Cours 3

# 4 Lois a priori

Le choix des lois a priori est une étape fondamentale dans l'analyse bayésienne. Ce choix peut avoir différentes motivations. Les stratégies sont diverses. Elle peuvent se baser sur des expériences du passé ou sur un intuition, une idée que le particien a du phénomène aléatoire qu'il est en train de suivre. Elles peuvent être également motivées par des aspects calculabilité . Enfin, ces stratégies peuvent également tenir compte du fait qu'on ne sait rien par le truchement des lois non informatives.

## 4.1 Lois conjuguées

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

**Définition 9** – *Une famille $\mathcal{F}$ de lois sur $\Theta$ est dite conjuguée si, pour tout $\pi$ appartenant à cette famille, la loi $\pi(\theta/x)$ appartient également à celle-ci.*

Dans ce cas, le praticien induit directement la forme de son estimateur dès qu'il a choisi sa loi a priori.

**Exemples de lois conjuguées**

| $f(x/\theta)$ | $\pi(\theta)$ | $\pi(\theta/x)$ | $E(\theta/x)$ |
|---|---|---|---|
| $\mathcal{N}(\theta, \sigma^2)$ | $\mathcal{N}(\mu, \tau^2)$ | $\mathcal{N}\left(x/\sigma^2 + \mu/\tau^2\ ,\ [1/\sigma^2 + 1/\tau^2]^{-1}\right)$ | $x/\sigma^2 + \mu/\tau^2$ |
| $\mathcal{G}(n, \theta)$ | $\mathcal{G}(\alpha, \beta)$ | $\mathcal{G}(\alpha + n, \beta + x)$ | $(\alpha + n)/(\beta + x)$ |
| $\mathcal{B}(n, \theta)$ | $\text{Bêta}(\alpha, \beta)$ | $\text{Bêta}(\alpha + n, \beta + x)$ | $(\alpha + n)/(\alpha + n + \beta + x)$ |
| $\mathcal{P}(\theta)$ | $\mathcal{G}(\alpha, \beta)$ | $\mathcal{G}(\alpha + x, \beta + 1)$ | $(\alpha + x)/(\beta + 1)$ |
| $\mathcal{M}(\theta_1, \theta_2, ..., \theta_n)$ | $\mathcal{D}(\alpha_1, \alpha_2, ..., \alpha_n)$ | $\mathcal{D}(\alpha_1 + x_1, \alpha_2 + x_2, ..., \alpha_n + x_n)$ | |

Une loi conjuguée peut être déterminée en considérant la forme de la vraisemblance $f(x|\theta)$ et en prenant une loi a priori de la même forme que cette dernière vue comme une fonction du paramètre. Les lois a priori conjuguées obtenues par ce procédé sont dites **naturelles**.

**Exemples :**

- Considérons une loi Pareto de paramètres $(\alpha, a)$.

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

Supposons $a$ connu, $f(x|\theta) \propto \theta e^{\theta \log(a/x)}$, on pourrait donc prendre une loi a priori de type gamma.

- Dans le cas d'une loi binomiale négative de paramètre $(n, p)$,

$$P(X = x \mid p\ ) = C^x_{n+x-1}\, p^x (1 - p\ )^n,\ \ 0 < p < 1,\ \ x \in \mathbb{N}.$$

On voit clairement qu'une loi naturelle conjuguée sera une loi bêta puisque : $P(X = x \mid p\ ) \propto p^x (1 - p\ )^n$.

□

Etudions maintenant en détail le cas d'une famille de lois très importante : la famille des lois exponentielles.

## 4.2 Cas du modèle exponentiel

On rappelle tout d'abord la définition du modèle exponentielle.

**Définition 10** *On appelle famille exponentielle à s-paramètres, toute famille de loi de distribution $\{P_\theta\}$ dont la densité a la forme suivante :*

$$f(x|\theta) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x) \qquad (2)$$

*où $\eta_i(.)$ et $B(.)$ sont des fonctions du paramètre $\theta$ et $T_i(.)$ sont des statistiques.*

**Exemples :**

- Loi exponentielle :

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{\theta}\, e^{-x/\theta}\, \mathbb{1}_{[0,+\infty[}(x) \\ &= \exp\left\{-\frac{1}{\theta}\, x - \log\theta\right\} \mathbb{1}_{[0,+\infty[}(x). \end{aligned}$$

Ici, $s$ vaut 1, $\eta_1(\theta) = 1/\theta$, $T_1(x) = x$, $B(\theta) = \log\theta$ et $h(x) = \mathbb{1}_{[0,+\infty[}(x)$.

- Loi binomiale :

$$\begin{aligned} P(X = x \mid \theta) &= C^x_n \theta^x (1-\theta)^{n-x} \\ &= C^x_n \exp\{x\log\theta + (n-x)\log(1-\theta)\} \\ &= C^x_n \exp\{x\log[\theta/(1-\theta)] + n\log(1-\theta)\} \end{aligned}$$

On a $s = 1$, $\eta_1(\theta) = \log(\theta/(1-\theta))$, $T_1(x) = x$, $B(\theta) = n\log(1-\theta)$ et $h(x) = C^x_n$.

- Loi Gamma $f(x|\alpha,\beta) = (\beta^\alpha/\Gamma(\alpha))x^{\alpha-1}e^{-\beta x}$.
  En remarquant que : $x^{\alpha-1}e^{-\beta x} = \exp\{-\beta x + (\alpha-1)\log x\}$ et $s = 2$, on peut écrire : $\eta_1(\alpha,\beta) = -\beta$ , $\eta_2(\alpha,\beta) = \alpha - 1$, $T_1(x) = x$, $T_2(x) = \log x$,$B(\alpha,\beta) = \log(\Gamma(\alpha)/\beta^\alpha$.

$\square$

Il est classique d'écrire le modèle exponentiel sous la forme dite **canonique** en reparamétrisant : $\eta_i(\theta) \equiv \theta_i$.

$$f(x \mid \theta) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(x) - A(\theta)\right] h(x).$$

On a le résultat suivant qui donne la forme des *lois naturelles conjuguées* dans le cas du modèle exponentiel.

**Proposition 4** – *Soit $f(x \mid \theta)$ appartenant à une famille exponentielle. Alors une famille de loi a priori conjuguée pour $f(x \mid \theta)$ est donnée par :*

$$\pi(\theta \mid \mu, \lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda A(\theta))$$

*où $K(\mu,\lambda)$ est une constante de normalisation.*
*Et la loi a posteriori est de la forme :*

$$\pi(\theta \mid x) \propto \exp((\mu + x)\theta - (\lambda+1)A(\theta)).$$

**Exemple :** Considérons le modèle :

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

Il s'agit d'une loi logistique. Elle appartient bien à la famille exponentielle.
On a :

$$P(X = x) = \exp[(\theta - \beta)x - \log(1 + e^{\theta-\beta})]$$

et la représentation : $h(x) = 1$, $\theta = [\theta \quad \beta]'$, $T(x) = [x \quad -x]'$
et $A(\theta,\beta) = \log(1 + e^{\theta-\beta})$.
En appliquant le théorème, on obtient une loi a priori de la forme :

$$\pi(\theta,\beta \mid \mu_1,\mu_2,\lambda) \propto \exp\{[\theta \quad \beta][\mu_1 \quad \mu_2]' - \lambda A(\theta,\beta)] = \frac{e^{\mu_1\theta+\mu_2\beta}}{(1 + e^{\theta-\beta})^\lambda}.$$

On remarquera que cette loi est impropre. La loi a posteriori aura la forme suivante :

$$\pi(\theta,\beta \mid x) \propto \frac{e^{(\mu_1+x)\theta+(\mu_2+x)\beta}}{(1 + e^{\theta-\beta})^{\lambda+1}}.$$

**Proposition 5** – *Soit $f(x \mid \theta)$ appartenant à une famille exponentielle. Si $\theta \sim \exp\{\theta x_0 - \lambda A(\theta)\},\ x_0 \in \mathfrak{X}$,*
*Alors*

$$E^{\pi}(\mu(\theta)) = E^{\pi}[\nabla A(\theta)] = \frac{x_0}{\lambda}$$

*où $\mu(\theta)$ est la moyenne de $f(x|\theta)$.*
*Si $x_1, x_2, \cdots, x_n$ iid de loi $f(x|\theta)$,*

$$E^{\pi}(\mu(\theta)|x_1, \cdots, x_n) = \frac{x_0 + n\bar{x}}{\lambda + n}.$$

La preuve de ce théorème s'appuie sur les propriétés de l'emv.

### 4.3 Lois impropres

La loi a priori peut être impropre i.e. $\displaystyle\int_{\Theta} \pi(\theta) d\theta = +\infty$. Ce choix de type de loi n'a donc plus d'intérêt que calculatoire et s'interprète difficilement. Nous verrons par la suite que la construction de lois non informative peut conduire à des lois a priori de ce type.

**Exemple :** Considérons la loi $\mathbb{1}_{\mathbb{R}^+}(x)$, loi uniforme sur $\mathbb{R}^+$. Supposons $f(x|\lambda) = \lambda \exp\{-\lambda x\}$.
La loi a posteriori est : $\pi(\lambda|x) = \lambda \exp\{-\lambda x\} / \int_0^{+\infty} \lambda e^{-\lambda x}$.
Le dénominateur est égal à $\Gamma(2)/x^2$ d'où $\pi(\lambda|x) = [x^2/\Gamma(2)]\lambda \exp\{-\lambda x\}$, une loi gamma de paramètre $(2, x)$.

### 4.4 Lois non informatives

Une loi non informative est une loi qui porte une information sur le paramètre à estimer dont le poids dans l'inférence est réduit. Certains auteurs la définissent également comme une loi a priori qui ne contient aucune information sur $\theta$ ou encore comme une loi qui ne donne pas davantage de poids à telle ou telle valeur du paramètre.
Par exemple, supposons $\Theta$ un ensemble fini de taille $q$, une loi a priori non informative pourra être une loi de la forme :

$$P(\theta_i) = 1/q.$$

On a équiprobabilité, les valeurs possibles de $\theta$ se voit attribuer le même poids.

**La règle de Jeffreys** – Une méthode proposée par Jeffreys (1961) permet de fabriquer des lois a priori non informative. Cette méthode utilise l'information de Fischer : $I(\theta)$. L'argument pourrait être le suivant. $I(\theta)$ représente une mesure de la quantité d'information sur $\theta$ contenue dans l'observation. Plus $I(\theta)$ est grande, plus l'observation apporte de l'information. Il semble alors naturel de favoriser (au sens rendre plus probable suivant $\pi(\theta)$), les valeurs de $\theta$ pour lesquels $I(\theta)$ est grande ; ce qui minimise l'influence de la loi a priori au profit de l'observation.

Le choix de ce type de loi conduit ainsi souvent à des estimateurs classiques du type maximum de vraisemblance.
La règle de Jeffreys consiste donc à considérer des lois a priori de la forme :

$$\pi(\theta) = C\sqrt{I(\theta)} \quad \text{où} \quad I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta)\right],$$

dans le cas unidimensionnel.

**Rappels sur l'information de Fisher –** Soit un $n$-échantillon $(X_1, \cdots, X_n)$ de loi $f(x \mid \theta)$. On a le résultat suivant. Sous certaines conditions de régularité, l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta$ est tel que, pour $n$ assez grand, $\sqrt{n}(\hat{\theta} - \theta)$ suit une loi normale centrée, de variance $1/I(\theta)$.

Autre résultat important : l'*inégalité de Cràmer-Rao.*
Soit $t_n$, un estimateur de $g(\theta)$, alors :

$$E(t_n - g(\theta))^2 \geq \frac{(g'(\theta) + b_n'(\theta))^2}{nI(\theta)},$$

où $b_n(\theta) = E(t_n) - g(\theta)$, est le biais.
Si $g(\theta) = \theta$ et si l'estimateur est sans biais, on a :

$$Var(t_n) \geq \frac{1}{nI(\theta)}.$$

Ces résultats illustrent l'idée que l'information de Fisher se compare à une variance.

□

**Exemple :** Soit $f(x \mid \theta) = \lambda \exp\{-\lambda x\}\mathbb{1}_{[0,+\infty[}(x)$. On calcule l'information de Fischer :

$$\frac{\partial}{\partial\lambda}\log f(x|\lambda) = \frac{1}{\lambda} - x, \quad \frac{\partial^2}{\partial\lambda^2}\log f(x \mid \lambda) = -\frac{1}{\lambda^2}$$

d'où $\pi(\lambda) \propto 1/\lambda$.

□

Ce type de construction conduit très souvent à des lois impropres, des lois telles que : $\int_\Theta \pi(\theta)d\theta = +\infty$.
On appelle également ces lois des *quasi a priori.*
Remarquons que l'on peut normaliser ces lois (à condition de se donner de l'information sur $\theta$, ce qui est, dans ce cadre non informatif, certes paradoxale).

**Exemple (suite) :** $\pi(\lambda) \propto 1/\lambda$, si on prend $\lambda \in [a,b]$ ; on cherche $C$ tel que : $C \int_a^b \pi(\theta) d\theta = 1$ et $\int_a^b 1/\lambda d\lambda = \log b - \log a$.
Ainsi, on pourra prendre : $\pi(\lambda) = [\log(b/a)]^{-1} \mathbb{1}_{[a,b]}(\lambda)$.

□

Dans le cas multidimensionnel, $\theta = (\theta_1, \cdots, \theta_q)$, on peut considérer des lois de la forme :

$$\pi(\theta) = [d\acute{e}tI(\theta)]^{1/2}$$

où $I(\theta)$ est la matrice d'information de Fischer :

$$I(\theta) \mid_{i,j} = E\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x \mid \theta)\right], \ 1 \leq i, j \leq q.$$

Remarquons que dans le cadre de la théorie du maximum de vraisemblance, le déterminant de cette quantité représente ce qu'on appelle la variance généralisée.

**Exemple :** Dans le cas de la loi de Gauss de paramètres $(\theta, \sigma^2)$,

$$f(x \mid \theta, \sigma^2) \propto \frac{1}{\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}$$

On a :

$$\log f(x \mid \theta, \sigma^2) \propto -\frac{1}{2} \log \sigma^2 - \frac{(x-\theta)^2}{2\sigma^2}$$

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

La matrice d'information de Fisher s'obtient en calculant, l'espérance mathématique des dérivées secondes. $E(x-\theta) = 0$ et $E[(x-\theta)^2] = \sigma^2$.
On a donc :

$$I(\theta, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^2 \end{pmatrix}$$

Un loi informative est donc de la forme : $\pi(\theta) \propto 1/\sigma^2$.

□