Master 2 Biostatistique – UE STA305
Travaux Dirigés

Boris Hejblum

Exercice 1

Les variables aléatoires $Y_i, i = 1, \ldots, n$ sont indépendantes et identiquement distribuées (iid) suivant une loi Normale de paramètres θ et σ^2. La densité de la loi de Normale est : $f_{\theta,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$. On considérera σ^2 connu.

1. Écrire la vraisemblance et la log-vraisemblance de l'échantillon $y_i, i = 1, \ldots, n$, en faisant apparaître $\bar{y}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i$ sous la forme $\left(\theta - \bar{y}_{(n)}\right)^2$. Attention : on rappelle qu'une somme de nombre au carré n'est pas égale au carré de la somme de ces nombres...
2. Écrire la dérivée première et seconde de la log-vraisembance par rapport à θ et l'information de Fisher pour θ.
3. Quel est la loi *a priori* de Jeffrey pour θ? Est-ce qu'il définit densité propre ou impropre ?
4. En prenant cette loi *a priori*, écrire le numérateur de la loi *a posteriori* de θ. En déduire la distribution *a posteriori* de θ.
5. On observe un deuxième échantillon $\{y_i\}, i = n+1, \ldots, 2n$ iid 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 iid 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 $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ d'une suite de variables aléatoires iid $\{X_i\}_{i=1,\ldots,n}$ réelles et supérieures à 1, dont la loi P_θ est supposée connue à un paramètre $\theta > 0$ près. Cette loi P_θ est une loi continue, appelée loi de Pareto de paramètres $(\theta + 1, 1)$ dont la densité est définie, pour $x > 1$, par :

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

1. L'*a priori* utilisé pour θ est une loi exponentielle de paramètre 1, dont la fonction de densité s'écrit : $g(\theta) = e^{-\theta}$. 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 :

$$exp(-\theta)\,(\theta+1)^n \left(\prod_{i=1}^{n} x_i^{-\theta}\right) \qquad ; \quad \theta > 0$$

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 3

On considère les réalisations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ d'une suite de variables aléatoires *iid* $\{X_i\}_{i=1,\ldots,n}$ suivant une loi exponentielle de paramètre λ : $\mathcal{E}(\lambda)$, où $\lambda > 0$ est inconnu. On prend comme loi *a priori* sur λ la loi Gamma $\mathcal{G}(\alpha, \beta)$ dont la densité s'écrit :

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

1. Ecrire est le modèle bayésien associé.

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

Exercice 4

Propriétés utiles :

- La densité de probabilité de la loi Beta de paramètres $a > 0$ et $b > 0$ évaluée en θ est donnée par

$$\text{Beta}(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

- Soit θ une variable aléatoire suivant une loi Beta de paramètres a et b. On a alors

$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

On souhaite estimer la probabilité de contracter une maladie M dans l'hopital A. On dispose pour cela de 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 $\{Y_i^A\}_{i=1,\ldots,n_A}$ sont *iid* conditionnellement à θ_A.

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 $p(y_{n_A+1}^A = 1|\boldsymbol{y}^A)$ qu'un nouveau patient n_A+1 contracte la maladie sachant $\boldsymbol{y}^A = \{y_1^A, \ldots, y_{n_A}^A\}$.

6. On dispose maintenant des données $y_1^B, \ldots, y_{n_B}^B$ de contraction de la maladie pour n_B patients d'un second hôpital B. On note θ_B la probabilité que le patient i de l'hôpital B ait contracté la maladie, et l'on suppose toujours l'indépendance conditionnellement à θ_B. 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:

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

   où $U([0,1])$ dénote la distribution 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_{10} de l'hypothèse H_1 par rapport à l'hypothèse H_0, qui se définie comme le ratio des probabilités à posteriori :

   $$B_{10} = \frac{p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_1)}{p(\boldsymbol{y}^A, \boldsymbol{y}^B|H_0)}$$

Exercice 5

Une personne effectue un test afin de détecter si elle est porteuse d'un virus A, potentiellement mortel, présent dans 0.1% de la population. Le test a les propriétés suivantes:

- Si la personne est porteuse du virus, le test sera positif dans 99% des cas

- Si la personne n'est pas porteuse du virus, le test sera négatif dans 95% des cas

Le résultat du test est positif.

1. Quelle est probabilité que la personne soit porteuse du virus?
2. Suite à un test positif, la personne a le choix de suivre ou non un traitement thérapeutique. Si elle est effectivement porteuse du virus, ce traitement rallongera sa durée de vie de 6 mois. Qu'elle soit ou non porteuse du virus, ce traitement est assez lourd, et l'on considère qu'il diminue de 1 mois la durée de vie. Déterminer le coût moyen de la décision de suivre ce traitement.

Exercice 6

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 d'une loi de Pareto de paramètres $\lambda = 2$ et $k = 5$. La densité de la loi de Pareto est la suivante : $f(x) = k\frac{\lambda^k}{x^{k+1}}$).
2. Grâce à ce premier algorithme nous pouvons simuler un n-échantillon iid $\boldsymbol{x} = x^(1), ...x^(n)$ suivant une loi de Pareto de paramètres $\lambda = 2$ et $k = 5$. 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 : $\pi(k) = \frac{1}{200}e^{-\frac{k^2}{2*100^2}}\mathbb{1}_{k\in]0,\infty[}$. 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 apparait en prenant pour loi instrumentale la loi *a priori* du paramètre? Comment appelle-t-on ce phénomène ?
4. Proposer un algorithme de Métropolis-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.