Cours 4

4.5 Lois a priori d'entropie maximum

L'entropie est une grandeur bien connue des physiciens comme étant une mesure du *désordre*. Dans le cadre de la statistique, elle mesure la quantité d'incertitude inhérente à la loi de probabilité.

Définition 11 – *Soit Θ espace de paramètres discret, soit π une probabilité sur Θ. L'entropie de π – notée $\mathcal{E}_q(\pi)$ – est définie par :*

$$\mathcal{E}_q(\pi) = -\sum_{\Theta} \pi(\theta_i) \log \pi(\theta_i).$$

Convention : si $\pi(\theta_i) = 0$ alors $\pi(\theta_i) \log \pi(\theta_i) = 0$

Exemple : On considère : $\Theta = \{\theta_1, \cdots, \theta_q\}$.
On pose : $\pi(\theta) = 1$ si $\theta = \theta_k$ et $\pi(\theta) = 0$ pour tout $theta = \theta_i$, $i \neq k$.
La loi a priori donne exactement la valeur de θ ; l'incertitude est donc ici nulle et $\mathcal{E}_q(\pi) = 0$.
Si maintenant $\pi(\theta_i) = 1/q$ pour tout i alors

$$\mathcal{E}_q(\pi) = -\sum_{i=1}^{q} 1/q \log(1/q) = \log q.$$

Ce qui correspond à l'incertitude la plus grande. En effet, on peut montrer que : $\mathcal{E}_q(\pi) \leq \log q$ pour tout π non dégénéré. π est appelée la *loi d'entropie maximum.*
Supposons que l'on dispose d'informations a priori concernant θ telles que l'on puisse écrire :

$$E^{\pi}[g_k(\theta)] = \sum_{i=1}^{q} g_k(\theta_i)\pi(\theta_i) = \mu_k \ , \quad k = 1, \cdots, m. \tag{3}$$

Une loi a priori d'entropie maximum est une solution d'un problème d'optimisation sous la contrainte (**??**) :

$$\overline{\pi} = \underset{\pi}{Argmax}\ \mathcal{E}_q(\pi).$$

Si π est propre ($\sum \pi(\theta_i) = 1$), on montre que la solution est de la forme :

$$\overline{\pi}(\theta_i) = \frac{\exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}}{\sum_{\Theta} \exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}}.$$

Exemple : $\Theta = \mathbb{N}$. Supposons $E^{\pi}(\theta) = 5$, $m = 1$. On a : $g_1(\theta) = \theta$ et $\mu_1 = 5$. Supposons $\lambda_1 < 0$,

$$\bar{\pi}(\theta) = \frac{e^{\lambda_1 \theta}}{\sum_{\theta=0}^{+\infty} e^{\lambda_1 \theta}} = (1 - e^{\lambda_1})(e^{\lambda_1})^{\theta}.$$

On reconnaît une loi géométrique.
On détermine alors λ_1 par :

$$E^{\overline{\pi}}(\theta) = \frac{e^{\lambda_1}}{1-e^{\lambda_1}} = 5 \Longleftrightarrow e^{\lambda_1} = 5/6 \Longleftrightarrow \lambda_1 = \log(5/6)$$

$\overline{\pi}$ est donc une loi géométrique de paramètre $5/6$.

□

Si Θ est continu, on peut proposer la définition suivante de l'entropie :

$$\mathcal{E}(\pi) = -E^{\pi}\left[\log \frac{\pi(\theta)}{\pi_0(\theta)}\right] = -\int \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$$

où $\pi_0(\theta)$ est la loi a priori non informative naturelle pour le problème.
Comme précédemment, si on dispose d'information a priori du type :

$$E^{\pi}[g_k(\theta)] = \int_{\Theta} g_k(\theta)\pi(\theta)d\theta = \mu_k \ , \quad k = 1, \cdots, m,$$

la loi a priori du maximum d'entropie est alors donnée par :

$$\overline{\pi}(\theta_i) = \frac{\pi_0(\theta)\exp\left\{\sum_{k=1}^m \lambda_k g_k(\theta)\right\}}{\int_{\Theta}\pi_0(\theta)\exp\left\{\sum_{k=1}^m \lambda_k g_k(\theta)\right\} d\theta},$$

où λ_k sont des constantes obtenues par la contrainte.

Exemple : $\Theta = \mathbb{R}$. Supposons que θ est un paramètre de position. La loi a priori naturel non informative est alors $\pi_0(\theta) = 1$.
Si on fixe les valeurs de la moyenne et de la variance de la loi a priori à (μ, σ^2), on a $E^{\pi}(\theta) = \mu$ et $g_1(\theta) = \theta$, $\mu_1 = \mu$, et $E^{\pi}[(\theta-\mu)^2] = \sigma^2$ et $g_2(\theta) = (\theta-\mu)^2$, $\mu_2 = \sigma^2$.
La loi du maximum d'entropie est alors :

$$\bar{\pi}(\theta) = \frac{\exp\{\lambda_1\theta + \lambda_2(\theta-\mu)^2\}}{\int_{\Theta}\exp\{\lambda_1\theta + \lambda_2(\theta-\mu)^2\}d\theta}.$$

Calculons λ_1, λ_2. On montre sans difficultés que $\lambda_1\theta + \lambda_2(\theta-\mu)^2 \propto \lambda_2[\theta - (\mu - \lambda_1/2\lambda_2)]^2$.
$\overline{\pi}(\theta)$ est donc une loi normale de paramètres $[(\mu - \lambda_1/2\lambda_2); -1/2\lambda_2]$.
On cherche alors λ_1 et λ_2 tels que : $\mu - \lambda_1/2\lambda_2 = \mu$ et $\sigma^2 = -1/2\lambda_2$. Il vient donc : $\lambda_1 = 0$ et $\lambda_2 = -1/2\sigma^2$ d'où $\bar{\pi}(\theta)$ est un loi normale de paramètres (μ, σ^2).

4.6 Modèle hiérarchique

Le choix de $\pi(\theta)$ dans sa forme est une chose. Ce choix s'accompagne également d'un choix de valeurs pour les paramètres de cette distribution. On peut parvenir à fonder ce choix sur des

considérations pratiques s'appuyant sur la nature de l'expérience qui génère l'observation.
On peut alors pousser le paradigme bayésien en considérant des lois a priori sur les paramètres – appelés *hyperparamètres* – fabriquant ainsi un *modèle hierarchique.*
Le modèle se caractérise par la donnée d'une distribution de l'observation $f(x|\theta)$, la donnée d'une loi a priori $\pi(\theta|\theta_1)$ sur θ et la donnée d'une loi a priori sur θ_1 : $\pi(\theta_1)$.

Exemple : Soit $X \mid \theta \sim \mathcal{P}(\theta)$. Considérons une loi exponentielle a priori de paramètre λ sur θ. La démarche hiérarchique nous conduit à considérer alors une loi a priori sur λ ; On peut prendre par exemple une loi exponentielle de paramètre ξ.
On a donc les lois suivantes : $\pi(\theta \mid \lambda) = \lambda\ e^{-\lambda\theta}$ et $\pi(\lambda) = \xi\ e^{-\xi\lambda}$.

□

On peut bien évidemment continuer à emboîter la démarche bayésienne.
D'une manière générale, on donne la définition suivante :

Définition 12 – *On appelle modèle bayésien hiérarchique, un modèle statistique bayésien où la loi a priori est décomposée en distributions conditionnelles :*

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

Les paramètres θ_i, $i = 1, \cdots, k$, sont appelés hyperparamètres.

L'analyse de ce type de modèle rejoint le cas standard.
On cherche à calculer la loi a posteriori $\pi(\theta \mid x)$ et cette loi s'obtient en appliquant le théorème de Bayes et en remarquant que la loi a priori s'écrit :

$$\pi(\theta) = \int_{\Theta_1\times\cdots\times\Theta_k} \pi(\theta \mid \theta_1)\ \pi(\theta_1 \mid \theta_2)\ \cdots\ \pi(\theta_{k-1} \mid \theta_k)\ \pi(\theta_k) d\theta_1 \cdots d\theta_k.$$

Considérons le cas d'un modèle hiérarchique simple : $X \sim f(x \mid \theta)$,
$\theta \sim \pi(\theta \mid \theta_1)$ et $\pi(\theta_1)$.
La loi a posteriori s'écrit :

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

On a :

$$\pi(\theta) = \int_{\Theta_1} \pi(\theta \mid \theta_1)\pi(\theta_1) d\theta_1$$

La predictive aura donc pour expression :

$$f(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta) d\theta = \int_{\Theta}\int_{\Theta_1} f(x \mid \theta)\pi(\theta \mid \theta_1)\pi(\theta_1) d\theta_1 d\theta.$$

Autrement dit :

$$\pi(\theta \mid x) = \frac{\int_{\Theta_1} f(x \mid \theta)\pi(\theta \mid \theta_1)\pi(\theta_1)d\theta_1 d\theta}{\int_{\Theta}\int_{\Theta_1} f(x \mid \theta)\pi(\theta \mid \theta_1)\pi(\theta_1)d\theta_1 d\theta}. \quad (4)$$

L'intérêt essentiel de la hiérarchisation est de lever certaines difficultés calculatoires pour obtenir la loi a posteriori. En effet, dans certaine situation la loi a priori conduit à des lois a posteriori difficile à "manipuler". Utiliser la décomposition en loi conditionnelle peut permettre de lever ces difficultés.

Dans le cas d'une hiérarchisation simple, on peut énoncer le résultat suivant :

Proposition 6 – *Si :* $\theta \mid \theta_1 \sim \pi(\theta \mid \theta_1)$ *et* $\theta_1 \sim \pi(\theta_1)$.
La loi a posteriori de θ *sachant* x *est :*

$$\pi(\theta \mid x) = \int_{\Theta_1} \pi(\theta \mid \theta_1, x)\pi(\theta_1 \mid x)d\theta_1,$$

$$\textit{avec} \quad \pi(\theta \mid \theta_1, x) = \frac{f(x \mid \theta)\pi(\theta \mid \theta_1)}{f(x \mid \theta_1)} \quad \textit{où} \quad f(x \mid \theta_1) = \int_{\Theta} f(x \mid \theta)\pi(\theta \mid \theta_1)d\theta$$

$$\textit{etavec} \quad \pi(\theta_1 \mid x) = \frac{f(x \mid \theta_1)\pi(\theta_1)}{f(x)} \quad \textit{où} \quad f(x) = \int_{\Theta_1} f(x \mid \theta_1)\pi(\theta_1)d\theta_1.$$

Preuve : En remplaçant $\pi(\theta \mid \theta_1, x)$ et $\pi(\theta_1 \mid x)$ par leur expression sous l'intégrale, il vient :

$$\begin{aligned}
\pi(\theta \mid x) &= \int_{\Theta_1} \frac{f(x \mid \theta)\pi(\theta \mid \theta_1)}{f(x \mid \theta_1)} \cdot \frac{f(x \mid \theta_1)\pi(\theta_1)}{f(x)} d\theta_1 \\
&= \int_{\theta_1} \frac{f(x \mid \theta)\pi(\theta \mid \theta_1)\pi(\theta_1)}{f(x)} d\theta_1 \\
&= \frac{f(x \mid \theta)}{f(x)} \int_{\Theta_1} \pi(\theta \mid \theta_1)\pi(\theta_1)d\theta_1 \\
&= \frac{f(x \mid \theta)\pi(\theta)}{f(x)}
\end{aligned}$$

□