

Master 2 Biostatistiques – UE Bayes (STA305)

Sujet d'examen

Boris Hejblum

17 décembre 2013

Exercice 1 (Durée conseillée : 1h)

Les observations $y_i, i = 1, \dots, n$ 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 $y_i, i = 1, \dots, n$
3. Écrire la log-vraisemblance et sa dérivée première et seconde par rapport à θ .
4. Écrire l'information de Fischer pour θ
5. Quel est la loi *a priori* de Jeffrey pour θ ; est-ce une densité propre ou 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
7. Si en fait on ne connaît pas τ , comment pourrait-t-on faire ?
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.
9. Que peut-on dire de la distribution *a posteriori* de θ quand n est grand ?

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 .

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.
4. Quel résultat théorique garantit la convergence de l'algorithme de Metropolis-Hastings ? Expliquer brièvement.

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 3.

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)`.
 - (b) Donner l'écriture mathématiques du modèle bayésien correspondant.
 - (c) Comment appelle-t-on le type d'*a priori* suggéré ci-dessus pour les paramètres de régression ? Pourquoi sont-ils conseillés, quelles propriétés de leurs distributions nous intéressent dans ce modèle ? Quel(s) en est (sont) l'(les) avantages ?
 - (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 3.
2. Nous donnons certaines sorties obtenues via les packages  `rjags` et `coda` (page 4 – ne pas considérer la table 2 page 5 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 : $n_{\text{epaisseur}}=1$.
 - (a) Quel est l'intérêt d'échantillonner simultanément plusieurs chaînes ?
 - (b) Quelle est la longueur de la phase de chauffe utilisée (notée n_{chauffe}) ? Quelle est la longueur de la phase d'échantillonnage utilisée (notée $n_{\text{echantillonnage}}$) ?
 - (c) Commentez chaque diagnostic de convergence pour β_{age} , β_{sexe} , et σ . Si besoin, proposez une solution pour améliorer les résultats. Justifiez quels diagnostics et quels arguments vous poussent à chaque conclusion.
 - (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_{chauffe} , $n_{\text{echantillonnage}}$, $n_{\text{epaisseur}}$? Justifier.

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

3. L'analyse est menée de nouveau avec un meilleur calibrage. Les résultats sont désormais interprétables et présentés dans la table 2 page 5.
 - (a) Estimer l'effet du sexe sur la réponse immunitaire en minimisant la fonction de coût quadratique. Justifier.
 - (b) L'âge a-t-il un impact significatif sur la réponse immunitaire suite au vaccin ? Justifier.

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.

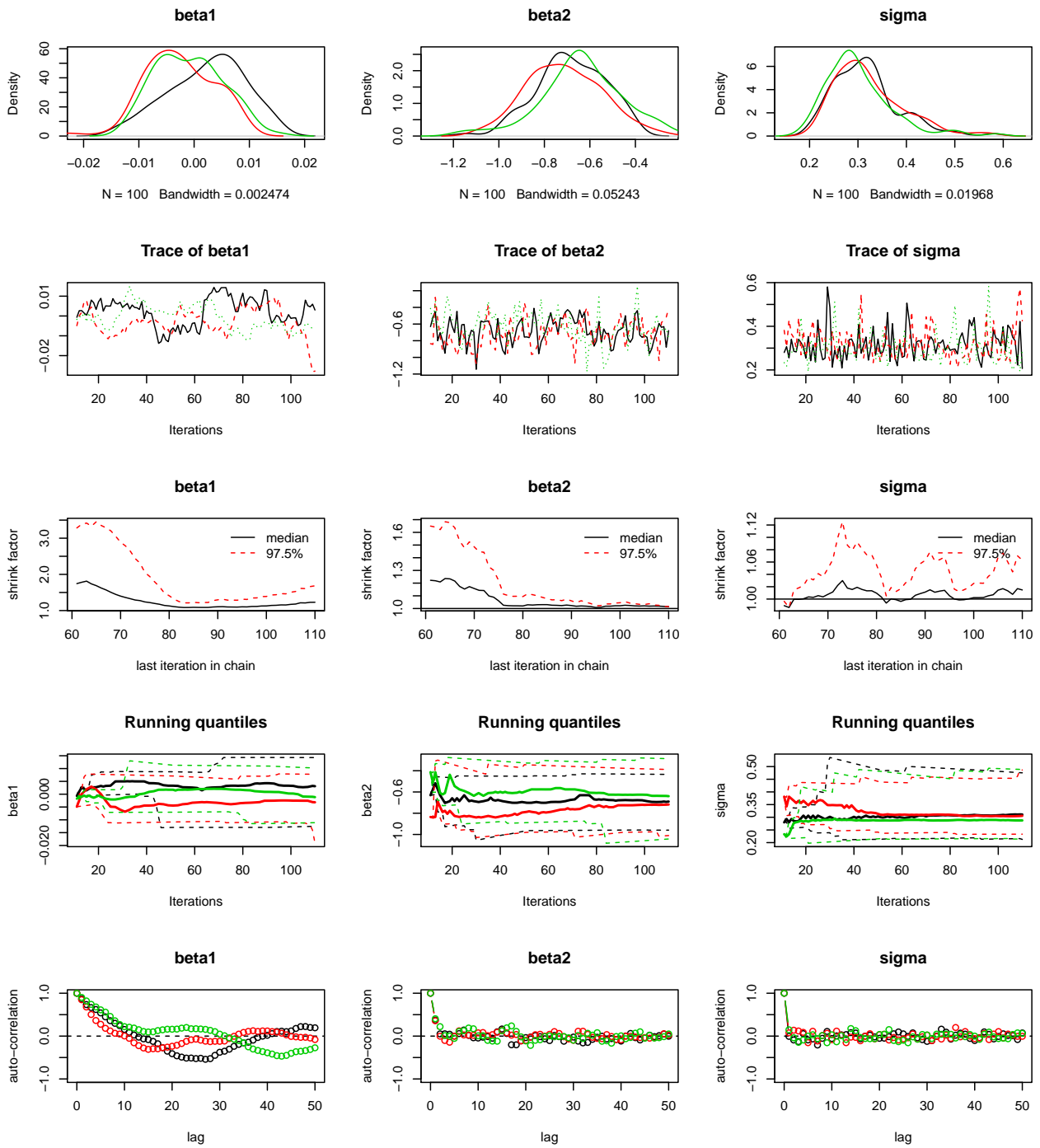


FIGURE 1 – Sorties graphiques pour l'effet des variables âge et sexe (coefficient β_1 et β_2 respectivement) et l'écart-type des réponses observées (σ)

Noeud	moyenne	écart-type	2.5%	médiane	97.5%
β_{age}	-1.4E-4	0.0095	-0.019	-1.6E-4	0.019
β_{sexe}	-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 via le package `coda` 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)