Chapitre 8

Calcul bayésien

Nous avons vu qu'un estimateur bayésien $\delta(\boldsymbol{x})$ associé à une fonction de vraisemblance $l(\theta|\boldsymbol{x})$, à une fonction coût $C(\theta, \delta)$ et à une loi *a priori* $p(\theta)$ est celui qui minimise le coût moyen *a posteriori*,

$$\bar{C}(\delta) = \int_\Theta C(\theta, \delta) p(\theta|\boldsymbol{x}) \, \mathrm{d}\theta; \tag{8.1}$$

et l'estimé optimale $\widehat{\theta}$ s'obtient par

$$\widehat{\theta} = \delta(\boldsymbol{x}) = \arg \min_{\delta \in \mathcal{D}} \{ \bar{C}(\delta) \}$$

En pratique deux difficultés peuvent apparaître dans ce problème :

- le calcul explicite de la loi *a posteriori*

  $$p(\theta|\boldsymbol{x}) = \frac{1}{Z} \, l(\theta|\boldsymbol{x}) p(\theta) \propto g(\theta|\boldsymbol{x})$$

  peut être impossible. En général, c'est surtout le calcul du constant de normalisation Z est difficile, car cela nécessite le calcul de l'intégrale

  $$Z = \int_\Theta l(\theta|\boldsymbol{x}) p(\boldsymbol{\theta}) \, \mathrm{d}\theta$$

  qui admet rarement une solution explicite.
- Même quand $p(\theta|\boldsymbol{x})$ est connue, le calcul du coût moyen et de sa minimisation nécessite une puissance de calcul considérable.

Le premier point peut sembler mineur car minimiser (8.1) revient à minimiser

$$\bar{C}(\delta) = \int_\Theta C(\theta, \delta) \, l(\theta|\boldsymbol{x}) \, p(\boldsymbol{\theta}) \, \mathrm{d}\theta; \tag{8.2}$$

qui, apparement ne nécessite pas le calcul du constant Z. Mais en fait, même quand on utilise les fonctions coûts classiques on est conduit à des estimateurs exprimés en fonction de la loi $p(\theta|\boldsymbol{x})$. Par exemple le coût quadratique nous conduit à un estimateur qui correspond au calcul de la moyenne *a posteriori*

$$\begin{aligned} \delta(\boldsymbol{x}) &= \int_\Theta \theta p(\theta|\boldsymbol{x}) \, \mathrm{d}\theta \\ &= \frac{\displaystyle\int_\Theta \theta l(\theta|\boldsymbol{x}) p(\boldsymbol{\theta}) \, \mathrm{d}\theta}{\displaystyle\int_\Theta l(\theta|\boldsymbol{x}) p(\boldsymbol{\theta}) \, \mathrm{d}\theta}, \end{aligned}$$

qui nécessite encore le calcul de Z. Dans tous les cas on est amené soit à calculer un intégrale simple soit à calculer le rapport de deux intégrales simples. Dans les paragraphes qui suivent nous allons considérer un certain nombre de techniques utilisées pour ce type de calcul.

8.1 Intégration numérique

Le principe de ces méthodes est basée sur l'approximation suivante :

$$I = \int_a^b f(x)\,\mathrm{d}x \simeq \sum_{i=1}^n w_i f(x_i)$$

Lorsqu'il est possible de définir les limites $[a, b]$ de l'intégrale cette méthode peut être utilisée. Tous l'art est ensuite dans le choix des coéfficients w_i et les points x_i.
Lorsque ces limites sont infinies, et sous certaines conditions de régularités de la fonction $f(x)$ on peut utiliser l'approximation suivante :

$$I = \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}x^2\right] f(x)\,\mathrm{d}x \simeq \sum_{i=1}^n w_i f(x_i),$$

avec

$$w_i = \frac{2^{n-1}\, n! \sqrt{n}}{n^2 [H_{n-1}(x_i)]^2}$$

et où x_i est le $i$ème zéro du polynôme de Hermite de degré n, $H_n(x)$. Cette méthode que l'on appele *méthode de Gausse–Hermite* choisit ainsi les points x_i et les poids w_i tels que l'approximation devient exacte si $\exp\left[-\frac{1}{2}x^2\right] f(x)$ est un polynôme de degré inférieur ou égale à $2n - 1$.
On peut utiliser d'autres bases orthogonales pour approcher les intégrales. Mais, en générale, ces méthodes exigent que la fonction $f(x)$ soit suffisamment réguliere (voir Abramowitz et Stegun 1964).
Souvent on connaît la loi *a posteriori* $p(\theta|\boldsymbol{x})$ à une constante prés, *i.e.*;

$$p(\theta|\boldsymbol{x}) \propto g(\theta)$$

et on doit calculer

$$\mathrm{E}\,[h(\theta)] = \frac{\displaystyle\int h(\theta)g(\theta)\,\mathrm{d}\theta}{\displaystyle\int g(\theta)\,\mathrm{d}\theta}.$$

Lorsque cette loi est proche d'une loi gaussienne $\mathbf{N}(\theta|m, v)$ on peut approximer la dénominateur par

$$\int g(\theta)\,\mathrm{d}\theta \simeq v^{\frac{1}{2}} \sum_{i=1}^n w_i g(\theta_i),$$

où $\theta_i = m + v^{\frac{1}{2}} x_i$ et où x_i et w_i sont les points et les poids de la méthode de Gausse-Hermite. Lorsque $h(\theta)$ est une fonction polynômiale, on peut utiliser ces mêmes x_i et w_i pour approximer le nominateur pour obtenir

$$\mathrm{E}\,[h(\theta)] = \frac{\displaystyle\int h(\theta)g(\theta)\,\mathrm{d}\theta}{\displaystyle\int g(\theta)\,\mathrm{d}\theta}.$$

A COMPLETER

8.2 Réduction de la dimension de l'intégration

Si la loi *a posteriori*

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto g(\boldsymbol{\theta})$$

le calcule de

$$\mathrm{E}\left[h(\boldsymbol{\theta})\right] = \frac{\displaystyle\int h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\displaystyle\int g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}.$$

nécessite des intégrations de dimension du vecteur $\boldsymbol{\theta}$.
Supposons que $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$. On a alors les relations suivantes :

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = p(\boldsymbol{\phi}, \boldsymbol{\psi}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\phi}, \boldsymbol{\psi})\, p(\boldsymbol{\phi}|\boldsymbol{\psi})\, p(\boldsymbol{\psi})}{p(\boldsymbol{x})}$$

$$\mathrm{E}\left[h(\boldsymbol{\theta})|\boldsymbol{x}\right] = \mathrm{E}\left[\mathrm{E}\left[h(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\psi})\right]|\boldsymbol{x}\right].$$

Supposons que l'on dispose de résultats analytique pour

$$\mathrm{E}\left[h(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\psi})\right] = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\psi})\,\mathrm{d}\boldsymbol{x} = h_1(\boldsymbol{\psi}),$$

alors on a

$$\mathrm{E}\left[h(\boldsymbol{\theta})|\boldsymbol{x}\right] = \mathrm{E}\left[h_1(\boldsymbol{\psi})|\boldsymbol{x}\right] = \frac{\displaystyle\int h_1(\boldsymbol{\psi})g_1(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi}}{\displaystyle\int g_1(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi}},$$

où $g_1(\boldsymbol{\psi})$ est une fonction proportionnelle à la loi marginale $p(\boldsymbol{\psi}|\boldsymbol{x})$, *i.e.*;

$$p(\boldsymbol{\psi}|\boldsymbol{x}) = \int p(\boldsymbol{\psi}, \boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \propto g_1(\boldsymbol{\psi}).$$

Sachant que la dimension du vecteur $\boldsymbol{\psi}$ est inférieure à celle du $\boldsymbol{\theta}$, on a réduit la dimension du problème.

Exemple 17

$$\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2 \sim \mathbf{N}_k(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2\boldsymbol{I}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\theta})^t(\boldsymbol{x}-\boldsymbol{\theta})\right]$$

$$\boldsymbol{\theta}|\alpha, \tau \sim \mathbf{ST}_k(\boldsymbol{\theta}|\alpha, (0), \tau^2\boldsymbol{I}) = \frac{\Gamma((\alpha+n)/2)}{(\alpha\pi)^{n/2}\Gamma(\frac{\alpha}{2})}\left(1+\frac{\boldsymbol{\theta}^t\boldsymbol{\theta}}{\alpha\tau^2}\right)^{-(\alpha+n)/2}$$

$$\boldsymbol{\theta}|\boldsymbol{x} \propto g(\boldsymbol{\theta}) = \left(1+\frac{\boldsymbol{\theta}^t\boldsymbol{\theta}}{\alpha\tau^2}\right)^{-(\alpha+n)/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^t(\boldsymbol{x}-\boldsymbol{\theta})\right]$$

Si on choisit comme estimateur de $\boldsymbol{\theta}$ la moyenne *a posteriori*

$$\delta(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{x}\right] = \frac{\displaystyle\int \boldsymbol{\theta}g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\displaystyle\int g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}$$

qui n'est pas facile à calculer. Il est cependant possible de réduire le calcul de $\delta(\boldsymbol{\theta})$ à celui d'un intégrale simple si on remarque que l'on a

$$\begin{aligned} \boldsymbol{\theta}|z &\sim \mathbf{N}_k(\boldsymbol{\theta}|\mathbf{0}, \tau^2 z\boldsymbol{I}), \\ z &\sim \mathbf{IGam}(z\left|\frac{\alpha}{2}, \frac{\alpha}{2}\right.), \end{aligned}$$

où z est une variable auxiliaire. Travaillant conditionnellement à z, on a

$$\boldsymbol{\theta}|\boldsymbol{x} \sim \mathbf{N}_k\left(\boldsymbol{\theta}|\frac{1}{1+\tau^2 z}\boldsymbol{x}, \frac{\tau^2 Z}{1+\tau^2 z}\boldsymbol{I}\right)$$

et, comme

$$\pi(z|\boldsymbol{x}) \propto g(z) = (1+\tau^2 z)^{-k/2} \exp\left[-\frac{1}{2(1+\tau^2 z}\boldsymbol{x}^t\boldsymbol{x}\right] \pi(z)$$

il vient

$$\delta(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{x}\right] = \int_0^\infty \mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{x}, z\right] \pi(z|\boldsymbol{x})\,\mathrm{d}z = \frac{\displaystyle\int \frac{1}{1+\tau^2 z}\boldsymbol{x} g(z)\,\mathrm{d}z}{\displaystyle\int g(z)\,\mathrm{d}z} = \boldsymbol{x}\frac{\displaystyle\int \frac{1}{1+\tau^2 z} g(z)\,\mathrm{d}z}{\displaystyle\int g(z)\,\mathrm{d}z};$$

l'estimateur $\delta(\boldsymbol{\theta})$ s'exprime donc bien en fonction des intégrales simples.

Mais les décompositions comme celle-ci ne sont pas toujours possibles. Cependant, dans le cas des lois de la famille exponentielles naturelle, on a le résultat suivant :

Lemme 0.1 Soit $p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t\boldsymbol{x} - \phi(\boldsymbol{\theta})\right]$, densité d'une famille exponentielle. Pour toute loi *a priori* $p(\boldsymbol{\theta})$, la moyenne *a posteriori* est donnée par

$$\delta(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{x}\right] = \nabla \log m(\boldsymbol{\theta}) - \nabla \log h(\boldsymbol{x}),$$

où $m(\boldsymbol{\theta})$ est la loi marginale

$$m(\boldsymbol{\theta}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x}$$

8.3 Approximations gaussiennes

$$g(\boldsymbol{\theta}) \simeq g(\boldsymbol{m}) \exp\left[\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{m})^t\boldsymbol{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{m})\right]$$

où $\boldsymbol{m}$ est la mode

$$\boldsymbol{m} = \arg\max_{\boldsymbol{\theta}} \left\{g(\boldsymbol{\theta})\right\},$$

et $\boldsymbol{V}$ est la dispersion modale

$$\boldsymbol{V} = \left(-\frac{\partial^2 \log g(\boldsymbol{\theta})}{\partial\theta_i \partial\theta - j}\right)^{-1}_{\boldsymbol{\theta}=\boldsymbol{m}},$$

et on a

$$\int g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \simeq g(\boldsymbol{m})(2\pi)^{\frac{n}{2}}|\boldsymbol{V}|^{\frac{1}{2}},$$

où n est la dimension du vecteur $\boldsymbol{\theta}$.
Il est évident que cette approximation est valable seulement si $g(\boldsymbol{\theta})$ est une fonction unimodale de $\boldsymbol{\theta}$ ou au moins si elle a une mode dominante à $\boldsymbol{\theta} = \boldsymbol{m}$. On a alors

$$\mathrm{E}\left[\boldsymbol{\theta}\right] \simeq \boldsymbol{m}, \quad \mathrm{Var}\left[\boldsymbol{\theta}\right] \simeq \boldsymbol{V}.$$

Si la fonction $g(\boldsymbol{\theta})$ est multimodale on peut l'approximer par une mélange de gaussiennes

$$g(\boldsymbol{\theta}) \simeq \sum_{i=1}^{p} g(\boldsymbol{m}_i) \exp\left[\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{m}_i)^t \boldsymbol{V}_i^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_i)\right]$$

où $\boldsymbol{m}_1, \cdots, \boldsymbol{m}_p$ sont les modes et $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_p$ sont les matrices de dispersion modales correspondantes. Cette approximation est valable si les modes sont bien séparés les unes aux autres. On a alors

$$\int g(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \simeq (2\pi)^{\frac{n}{2}} \sum_{i=1}^{p} g(\boldsymbol{m}_i)|\boldsymbol{V}_i|^{\frac{1}{2}} = \sum_{i=1}^{p} s_i,$$

et

$$\mathrm{E}\left[\boldsymbol{\theta}\right] \simeq \sum_{i=1}^{p} w_i \boldsymbol{m}_i = \bar{\boldsymbol{m}}$$

$$\mathrm{Var}\left[\boldsymbol{\theta}\right] \simeq \sum_{i=1}^{p} w_i \left(\boldsymbol{V}_i + (\boldsymbol{m}_i - \bar{\boldsymbol{m}})(\boldsymbol{m}_i - \bar{\boldsymbol{m}})^t\right),$$

où

$$w_i = s_i / \sum_{i=1}^{p} s_i.$$

8.4 Méthodes de Monte-Carlo

8.5 Approximation de Laplace

8.6 Méthodes de Monte-Carlo bayésien