Master 2 Biostatistiques - UE Bayes

Daniel Commenges, François Caron, Mélanie Prague, Rodolphe Thiébaut

13 janvier 2012 - 3h

Les exercices sont indépendants les uns des autres.

Exercice 1 (Durée conseillée : 1 heure)

Les observations $Y_i, i = 1, \ldots, n$ sont indépendantes et identiquement distribuées suivant une loi exponentielle de paramètre $\theta > 0$. La densité de la loi exponentielle est : $f_Y^\theta(y) = \theta e^{-\theta y}$.

(a) Écrire la vraisemblance de l'échantillon $Y_i, i = 1, \ldots, n$.

(b) Écrire la logvraisemblance et sa dérivée première et seconde par rapport à θ.

(c) Écrire l'information de Fischer pour θ

(d) Quel est le prior de Jeffrey pour θ; est-ce une densité propre ou impropre ?

(e) En prenant ce prior, écrire le numérateur de la loi a posteriori de θ

(f) La loi Gamma a une densité $f(x; \alpha, \beta)$ proportionnelle à $x^{\alpha-1}e^{-\beta x}$. Quand α est grand la distribution est proche d'une distribution normale. Montrer que la distribution a posteriori de θ est une loi Gamma. Quels en sont les paramètres ?

(g) Que peut-on-dire de la distribution a posteriori de θ quand n est grand ?

Exercice 2 (Durée conseillée : 30 min)

On considère le même modèle que dans l'exercice 1. Les observation Y_i, $i = 1, \ldots, n$ sont ainsi indépendantes et identiquement distribuées selon une loi exponentielle de paramètre θ.

On souhaite tester les hypothèses suivantes

$$H_0 : \theta = 1 \text{ vs } H_1 : \theta \sim \text{Exp}(b)$$

où Exp(b) désigne la loi exponentielle de paramètre $b > 0$.

(a) Déterminer le facteur de Bayes B_{10} de H_1 par rapport à H_0.

On utilisera la propriété suivante

$$\int_0^\infty \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) dx = 1$$

(b) On considère que l'on dispose d'un jeu de données pour lequel

$$\log_{10} B_{10} = 0.1$$

Quelle conclusion pouvez-vous apporter ?

Exercice 3 (Durée conseillée : 15 min)

On considère le même modèle que dans l'exercice 1. On fait maintenant l'hypothèse que le paramètre inconnu θ peut prendre uniquement deux valeurs θ_1 et θ_2. Après obtention d'un jeu de données, on a les distributions a posteriori suivantes

$$P(\theta = \theta_1 | Y_1, \ldots, Y_n) = 0.2$$
$$P(\theta = \theta_2 | Y_1, \ldots, Y_n) = 0.8$$

On considère la fonction de coût $L(\theta, d)$ définie par

$L(\theta, d)$	$d = \theta_1$	$d = \theta_2$
$\theta = \theta_1$	0	10
$\theta = \theta_2$	1	0

(a) Déterminer le coût moyen a posteriori pour chaque décision $d = \theta_1$ et $d = \theta_2$

(b) En déduire l'estimateur bayésien associé à cette fonction de coût et cette loi a posteriori.

Exercice 4 (Durée conseillée : 45 min)

On considère une variable aléatoire réelle X, 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 et la fonction de répartition sont définies par, pour $x > 1$:

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

(a) L'a priori utilisé pour θ est une loi exponentielle de paramètre 1 (voir Exercice 1 pour la définition). Montrer que la densité de la loi a posteriori de $\theta | X_1, ..., X_n$, notée $p(\theta | X_1, ..., X_n)$, est proportionnelle à :

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

(b) Proposer un algorithme de Métropolis-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.

(c) Quel résultat théorique garantit sa convergence ? Expliquer brièvement.

Exercice 5 (Durée conseillée : 30 min)

La réponse à chaque question doit être courte (5 lignes maximum).

Nous vous présentons les résultats de l'analyse WinBUGS des données de McGilchrist et Aisbett (1991) pour l'analyse du temps de première et seconde infection ($j = 1, 2$) du rein chez des patients ($i = 1...38$) sous dialyse en utilisant un modèle de Cox avec un paramètre de fragilité pour chaque individu (effet aléatoire individuel). Nous avons analysé les données en supposant une fonction paramétrique de Weibull pour la fonction de survie en incluant un terme additif d'effet aléatoire b_i pour chaque patient comme suit :

$t_{ij} \sim Weibull(r, t_{ij}) \quad i = 1, ..., 38; \quad j = 1, 2$
$log(t_{ij}) = \alpha + \beta_{age} AGE_{ij} + \beta_{sex} SEX_i + \beta_{disease1} DISEASE1_i + \beta_{disease2} DISEASE2_i + \beta_{disease3} DISEASE3_i + b_i$
$b_i \sim Normal(0, \tau)$

Il y a une variable explicative continu : l'âge (AGE_{ij}) à l'apparition d'une infection (première ou deuxième). Il y a deux covariables qualitatives le sexe SEX_i (ref='Femme'=0) dépendant de l'individu et le type de la maladie rénale ($DISEASEk_i$, k=1,2 ou 3) préalablement dichotomisée. Les temps ne sont observée qu'à partir d'un temps "t.cen" dépendant de l'individu et du numéro de récidive.

(a) Nous donnons le code en langage WinBUGS :

```
1.  model {
2.      for (i in 1:N) {
3.          for (j in 1:M) {
4.              t[i,j] ~ dweib(r,mu[i,j]) I(t.cen[i,j],);
5.              log(mu[i,j]) <- alpha + beta.age*age[i,j] + beta.sex *sex[i]  + beta.dis[disease[i]] + b[i];
6.          }
7.      # Random effects:
8.          b[i] ~ dnorm(0.0, tau)
9.      }
10.     # Priors:
11.       alpha ~ dnorm(0.0, 0.0001);
12.       beta.age ~ dnorm(0.0, 0.0001);
13.       beta.sex ~ dnorm(0.0, 0.0001);
14.       for(k in 2 : 4) {
15.           beta.dis[k] ~ dnorm(0.0, 0.0001);
16.       }
17.       tau ~ dgamma(1.0E-3, 1.0E-3);
18.       r ~ dgamma(1.0, 1.0E-3);
19.   }
```

(1) Ligne 4, que signifie le "I(t.cen[i,j],)".

(2) Ligne 5, Donner le type de structure de données (scalaire, vecteur ...) de chacune des variables suivantes : beta.age, age et sex.

(3) Ligne 10 à 18 commenter le choix des a prioris.

(b) Nous donnons certaines sorties de Winbugs (page 4), uniquement pour le paramètre β_{age}, pour ce modèle, les données correspondantes et deux jeux de valeurs initiales pour une phase de chauffe de 1000 réalisations et une phase d'échantillonnage de 2000 réalisations avec une épaisseur (thin) de 1.

(1) Quel est l'intérêt d'échantillonner simultanément sur deux chaînes ?

(2) Interprétez rapidement chaque résultat (une phrase par résultat).

(3) Pouvez-vous conclure à la convergence ? Comment améliorer les résultats ? Justifiez quel diagnostic vous pousse à chaque conclusion.

(c) 3. L'analyse est menée de nouveau avec un meilleur calibrage. Les résultats sont désormais interprétables (Table 1 page 4).

(1) Donner un sens aux estimations pour β_{sex}.

(2) Donner un sens aux estimations pour β_{age}.

(3) Donner un sens aux estimations pour r.

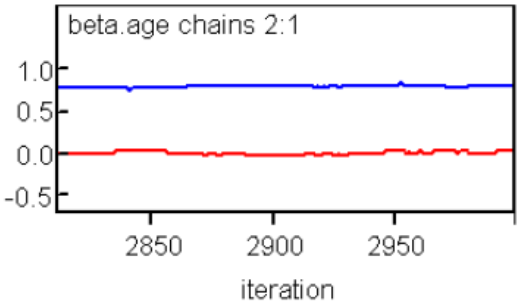

Figure 1: Trace dynamique des chaînes pour β_{age}

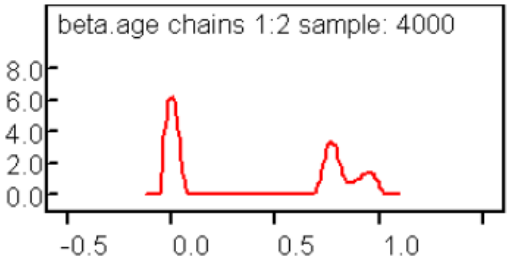

Figure 2: Densité de l'a posteriori de β_{age}

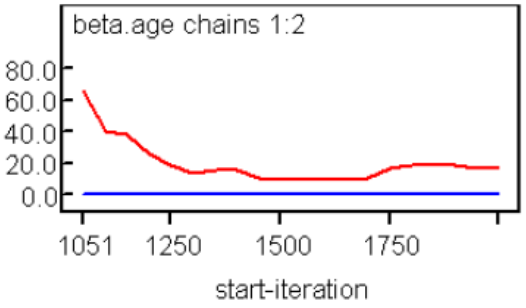

Figure 3: Statistiques de Gelman et Rubin pour β_{age}

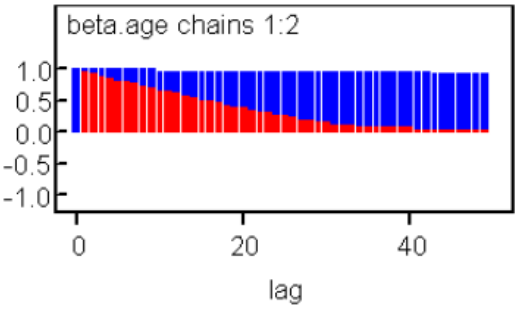

Figure 4: Autocorrélations pour β_{age}

Noeud	moyenne	écart-type	2.5%	médiane	97.5%
α	-4.54	0.91	-6.56	-4.46	-2.99
β_{age}	0.0035	0.015	-0.03	0.003	0.03
β_{sex}	-1.98	0.542	-3.18	-1.93	-1.05
r	1.21	0.18	0.92	1.19	1.61

Table 1: Sorties statistiques WinBUGS