TP « Algo et Stat »

A rendre avant le lundi 8 février

Partie I: Estimation d'une fréquence dans un modèle « Beta-Binomial »

Dans cette première partie, on s'intéresse à l'estimation bayésienne d'un modèle à un paramètre. On souhaite estimer la probabilité p de naissance d'une fille dans une population donnée. On dispose pour cela d'un échantillon de $n = 20$ individus, parmi lesquels $y = 9$ sont des filles.

1) Comment modéliseriez-vous de façon simple la distribution du nombre de filles Y dans une population de n individus, sachant la probabilité $p = P$r(un nouveau né est une fille). Quelles sont les hypothèses nécessaires à votre choix ?

2) On peut choisir pour la loi a priori du paramètre p la loi uniforme sur [0,1]. Il s'agit d'une loi a priori non informative, car elle donne le même poids à toutes les valeurs que peut prendre le paramètre. Décrire la loi a posteriori du paramètre p : forme analytique, histogramme et densité.

3) Rappeler le principe de l'algorithme de Metropolis-Hasting. Implémenter une version de cet algorithme pour simuler la loi a posteriori de p. Fournir une estimation ponctuelle du paramètre p (moyenne conditionnelle, médiane et mode a posteriori). Donner un intervalle de crédibilité à 95% de ce paramètre.

4) Représenter graphiquement l'estimation de la loi a posteriori du paramètre p suivant l'algorithme de Metropolis-Hasting. Superposer la « vraie » loi a posteriori de p (faire varier le nombre de pas de votre algorithme de Metropolis-Hasting).

5) Prédiction. On observe une nouvelle naissance. Calculer, sachant y, la probabilité que cette nouvelle naissance soit une fille. Vérifier à l'aide d'une simulation.

Nous avons vu dans cette première partie l'estimation bayésienne d'un seul paramètre unidimensionnel. Cependant, pour des modèles plus complexes, nous serons souvent amenés à estimer plusieurs paramètres simultanément. C'est l'objet de la seconde partie de ce TP.

Partie II: Modèle hiérarchique

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 les hyperparamètres.

L'objectif de cette partie est d'estimer le risque de développer une tumeur dans un groupe de rats, à travers un modèle hiérarchique. Nous disposons pour cela de 71 groupes de rats, dont le nombre d'individus varie entre 14 et 52. On note θ_i, i variant de 1 à 71, la probabilité qu'un rat du groupe i développe une tumeur. On note θ le vecteur des θ_i. Dans chaque 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.

Cette fois, on ne souhaite pas utiliser une loi a priori non informative pour θ. Pour des raisons de simplicité de calcul, on modélise chaque θ_i par une loi Beta(α,β). Les réels α et β sont appelés les hyperparamètres du modèle. Selon la valeur des hyperparamètres, la loi Beta peut prendre une grande variété de forme (dont la loi uniforme). Cette paramétrisation de θ est donc flexible. La figure 1 représente la structure de ce modèle hiérarchique.

Il reste maintenant à spécifier α et β. Afin d'effectuer un traitement bayésien complet du modèle (c'est-à-dire apprendre la distribution a posteriori de tous les paramètres du modèle), il faut spécifier une distribution a priori pour α et β. Comme nous n'avons pas d'information a priori sur la distribution du taux θ, il est naturel de spécifier une loi a priori non informative pour α et β. Une loi uniforme sur le couple (α, β) ne convient pas, car elle conduit à une loi a posteriori impropre (non intégrable). Pour des raisons que nous ne détaillons

pas ici, il est raisonnable de choisir une loi a priori uniforme sur le couple $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$.

Afin d'apprendre la distribution a posteriori de tous les paramètres du modèle, une méthode est de programmer une chaîne de Markov par méthode de Monte-Carlo (MCMC). Il s'agit de mettre à jour les paramètres du modèle de façon séquentielle, pour un nombre fixé de pas. Voici une description synthétique de l'algorithme MCMC à implémenter:

a – Partir de valeurs arbitraires α_0 , β_0 et θ_0 et fixer un nombre de pas N. Pour t:=1..N, répéter:

b – Mettre à jour le couple de paramètres courant $(\alpha, \beta)_t$ à partir de la valeur précédente $(\alpha, \beta)_{t-1}$ avec un pas de Métropolis-Hasting.

c – Mettre à jour le vecteur de paramètres courant θ_t à partir de la valeur précédente θ_{t-1} avec un pas d'échantillonnage de Gibbs.

d – On choisit un nombre B < N et on conserve les θ_t et les $(\alpha, \beta)_t$ pour t > B (B pour Burn-in).

Si N et B sont suffisamment grands, et si les mises à jours sont faites correctement, les $(\theta_t, (\alpha, \beta)_t)$ pour t > B seront de loi Prob($\theta, \alpha, \beta \mid y$).

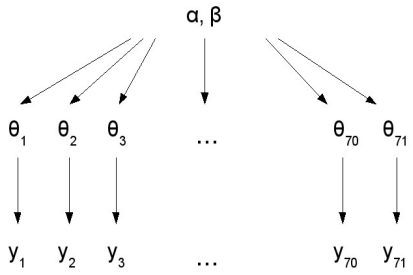

Figure 1: *Structure du modèle hiérarchique*

1) Récupérer le fichier *rats.asc* à l'adresse http://www.stat.columbia.edu/~gelman/book/data/ qui contient les données n_i et y_i.

2) Calculer analytiquement le terme général de la densité Prob(α, β) par un changement de variable, puis Prob($y \mid \theta$) et Prob($\theta \mid \alpha, \beta, y$) (utiliser la formule de Bayes).

3) Calculer le ratio de Metropolis pour la mise à jour du couple de paramètres (α, β).

4) Rappeler brièvement le principe de l'échantillonnage de Gibbs. Décrire son application à la mise à jour du paramètre θ.

5) Implémenter l'algorithme MCMC complet. Donner une estimation ponctuelle (moyenne conditionnelle, médiane et mode a posteriori) des paramètres θ_i. Calculer les intervalles de crédibilité à 95% des θ_i.

6) Comparer ces estimations avec les estimations de maximum de vraisemblance (*yi/ni*). Commenter.