$

où  $\lambda$  est déterminé tel que

$$P(\theta \in I^*) = \int_{I^*} \pi(\theta) d\theta = 0.95.$$

**Situation 3 :**  $\theta \in \mathbb{R} = ] - \infty, +\infty[$

Dans ce cas, on choisit comme loi a priori la loi normale  $\mathcal{N}(a, b)$  avec  $a = \mathbb{E}(\theta)$  et  $b = Var(\theta)$ .

### 3.6 Le poids de l'a priori dans la réponse bayésienne

Considérons le modèle bayésien suivant :

$$\begin{cases} X|\theta \sim \text{Bernoulli}(\theta) \\ \theta \sim \text{Be}(a, b) \end{cases}$$

L'information donnée par l'a priori et celle contenue dans les observations peuvent se combiner pour produire la réponse bayésienne.

Dans le cas de ce modèle, la loi a posteriori est donnée par

$$\theta|\underline{x} \sim \text{Be}(a + n\bar{x}, b + n - n\bar{x}).$$

L'estimateur bayésien obtenu relativement à la perte quadratique est alors

$$\mathbb{E}(\theta|\underline{X}) = \frac{a + n\bar{X}}{a + b + n}.$$

En procédant à la reparamétrisation de la loi bêta à l'aide des paramètres  $\lambda$  et  $\mu$ , on obtient :

$$\begin{cases} a = \lambda\mu \\ b = \lambda(1 - \mu) \end{cases} \Rightarrow \begin{cases} \lambda = a + b \\ \mu = \frac{a}{a+b} \end{cases} \text{ avec } \mu = \mathbb{E}(\theta) = \theta^* \text{ et } Var(\theta) = \frac{\mu(1 - \mu)}{1 + \lambda}.$$

La réponse bayésienne est alors donnée par :

$$\mathbb{E}(\theta|\underline{X}) = \left(\frac{\lambda}{\lambda + \mu}\right) \mathbb{E}(\theta) + \left(\frac{n}{\lambda + n}\right) \bar{X}. \tag{3.2}$$

**Interprétation :**

- L'estimation bayésienne de  $\theta$  apparaît comme la moyenne pondérée de  $\bar{X}$  (estimateur de  $\theta$  par la méthode MLE) et de la moyenne a priori  $\mathbb{E}(\theta)$ .
- A partir de l'équation (3.2), on peut considérer que le poids de  $\bar{X}$  est la taille de l'échantillon alors que le poids de  $\mathbb{E}(\theta)$  est la précision de l'a priori  $\lambda$ .
- Géométriquement,  $\mathbb{E}(\theta|\underline{x})$  est le barycentre des points de coordonnées  $\mathbb{E}(\theta)$  et  $\bar{x}$  affectés des coefficients  $\frac{\lambda}{\lambda+\mu}$  et  $\frac{n}{\lambda+n}$ , respectivement.

**Cas particuliers :**

- Si  $\lambda = n$ , l'estimation bayésienne de  $\theta$  se situe exactement au milieu de l'intervalle  $[\mathbb{E}(\theta), \bar{x}]$ .
- Si  $\lambda > n$ , cette estimation est plus proche de  $\mathbb{E}(\theta)$  que de  $\bar{x}$ .
- Si  $\lambda < n$ , cette estimation est plus proche de  $\bar{x}$  que de  $\mathbb{E}(\theta)$ .

**Cas limites :**

Pour  $n$  fixé et  $\mu = \mathbb{E}(\theta)$  donné, nous avons les deux cas limites suivants :

- $\lambda \rightarrow 0$  : Dans ce cas,  $Var(\theta) = \mu(1 - \mu)$  (valeur maximale), le poids de l'a priori est nul et l'estimation bayésienne tend vers la réponse classique  $\mathbb{E}(\theta|\underline{x}) \rightarrow \bar{X}$  (situation non informative).
- $\lambda \rightarrow +\infty$  : Dans ce cas,  $Var(\theta) = 0$  (valeur minimale) et le poids des données est nul  $\mathbb{E}(\theta|\underline{x}) \rightarrow \mathbb{E}(\theta)$  indépendante des observations (situation extrêmement informative).

On peut également regarder le cas limite  $n \rightarrow +\infty$  en considérant  $\lambda$  et  $\mu$  fixés. Dans ce cas, le poids de l'a priori devient négligeable et la réponse bayésienne coïncide avec la réponse classique  $\bar{X}$  (obtenue par le MLE).

*Remarque 3.6.1.* Dans le cadre d'une étude de sensibilité de la réponse bayésienne à la loi a priori, on comparera toujours l'estimateur bayésien obtenu pour la loi a priori informative avec celui correspondant à la loi a priori non informative, et ce afin d'apprécier le poids de l'a priori dans la réponse bayésienne informative.

**Exemple 3.6.1.** Considérons un échantillon  $X_1, \dots, X_n$  issu d'une v.a.  $X$  de loi Bernoulli  $\mathcal{B}(\theta)$  et une loi a priori pour  $\theta$  de type bêta  $\mathcal{Be}(\alpha, \beta)$ .

- Pour  $\alpha = \beta = 1$ , nous avons un a priori Uniforme.
- Pour  $\alpha = \beta > 1$ , nous avons une courbe en cloche.
- Pour  $\alpha = \beta < 1$ , nous avons une courbe en U.

La moyenne  $\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$ , la variance  $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{(\alpha+\beta)} \times \frac{\beta}{\alpha+\beta} \times \frac{1}{\alpha+\beta+1}$  et le mode est donné par  $\frac{\alpha-1}{\alpha+\beta-2}$ , pour  $\alpha > 1$  et  $\beta > 1$ .

L'estimateur de bayes de  $\theta$ , relativement à la perte quadratique, est alors

$$\mathbb{E}(\theta|\underline{x}) = \frac{\alpha + k}{\alpha + \beta + n} = \frac{\alpha}{(\alpha + \beta)} \times \frac{(\alpha + \beta)}{\alpha + \beta + n} + \frac{n}{(\alpha + \beta + n)} \times \frac{k}{n}, \text{ où } k = \sum_i x_i.$$

On voit bien que l'espérance a posteriori du paramètre est une combinaison convexe de l'estimateur du maximum de vraisemblance  $\hat{\theta}_{MLE} = \frac{\sum_i X_i}{n} = \bar{X}$  et de l'espérance a priori  $\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$ . Il se rapproche asymptotiquement de l'estimateur du maximum de vraisemblance.

La variance de la distribution a posteriori est donnée par

$$Var(\theta|\underline{x}) = \hat{\theta}_{MLE}(1 - \hat{\theta}_{MLE}) \frac{1}{(\alpha + \beta + n)}$$

qui décroît en  $1/n$ . Nous avons une distribution de plus en plus piquée autour de  $\theta_{MLE}$  et les intervalles de confiance de plus en plus serrés autour de l'estimateur lorsque le nombre d'observations augmente.

### 3.7 Sensibilité de la réponse bayésienne à la loi a priori

Le choix de la famille de probabilité retenue pour la loi a priori et le choix des valeurs numériques communiquées par l'Expert, pour déterminer les hyperparamètres de la loi a priori, affectent l'estimation bayésienne du paramètre. Dans la pratique, on se contente d'examiner dans quelle mesure une petite perturbation des hyperparamètres modifie l'estimation bayésienne du paramètre "étude de sensibilité de la réponse bayésienne à la loi a priori".

On comparera toujours l'estimateur bayésien obtenu en utilisant une la loi a priori informative avec celui correspondant à la loi a priori non informative, afin d'apprécier le poids de l'a priori dans la réponse bayésienne informative.

# Méthodes de simulation de Monte-Carlo

L'estimation bayésienne étant la décision minimisant le coût a posteriori, dans la pratique ceci peut être rendu difficile pour principalement deux raisons :

- le calcul explicite de  $\pi(\theta/\underline{x})$  peut être impossible ;
- même si  $\pi(\theta/\underline{x})$  est connu, la minimisation du coût a posteriori nécessite un temps de calcul coonsidérable. En particulier lorsque l'espace des paramètres et l'espace des décisions sont de grandes dimensions.

Une solution à ce problème est d'utiliser des approximations de la loi a posteriori  $\pi(\theta/\underline{x})$  et des intégrales correspondantes. Cela permettra de construire une approximation du coût a posteriori pour une décision arbitraire et de l'utiliser dans une méthode de minimisation classique.

Les méthodes d'approximation analytique sont en général fondées sur des considérations asymptotiques. Les plus courantes pour approximer les inférences a posteriori sont : l'approximation normale et l'approximation de Laplace. Ces méthodes donnent uniquement le comportement asymptotique de la densité a posteriori. On retrouve aussi les méthodes numériques d'intégration et d'optimisation qui peuvent être intéressantes pour résoudre les problèmes d'implémentation pratique de l'analyse bayésienne (voir [7], [13] et [20]).

Une autre alternative est donnée par les méthodes de simulation. On cite les méthode de simulation de Monte-carlo et les méthodes MCMC, développées ci-dessus.

## 4.1 Méthode de Monte Carlo

### Principe de la méthode

Soit  $\theta \in \mathbb{R}$  de densité  $f(\theta)$ . On souhaite calculer  $\mathbb{E}(g(\theta))$  où  $g(\theta)$  est un réel.

Le principe de la méthode de simulation de Monte-Carlo est de simuler une suite de v.a.  $\theta_1, \theta_2, \dots, \theta_L$  i.i.d. selon la loi de  $\theta$ , puis approcher  $\mathbb{E}(g(\theta))$  à l'aide de la loi forte des grands nombres par

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \text{ lorsque } L \rightarrow \infty.$$

### Contexte bayésien

On veut calculer  $\mathbb{E}(g(\theta/x))$  où  $g$  est une fonction d'intérêt.

On a

$$\mathbb{E}(g(\theta/x)) = \int_{\Theta} g(\theta)\pi(\theta/x) d\theta,$$

où  $\pi(\theta/x) = \frac{f(x/\theta)\pi(\theta)}{m(x)}$  et  $m(x) = \int_{\Theta} f(x/\theta)\pi(\theta) d\theta$ .

D'où,

$$\mathbb{E}(g(\theta/x)) = \frac{\int_{\Theta} g(\theta)f(x/\theta)\pi(\theta) d\theta}{\int_{\Theta} f(x/\theta)\pi(\theta) d\theta}.$$

Ce qui revient à approcher l'intégrale  $\int_{\Theta} g(\theta)f(x/\theta)\pi(\theta) d\theta$  par la méthode de simulation de Monte-Carlo.

- On peut simuler  $\theta_1, \theta_2, \dots, \theta_L$  selon la fonction d'importance  $\pi(\theta)$ . L'intégrale peut alors être approchée par

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i)f(x/\theta_i), \quad L \rightarrow \infty.$$

- On peut simuler  $\theta_1, \theta_2, \dots, \theta_L$  selon la fonction d'importance  $f(x/\theta)\pi(\theta)$ . L'intégrale peut alors être approchée par

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \quad L \rightarrow \infty.$$

- On peut simuler  $\theta_1, \theta_2, \dots, \theta_L$  selon la fonction d'importance  $\pi(\theta/x)$ . La quantité  $\mathbb{E}(g(\theta/x))$  peut alors être approchée, à l'aide de la loi forte des grands nombres, par

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \quad L \rightarrow \infty.$$

De plus, si la variance a posteriori  $Var(g(\theta)/x)$  est finie le Théorème Central-Limite (TCL) s'applique à la moyenne  $\mathbb{E}(g(\theta/x)) \simeq \frac{1}{L} \sum_{i=1}^L g(\theta_i)$  qui est asymptotiquement normale de variance  $\frac{Var(g(\theta)/x)}{L}$ . Des régions de confiance peuvent alors être construites à partir de cette approximation. L'ordre de grandeur est de  $\frac{1}{\sqrt{L}}$ , quelque soit la dimension du problème, contrairement aux méthodes numériques.

## 4.2 Méthode de Monte Carlo par Chaînes de Markov (méthode MCMC)

C'est une méthode de simulation très puissante. Elle propose une structure universelle de simulation facilement implémentable quelque soit la complexité et la dimension de la densité à simuler. Elle permet d'approcher la génération de v.a. de loi a posteriori  $\pi(\theta/x)$  lorsque celle-ci ne peut pas être simulée directement. Plusieurs types d'inférences statistiques sont possibles à partir d'échantillons produits par des méthodes MCMC comme l'estimation des moments, de modes ou de densités marginales, l'inférence prédictive, l'analyse de sensibilité, ... [5].

### Principe de la méthode

Le principe de la méthode MCMC est de générer une chaîne de Markov (CM)  $(\theta^{(l)}, l \geq 0)$ , avec  $\dim \theta^{(l)} = \dim \theta$ , qui suit une loi asymptotiquement distribuée selon la loi a posteriori de  $\theta$ , puis d'approcher la moyenne a posteriori à l'aide du théorème ergodique,

$$\frac{1}{L} \sum_{l=1}^L g(\theta^{(l)}) \xrightarrow{L \rightarrow +\infty} \mathbb{E}(g(\theta/x)).$$

Il existe deux algorithmes pour produire une telle chaîne de markov : "Algorithme de Gibbs" et "Algorithme de Metropolis-Hastings"(MH). L'algorithme de Metropolis-Hastings est considéré comme l'algorithme de base d'une grande partie des algorithmes MCMC. L'échantillonneur de Gibbs est considéré

comme un cas particulier de l'algorithme (MH) mais a des motivations méthodologiques et historiques fondamentalement différentes.

### 4.2.1 Algorithme de Metropolis-Hastings

Les méthodes (MH) ont l'avantage de ne requérir qu'une connaissance limitée de la densité à simuler en autorisant une grande liberté dans leur implémentation. Cet algorithme a été proposé par Metropolis en 1953 [10] pour le traitement de problèmes de physique, puis généralisé par Hastings en 1970 [18] et [4].

Cet algorithme permet de fabriquer une chaîne de Markov dont on s'est donné la loi stationnaire  $\pi$ . Sa mise en oeuvre présente l'avantage de ne nécessiter la définition de  $\pi$  qu'à une constante près.

Pour une densité  $\pi(\theta/x)$  connue à une constante près et une loi conditionnelle  $q(\cdot/\cdot)$ , appelée loi de proposition choisie symétrique au sens où  $q(\theta'/\theta) = q(\theta/\theta')$ , l'algorithme génère une CM comme suit : Partant d'un vecteur initial  $\theta^{(0)}$  arbitraire, A la  $p^{\text{ième}}$  étape, disposant du vecteur  $\theta^{(p)} = (\theta_1^{(p)}, \dots, \theta_m^{(p)})$ ,

- Générer  $\tilde{\theta}$  selon la loi  $q(\tilde{\theta}/\theta^{p-1})$  ( $\theta$  est choisi comme paramètre de position),
- Calculer la probabilité d'acceptation  $\alpha = \min\{1; \frac{\pi(\tilde{\theta}/x)q(\theta^{p-1}/\tilde{\theta})}{\pi(\theta^{p-1}/x)q(\tilde{\theta}/\theta^{p-1})}\}$ .
- Prendre

$$\theta^p = \begin{cases} \tilde{\theta}, & \text{avec probabilité } \alpha \\ \theta^{p-1}, & \text{avec probabilité } 1 - \alpha. \end{cases}$$

*Remarque 4.2.1.* On peut autoriser un nombre infini de lois de proposition, produisant toutes une CM convergeantes vers la loi d'intérêt.

### 4.2.2 Echantillonnage de Gibbs

L'échantillonneur de Gibbs introduit par S. Geman et al. en 1984 dans le cadre de la restauration d'images et a été ensuite généralisé par M. Tanner et al. en 1987 et par A.E. Gelfand et al. en 1990 en statistique appliquée. Le principe repose sur une décomposition d'une loi  $f(x)$  suivant des lois conditionnelles.

Considérons le cas élémentaire  $f(x, y)$ . On suppose que  $f(x/y)$  et  $f(y/x)$  sont disponibles. On peut alors générer une séquence de Gibbs de la manière suivante :

Partant d'une valeur  $x_0$ , on génère  $y_0$  suivant  $\pi(\cdot/x_0)$ , puis  $x_1$  suivant  $\pi(\cdot/y_0)$ , puis  $y_1$  suivant  $\pi(\cdot/x_1)$  et ainsi de suite.

Après  $M$  itérations de ce schéma, il vient une séquence  $(x_0, y_0, x_1, y_1, \dots, x_M, y_M)$ . Pour  $M$  assez grand,  $x_M$  est une réalisation de  $X$ .

**Exemple 4.2.1.** Considérons une loi jointe de la forme :

$$f(x, y) \propto C_n^x y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, \dots, n, \quad 0 \leq y \leq 1.$$

On remarque que l'on peut proposer une loi binomiale de paramètres  $(n, y)$  pour  $f(x/y)$  et une loi bêta de paramètres  $(x + \alpha, n - x + \beta)$  pour  $f(y/x)$ .

Bien que dans le cas de cet exemple la loi marginale  $f(x)$  est accessible,

$$f(x) \propto \int_0^1 f(x, y) dy \propto C_n^x \frac{\gamma(x + \alpha)\gamma(n - x + \beta)}{\gamma(\alpha + \beta + n)},$$

on peut toujours appliquer l'algorithme de Gibbs pour obtenir des réalisations de  $f(x)$  en simulant alternativement une réalisation  $x^*$  d'une binomiale de paramètres  $(n, y^*)$  où  $y^*$  est la valeur courante de  $y$  obtenu à l'étape précédente, puis une nouvelle réalisation  $x$  d'une loi bêta de paramètre  $(x^* + \alpha, n - x^*)$ .

Dans le cadre bayésien, l'algorithme de Gibbs va permettre d'obtenir une réalisation du paramètre  $\theta = (\theta_1, \dots, \theta_L)$  suivant la loi a posteriori  $\pi(\theta/x)$  dès que l'on est capable d'exprimer les lois conditionnelles  $\pi(\theta_i/\theta_j; x), j \neq i$ .

L'échantillonnage de Gibbs consiste à :

Partant d'un vecteur initial  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_L^{(0)})$  où  $L \geq 2$  est la dimension de  $\theta^{(0)}$ .

A la  $(p+1)$ ème étape, disposant du vecteur  $\theta^{(p)} = (\theta_1^{(p)}, \dots, \theta_L^{(p)})$ ,

- simuler  $\theta_1^{(p+1)} \sim \pi(\theta_1/\theta_2^{(p)}, \dots, \theta_L^{(p)}; x)$ ,
- simuler  $\theta_2^{(p+1)} \sim \pi(\theta_2/\theta_1^{(p+1)}, \theta_3^{(p)}, \dots, \theta_L^{(p)}; x)$ ,
- ⋮
- simuler  $\theta_L^{(p+1)} \sim \pi(\theta_L/\theta_1^{(p+1)}, \dots, \theta_{L-1}^{(p+1)}; x)$ .

Les itérations successives de cet algorithme génèrent successivement les états d'une chaîne de Markov  $\{\theta^{(p)}, p > 0\}$  à valeurs dans  $\mathbb{N}^{\otimes L}$ .

Cette chaîne admet une mesure invariante qui est la loi a posteriori. Pour un nombre d'itérations suffisamment grand, le vecteur  $\theta$  obtenu peut donc être considéré comme étant une réalisation de la loi a posteriori.

### 4.3 Mise en oeuvre des méthodes de simulation de Monte-Carlo

#### Exemple 1 : Étude bayésienne de l'efficacité d'un traitement

Soit  $n = 20$  patients atteints d'une même maladie  $M$  et suivent tous le même traitement  $T$ . L'état de chaque patient est déterminé régulièrement lors des visites médicales. On note  $x_{(i,t)}$  l'état du patient  $i$  lors de la  $t$ ème visite tel que  $x_{(i,t)} = a$ , si le patient présente les symptômes de la maladie et  $x_{(i,t)} = b$ , sinon. Soient  $\alpha$  et  $\beta$  les probabilités de transition de  $a$  vers  $a$  et de  $b$  vers  $b$ , respectivement. Pour  $i$  fixé, les v.a.  $X_{(i,t)}$  sont markoviennes.

On note  $\theta = (\alpha, \beta)$  le paramètre inconnu du modèle. Le suivi du patient  $i$  est noté  $x_i$ . Les  $X_i$  sont supposés indépendants (conditionnellement à  $\theta$ ) et suivent tous la même loi. On suppose de plus que les paramètres  $\alpha$  et  $\beta$  sont a priori indépendants et on adopte les lois uniformes comme lois a priori.

On note  $w$  la probabilité qu'un patient soit guéri à l'issue du traitement (on pratique c'est le cas si le patient ne présente plus les symptômes de la maladie lors de la quatrième visite). On considère que le traitement est efficace si l'estimation de  $w$  dépasse 0.95. Il s'agit de calculer  $\mathbb{E}(w/x)$ .

On note  $N(r, s)$  le nombre total de transitions de  $r$  vers  $s$  entre les temps  $t$  et  $t+1$  comptés sur le tableau de données  $x$ .

On peut établir facilement que  $\alpha/x \sim \mathcal{B}e(1 + N(a, a), 1 + N(a, b))$  et  $\beta/x \sim \mathcal{B}e(1 + N(b, b), 1 + N(b, a))$ .

A partir de la relation  $\pi(\alpha, \beta/x) = \pi(\alpha/x)\pi(\beta/x)$ , on obtient que pour simuler  $\theta$  selon la loi a posteriori, il suffit de simuler les deux lois Bêta définies ci-dessus de façon indépendantes. Il n'y a donc pas besoin d'utiliser les algorithmes de Gibbs et de Hastings puisque les lois a posteriori sont complètement connues.

### Exemple 2 : Cas de la loi Normale

Considérons une v.a.  $X$  de loi normale  $\mathcal{N}(\mu, \sigma^2)$  où  $\mu$  et  $\sigma^2$  sont inconnus. La densité de  $X$  s'écrit :

$$f(x/\mu, \sigma^2) \propto \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

On remarque que

$$\pi(\mu/\sigma^2, x) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

c'est donc une loi normale de paramètres  $x$  et  $\sigma^2$ .

D'un autre côté on a,

$$\begin{aligned} \pi(\sigma^2/\mu, x) &\propto \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ &\propto (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2}(\sigma^2)^{-1}(x-\mu)^2} \end{aligned}$$

on obtien que  $\sigma^2/(\mu, x)$  est une Inverse Gamma de paramètres  $\frac{1}{2}$  et  $\frac{1}{2}(x - \mu)^2$ .

Les lois conditionnelles étant connues, on obtient qu'on peut appliquer l'algorithme de Gibbs permettant de simuler les deux lois conditionnelles (Normale et Inverse Gamma) définies ci-dessus, pour simuler  $\theta = (\mu, \sigma^2)$  selon la loi a posteriori.

## Exercices de TD

### 5.1 Énoncé des exercices

**Exercice 1.**

Soit  $X$  une v.a. de loi de probabilité dépendant d'un paramètre inconnu  $\theta$ .

Déterminer la loi de probabilité a posteriori de  $\theta$  et la loi marginale de  $X$  dans les cas suivants :

1.  $\begin{cases} X/\theta \sim \mathcal{B}(n, \theta), \theta \in [0, 1] \\ \theta \sim \mathcal{B}e(\alpha, \beta). \end{cases}$
2.  $\begin{cases} X/\theta \sim \mathcal{N}(0, \theta^2), \theta > 0 \\ 1/\theta^2 \sim \mathcal{G}(1, 2). \end{cases}$

**Exercice 2.**

Soit  $X$  une v.a. de loi de probabilité Binomiale  $\mathcal{B}(2, \theta), \theta \in [0, 1]$ .

Considérons les règles de décision définies par :

$$\delta_1(x) = (x + 1)/4 \text{ et } \delta_2(x) = x/2,$$

où  $x$  est une observation de  $X$ .

1. Calculer le risque quadratique associé à la décision  $\delta_1$ .
2. Pour quelles valeurs de  $\theta$ , le risque obtenu est maximal ?
3. Calculer le risque fréquentiste associé à  $\delta_2$ , relativement à la perte  $l$  définie par

$$l(\theta, \delta) = (\theta - \delta)^2 / [\theta(1 - \theta)].$$

4. Déterminer le risque bayésien relativement à la loi a priori Uniforme sur  $[0, 1]$  et à la décision  $\delta_2$ .

**Exercice 3.**

Pour un faisceau de lignes d'un central téléphonique, le nombre d'arrivées d'appels pendant l'unité de temps suit une loi de Poisson de paramètre inconnu  $\theta > 0$ . La durée  $T$  séparant deux arrivées successives d'appels téléphoniques sur ce faisceau suit alors une loi exponentielle de paramètre  $\theta$ . Soit  $f_T(t) = \theta e^{-\theta t}, t > 0$  la densité de la v.a.  $T$ .

On désire estimer le paramètre inconnu  $\theta$  à l'aide de l'observation de  $n$  durées  $t_i, i = 1, \dots, n$  séparant des arrivées successives d'appels téléphoniques sur ce faisceau.

1. Dans un premier temps on suppose qu'aucune information a priori n'est disponible sur  $\theta$  (à part  $\theta > 0$ ). Déterminer à partir des observations  $t_i$  l'estimateur du maximum de vraisemblance de  $\theta$ , noté  $\hat{\theta}_{MV}$ .
2. On sait que pour l'ensemble des faisceaux téléphoniques, le nombre moyen  $\theta$  d'arrivée d'appels téléphoniques pendant l'unité de temps est distribué suivant une loi exponentielle de paramètre  $\lambda$  connu.

Soit  $g(\theta) = \lambda e^{-\lambda\theta}$ ,  $\theta > 0$  la densité de  $\theta$ , représentant la loi a priori de  $\theta$ .

Déterminer à partir des observations  $t_i$  les estimateurs de la moyenne a posteriori  $\hat{\theta}_{MMSE}$  et du maximum a posteriori  $\hat{\theta}_{MAP}$ .

3. Montrer que les estimateurs  $\hat{\theta}_{MV}$ ,  $\hat{\theta}_{MMSE}$  et  $\hat{\theta}_{MAP}$  sont équivalents lorsque  $n$  est grand. Interpréter ce résultat.

#### Exercice 4.

Soit  $X$  une v.a. indiquant si une expérience est un succès (1) ou un échec (0) et dont la distribution est une loi de Bernoulli de paramètre  $\theta$ . On sait que la distribution a priori de  $\theta$  est une loi Uniforme  $\mathcal{U}_{[0,1]}$ . On a observé un succès en trois essais.

1. Calculer l'estimateur bayésien  $\hat{\theta}$  si la fonction de perte choisie est l'erreur quadratique.

2. Trouver l'estimation bayésienne de la probabilité que  $\theta$  se trouve entre 0.2 et 0.4.

#### Exercice 5.

On suppose que  $X/\beta$  suit une loi exponentielle de paramètre  $\beta$  et que la distribution a priori de  $\beta$  est une Gamma  $\mathcal{G}(2, 3)$ . On a observé l'échantillon aléatoire suivant  $\{6, 11, 8, 13, 9\}$ .

1. Calculer l'estimateur bayésien de  $\beta$  si la fonction de perte est l'erreur quadratique.

2. Reprendre la question 1. pour la fonction de perte valeur absolue de l'erreur.

On fournit les valeurs suivantes :

$$\begin{aligned} \Gamma(7; 4.734) &= 0.2; & \Gamma(7; 5.411) &= 0.3; & \Gamma(7; 6.039) &= 0.4; \\ \Gamma(7; 6.670) &= 0.5; & \Gamma(7; 7.343) &= 0.6. \end{aligned}$$

#### Exercice 6.

Soit  $(X_1, X_2, \dots, X_n)$  un  $n$ -échantillon de loi Bernoulli de paramètre inconnu  $\theta$ ,  $\theta \in (0, 1)$ . On considère la loi a priori de  $\theta$ , définie par :

$$\pi(\theta) = [\theta(1 - \theta)]^{-1} 1_{]0,1[}(\theta) \text{ (appelée loi de Haldane).}$$

1. Calculer l'estimateur du Maximum de vraisemblance pour  $\theta$ , noté  $\hat{\theta}_{MV}$ .

2. Montrer que  $\pi(\theta)$  est une loi a priori impropre.

3. Calculer l'estimateur de Bayes Généralisé de  $\theta$ , noté  $\hat{\theta}_G$ , relativement au risque quadratique.

4. A quelle condition est-il défini ?

5. Comparer les deux estimateurs obtenus.

#### Exercice 7.

1. Déterminer une loi a priori conjuguée naturelle pour la loi normale  $\mathcal{N}(\mu, \sigma^2)$ , avec  $\mu$  connu.

2. Montrer que la loi Gamma  $\mathcal{G}(\alpha, \beta)$ ,  $\alpha, \beta > 0$  appartient à la famille Exponentielle.

3. Ecrire la loi Gamma sous sa forme canonique et donner une famille de lois a prioris conjuguées naturelles et la loi a posteriori correspondante.

#### Exercice 8.

Soit  $X_1, X_2, \dots, X_n$  un  $n$ -échantillon issu d'une v.a.  $X$ . Calculer la loi non informative de Jeffreys pour le modèle Gaussien  $\mathcal{N}(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$  et  $\sigma \in \mathbb{R}^+$ , dans les cas suivants :

1.  $\theta = \mu$  et  $\sigma^2$  connu.

2.  $\theta = \sigma^2$  et  $\mu$  connu.

3.  $\theta = (\mu, \sigma^2)$ ,  $\mu$  et  $\sigma^2$  sont inconnus.

#### Exercice 9.

Considérons le problème statistique défini par :

$$\begin{cases} X/\theta & \sim & \mathcal{P}(\theta) \\ \theta & \sim & \mathcal{G}(a, b). \end{cases}$$

1. Calculer l'estimateur bayésien  $T$  de  $\theta$ , relativement à la perte quadratique.
2. En procédant à la reparamétrisation de la loi Gamma à l'aide des paramètres  $\lambda$  et  $\mu$ , à définir, écrire la réponse bayésienne en fonction de  $\lambda$  et  $\mu$ .
3. Que représente la réponse bayésienne en fonction de  $\theta$  ?
4. Etudier les cas limites :
  - a.  $\lambda \rightarrow 0$  et  $\lambda + \infty$ , pour  $n$  fixé.
  - b.  $n \rightarrow \infty$ , pour  $\lambda$  et  $\mu$  fixés.

## 5.2 Éléments de réponses

### Exercice 1.

1. On a  $X/\theta \sim \mathcal{B}(n, \theta)$ ,  $\theta \in [0, 1] \Rightarrow P(X = x/\theta) = C_n^x \theta^x (1 - \theta)^{n-x}$ ,  $x \in \{0, 1, \dots, n\}$   
 et  $\theta \sim \mathcal{Be}(\alpha, \beta) \Rightarrow \pi(\theta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{\mathcal{B}(\alpha, \beta)}$ ,  $\theta \in (0, 1)$ .

La loi a posteriori est alors  $\pi(\theta/x) = \frac{P(X=x/\theta)\pi(\theta)}{m(x)}$ , où

$$m(x) = \int_0^1 P(X = x/\theta)\pi(\theta) d\theta = \frac{C_n^x}{\mathcal{B}(\alpha, \beta)} \int_0^1 \theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1} d\theta = \frac{C_n^x}{\mathcal{B}(\alpha, \beta)} \mathcal{B}(x+\alpha, n+\beta-x).$$

D'où,

$$\pi(\theta/x) = \frac{\theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1}}{\mathcal{B}(x+\alpha, n+\beta-x)} \text{ c'est à dire, } \theta/x \rightarrow \mathcal{Be}(x+\alpha, n+\beta-x).$$

2. On a  $X/\theta \sim \mathcal{N}(0, \theta^2)$ ,  $\theta > 0 \Rightarrow f(x/\theta) = \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2}x^2/\theta^2}$ ,  $x \in \mathbb{R}$  et  
 $1/\theta^2 \sim \mathcal{G}(1, 2) \Rightarrow \pi(\theta) = \frac{4}{\theta^3} e^{-2/\theta^2}$ ,  $\theta > 0$ .

La loi marginale de  $X$  est alors  $m(x) = \int_0^{+\infty} f(x/\theta)\pi(\theta) d\theta = \frac{4}{\sqrt{2\pi}} \int_0^{+\infty} \frac{1}{\theta^4} e^{-\frac{1}{\theta^2}(\frac{x^2}{2}+2)} d\theta$ .

En utilisant le changement de variable  $\alpha = \frac{1}{\theta^2}(\frac{x^2}{2} + 2)$  on obtient

$$m(x) = \frac{2}{\sqrt{2\pi(\frac{x^2}{2} + 2)}} \int_0^{+\infty} \alpha^{1/2} e^{-\alpha} d\alpha = \frac{2\Gamma(3/2)}{\sqrt{\pi(x^2 + 4)}}.$$

D'où, la loi a posteriori de  $\theta$  est donnée par

$$\pi(\theta/x) = \frac{\sqrt{2(x^2 + 4)}}{\Gamma(3/2)} \theta^{-4} e^{-\theta^{-2}(\frac{x^2}{2}+2)}, \theta > 0, x \in \mathbb{R}.$$

### Exercice 2.

1. On a  $R(\theta, \delta_1) = \mathbb{E}[(\theta - \delta_1)^2] = \mathbb{E}[(\theta - \frac{X+1}{4})^2] = \frac{1}{16}\mathbb{E}(X^2) + (\frac{1}{8} - \frac{\theta}{2})\mathbb{E}(X) + \theta^2 - \frac{\theta}{2} + \frac{1}{16}$ .  
 Or,  $X \sim \mathcal{B}(2, \theta) \Rightarrow \mathbb{E}(X) = 2\theta$  et  $Var(X) = 2\theta(1-\theta)$ . D'où,

$$R(\theta, \delta_1) = \frac{1}{8}(\theta^2 - \theta + \frac{1}{2}).$$

2. On a  $\sup_{\theta} R(\theta, \delta_1) = \sup_{\theta} (\frac{1}{8}(\theta^2 - \theta + \frac{1}{2}))$ . De plus,

$$\frac{\partial R(\theta, \delta_1)}{\partial \theta} = 0 \Rightarrow \theta = 1/2 \text{ et } \frac{\partial^2 R(\theta, \delta_1)}{\partial \theta^2} = \frac{1}{4} > 0.$$

D'où,  $R(\theta, \delta_1)$  est concave. Le risque est alors minimale pour  $\theta = 1/2$  et vaut  $\frac{1}{32}$ .

On a  $\theta$  est dans l'intervalle  $(0, 1)$ . On obtient  $R(0, \delta_1) = \frac{1}{16}$  et  $R(1, \delta_1) = \frac{1}{16}$ .

D'après le tableau de variation, le risque est maximum et vaut  $\frac{1}{16}$  pour  $\theta \in \{0, 1\}$ .

3. Le risque bayésien est égale à  $r(\pi, \delta_2) = \int_{\Theta} R(\theta, \delta_2)\pi(\theta) d\theta = 1/2$ .

### Exercice 3.

1. On a  $\hat{\theta}_{MV} = \arg_{\theta} \max L(\underline{t}, \theta) = \arg_{\theta} \max \ln L(\underline{t}, \theta)$  avec  $\underline{t} = (t_1, t_2, \dots, t_n)$ .

La vraisemblance  $L(\underline{t}, \theta) = \prod_{i=1}^n f_T(t_i) = \theta^n e^{-\theta \sum_i t_i}$  et  $\ln L(\underline{t}, \theta) = n \ln \theta - \theta \sum_i t_i$ .

De plus,  $\frac{\partial \ln L(\underline{t}, \theta)}{\partial \theta} = \frac{n}{\theta} - \sum_i t_i = 0 \Rightarrow \theta = \frac{n}{\sum_i t_i}$  et  $\frac{\partial^2}{\partial \theta^2} \ln L(\underline{t}, \theta) = -\frac{n}{\theta^2} < 0$ . D'où

$$\hat{\theta}_{MV} = \frac{n}{\sum_i T_i}.$$

2. On a  $\pi(\theta/\underline{t}, \lambda) \propto L(\underline{t}, \theta)\pi(\theta)$  avec  $\pi(\theta) = \lambda e^{-\lambda\theta}$ ,  $\theta > 0$ . D'où,

$$\pi(\theta/\underline{t}, \lambda) \propto \theta^n e^{-\theta(\sum_i t_i + \lambda)} \text{ c'est à dire, } \theta/\underline{t} \sim \mathcal{G}(n+1, \sum_i t_i + \lambda).$$

On obtient donc  $\hat{\theta}_{MMSE} = \mathbb{E}(\theta/\underline{t}) = \frac{n+1}{\sum_i T_i + \lambda}$ .

D'un autre côté on a,

$$\hat{\theta}_{MAP} = \arg_{\theta} \max \pi(\theta/\underline{t}) = \arg_{\theta} \max \ln \pi(\theta/\underline{t}) \text{ et } \pi(\theta/\underline{t}) = \frac{\theta^n (\sum_i t_i + \lambda)^{n+1}}{\Gamma(n+1)} e^{-\theta(\sum_i t_i + \lambda)}.$$

De plus,  $\frac{\partial \ln \pi(\theta/\underline{t})}{\partial \theta} = 0 \Rightarrow \theta = \frac{n}{\sum_i t_i + \lambda}$  et  $\frac{\partial^2}{\partial \theta^2} \ln \pi(\theta/\underline{t}) = -\frac{n}{\theta^2} < 0$ . D'où

$$\hat{\theta}_{MAP} = \frac{n}{\sum_i T_i + \lambda}.$$

3. On a  $\begin{cases} \hat{\theta}_{MV} = \frac{n}{\sum_i T_i} = \frac{1}{\bar{T}}, \quad \bar{T} = \frac{1}{n} \sum_i T_i \\ \hat{\theta}_{MMSE} = \frac{n+1}{\sum_i T_i + \lambda} = \frac{1}{\frac{\sum_i T_i + \lambda}{n+1}} \sim \frac{1}{\bar{T}} \text{ lorsque } n \rightarrow \infty, \hat{\theta}_{MV}, \hat{\theta}_{MMSE} \text{ et } \hat{\theta}_{MAP} \text{ sont équivalents.} \\ \hat{\theta}_{MAP} = \frac{n}{\sum_i T_i + \lambda} = \frac{1}{\frac{\sum_i T_i + \lambda}{n}} \sim \frac{1}{\bar{T}} \end{cases}$

### Exercice 4.

1. On a  $\theta \sim \mathcal{U}_{[0,1]}$ . On pose  $x = 1$  pour le succès et  $x = 0$  pour l'échec, alors

$P(X = 1) = 1 - P(X = 0) = \theta$ . On obtient  $\pi(\theta/x_1, x_2, x_3) \propto \theta^{\sum_i x_i} (1-\theta)^{3-\sum_i x_i} = \theta(1-\theta)^2$ .

D'où,  $\theta/x \sim \mathcal{Be}(2, 3)$ .

La moyenne a posteriori est alors  $\hat{\theta}_{MMSE} = 2/5 = 0.4$ .

2.  $P(0.2 < \theta < 0.4/\underline{X} = \underline{x}) = \int_{0.2}^{0.4} \pi(\theta/\underline{x}) d\theta = 0.344$ .

### Exercice 5.

1. On a  $\pi(\beta/\underline{x}) \propto \beta^6 e^{-\beta(\sum_i x_i + 3)}$ . D'où  $\beta/\underline{x} \sim \mathcal{G}(7, 50)$  et  $\hat{\beta} = \hat{\beta}_{MMSE} = 0.14$ .

2. Dans cette question  $\hat{\beta}$  est la médiane a posteriori, donc vérifiant la relation  $P(\beta \leq \hat{\beta}/\underline{X} = \underline{x}) = 0.5$ .

A partir des valeurs données on a,  $50\hat{\beta} = 6.670$ , ce qui donne  $\hat{\beta} = 0.1334$ .

### Exercice 6.

1.  $\hat{\theta}_{MV} = \arg \max \ln L(\underline{x}, \theta)$ , avec  $\ln L(\underline{x}, \theta) = \sum_i x_i \ln \theta + (n - \sum_i x_i) \ln(1 - \theta)$ .

De plus on a,  $\frac{\partial \ln L(\underline{x}, \theta)}{\partial \theta} = 0 \Rightarrow \theta = \frac{\sum_i x_i}{n}$  et  $\frac{\partial^2}{\partial \theta^2} \ln L(\underline{x}, \theta) = -[\frac{\sum_i x_i}{\theta^2} + \frac{n - \sum_i x_i}{(1-\theta)^2}] < 0$ . D'où  $\hat{\theta}_{MV} = \bar{X}$ .

2. On a  $\int_{\Theta} \pi(\theta) d\theta = \ln \frac{\theta}{(1-\theta)} \Big|_0^1 = +\infty$ , alors  $\pi(\theta)$  est une loi a priori impropre.

3. On a  $\pi(\theta/\underline{x}) \propto \theta^{\sum_i x_i - 1} (1-\theta)^{n - \sum_i x_i - 1}$ ,  $\theta \in (0, 1)$ , d'où  $\theta/\underline{x} \sim \mathcal{Be}(\sum_i x_i, n - \sum_i x_i)$ .

On obtient donc,  $\hat{\theta}_G = \bar{X}$ .

4. L'estimateur de bayes généralisé est défini tant que l'intégrale  $\int_{\Theta} L(\underline{x}, \theta)\pi(\theta) d\theta$  est finie.

On a  $\int_{\Theta} L(x, \theta) \pi(\theta) d\theta = \int_0^1 \theta^{\sum_i x_i - 1} (1 - \theta)^{n - \sum_i x_i - 1} d\theta = \mathcal{B}(\sum_i x_i, n - \sum_i x_i) < \infty$ . D'où,  $\hat{\theta}_G$  est bien défini.

5.  $\hat{\theta}_{MV} = \hat{\theta}_G$ .

### Exercice 7.

1. On a  $f(x, \mu/\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \propto \frac{1}{\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ .

On pose,  $\theta = \sigma^2$  on obtient  $f(x/\theta) \propto (\theta^{-1})^{1/2} e^{-\frac{1}{2}(\theta^{-1})(x-\mu)^2}$ . D'où,  $\theta^{-1} \sim \mathcal{G}(3/2, \frac{1}{2}(x-\mu)^2)$ .

Dans ce cas,  $\theta \sim \mathcal{IG}(3/2, \frac{1}{2}(x-\mu)^2)$ .

2. On a  $f(x/\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \exp(-\beta x + (\alpha-1) \ln x - \ln(\frac{\Gamma(\alpha)}{\beta^\alpha}))$ ,  $x > 0$ . Dans ce cas,  $s = 2$ ,  $T_1(x) = x$ ,  $T_2(x) = \ln x$ ,  $\eta_1(\alpha, \beta) = -\beta$ ,  $\eta_2(\alpha, \beta) = \alpha - 1$ ,  $B(\alpha, \beta) = \ln(\Gamma(\alpha)/\beta^\alpha)$  et  $h(x) = 1$ . La loi Bêta appartient donc à la famille exponentielle.

3. Forme canonique : On a  $f(x/\alpha, \beta) = \exp\{-\beta x + \alpha \ln x - \ln(\frac{\Gamma(\alpha)}{\beta^\alpha})\} \frac{1}{x}$ .

On pose  $\theta = {}^t(\alpha, \beta)$ ,  $T = {}^t(\ln x, -x)$ ,  $A(\theta) = \ln(\Gamma(\alpha)/\beta^\alpha)$  et  $h(x) = \frac{1}{x}$ . Dans ce cas,

$$f(x/\theta) = \exp\left\{\sum_{i=1}^s \theta_i T_i(x) - A(\theta)\right\} h(x), \text{ avec } s = 1.$$

Loi a priori conjuguée naturelle est donnée par,

$$\pi(\theta/\lambda, \mu) \propto \exp\{\theta\mu - \lambda A(\theta)\} \propto \exp\left\{(\alpha, \beta) \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} - \lambda \ln\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\right)\right\} \propto \frac{\beta^{\alpha\lambda} e^{\alpha\mu_1 + \beta\mu_2}}{(\Gamma(\alpha))^\lambda}.$$

La loi a posteriori correspondante est alors :

$$\begin{aligned} \pi(\theta/x) &\propto \exp\{(\mu + x)\theta - (\lambda + 1)A(\theta)\} \\ &\propto \exp\{(\mu_1 + x)\alpha + (\mu_2 + x)\beta - (\lambda + 1)\ln(\Gamma(\alpha)/\beta^\alpha)\} \\ &\propto \frac{e^{(\mu_1+x)\alpha + (\mu_2+x)\beta}}{\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\right)^{(\lambda+1)}}. \end{aligned}$$

### Exercice 8.

1. On a  $\theta = \mu$ ,  $\sigma$  est connu et  $f(x/\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}$ ,  $x \in \mathbb{R}$ .

D'où,  $\ln(f(x/\theta)) = -\ln(\sqrt{2\pi}\sigma) - \frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2$ ,  $\frac{\partial \ln(f(x/\theta))}{\partial \theta} = \frac{1}{\sigma^2}(x - \theta)$  et  $\frac{\partial^2}{\partial \theta^2} \ln(f(x/\theta)) = -\frac{1}{\sigma^2}$ .

On obtient  $I_X(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial \theta^2} \ln(f(x/\theta))\right) = 1/\sigma^2$ .

La loi a priori de Jeffreys est alors donnée par  $\pi_J(\theta) \propto 1$ ,  $\theta \in \mathbb{R}$ .

2. On a  $\theta = \sigma$ ,  $\mu$  connu et  $f(x/\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\theta}\right)^2}$ ,  $x \in \mathbb{R}$ . On obtient,

$\ln(f(x/\theta)) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \frac{1}{2\theta^2}(x - \mu)^2$ ,  $\frac{\partial \ln(f(x/\theta))}{\partial \theta} = -\frac{1}{\theta} + \frac{(x-\mu)^2}{\theta^3}$  et  $\frac{\partial^2}{\partial \theta^2} \ln(f(x/\theta)) = \frac{1}{\theta^2} - 3\frac{(x-\mu)^2}{\theta^4}$ . Le calcul de l'information de Fisher nous donne  $I_X(\theta) = 2/\theta^2$ . La loi a priori de Jeffreys est alors  $\pi_J(\theta) \propto 1/\theta$ ,  $\theta \in \mathbb{R}^+$ .

3. On a  $\theta = (\mu, \sigma^2)$  et  $f(x/\theta) \propto \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ ,  $x \in \mathbb{R}$ . D'où,  $\ln(f(x/\theta)) \propto -\frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2$ .

De plus  $\frac{\partial}{\partial \mu} \ln(f(x/\theta)) = \frac{x-\mu}{\sigma^2}$ ,  $\frac{\partial^2}{\partial \mu^2} \ln(f(x/\theta)) = -\frac{1}{\sigma^2}$  et  $\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln(f(x/\theta)) = -\frac{(x-\mu)}{\sigma^4}$ .

On calcule aussi  $\frac{\partial}{\partial \sigma^2} \ln(f(x/\theta)) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x - \mu)^2$  et  $\frac{\partial^2}{\partial (\sigma^2)^2} \ln(f(x/\theta)) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$ .

D'où la matrice d'information de Fisher  $I(\theta) = (I(\theta)_{i,j})_{i,j} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$ .

On obtient une loi a priori de Jeffreys  $\pi_J(\theta) \propto \frac{1}{\sigma^3}$ ,  $\mu \in \mathbb{R}$  et  $\sigma \in \mathbb{R}^+$ .

**Exercice 9.**

1. On a  $\hat{\theta}_{MMSE} = \frac{n\bar{X}+a}{n+b}$ .

2. On utilise la reparamétrisation de la loi Gamma suivante :  $\begin{cases} \lambda = b \\ \mu = a/b. \end{cases}$

On obtient donc  $\begin{cases} a = \lambda\mu \\ b = \lambda. \end{cases}$  avec  $\mu = \mathbb{E}(\theta)$  et  $Var(\theta) = \mu/\lambda$ .

La réponse bayésienne est alors donnée par  $\mathbb{E}(\theta/\underline{x}) = \frac{n}{\lambda+n}\bar{x} + \frac{\lambda}{\lambda+n}\mathbb{E}(\theta)$ .

3.  $\mathbb{E}(\theta/\underline{x})$  est la moyenne pondérée de  $\bar{X}$  et de  $\mathbb{E}(\theta)$ . Géométriquement,  $\mathbb{E}(\theta/\underline{x})$  est le barycentre des points des coordonnées  $\mathbb{E}(\theta)$  et  $\bar{x}$  affectés des coefficients  $\frac{\lambda}{\lambda+n}$  et  $\frac{n}{n+\lambda}$ , respectivement.

4. Cas limites : pour  $n$  fixé on deux situations,

1. si  $\lambda \rightarrow 0$   $Var(\theta) \rightarrow +\infty$ , le poids de l'a priori est nul et  $\mathbb{E}(\theta/\underline{x}) \rightarrow \bar{x}$  (l'estimation bayésienne tend vers l'estimation classique ; on est dans un cas non informative).
2. Si  $\lambda \rightarrow +\infty$   $Var(\theta) = 0$  (minimale), le poids des données est nul et  $\mathbb{E}(\theta/\underline{x}) \rightarrow \mathbb{E}(\theta)$  (l'estimation bayésienne est indépendante des données ; on est dans un cas extrêmement informatif).

Pour  $\lambda$  et  $\mu$  fixés, si  $n \rightarrow +\infty$  on a,

$$\mathbb{E}(\theta/\underline{x}) \rightarrow \bar{x} \text{ (convergence vers l'estimation classique).}$$

**Examen de Statistique Bayésienne. Année 2014.**  
**Durée 2H.**

**Exercice 1.**

Soit  $X_1, X_2, \dots, X_n$  un  $n$ -échantillon issu d'une v.a.  $X$  de densité définie par :

$$f_{\theta}(x) = (k + 1)\theta^{-(k+1)}x^k 1_{[0,\theta]}(x)$$

où  $k \in \mathbb{N}$  donné et  $\theta > 0$  inconnu. On note  $m = n(k + 1)$ .

1. Montrer que l'estimateur du Maximum de vraisemblance pour le paramètre inconnu  $\theta$  est  $\hat{\theta} = \max_i X_i$ .

2. Calculer la fonction de répartition de  $\hat{\theta}$  en fonction de  $m$ .

3. Calculer  $E(\hat{\theta})$ .

4.  $\hat{\theta}$  est-il sans biais ?

5. Calculer le risque quadratique de  $\lambda\hat{\theta}$  où  $\lambda$  est un réel donné.

6. Trouver la valeur de  $\lambda$  telle que :

a.)  $\lambda\hat{\theta}$  soit sans biais.

b.)  $\lambda\hat{\theta}$  soit de risque minimal.

Calculer le risque obtenu dans chaque cas.

7. On se donne une loi a priori de  $\theta$  de densité

$$\pi_{\delta}(\theta) = \delta\theta^{-2}1_{[\delta,+\infty[}(x), \quad \delta > 0 \text{ fixé.}$$

a.) Montrer que la loi a posteriori est une Pareto dont il faut donner ses paramètres.

b.) Déterminer l'estimateur de Bayes de  $\theta$ , noté  $\hat{\theta}_{MMSE}$ , relativement à la perte quadratique.

c.) Quel est son comportement lorsque  $\delta \rightarrow 0$ .

**[Rappel :** la densité d'une v.a.  $X$  de loi Pareto( $\alpha, \beta$ ) est  $f(x) = \frac{\beta\alpha^{\beta}}{x^{\beta+1}}1_{x \geq \alpha}$ . Dans ce cas on a,  $E(X) = \frac{\beta\alpha}{\beta-1}$ .]

**Exercice 2.**

Soit  $X$  une v.a. discrète de loi Binomiale de paramètres  $n$  et  $p$ ,  $p \in ]0, 1[$ .

1. Donner la loi non informative de Jeffrey pour  $p$ .

2. Considérons une loi a priori impropre pour  $p$ , dite de Haldane, définie par

$$\pi(p) = [p(1-p)]^{-1}, \quad p \in ]0, 1[.$$

Déterminer l'estimateur de Bayes généralisé pour  $p$ , relativement à la perte quadratique.

Bon courage...

Université de Béjaia  
Faculté des Sciences Exactes  
Département des Mathématiques.  
Master 1 SAD.

**Examen de Statistique Bayésienne.**  
**Année 2015. Durée : 2h.**

**Exercice 1.**

Soit  $X_1, X_2, \dots, X_n$  un  $n$ -échantillon issu d'une v.a.  $X$  de loi de Poisson de paramètre inconnu  $\theta$  ayant pour loi a priori  $\pi(\theta)$  de type  $\mathcal{G}(\alpha, \beta)$ ,  $\alpha, \beta > 0$ .

1. Montrer que la loi Gamma est une loi conjuguée naturelle de la loi de Poisson.
2. Déterminer l'estimateur bayésien  $T_n$  de  $\theta$ , relativement à la perte quadratique.
3. Calculer le risque quadratique de  $T_n$ .
4. Calculer le risque bayésien de  $T_n$ , relativement à la loi a priori  $\pi(\theta)$ .

**Exercice 2.**

Soit  $X_1, X_2, \dots, X_n$  un  $n$ -échantillon issu d'une v.a.  $X$  de loi Paréto de paramètre  $\theta$ ,  $\theta > 0$ , de densité

$$f(x, \theta) = \frac{\theta}{x^{\theta+1}} 1_{[1, +\infty[}(x).$$

1. Calculer l'estimateur  $\hat{\theta}_{MLE}$  de  $\theta$  par la méthode de maximum de vraisemblance.
2. Calculer la loi a priori non informative de Jeffreys pour  $\theta$  et déterminer dans ce cas l'estimateur de bayes  $\hat{\theta}_J$ , relativement à la perte quadratique.

Supposons dans la suite que nous disposons d'une information a priori sur le paramètre  $\theta$  résumée dans la densité a priori

$$\pi(\theta) = \lambda e^{-\lambda\theta} 1_{[0, +\infty[}(\theta).$$

3. Montrer que la densité à posteriori de  $\theta$  est la densité d'une loi Gamma dont on déterminera les paramètres.
4. Déterminer les estimateurs  $\hat{\theta}_{MAP}$  et  $\hat{\theta}_{MMSE}$  de  $\theta$ .
5. Comparer les estimateurs  $\hat{\theta}_{MLE}$ ,  $\hat{\theta}_{MAP}$  et  $\hat{\theta}_{MMSE}$  pour une taille d'échantillon  $n$  assez grande.

Bon courage...

# Bibliographie

- [1] T. Bayes, An Essay towards Solving a problem in the Doctrine of chance. *Phil. Trans. Roy. Soc.*, 53 : 370-418, (1763).
- [2] J. Berger, Statistical Decision Theory and Bayesian Analysis. *Springer-Verlag, New York, 2 Edition.*, (1985).
- [3] J.M. Bernardo and A.F.M. Smith, Bayesian Theory, *Wiley Series in Probability and Mathematical statistics*, (1994).
- [4] S. Chib and E. Greenberg, Understanding the Metropolis Hastings algorithm, *The American Statistician*, 49 :327-335, (1995).
- [5] S. Chib and E. Greenberg, Markov Chain Monte Carlo Simulation Methods in Econometrics, *Econometric Theory*, 12 :409-413, (1996).
- [6] R. Delyon, Estimation non paramétrique, *Cours de Master2, IRMAR, Université Rennes I, France*, (2015).
- [7] R. Fletcher, Practical Methods of Optimisation. *John Wiley and Sons, 2 edition*, (1987).
- [8] A.E. Gelfand and A.F.M. Smith, Sampling-based Approaches to calculating Marginal Densities. *JAMSTAT*, 85(410) :398-409, (1990).
- [9] S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions and the Bayesian Resoration of Images. *IEEE Trans. on Pattern Anal.Mach.Intell.*, 6 :721-740, (1984).
- [10] W. Hastings, Monte Carlo Sampling Methods using Markov Chains and their Applications, *Biometrika*, 57 :97-109, (1970).
- [11] H. Jeffreys, Theory of Probability. *Oxford University Press, 3rd edition*, (1961).
- [12] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equations of State Calculations by Fast Computing Machine, *J. Chem. Phys.*, 21 :1087-1093, (1953).
- [13] S.J. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes : The Art of Scientific Computing. *Combridge University Press*, (1986).
- [14] A.P. Quinn, Bayesian Point Inference in Signat Processing. *PHD Thesis, Combridge University Engineering Dept.*, (1992).
- [15] H. Raiffa and R. Schlaifer, Applied Statistical Decision Theory. *Boston Massachussetts : Hawaed Business School*, (1961).
- [16] C.P. Robert, L'analyse Statistique Bayésienne. *Economica*, (1992).
- [17] C.P. Robert, The Bayesian Choice. *Springer-Verlag*, (1994).

- [18] C.P. Robert, Méthodes de Monte-Carlo par Chaînes de Markov. *Economica*, (1996).
- [19] C.P. Robert, Le choix bayésien. Principes et pratique. *Ed. Springer-Verlag, France, Paris.*, (2006).
- [20] A.F.M. Smith, A.M. Skene, J.E.H. Shaw, J.C. Naylor and M. Dransfield, The implementation of the Bayesian Paradigm. *Communications in Statistics*, 14 : 1079-1102, (1985).
- [21] M. Tanner and W. Wong, The calculation of Posterior Distributions by Data Augmentation, *Journal of the American statistical Association*, 82 :528-550, (1987).

## Sites Internet

- Statistique Bayésienne et algorithmes MCMC. Cours IMAT (Master1) de J. Dupuis, Septembre 2007.  
[http://www.math.univ-toulouse.fr/dupuis/bayes\\_mim08.pdf](http://www.math.univ-toulouse.fr/dupuis/bayes_mim08.pdf)
- Statistique Bayésienne. Cours destiné au Master2 Ingénierie mathématique option : Statistique de Anne Philippe, 2007.  
<http://www.math.sciences.univ-nantes.fr/philippe/download/cours-bayes-2007.pdf>
- Cours et exercices de Anne Philippe.  
<http://www.math.sciences.univ-nantes.fr/philippe/Enseignement.html>
- Statistique Bayésienne. Support de Cours. Rédigé par Mathias André et Alexis Eidelman.  
[http://www.crest.fr/ckfinder/userfiles/files/pageperso/mandre/SB\\_cours.pdf](http://www.crest.fr/ckfinder/userfiles/files/pageperso/mandre/SB_cours.pdf)
- Bayesian Statistics. Cours de Christian P. Robert, Université Paris-Dauphine et CREST-INSEE, 2006.  
<http://www.ceremade.dauphine.fr/xian/coursBC.pdf>