

## ECN option mathématiques Parcours S2D

Statistique Bayésienne.

Anne PHILIPPE  
Université de Nantes  
Laboratoire de Mathématiques Jean Leray

### EXERCICE 1. MODÈLE DE POISSON ET JAGS

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

1) Récupérer le fichier [http://www.math.sciences.univ-nantes.fr/~philippe/Enseignement\\_files/poisson.zip](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 chaîne de Markov simulée sont stockées dans la matrice `samp[[1]]`. les chaînes 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 chaînes, les valeurs de la *i* ème chaînes sont stockées dans `samp[[i]]`

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

**Commande R .**

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

- 4) On compare l'évolution des moyennes empiriques des 10 chaînes

**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")
```

- 5) Calculer le critère de Gelman. (`gelman.diag`).
- 6) 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)
```

- 7) A partir de chacune des 10 chaînes, donner une estimation du plus court intervalle de crédibilité. (`HPDinterval`)

## EXERCICE 2. MODÈLE DE GAUSSIEN POUR COMBINER DES MESURES

Le fichier [http://www.math.sciences.univ-nantes.fr/~philippe/Enseignement\\_files/Mcombine.zip](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 terrestre (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)

2.1. **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 connaissances sur l'évolution du CM en France, on construit la loi a priori sur  $\theta$ . 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  $\theta$  et calculer une région HPD de niveau 95%
- 3) Représenter la densité de la loi a posteriori de  $\theta$  et ajouter les données sur le graphique en utilisant la commande `rug(x)`

2.2. **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  $\theta$  et la loi sur  $\sigma^2$  est la loi de shrinkage (une loi non informative) définie par

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

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

2.3. **Modèle 3.** On peut generaliser le modèle hierarchique en supposant que les variances des  $\mu_i$  sont différentes :

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

où les  $\sigma_i$  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  $\theta$  et calculer une région HPD de niveau 95%
- 4) Donner une estimation des  $\sigma_i^2$ .
- 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  $\theta$  et ajouter les données.
- 7) Identifier sur le graphiques les données pour lesquelles l'estimation de  $\sigma_i$  prend des grandes valeurs. Commenter le résultat.

2.4. **Comparaison.**

- 1) Comparer les estimations de  $\theta$  et les régions HPD obtenues pour les trois modèles.
- 2) Conclure