MPA - Ensimag 2A **2012-13**

TPs de MPA

Les TPs peuvent être effectués en binôme, et donnent lieu à un compte-rendu noté à rendre **par courrier électronique à son enseignant de TD avant le lundi 3 décembre à 19h.** Toute journée de retard dans la réception du courrierélectronique par l'enseignant de TD concerné entraînera un décompte de 4 points sur la note finale. Le compte-rendu ne dépassera pas 15 pages incluant les formules, figures, tableaux et les principales commandes de `R` (format doc ou PDF acceptés). Le barème de notation prend en compte l'exactitude, la qualité de la présentation, les commentaires et la discussion des résultats. Les légendes des figures et de leurs axes devront être explicites sans référence systématique au texte. La description des algorithmes devra garantir que l'on puisse les programmer en `R` sans ambiguité. La qualité de la rédaction, en français ou en anglais, sera aussi notée.

Exercice 1. Modèle de Poisson. On souhaite estimer le nombre moyen d'occurrences d'un phénomène donné, correspondant par exemple au nombre de clics journaliers sur un type de produit spécifique dans un site de vente en ligne. Pour cela, on dispose de $n = 19$ observations entières positives ou nulles, notées $y_1, \ldots, y_n$.

y :	9	5	7	15	9	0	22	1	9	11	9	13	12	11	2	1	26	7	0

1. On suppose les observations indépendantes et de loi de Poisson de paramètre $\theta > 0$. Déterminer la vraisemblance du paramètre dans ce modèle.
2. On suppose que la loi a priori est non informative
$$p(\theta) \propto \frac{1}{\theta}, \quad \theta > 0$$
Déterminer la loi a posteriori du paramètre θ. Quelle est l'espérance de la loi a posteriori ?
3. Rappeler le principe de l'algorithme de Metropolis-Hasting. Ecrire dans un programme en `R` une version de cet algorithme pour simuler la loi a posteriori du paramètre θ. On pourra, par exemple, choisir une loi instrumentale exponentielle.
4. Fournir une estimation ponctuelle du paramètre θ (moyenne et médiane a posteriori). Donner un intervalle de crédibilité à 95% pour ce paramètre.
5. Représenter graphiquement l'histogramme de la loi a posteriori du paramètre θ obtenu suivant l'algorithme de Metropolis-Hasting. Superposer la loi obtenue à la question 2.
6. Déterminer la loi prédictive *a posteriori* d'une nouvelle donnée $\tilde{y}$. Ecrire un algorithme de simulation de cette loi et comparer l'histogramme des résultats simulés aux données. Quelles critiques pouvez-vous faire du modèle proposé pour les données et le paramètre θ.

Exercice 2. Modèle hiérarchique, effet aléatoire et *shrinkage*. Dans un modèle hiérarchique, les observations sont modélisées conditionnellement à certains paramètres, et ces paramètres sont eux-mêmes décrits par des lois de probabilités dépendant d'autres paramètres, appelés *hyper-paramètres*.

L'objectif de cet exercice est d'estimer le risque pour un groupe de rats de décéder d'une tumeur cancéreuse à l'issue d'un traitement thérapeutique spécifique (les rats ont été sélectionnés pour développer la tumeur). Nous disposons pour cela de 71 groupes de rats, dont le nombre d'individus varie entre 14 et 52. Pour tenir compte de différences dans les échantillons considérés, nous souhaitons utiliser un modèle hiérarchique pour le risque. On note θ_i , i variant de 1 à 71, la probabilité qu'un rat du groupe i développe une tumeur, et on considère le vecteur $\theta = (\theta_i)$. Dans un groupe de rats de taille n_i, on modélise le nombre y_i de rats avec une tumeur par une loi binomiale de paramètres n_i et θ_i.

A priori, on modélise θ_i par une loi Beta de paramètres α et β. Les réels positifs α et β sont les hyperparamètres du modèle. La figure ci dessous représente la structure du modèle hiérarchique que l'on souhaite ajuster aux données.

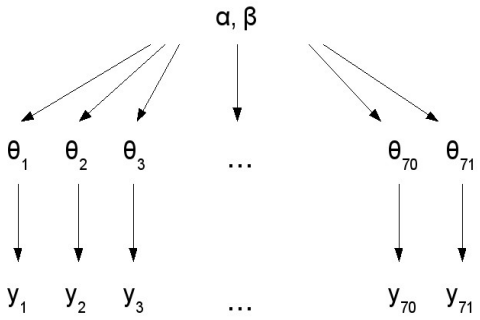

Figure 1: *Structure du modèle hiérarchique*

Il reste maintenant à spécifier les paramètres α et β. Afin d'effectuer un traitement bayésien du modèle, c'est-à-dire apprendre la loi a posteriori jointe de tous les paramètres du modèle, il est important ded spécifier une loi a priori pour α et β. Comme nous n'avons pas d'information sur l'ensemble des paramètres, il est naturel de chercher à décrire une loi a priori non informative. Une loi uniforme ne convient pas, car elle conduit à une loi a posteriori non intégrable. Pour des raisons que l'on admettra ici, il est raisonnable de choisir une loi a priori qu'on voudra uniforme pour le couple

$$(\mu, \nu) = \left(\frac{\alpha}{\alpha + \beta}, \frac{1}{\sqrt{\alpha + \beta}} \right).$$

Afin d'apprendre la loi a posteriori de tous les paramètres du modèle, nous souhaitons programmer une méthode de Monte-Carlo par chaîne de Markov hybride. Il s'agit

de mettre à jour les paramètres du modèle de manière séquentielle, pour un nombre fixé de pas. Voici une description synthétique de la méthode de MCMC à programmer :

Etape 1. Partir de valeurs arbitraires α_0 , β_0 et θ_0 et fixer un nombre de pas pour l'algorithme, N. Pour $t = 1, \ldots, N$, répéter les opérations suivantes :

Etape 2. Mettre à jour les paramètres α_t , β_t à partir de leur valeur au pas $t-1$ à l'aide de l'algorithme de Métropolis-Hasting.

Etape 3. Mettre à jour le vecteur de paramètres θ_t à partir de la valeur précédente θ_{t-1} par échantillonage de Gibbs.

Etape 4. Après N balayages de l'ensemble des paramètres, choisir un nombre $b < N$ et conserver les valeurs de θ_t pour $t > b$ (b pour la période de *burnin*).

1. Télécharger le fichier de données `rats.asc` à l'adresse suivante

   `http://www.stat.columbia.edu/~gelman/book/data/`

2. Calculer analytiquement le terme général de la densité $p(\alpha, \beta)$ par un changement de variables. Nous pouvons remarquer que cette densité est impropre (la densité n'est pas intégrable).
3. Donner l'expression de la vraisemblance $p(y|\theta)$ et de la loi conditionnelle $p(\theta|\alpha, \beta, y)$.
4. Calculer le rapport de Metropolis permettant la mise à jour du couple d'hyper-paramètres (α, β) (étape 2). Pour cette question, on pourra choisir librement la loi de transition instrumentale servant à proposer de nouvelles valeurs des paramêtres.
5. Rappeler brièvement le principe de l'échantillonnage de Gibbs. Décrire l'algorithme de mise à jour du paramètre θ (étape 3).
6. Programmer l'algorithme MCMC dans le langage `R`. Donner une estimation ponctuelle (moyenne conditionnelle, médiane et mode a posteriori) pour chacun des paramètres θ_i. Calculer les intervalles de crédibilité à 95% de ces paramètres .
7. Comparer ces estimations avec les estimations de maximum de vraisemblance $\hat{\theta}_i = y_i/n_i$. On pourra trier les valeurs y_i/n_i par ordre croissant, et afficher les moyennes a posteriori des lois marginales (les θ_i) en utilisant le même ordre. Commenter le phénomène de régularisation des risques observé dans cette courbe.