UNIVERSITÉ DE NANTES

Master Professionnel Ingénierie Mathématique

Master professionnel II : Ingénierie mathématique : Option Statistique

Statistique Bayésienne.

Anne Philippe
Université de Nantes
Laboratoire de Mathématiques Jean Leray

Fiche 8. Algorithmes MCMC et programmation avec JAGS

Exercice 1. Algorithme de Gibbs

On considère la loi de densité

$$f(x,y) = Ce^{-x-y-xy}\,\mathbb{I}_{\mathbb{R}_+}(x)\,\mathbb{I}_{\mathbb{R}_+}(y)$$

où C est une constante positive (que l'on ne cherche pas à calculer)

1) Calculer la loi conditionnelle de X sachant Y, puis de Y sachant X
2) Proposer un algorithme MCMC pour simuler une chaîne de Markov de loi invariante f
3) Programmer cet algorithme.
4) Représenter une réalisation de la chaine de Markov simulée.
5) Tracer sur un même graphique
   — la moyenne empirique de l'échantillon $(X_i^{x,y})_{i\in\{1,\dots,n\}}$ en fonction de n, pour différentes valeurs initiales (x,y) de la chaîne de Markov.

Exercice 2. Modèle linéaire

On observe (x_i, Y_i) $i = 1, ..., n$. On suppose que les variables aléatoires Y_i sont indépendantes et

$$Y_i = a + bx_i + \varepsilon_i$$

où les variables aléatoires ε_i sont iid suivant la loi gaussienne de moyenne zéro et de variance inconnue σ^2. Les paramètres à estimer sont a, b, σ^2. Les variables x_i sont déterministes.

1) Calculer la loi de Jeffrey, puis la loi a posteriori associée
2) Peut-on calculer explicitement les estimateurs de Bayes des paramètres a, b, σ^2 ?
3) Proposer un algorithme MCMC pour approcher les estimateurs de Bayes des paramètres (a, b, σ^2).

Exercice 3. Modèle de Poisson et JAGS

On dispose de N observations iid suivant la loi de Poisson de paramètre τ. On choisit comme loi a priori sur τ la loi exponentielle de paramètre a fixé.

1) Quelle est la loi a posteriori ?
2) Quel est l'estimateur de Bayes sous cout L^2 ?

3) Donner une approximation du plus court intervalle de crédibilité au niveau 95% en utilisant la fonction `qgamma`.

4) Donner une approximation du plus court intervalle de crédibilité au niveau 95% à partir d'un échantillon simulé de variables iid en suivant la loi a posteriori.

5) Ecrire le DAG de ce modèle.

6) Récupérer le fichier `http://www.math.sciences.univ-nantes.fr/~philippe/Enseignement_files/poisson.zip` qui contient les codes R& jags

- Le modèle JAGS s'écrit

JAGS (`modelPoisson.R`).

```
model
{
  a<- 8
  for( i in 1 : N ) {  x[i] ~ dpois(tau) }
  tau ~ dgamma(1,a)
}
```

- Pour exécuter JAGS et donc simuler une chaîne de Markov de loi limite la loi a posteriori, on exécute les commandes suivantes, dont le code est dans le fichier `Poisson.R`,

Commande R .

```
library(rjags)
#nombre de données
N <- 10
#x contient les données
x = c(5,  1,  5,   14,  3,19,  1,   1,4,22)
#
jags <- jags.model('modelPoisson.R',
                   data = list('x' = x,'N' = N),
                   n.chains = 1)

# période de "chauffe"
> update(jags, 1000)

# On va stocker dans samp  la chaîne de markov (10000 itérations)
# pour le  paramètre  tau

>samp = coda.samples(jags,c('tau'),10000)
```

- Exploitation de la chaîne simulée qui est stockée dans la variable `samp`.

Commande R .

```
#moments/quantiles
summary(samp)
#trajectoire et estimation de la densité a posteriori
plot(samp)
```

Remarque :
Dans l'objet `samp` les valeurs de la chaine de Markov simulée sont stockées dans la matrice `samp[[1]]`. les chaines des différents paramètres sont stockées en colonne, rangée par ordre alphabétique. pour cet exemple il y a un seul paramètre stocké dans la colonne 1 `samp[[1]][,1]`.
Lorsque l'on lance plusieurs chaines, les valeurs de la i ème chaines sont stockées dans `samp[[i]]`

7) Comparer les résultats numériques avec les résultats théoriques
   — Loi a posteriori
   — estimateur de Bayes
8) Pour contrôler la convergence des chaînes vers la loi limite, relancer le code avec 10 chaînes en parallèle.

Commande R .

le nombre de chaines est fixé par le paramètre `n.chains` de la fonction `jags.model`

9) On compare l'évolution des moyennes empiriques des 10 chaines

Commande R .

```
#moyenne empirique
plot(cumsum(samp[[1]][,1])/1:10000, type="l")
for (i in 2:10)
lines(cumsum(samp[[i]][,1])/1:10000, type="l")
```

10) Calculer le critère de Gelman. Commenter.
11) Comparer graphiquement les estimations de la densité a posteriori sur les différentes chaînes.

Commande R .

```
plot(density(samp[[1]][,1]), type="l")
for (i in 2:10)
lines(density(samp[[i]][,1]), type="l", col=i)
```

12) Superposer les estimations avec la densité théorique.
13) A partir de chacune des 10 chaines, donner une estimation du plus court intervalle de crédibilité. Comparer avec les résultats des questions 3-4

Exercice 4. Modèle de Gaussien pour combiner des mesures

Le fichier `http://www.math.sciences.univ-nantes.fr/~philippe/Enseignement_files/Mcombine.zip`
contient des codes R, des modèles jags et les données.

On veut estimer l'intensité du champ magnétique terreste (CMT) à partir de mesures effectuées sur des briques prélevées sur un mur de ND sous terre au Mt St Michel.

On suppose que les mesures sont gaussiennes indépendantes et de variances connues. Les données sont dans les fichiers `intensityCMT.dat` (mesure) et `SDintensityCMT.dat` (écart type)

4.1. **Modèle 1.** Le modèle le plus simple est

$$X_i \sim \mathcal{N}(\theta, s_i^2)$$

avec s_i est connu (`SDintensityCMT.dat`)

A partir des connaissance sur l'évolution du CM en France, on construit la loi a priori sur θ. On choisit une loi uniforme sur $[0, 200]$

Le modèle JAGS s'écrit

JAGS (`modelGauss0.R`).

```
model{
  for( i in 1 : N ) {  x[i] ~ dnorm(theta,prec[i]) }
theta ~ dunif(0,200)
}
```

Attention les paramètres de la loi normale sont la moyenne et l'inverse de la variance (aussi appelée précision).

1) Ecrire le DAG de ce modèle.
2) Donner une estimations de θ et calculer une région HPD de niveau 95%
3) Représenter la densité de la loi a posteriori de θ et ajouter les données sur le graphique en utilisant la commande `rug(x)`

4.2. **Modèle 2.** On fait évoluer ce modèle en ajoutant une structure hierarchique. On suppose que les vraies mesures du CMT ne sont pas identiques. Autrement dit

$$X_i \sim \mathcal{N}(\mu_i, s_i^2)$$

$$\mu_i \sim \mathcal{N}(\theta, \sigma^2)$$

On conserve la même loi sur θ et la loi sur σ^2 est la loi de shrinkage (une loi non informative) définie par

$$\sigma^2 = u/(1-u) \qquad u \sim U(0,1)$$

Le modèle JAGS s'écrit

JAGS (`modelGauss1.R`).

```
model{
  for( i in 1 : N ) {  x[i] ~ dnorm(mu[i],prec[i]) }
  for( i in 1 : N ) {  mu[i] ~ dnorm(theta,p) }

   p <- u/(1-u)
   u ~ dunif(0,1)
   theta ~ dunif(0,200)
   sigma2<- 1/p
}
```

1) Ecrire le DAG de ce modèle.
2) Donner une estimations de θ et calculer une région HPD de niveau 95%
3) Représenter la densité de la loi a posteriori de θ et ajouter les données.
4) Donner une estimation de σ^2.
5) Le modèle précédent correspond à ce modèle en supposant la variance σ^2 connue et égale à zéro. Au vu de l'estimation de σ^2 cette hypothèse du modèle 1 était-elle justifiée ?

4.3. **Modèle 3.** On peut generaliser le modèle hierarchique en supposant que les variances des μ_i sont différentes :

$$\mu_i \sim \mathcal{N}(\theta, \sigma_i^2)$$

où les σ_i sont iid suivant la loi de shrinkage.

1) Ecrire le DAG du modèle.
2) Adapter le code du fichier `modelGauss1.R`.
3) Donner une estimations de θ et calculer une région HPD de niveau 95%
4) Donner une estimation des σ_i^2.
5) Est ce que l'hypothèse du modèle précédent où toutes les variances étaient supposées égales vous semble justifiée ?
6) Représenter la densité de la loi a posteriori de θ et ajouter les données.
7) Identifier sur le graphiques les données pour lesquelles l'estimation de σ_i prend des grandes valeurs. Commenter le résultat.

4.4. **Comparaison.**

1) Comparer les estimations de θ et les régions HPD obtenues pour les trois modèles.
2) Conclure

Exercice 5. Regression non linéaire

La part chauffage dans la consommaton d'électricité s'écrit de la forme

$$Conso_t = f_t(Temp_t) + \varepsilon_t$$

avec

$$f_t(Temp_t) = (\alpha - \beta 1_{we}(t))\ (u - Temp_t)^+$$

où

- les données sont journalières,
- 1_{we} vaut 0 du lundi au vendredi et 1 les jours de week-end, samedi et dimanche,

— les (ε_t) sont des variables indépendantes identiquement distribuées gaussiennes centrées et de variance inconnue σ^2.
— $(u - Temp_t)^+ = max(0, u - Temp_t)$ vaut zéro lorsque la température dépasse le seuil u

Définition de la loi a priori : les variables sont indépendantes et les lois marginales vérifient
— Les lois a priori sur α et β ont une variance très grande : $\Gamma(0.04, 0.01)$ et $\Gamma(0.01, 0.01)$
— La loi sur σ^2 est la loi impropre $1/\sigma^2$
— La loi sur u est faiblement informative $\mathcal{N}(15, 1)$

1) Récupérer les données suivantes

```
http://www.math.sciences.univ-nantes.fr/~philippe/lecture/edf
load("edf")
```

2) À partir de quelques représentations graphiques, justifier le choix du modèle. (Indication : tracer la consommation en fonction de la température)
3) Ecrire le DAG de ce modèle
4) Écrire le code JAGS de ce modèle
5) Simuler une chaine de Markov dont la loi stationnaire est la loi a posteriori
6) Calculer une approximation de la loi a posteriori
7) Calculer une approximation de l'estimateur de bayes sous coût L^2, puis L^1
8) Calculer une approximation de la région de confiance bayésienne à 95 %, puis de la région hpd.
9) Vérifier la convergence de l'algo MCMC en utilisant le critère de Gelman et Rubin.
10) On suppose maintenant que l'on dispose d'une information a priori sur u : u suit une loi normale $\mathcal{N}(15, 1/10)$

    Reprendre les questions précédentes et comparer les résultats.