Examen 2018 : Modélisation et statistique bayésienne computationnelle – Corrigé –

13 avril 2018

L'examen dure 3h et est noté sur 30. Tous les supports de cours sont autorisés. Il est attendu un code R commenté a minima pour les parties computationnelles, et un support papier peut être utilisé pour la partie formelle. Lisez bien tout le document, certaines questions peuvent être traitées formellement indépendamment du reste de l'exercice qui les contient.

1 Fonction de coût (9 pts)

Soit $\theta \in \Theta = \mathbb{R}$ le paramètre d'un modèle, sur lequel on dispose d'une loi *a priori* $\pi(\theta)$ et de données $x_1, \ldots, x_n$. On suppose que la loi a posteriori de densité $\pi(\theta|x_1, \ldots, x_n)$ est propre et telle que $\mathrm{E}_\pi[\exp(k\theta)|x_1, \ldots, x_n] < \infty$ pour tout $k \in \mathbb{R}$. On considère la fonction de coût pour l'estimation δ de θ définie sur $\mathbb{R}$ par

$$L_a(\theta, \delta) \quad = \quad \exp(a(\theta - \delta) - a(\theta - \delta) - 1$$

où a est un réel.

1. Montrer que $L_a(\theta, \delta) \geq 0$ pour tout $\theta \in \Theta$ et pour tout a et qu'elle est convexe en θ ; représenter cette fonction de coût comme une fonction de $(\theta - \delta)$ lorsque $a = \{0.1, 0.5, 1, 2\}$.
2. On suppose que $a > 0$. À quelles conditions cette fonction pénalise-t-elle les coûts de sous-estimation et de surestimation de θ de façon similaire ? Au contraire, à quelles conditions cette fonction pénalise-t-elle les coûts de sous-estimation et de surestimation de θ de façon très dissymétrique ?
3. On suppose que $a \neq 0$. Donner l'expression de l'estimateur de Bayes $\hat{\delta}_a$ sous cette fonction de coût.
4. Supposons que les données sont issues de $\mathcal{N}(\theta, 1)$ et que $\pi(\theta) \propto 1$; donnez l'estimateur de Bayes associé.

Réponses.

1. Avec

$$\frac{\partial L_a}{\partial \delta}(\theta, \delta) \quad = \quad a\,(1 - \exp(a(\theta - \delta))$$

qui s'annule en $\delta = \theta$, et

$$\frac{\partial^2 L_a}{\partial \delta^2}(\theta, \delta) \quad = \quad a^2 \exp(a(\theta - \delta) \;\geq\; 0,$$

on a clairement que $\delta \to L_a(\theta, \delta)$ est convexe et de minimum 0 en $\delta = \theta$. Un petit tableau de variations peut achever de nous en convaincre. Un code R minimal pour représenter le comportment de la fonction est le suivant :

```
f <- function(a) {
    curve(exp(a*x)-a*x-1, xlim=c(-10,10))
}

par(mfrow=c(2,2))
f(0.1)
f(0.5)
f(1)
f(2)
```

2. Pour $a > 0$, $L_a(\theta, \delta)$ se comporte comme une fonction linéaire pour des grandes valeurs négatives de l'écart $\theta - \delta$, soit pour des surestimations de θ. Elle se comporte comme une fonction exponentielle pour des grandes valeurs positives de l'écart $\theta - \delta$, soit pour des sous-estimation de θ. Elle pénalise donc bien plus fortement les sous-estimations de θ que les surestimations de θ. Elle se comporte similairement comme $a(\theta - \delta)^2$ pour $\delta \to \theta$ (à gauche comme à droite). On en déduit que cette fonction de coût est appropriée dans les cas où les petites erreurs de sous-estimation et de surestimation ne provoquent pas un coût très différent, mais où les grandes erreurs amènent à des coûts très différents.

3. L'estimateur de Bayes est défini par

$$\hat{\delta}_a \quad = \quad \arg\min_{\delta} \underbrace{\int_{\Theta} L_a(\theta, \delta) \pi(\theta | x_1, \ldots, x_n)\, d\theta}_{J(\delta)}.$$

Comme la fonction de coût est convexe, l'estimateur est donc défini comme la valeur de δ qui annule la dérivée du terme $J(\delta)$. Alors

$$\begin{aligned} J'(\hat{\delta}) = 0 \quad &\Leftrightarrow \quad \int_{\Theta} \frac{\partial L_a}{\partial \delta}(\theta, \hat{\delta}) \pi(\theta | x_1, \ldots, x_n)\, d\theta = 0, \\ &\Leftrightarrow \quad \exp(-a\hat{\delta}) \int_{\Theta} \exp(a\theta) \pi(\theta | x_1, \ldots, x_n)\, d\theta = 1, \end{aligned}$$

le terme de droite étant bien défini car on suppose $\mathrm{E}_\pi[\exp(k\theta)|x_1,\ldots,x_n] < \infty$ pour tout $k \in \mathbb{R}$. Il vient alors (avec $a \neq 0$)

$$\hat{\delta} = \frac{1}{a}\log\int_\Theta \exp(a\theta)\pi(\theta|x_1,\ldots,x_n)\,d\theta. \quad (1)$$

4. Avec $x_1,\ldots,x_n \sim \mathcal{N}(\theta,1)$ et $\pi(\theta) \propto 1$, il vient

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

et donc $\pi(\theta|x_1,\ldots,x_n)$ est la densité de la loi $\mathcal{N}(\bar{x}_n, 1/n)$. En appliquant (1), on déduit alors

$$\begin{aligned}\hat{\delta} &= \frac{1}{a}\log\int_\Theta \frac{n}{2\pi}\exp\left(a\theta - \frac{n}{2}\left\{\theta-\bar{x}_n\right\}^2\right)\,d\theta,\\ &= \frac{1}{a}\log\int_\Theta \frac{n}{2\pi}\exp\left(-\frac{n}{2}\theta^2 - \frac{n}{2}\bar{x}_n^2 + 2\frac{n}{2}\theta(\bar{x}_n + a/n)\right)\,d\theta,\\ &= \frac{1}{a}\log\exp\left(-\frac{n}{2}\bar{x}_n^2 + \frac{n}{2}(\bar{x}_n+a/n)^2\right)\int_\Theta \frac{n}{2\pi}\exp\left(-\frac{n}{2}\left\{\theta-(\bar{x}_n+a/n)\right\}^2\right)\end{aligned}$$

On reconnaît dans le terme intégral la densité d'une loi $\mathcal{N}(\bar{x}_n + a/n, 1/n)$. Avec $\Theta = \mathbb{R}$, cette intégrale vaut donc 1, et

$$\begin{aligned}\hat{\delta} &= \frac{1}{a}\left(-\frac{n}{2}\bar{x}_n^2 + \frac{n}{2}(\bar{x}_n+a/n)^2\right),\\ &= \bar{x}_n + a/2n.\end{aligned}$$

Remarque. Cette fonction de coût alternative aux fonctions classiques (coûts absolu, quadratique...) est dite LINEX (*linear-exponential*) et a été introduite par Varian en 1974 puis très utilisée par Zellner en 1986.

2 Élicitation d'*a priori* non informatif (7 pts)

On considère le problème suivant

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

où les x_i sont indépendants.

1. Quelle est la densité jointe des données $x_1,\ldots,x_n$?
2. Calculer la matrice d'information I de Fisher pour ce jeu de données
3. En déduire la mesure *a priori* de Jeffreys $\pi^J(\theta)$ pour $\theta = (\mu_1,\ldots,\mu_n,\sigma)$
4. Que peut-on dire de $\pi^J(\sigma^2|x_1,\ldots,x_n,\mu_1,\ldots,\mu_n)$? Est-ce une loi vue en cours ?

Réponses.

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

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

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

$$I = -\mathrm{E}\begin{bmatrix} \frac{\partial^2}{\partial\theta_{i_1}^2}\log f(x|\theta) & \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_2}}\log f(x|\theta) & \ldots & \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_d}}\log f(x|\theta) \\ \frac{\partial^2}{\partial\theta_{i_1}\theta_{i_2}}\log f(x|\theta) & \frac{\partial^2}{\partial\theta_{i_2}^2}\log f(x|\theta) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

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

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

et

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

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

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

et donc

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

3. En utilisant (2), la loi *a posteriori* s'écrit sous une forme condensée comme

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

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

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

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

3 Élicitation et calcul bayésien pour un problème de Gumbel (14 pts)

La loi de Gumbel, de fonction de répartition

$$P(X<x|\theta) \quad = \quad \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \quad \text{avec } \sigma>0,\ \mu\in I\!R \text{ et } x\in I\!R$$

et $\theta=(\mu,\sigma)$, est souvent utilisée en météorologie pour modéliser le comportement d'un échantillon de *maxima* d'une variable environnementale. Son espérance vaut $\mathrm{E}[X|\theta]=\mu+\sigma\gamma$ où γ est la constante d'Euler. On suppose connaître un échantillon de données de pluies (en mm) $\mathbf{x_n}=(x_1,\ldots,x_n)$ suivant cette loi. Elles sont fournies dans la table 1 et correspondent aux années 1987 à 2013. Par ailleurs on dispose d'une expertise *a priori* qui s'exerce sur la loi *a priori* prédictive de X, et est spécifiée statistiquement sous la forme $P(X<75)=25\%$, $P(X<100)=50\%$, $P(X<150)=75\%$.

107.6	72.4	204.5	83.8	142	95.5	316.1	177.9	87.3
81.9	109.1	89.5	150.7	122.1	98.2	113.2	104.4	66.9
136.4	275.4	125	199.8	51.2	75	168.2	106	72.8

TABLE 1 – Données de pluviométrie extrême.

On considère la mesure *a priori*

$$\pi(\mu,\sigma) \quad \propto \quad \sigma^{-m}\exp\left(m\frac{(\mu-\bar{\tilde{x}}_m)}{\sigma}-\sum_{i=1}^{m}\exp\left\{-\frac{\tilde{x}_i-\mu}{\sigma}\right\}\right)$$

où les hyperparamètres $(m,\bar{\tilde{x}}_m,\tilde{x}_1,\ldots,\tilde{x}_m)$ correspondent respectivement à la taille d'un échantillon de données *a priori* (virtuelles), sa moyenne et les données elles-mêmes (supposées calibrables).

1. Ecrivez la densité de la loi *a posteriori* conditionnelle aux données réelles $\mathbf{x_n}$. La loi *a priori* est-elle conjuguée ?
2. Produisez un algorithme qui simule la loi *a priori* prédictive de X en fonction des hyperparamètres et estime les quantiles prédictifs *a priori*. En fixant $m=3$ et $(\tilde{x}_1,\tilde{x}_2)=(81,93)$, testez les valeurs de $\tilde{x}_3$ suivantes : 97,101,110,120. Quelle calibration vous semble la plus adéquate vis-à-vis de l'expertise *a priori* ?
3. Pour les calibrations des hyperparamètres précédentes, écrivez un algorithme qui produit un tirage de la loi *a posteriori* de θ ainsi qu'une représentation (densité empirique) de la loi *a posteriori prédictive* sur X. Comparez avec un histogramme des données $\mathbf{x_n}$.

4. **Cette question peut être traitée indépendamment du reste.** On pose à présent $\mu > 0$ et on cherche à définir une nouvelle loi *a priori* $\pi_2(\theta)$ par maximum d'entropie qui est telle que les contraintes linéaires suivantes soient respectées :

$$\begin{aligned} \mathrm{E}[X] &= 100, \\ \mathrm{E}_\pi[\log \sigma] &= 1. \end{aligned}$$

Formalisez et résolvez numériquement (possiblement graphiquement) le problème de maximum d'entropie en supposant que la mesure de référence est la mesure de Jeffreys $\pi_0(\theta) \propto \sigma^{-2}$ (valable pour le modèle de Gumbel). Sous quelles contraintes sur les multiplicateurs de Lagrange pouvez-vous trouver une loi jointe propre ? Celle-ci appartient-elle à une classe de lois connues ?

Rappel : *Si Y suit une loi gamma $\mathcal{G}(a, b)$, alors $E[Y] = \Psi(a) - \log(b)$ où Ψ est la fonction digamma (`digamma` en R).*

5. Adaptez le code produit à la question 3 pour produire un nouveau calcul *a posteriori*, en utilisant $\pi_2(\theta)$.

Réponses.

1. Sachant des données réelles $\mathbf{x_n}$, la loi *a posteriori* s'écrit

$$\begin{aligned} \pi(\mu, \sigma | \mathbf{x_n}) \quad \propto \quad & \sigma^{-m-n} \exp \left(\{m+n\} \frac{\left(\mu - \frac{m\bar{\tilde{x}}_m + n\bar{\mathbf{x_n}}}{m+n} \right)}{\sigma} \right. \\ & \left. - \sum_{i=1}^{m} \exp \left\{ -\frac{\tilde{x}_i - \mu}{\sigma} \right\} - \sum_{k=1}^{n} \exp \left\{ -\frac{x_k - \mu}{\sigma} \right\} \right). \end{aligned}$$

Elle est donc en effet conjuguée, car on retrouve la même forme que la loi *a priori*.

2. Pour simuler selon la loi *a priori* marginale, le plus simple est d'utiliser l'algorithme ci-dessous :
   (a) simuler μ_i, σ_i *a priori* ;
   (b) simuler X_i selon la loi de Gumbel en μ_i, σ_i.

   Pour réaliser la première simulation (la seconde peut être faite très facilement par inversion), on peut tenter de procéder de plusieurs façons : acceptation-rejet, échantillonnage d'importance, MCMC... En regardant la forme de la loi *a priori*, on privilégie l'approche par échantillonnage d'importance en utilisant (par exemple) une loi instrumentale de densité

$$g(\mu, \sigma) \equiv \mathcal{IG}_\sigma(m-1, m\bar{\tilde{x}}_m) \mathcal{E}_\mu(\lambda)$$

avec $m > 1$, où $\mathcal{IG}$ est une loi inverse gamma. Les poids d'importance s'écrivent alors (à un coefficient près)

$$\begin{aligned}\omega_k &= \frac{\pi(\mu_k,\sigma_k)}{g(\mu_k,\sigma_k)},\\ &\propto \exp\left(\mu_k\left[m/\sigma_k+\lambda\right]-\sum_{i=1}^{m}\exp\left\{-\frac{\tilde{x}_i-\mu_k}{\sigma_k}\right\}\right)\end{aligned}$$

où $(\mu_k,\sigma_k)\sim g(\mu,\sigma)$. Le logarithme de ces poids non normalisés est aisé à calculer, ce qui permet une approche numérique plus stable (en jouant éventuellement sur le λ). La valeur la plus adéquate était 110 (c'est elle qui permet un meilleur *matching* avec les requis de l'expertise).

3. La loi *a posteriori* étant connue explicitement, on peut utiliser le même type d'algorithme pour mener le calcul *a posteriori*.
4. On a

$$\mathrm{E}[X] = \mathrm{E}_\pi[\mathrm{E}[X|\theta]] = \mathrm{E}_\pi[\mu+\sigma\gamma]$$

Sous cette contrainte, et sous l'autre contrainte $\mathrm{E}_\pi[\log\sigma]=1$, la solution du problème classique de maximisation d'entropie est

$$\begin{aligned}\pi_2(\theta) &\propto \pi_0(\theta)\exp(\lambda_1(\mu+\sigma\gamma)+\lambda_2\log\sigma),\\ &\propto \sigma^{\lambda_2-2}\exp(\lambda_1\gamma\sigma)\exp(\lambda_1\mu).\end{aligned}$$

Avec $\mu>0$, pour avoir une loi *a posteriori* intégrable, il nous faut avoir $\lambda_1 = -\tilde{\lambda}_1 < 0$ et $\lambda_2 = \tilde{\lambda}_2+1$ avec $\tilde{\lambda}_2>0$. Dans ce cas, on reconnaît aisément un mélange de loi gamma $\mathcal{G}(\tilde{\lambda}_2,\gamma\tilde{\lambda}_1)$ pour σ, et de loi exponentielle $\mathcal{E}(\tilde{\lambda}_1)$ pour μ. Dans ce cas, on a

$$\begin{aligned}\mathrm{E}[X] &= \mathrm{E}_\pi[\mu+\sigma\gamma],\\ &= \frac{1}{\tilde{\lambda}_1}+\gamma\frac{\tilde{\lambda}_2}{\gamma\tilde{\lambda}_1},\\ &= \frac{1}{\tilde{\lambda}_1}(1+\tilde{\lambda}_2),\\ &= 100. \qquad (3)\end{aligned}$$

et

$$\mathrm{E}_\pi[\log\sigma] = \Psi(\tilde{\lambda}_2)-\log(\gamma\tilde{\lambda}_1) = 1.$$

Ce système de deux équations à deux inconnues peut se résoudre numériquement. D'après (3), on a

$$\tilde{\lambda}_1 = (1+\tilde{\lambda}_2)/100$$

et

$$\Psi(\tilde{\lambda}_2)-\log(1+\hat{\lambda}_2)+\log 100-1 = 0.$$

Il suffit de tracer la courbe de l'équation précédente pour obtenir

$$\begin{aligned} \tilde{\lambda}_2 &\simeq 0,3155, \\ \tilde{\lambda}_1 &\simeq 0,013155. \end{aligned}$$

5. On produit ici un algorithme MCMC, car on perd la propriété de conjugaison. Voici un lien vers un code R permettant de répondre à la question :

`http://www.lsta.upmc.fr/bousquet/coursM2-2018/calcul-MCMC-gumbel.r`