Master 2 Biostatistique – UE STA305
Travaux Dirigés

Exercice 1

Considérons des variables aléatoires $Y_1, \ldots, Y_n$ indépendantes et identiquement distribuées (i.i.d.) suivant une loi normale $\mathcal{N}(\theta, \sigma^2)$, dont la densité est définie pour tout $y \in \mathbb{R}$ par :

$$f_{\theta,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$$

On suppose que σ^2 est connu et le paramètre d'intérêt est donc le paramètre de moyenne θ.

1. Écrire le modèle bayésien considéré.
2. Écrire la vraisemblance et la log-vraisemblance de l'échantillon $(y_1, \ldots, y_n)$, en faisant apparaître $\overline{y}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i$ sous la forme $\left(\theta - \overline{y}_{(n)}\right)^2$.
3. Écrire les dérivées première et seconde de la log-vraisemblance par rapport à θ et l'information de Fisher $I(\theta)$.
4. Quel est la loi *a priori* de Jeffrey pour θ ? Est-ce qu'il définit une densité propre ou impropre ?
5. En prenant cette loi *a priori*, écrire le numérateur de la loi *a posteriori* de θ. En déduire la distribution *a posteriori* de θ.
6. On observe un deuxième échantillon $(y_{n+1}, \ldots, y_{2n})$ i.i.d. de même loi que le premier échantillon. Quelle est la distribution *a posteriori* de θ en prenant un *a priori* uniforme ? Faire le calcul de deux façons:
   (a) en considérant que l'on a un échantillon i.i.d. de taille $2n$;
   (b) en utilisant la distribution *a posteriori* obtenue pour le premier échantillon comme distribution *a priori* pour le second échantillon.

Exercice 2

On considère les réalisations $\{x_1, \ldots, x_n\}$ de n variables aléatoires i.i.d. $X_1, \ldots, X_n$ suivant une loi exponentielle $\mathcal{E}(\lambda)$, avec $\lambda > 0$, dont la densité est définie pour tout $x \geq 0$ par : $f_\lambda(x) = \lambda e^{-\lambda x}$. On prend comme loi *a priori* sur λ la loi Gamma(α, β) dont la densité est définie pour tout $x > 0$ par :

$$g_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

1. Écrire est le modèle bayésien associé.

2. Quelle est la loi *a posteriori* correspondante ?

Exercice 3

On considère les réalisations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ de n variables aléatoires i.i.d. $X_1, \ldots, X_n$, suivant une loi de Pareto$(\theta + 1, 1)$, avec $\theta > 0$, dont la densité est définie pour tout $x > 1$ par :

$$f_\theta(x) = \frac{\theta + 1}{x^{\theta+2}}$$

1. L'*a priori* utilisé pour θ est la loi exponentielle $\mathcal{E}(1)$, dont la fonction de densité est définie pour tout $x \geq 0$ par : $g(\theta) = e^{-\theta}$. Écrire le modèle bayésien associé.

2. Montrer que la densité de la loi *a posteriori* de $\theta|\boldsymbol{x}$, notée $p(\theta|\boldsymbol{x})$, est proportionnelle à :

$$\exp(-\theta)(\theta + 1)^n \prod_{i=1}^{n} x_i^{-\theta}$$

3. Proposer un algorithme de Metropolis-Hastings indépendant pour estimer la loi *a posteriori* de $\theta|X_1, ..., X_n$. On prendra comme loi instrumentale la loi *a priori* de θ. Expliciter l'estimateur Bayésien de θ construit pour le coût quadratique. Ne pas oublier de faire apparaître les calculs et la formule de la probabilité d'acceptation.

4. Quel résultat théorique garantit sa convergence ? Expliquer brièvement.

Exercice 4

Rappels sur la loi Beta :

- La densité de probabilité de la loi Beta(a, b) de paramètres $a > 0$ et $b > 0$ est définie pour tout $\theta \in [0, 1]$ par :
$$g_{a,b}(\theta) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}$$
- Si $X \sim$ Beta(a, b), alors $\mathbb{E}[X] = \frac{a}{a+b}$.

On souhaite estimer la probabilité de contracter une maladie M dans l'hôpital A. On dispose pour cela des données de n_A patients indiquant s'ils ont ou non contracté la maladie. On note $\boldsymbol{y}^A = (y_1^A, \ldots, y_{n_A}^A)$ l'échantillon observé de la variable binaire définie par :

$$y_i^A = \begin{cases} 1 & \text{si le patient } i \text{ a contracté la maladie} \\ 0 & \text{sinon} \end{cases}$$

On note $\theta_A \in [0, 1]$ la probabilité inconnue de contracter la maladie dans l'hôpital A et l'on suppose que les variables aléatoires $Y_1^A, \ldots, Y_{n_A}^A$ sont i.i.d.

1. Écrire la vraisemblance des données $p(\boldsymbol{y}^A|\theta_A)$.

2. On utilise une approche bayésienne, et l'on suppose que θ_A suit *a priori* une distribution uniforme sur l'intervalle $[0,1]$. Donner la forme de la densité *a posteriori* $p(\theta_A|\boldsymbol{y}^A)$. Montrer que celle-ci prend une forme paramétrique connue.

3. Cette densité *a posteriori* est-elle propre ? Pourquoi ?

4. Calculer la loi marginale des observations $f(\boldsymbol{y}^A)$.

5. Donner la probabilité $p(y^A_{n_A+1} = 1|\boldsymbol{y}^A)$ qu'un nouveau patient contracte la maladie, sachant $\boldsymbol{y}^A$.

6. On dispose maintenant des données $y^B_1, \ldots, y^B_{n_B}$ de contraction de la maladie pour n_B patients d'un second hôpital B. On note θ_B la probabilité qu'un patient de l'hôpital B ait contracté la maladie, et l'on suppose toujours l'indépendance des variables aléatoires associées. On souhaite tester l'hypothèse H_0 selon laquelle les taux de contraction de la maladie sont les mêmes dans les hôpitaux A et B, versus l'hypothèse H_1 que ces taux sont différents:

$$H_0 : \theta_B = \theta_A,\ \theta_A \sim \mathcal{U}([0,1]) \text{ v.s. } H_1 : \theta_A \sim \mathcal{U}([0,1]) \perp \theta_B \sim \mathcal{U}([0,1])$$

où $\mathcal{U}([0,1])$ désigne la loi uniforme sur l'intervalle $[0,1]$.
Écrire $p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_0)$ et $p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_1)$.

7. En déduire le facteur de Bayes $B_{1,0}$ de l'hypothèse H_1 par rapport à l'hypothèse H_0, qui se définit comme le ratio des vraisemblances marginales :

$$B_{1,0} = \frac{p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_1)}{p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_0)}$$

Exercice 5

Dans cet exercice, nous nous proposons de voir comment il est possible de simuler des réalisations d'une loi puis de vérifier qu'elles sont bien issues de cette loi en ré-estimant les paramètres.

1. Proposer un algorithme basé sur la méthode par inversion, permettant de simuler la réalisation d'un échantillon de taille n de loi de Pareto($\lambda = 2, k = 5$). La densité de la loi de Pareto est : $f_{\lambda,k}(x) = \frac{k\lambda^k}{x^{k+1}} \mathbb{1}_{x>\lambda}$.

2. Grâce à ce premier algorithme nous pouvons donc maintenant simuler un n -échantillon i.i.d. $\boldsymbol{x} = (x_1, \ldots, x_n)$ suivant une loi de Pareto$(2,5)$. Désormais, nous voulons vérifier que l'algorithme est valide et nous voulons ré-estimer le paramètre k ayant servi à simuler ces données. On suppose $\lambda = 2$ connu et fixé. Pour cela, nous allons appliquer des méthodes bayésiennes avec l'*a priori* suivant pour k : $\pi(k) = \frac{1}{200} e^{-\frac{k^2}{2\times 100^2}} \mathbb{1}_{k\in]0,\infty[}$. Écrire le modèle bayésien associé puis calculer la loi *a posteriori* de $k|\boldsymbol{x}$.

3. Expliquer brièvement la logique de l'acceptation/rejet en fonction de la loi instrumentale de proposition et de la loi que l'on veut échantillonner. Quelle simplification apparaît en prenant pour loi instrumentale la loi *a priori* du paramètre ? Comment appelle-t-on ce phénomène ?

4. Proposer un algorithme de Metropolis-Hastings indépendant pour échantillonner la loi *a posteriori* de $k|\boldsymbol{x}$. On prendra comme loi instrumentale la loi *a priori* de k.

5. Expliciter l'estimateur Bayésien $\hat{E}(k|X_1, ..., X_n)$ de k construit pour le coût quadratique.