Université de Bordeaux
M1 MAS
UE : Simulation stochastique et méthodes bayésiennes pour le traitement du signal
Année 2017-2018

TP5 - Modèles hiérarchiques bayésiens avec RJAGS

Les TP de cette UE sont inspirés de ceux proposés par François Caron, Jean-François Giovannelli, et Adrien Todeschini qui ont assuré cet enseignement pendant plusieurs années.

Dans ce TP, on propose d'illustrer l'intérêt des modèles hiérarchiques bayésiens par rapport à une approche fréquentiste. Ce TP est également l'occasion d'une première utilisation du package `rjags` pour décrire un modèle hiérarchique bayésien à partir du langage et logiciel JAGS.

1 Niveaux moyens des lycées

Soit Y_{ij} le niveau d'un étudiant i dans un lycée j, et soit N_j le nombre d'étudiants dans le lycée j, pour $j = 1, \ldots, m$. On considère le modèle à effets aléatoires suivant

$$Y_{ij} = \theta_j + \epsilon_{ij},\ \epsilon_{ij} \sim_{iid} \mathcal{N}(0, \sigma^2)$$

où θ_j est le niveau moyen dans le lycée j. On cherche à estimer les moyennes des lycées $\theta_1, \ldots, \theta_m$.

1.1 Approche fréquentiste

Pour chaque $1 \leq j \leq m$, on rappelle que l'estimateur du maximum de vraisemblance de θ_j est

$$\theta_j^{MV} = \frac{\sum_{i=1}^{N_j} Y_{ij}}{N_j},$$

et qu'un intervalle de confiance de niveau 95% est donné par

$$\theta_j^{MV} \pm 1.96 \frac{\sigma}{\sqrt{N_j}}.$$

Dans **R**, récupérer les données simulées contenues dans le fichier `lycees.RData` avec la commande `load`. Ce fichier contient :

- la matrice `Y` (niveaux des étudiants)
- le vecteur `N` (nombre d'étudiants par lycée)
- l'écart-type des effets aléatoires `sigma`
- le vecteur des vrais niveaux des `lycées theta.true`

Q. 1 *Observer les valeurs et tracer un boxplot des observations* `Y` *à partir du code ci-dessous.*

```
## Recuperation des donnees
load("lycees.RData")

m = length(N) # nombre de lycees

# observations : niveaux des etudiants
boxplot(Y)

# nombre d'etudiants par lycee
plot(N)
```

Q. 2 *Compléter le code ci-dessous pour calculer l'estimateur du maximum de vraisemblance et un intervalle de confiance à 95%. Tracer sur un même graphique les résultats et les vraies valeurs. Utiliser la fonction* `segments` *pour les intervalles de confiance. Observer les limites de cette approche.*

```
## Estimation par maximum de vraisemblance
# moyenne
theta.MLE = colMeans(Y, na.rm=TRUE)

# intervalles de confiance
theta.inf.MLE = ### .... ###
theta.sup.MLE = ### .... ###

# graphique
dev.new()
plot(theta.MLE, type="p", pch=16,
     ylim=range(theta.inf.MLE, theta.sup.MLE),
     xlab="lycee",
     ylab="niveau")
segments(1:m, theta.inf.MLE, 1:m, theta.sup.MLE)

points(theta.true, type="p", pch=18, col="green") # vraies valeurs
```

1.2 Approche bayésienne : modèle hiérarchique

On introduit maintenant le modèle multi-niveaux suivant :

- niveau 1 (vraisemblance) : $Y_{ij} = \theta_j + \epsilon_{ij},\ \epsilon_{ij} \sim_{iid} \mathcal{N}(0, \sigma^2)$,
- niveau 2 (distribution de la population) : $\theta_j|\phi \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$,
- niveau 3 (hyperparamètres) : a priori vague $\pi(\phi)$ sur le couple $\phi = (\mu_\theta, \sigma_\theta^2)$.

La distribution a posteriori

$$\pi(\theta_1, \ldots, \theta_m, \phi|Y) \propto \left(\prod_{j=1}^{m} \prod_{i=1}^{N_j} f(Y_{ij}|\theta_j) \right) \left(\prod_{j=1}^{m} \phi(\theta_j|\phi) \right) \pi(\phi)$$

ne peut être calculée de façon analytique mais peut être approchée par méthodes MCMC.

BUGS (`http://www.mrc-bsu.cam.ac.uk/bugs/`) est un logiciel ainsi qu'un langage pour décrire un modèle hiérarchique bayésien. Le logiciel implémente les méthodes MCMC de façon automatique (échantillonneur de Gibbs, a priori conjugué, Slice sampler, Metropolis-Hastings, ...) pour échantillonner les paramètres inconnus selon la loi a posteriori décrite en langage BUGS. Ce logiciel très simple d'utilisation a largement popularisé l'usage des méthodes MCMC parmis les praticiens.

JAGS (`http://mcmc-jags.sourceforge.net/`) est un clone de BUGS que nous allons utiliser via son interface **R** fournie par le package `rjags`. Le manuel utilisateur est disponible à l'adresse :
`http: //sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/jags_user_manual.pdf`

Q. 3 *Le modèle hiérarchique est décrit en langage BUGS dans le fichier* `lycees.bug`*. Lire et comprendre le fichier en notant que, par convention en langage BUGS, la loi normale* `dnorm` *est paramétrée par une moyenne et une précision* $\tau = \frac{1}{\sigma^2}$

Q. 4 *Charger le package* `rjags` *et compiler le modèle à l'aide de la fonction* `jags.model`*.*

Q. 5 *Générer 1000 échantillons MCMC pour la variable theta avec la fonction* `coda.samples`*.*

Q. 6 *Charger le package* `coda` *(outils d'analyse de convergence de chaînes MCMC). Pour l'une des composantes composantes* θ_j *:*

- *tracer les échantillons avec la fonction* `traceplot`*.*
- *tracer l'estimation des quantiles en fonction des itérations avec la fonction* `cumuplot`*.*

Ces premières itérations sont une étape de chauffe (burn in) de l'algorithme.

Q. 7 *Générer 5000 nouveaux échantillons MCMC pour les variables* $\theta_1, \ldots, \theta_m$*.*

Q. 8 *Calculer les statistiques avec la fonction summary. En extraire les moyennes empiriques et un intervalle de confiance à 95% basé sur les quantiles empiriques. Si* `s` *est la sortie de la fonction* `summary`*, les moyennes sont contenues dans* `s$statistics[,"Mean"]`*, les quantiles sont contenus dans* `s$quantiles[,"2.5%"]` *et* `s$quantiles[,"97.5%"]`*.*

Q. 9 *Tracer sur un même graphique les résultats MCMC, les résultats MLE et les vraies valeurs.*

2 Taux de cancer du sein dans les comtés de Caroline du Nord (MacLehose et al. 2007)

Soit Y_i le nombre de cas de cancer dans le i-ème compté, et soit N_i la population de chaque comté en 2004, pour $i = 1, \ldots, m$. On considère le modèle suivant :

$$Y_i \sim Poisson(N_i \lambda_i).$$

On souhaite estimer le taux de cancer λ_i par comté.

2.1 Approche fréquentiste

On rappelle que l'estimateur du maximum de vraisemblance pour la moyenne d'une loi de Poisson est donné par :

$$\hat{\lambda}_i^{MV} = \frac{Y_i}{N_i}$$

et qu'un intervalle de confiance pour λ_i de niveau $1-\alpha$ est donné par :

$$\frac{1}{2N_i}\chi^2(\frac{\alpha}{2}, 2Y_i) \leq \lambda_i \leq \frac{1}{2N_i}\chi^2(1-\frac{\alpha}{2}, 2Y_i+2)$$

où $\chi^2(p,k)$ est le quantile d'odre $p \in]0,1[$ d'une loi du chi2 à k degrés de liberté.

2.2 Approche bayésienne : modèle hiérarchique

On pose $\theta_i = \log(\lambda_i)$, et on introduit le modèle multi-niveaux suivant :
- niveau 1(vraisemblance) : $Y_i|\theta_i \sim Poisson(N_i \exp(\theta_i))$,
- niveau 2 (distribution de la population) : $\theta_i|\phi \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$,
- niveau 3 (hyperparamètres) : a priori vague $\pi(\phi)$ sur le couple $\phi = (\mu_\theta, \sigma_\theta^2)$.

Dans **R**, récupérer les données contenues dans le fichier `counties.txt` avec la commande : `counties = read.table('counties.txt', header=TRUE)`

La liste récupérée contient :
- la population de chaque comté : `counties$pop`
- le nombre de cas de cancer par comté : `counties$cases`

Q. 10 *Calculer l'estimateur du maximum de vraisemblance et un intervalle de confiance à 95% (utiliser la fonction* `qchisq`*), et trier les comtés par ordre décroissant en utilisant la fonction* `order`*. Tracer les résultats et observer les limites de cette approche.*

Q. 11 *Ecrire le modèle hiérarchique en langage BUGS. On utilisera la distribution de Poisson* `dpois`*, et on utilisera également la transformation du paramètre θ_i par la relation déterministe :* `lambda[i] <- exp(theta[i])`

Q. 12 *Reprendre la démarche de la partie précédente pour générer des échantillons MCMC de la loi a posteriori pour les $\lambda_1, \ldots, \lambda_m$ à partir du modèle hiérarchique décrit ci-dessus en langage BUGS.*

Q. 13 *Analyser la convergence des échantillons MCMC en simulant plusieurs chaînes indépendantes.*

Q. 14 *Etudier l'influence du choix de l'a priori $\pi(\phi)$ sur les résultats. Préciser le choix de $\pi(\phi)$ qui vous semble le plus adapté.*

Q. 15 *Proposer des estimateurs bayésiens pour $\lambda_1, \ldots, \lambda_m$ et les intervalles de confiance associés. Comparer graphiquement les résultats à ceux de l'approche fréquentiste. Interpréter les résultats obtenus. Quelle approche vous semble la plus intéressante ?*