Master 2 Biostatistiques – UE Bayes (STA305)
Exercices d'examen
Élements de correction

Boris Hejblum

17 décembre 2013

Exercice 1 (Durée conseillée : 1h)

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 : $f(y|\theta) = \theta e^{-\theta y}$.

1. Écrire le modèle bayésien
2. Écrire la vraisemblance de l'échantillon $y_i, i = 1, \ldots, n$
   Correction : $\prod_{i=1}^{n} \theta e^{-\theta y_i} = \theta^n e^{-\theta \sum_{i=1}^{n} y_i}$
3. Écrire la log-vraisemblance et sa dérivée première et seconde par rapport à θ.
   Correction : $L = n \log \theta - \theta y_i$; $L' = n/\theta - y_i$; $L'' = -n/\theta^2$
4. Écrire l'information de Fischer pour θ
   Correction : $I = -E(L'') = n/\theta^2$ (NB : espérance par rapport aux données)
5. Quel est la loi *a priori* de Jeffrey pour θ ; est-ce une densité propre ou impropre ?
   Correction : $\pi(\theta) = I = n/\theta \propto 1/\theta$: loi impropre.
6. En prenant pour loi *a priori* une loi exponentielle : $\theta \sim \tau e^{-\tau\theta}$, avec τ connu, écrire le numérateur de la loi *a posteriori* de θ
   Correction : $\pi(\theta|y_1, \ldots, y_n) \propto \theta^n e^{-\theta\left(\tau + \sum_{i=1}^{n} y_i\right)}$
7. Si en fait on ne connait pas τ, comment pourrait-t-on faire ?
   Correction : Approche par Bayes empirique : maximiser la vraisemblance marginale de τ. Ou modèle hiérarchique, c'est-à-dire donner un hyper-prior à τ.
8. 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 en donnant ses paramètres.
   Correction : La loi *a posteriori* est proportionnelle à $\theta^{\alpha-1}e^{-\beta\theta}$ avec $\alpha = n + 1$ et $\beta = \sum_{i=1}^{n} y_i + \tau$.
9. Que peut-on-dire de la distribution a posteriori de θ quand n est grand ?
   Correction : Quand n est grand, α est grand, et donc la distribution *a posteriori* est proche d'une normale.

Exercice 2 (Durée conseillée : 40 minutes)

1. Montrer que la fonction de répartition de la loi double-exponentielle de paramètre λ (correspondant à une loi de Laplace de paramètres 0 et $1/\lambda$), dont la densité s'écrit $g(x) = \frac{\lambda}{2}e^{-\lambda|x|}$, est :
$$G(x) = \begin{cases} \frac{1}{2}e^{\lambda x} & \text{si } x < 0 \\ 1 - \frac{1}{2}e^{-\lambda x} & \text{si } x \geq 0 \end{cases}$$

2. Proposer un algorithme basé sur la méthode par inversion, permettant de simuler la réalisation d'un échantillon de taille n d'une loi double-exponentielle de paramètre λ.
Correction :
$$G^{-1}(x) = \begin{cases} \dfrac{log\,(2x)}{\lambda} & \text{si } 0 < x < 0.5 \\ -\dfrac{log\,(2(1-x))}{\lambda} & \text{si } 0.5 \leq x < 1 \end{cases}$$
Pour i=1, ..., n :
1) échantillonner $u_i \sim \mathcal{U}_{[0,1]}$
2) $y_i := G^{-1}(u_i)$
$(y_1, \ldots, y_n)$ suit alors une loi double-exponentielle de paramètre λ.

3. Proposer maintenant un algorithme de Métropolis-Hastings indépendant pour échantillonner la loi Normale $\mathcal{N}(0,1)$ dont la densité s'écrit $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. On prendra comme loi de proposition la loi double-exponentielle de paramètre $\lambda = 1$. Expliciter la probabilité d'acceptation.
Correction : Faire pour i=1, ..., N :
1) Initialiser y_0
2) A l'itération i :
   - Simuler $\tilde{y} \sim DE(\lambda)$
   - Calcul de la probabilité d'acceptation α
   - Acceptation – Rejet de $\tilde{y}$
     — Simuler $u_i \sim U[0;1]$
     — $y_i = \begin{cases} \tilde{y} & \text{si } u_i \leq \alpha = min\left(1, \dfrac{g(\tilde{y})f(y_{i-1})}{g(y_{i-1})f(\tilde{y})}\right) = min\left(1, e^{\frac{{y_{i-1}}^2 - \tilde{y}^2}{2} + |y_{i-1}| - |\tilde{y}|}\right) \\ y_{i-1} & \text{sinon} \end{cases}$
3) i := i+1 retour à l'étape 2) tant que $i < N$

4. Quel résultat théorique garantit la convergence de l'algorithme de Metropolis-Hastings ? Expliquer brièvement.
Correction : Il s'agit du théorème ergodique. L'algorithme de Metropolis-Hastings échantillonne à l'aide d'une chaine de Markov dont la distribution stationnaire est la loi cible : la loi Normale centrée-réduite. Le théorème ergodique permet d'appliquer la loi des grands nombres aux réalisations de cette chaîne.

Exercice 3 (Durée conseillée : 1h20 minutes)

Nous nous intéressons ici à la réponse immunitaire de 14 patients suite à un vaccin contre la grippe. La réponse immunitaire y_i du patient i est mesurée par le *fold change* du taux d'anticorps dans le sang (il s'agit du rapport entre le taux d'anticorps mesuré après vaccination et celui mesuré avant vaccination). On dispose de 2 covariables : l'âge a_i et le sexe s_i des patients. Ces données sont présentées dans la table 1 page 6.

1. On veut faire une regression linéaire multiple afin d'expliquer la réponse vaccinale par l'âge et le sexe des patients.

   (a) Écrire le code de spécification du modèle en `BUGS` qui doit figurer dans le fichier externe .txt fourni à `JAGS`. Pour le(s) coefficient(s) de régression, vous utiliserez l'*a priori* suivant : `dnorm(0,1.0E-4)`. Pour le(s) paramètre(s) de précision, vous utiliserez l'*a priori* suivant : `dgamma(1.0E-4, 1.0E-4)`. **Correction :**

```
model{

    for(i in 1:N){
        y[i] ~ dnorm(mu[i], tausq)
        mu[i] <- beta1*a[i] + beta2*s[i]
    }

    beta0 ~ dnorm(0, 0.001)
    beta1 ~ dnorm(0, 0.001)
    tausq ~ dgamma(0.001, 0.001)

    sigma <- 1/sqrt(tausq)
}
```

   (b) Donner l'écriture mathématiques du modèle bayésien correspondant.

   (c) Comment appelle-t-on le type d'*a priori* suggéré ci-dessus pour le(s) paramètre(s) de régression ? Pourquoi sont-il conseillés, quelles propriétés de leurs distributions nous intéressent dans ce modèle ? Quel(s) en est(sont) l'(les) avantages ?

   Correction : Il s'agit d'*a priori* faiblement informatifs.
   C'est leur propriété de conjugaison avec la loi Normale nous intéresse ici.
   L'avantage principal de ces distributions est la connaissance analytique du posterior, ce qui permet d'accélérer les calculs.

   (d) Quelles variables doivent être fournies dans l'argument `data` de l'appel de la fonction `coda.samples()` ? Préciser l'argument correspondant aux données d'entrée à partir de la table 1 page 6.

   Correction :

```
list(
  N=14,
  y=c(1.98, 1.83, 2.79, 1.02, 1.91, 2.48, 2.22, 2.26, 2.25, 1.43,
      2.29, 2.16, 2.14, 1.80),
  a=c(49, 38, 34, 30, 31, 64, 41, 35, 35, 44, 33, 25, 45, 29),
  s=c(0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1)
)
```

2. Nous donnons certaines sorties de Winbugs (page 7 – ne pas considérer la table 2 page 8 pour l'instant). L'épaisseur d'échantillonnage (`thin` en anglais, qui désigne l'espacement entre les échantillons de la chaîne de markov utilisés dans l'échantillonnage final) est : $n_{epaisseur}$=1.

   (a) Quel est l'intérêt d'échantillonner simultanément plusieurs chaînes ?

   Correction : Cela permet d'évaluer la mélangeance des chaînes et donc de mieux pouvoir conclure à la convergence ou non des chaînes.

   (b) Quelle est la longueur de la phase de burn-in utilisée (notée $n_{burn-in}$) ? Quelle est la longueur de la phase d'échantillonnage utilisée (notée $n_{echantillonage}$) ?

   Correction : $n_{burn-in} = 10$ et $n_{echantillonage} = 100$

   (c) Commentez chaque diagnostic de convergence pour β_0,, β_1, et σ. Si besoin, proposez une solution pour améliorer les résultats. Justifiez quels diagnostics et quels arguments vous poussent à chaque conclusion.

   Correction : Les densités a posteriori ne sont pas très lisses : manifestement, il faut augmenter $n_{echantillonnage}$
   Les autocorrélogrammes de β_0 et β_1 sont globalement satisfaisants, celui concernant σ paraît plus agité, même s'il n'est pas pathologique.
   Les statistique de Gelmann & Rubin sont proches de 1 pour les 2 chaînes : satisfaisant.
   Les quantiles sont trop différents d'une chaîne à l'autre. Il faut augmenter $n_{burn-in}$
   Les traces sont confondus d'une chaine à l'autre : au delà de la 50ième observation, les chaînes semblent bien mélangées.

   (d) Dans l'état actuel, pouvez-vous conclure à la convergence ? Si ce n'est pas le cas, avec quel paramétrage proposeriez-vous de refaire tourner le modèle : $n_{burn-in}$, $n_{echantillonage}$, $n_{epaisseur}$? Justifier.

   Correction : Si la convergence semble ici atteinte en fin de chaîne, il faut clairement augmenter $n_{burn-in}$ et $n_{echantillonnage}$. On propose par exemple $n_{burn-in} = 100$ et $n_{echantillonnage} = 500$. Il semble également nécessaire d'augmenter $n_{epaisseur}$.

3. L'analyse est menée de nouveau avec un meilleur calibrage. Les résultats sont désormais interprétables et présentés dans la table 2 page 8.

   (a) Estimer l'effet du sexe sur la réponse immunitaire. Justifier.

   Correction : En utilisant l'estimateur minimisant la fonction de coût quadratique est la moyenne de la loi *a posteriori*. On estime donc l'effet du sexe à $\widehat{\beta_2}^{MMSE} = -0.69$

   (b) L'âge a-t-il un impact significatif sur la réponse immunitaire suite au vaccin ? Justifier.

   Correction : L'âge ne semble pas avoir d'impact sur la réponse immunitaire puisque son intervalle de crédibilité à 5% ([-0.019 ; 0.019]) inclut 0.

BONUS :

Les mesures de du taux d'anticorps ont été analysées par différents laboratoires d'analyses, induisant de la variabilité technique dans nos données. On décide de prendre en compte cette variabilité technique en ajoutant un effet aléatoire dans notre modèle de régression linéaire. Écrire le nouveau code de spécification du modèle.

Correction :

```
model{

    for(j in 1:L){
         for(i in 1:N){
             y[i,j] ~ dnorm(mu[i,j], tau)
             mu[i,j] <- beta0 + beta1*a[i] + beta2*s[i] + gamma0[j]
         }
         gamma0[j] ~ dnorm(0, tau.gamma0)
    }

    beta0 ~ dnorm(0, 0.0001)
    beta1 ~ dnorm(0, 0.0001)
    beta2 ~ dnorm(0, 0.0001)
    tau ~ dgamma(0.0001, 0.0001)
    tau.gamma0 ~ dgamma(0.0001, 0.0001)

    sigma <- 1/tau
    sigma.gamma0 <- 1/tau.gamma0
}
```

Patient (i)	FC[a] du taux d'anticorps (y_i)	Âge (a_i)	Sexe[b] (s_i)
1	1.98	49	0
2	1.83	38	1
3	2.79	34	0
4	1.02	30	1
5	1.91	31	1
6	2.48	64	0
7	2.22	41	0
8	2.26	35	0
9	2.25	35	0
10	1.43	44	1
11	2.29	33	0
12	2.16	25	0
13	2.14	45	0
14	1.80	29	1

[a] FC signifie *Fold Change*
[b] 0 désigne les femmes et 1 les hommes

TABLE 1 – Données de seroconversion chez des patients vaccinés contre la grippe

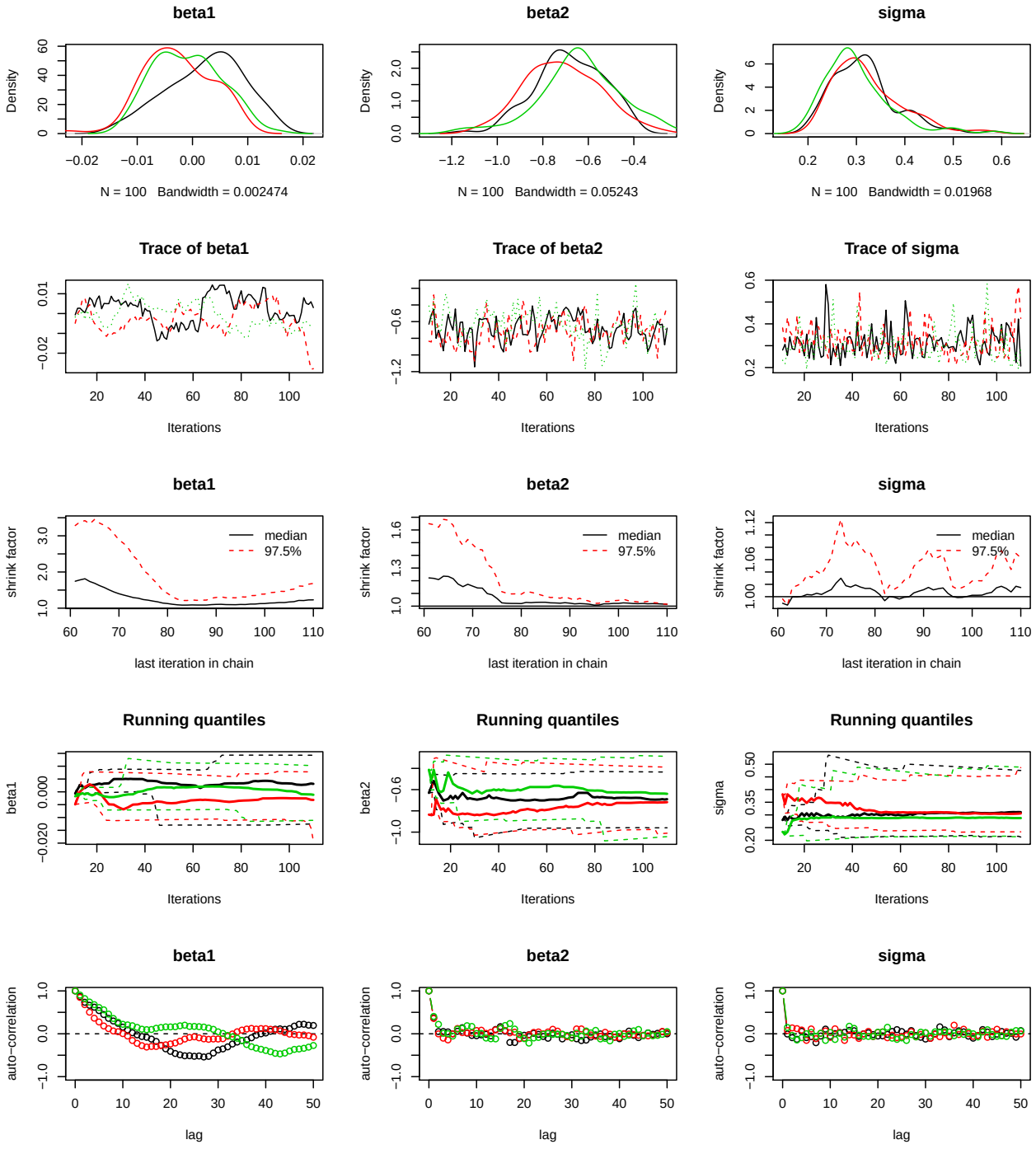

FIGURE 1 – Sorties graphiques pour l'effet des variables âge et sexe (coefficient `beta1` et `beta2` respectivement) et l'écart-type des réponses observées (`sigma`)

Noeud	moyenne	écart-type	2.5%	médiane	97.5%
β_0	-1.4E-4	0.0095	-0.019	-1.6E-4	0.019
β_1	-0.69	0.19	-1.074	-0.69	-0.31
σ	0.11	0.057	0.044	0.096	0.26

TABLE 2 – Sorties statistiques après recalibrage pour l'effet de la variable âge (coefficient β_{age}), celui la variable sexe (coefficient β_{sexe}) et l'écart-type des réponses observées (σ)