STT3010 (Statistique informatique) — Hiver 2024
Devoir 4

Instructions

Date limite de remise: 16 avril à 23h59

Matériel à remettre:

- Un fichier .R ou .Rmd que je pourrai exécuter sans modification afin de reproduire tous vos résultats. Le fichier doit commencer avec le choix d'un germe avec la fonction `set.seed`, être structuré, et être dûment commenté.

Modalités de remise:

- Par Moodle.

Consignes:

- Le devoir est individuel, donc chaque étudiant(e) remet son propre travail. Par contre, vous êtes encouragés à discuter sans toutefois partager vos solutions complètes.
- Vous pouvez emprunter des sections du code fourni dans mes démonstrations R, ou vous en inspirer.

Considérez à nouveau le modèle hiérarchique bayésien du Devoir 3. On se rappelle que la log densité *a posteriori* conjointe de tous les paramètres est donnée par

$$\begin{aligned}\log f(\theta_1, \ldots, \theta_k, \sigma^2, \nu, \tau^2 \mid \mathbf{X}) = K - \frac{N+8}{2} \log \sigma^2 - \frac{k}{2} \log \tau^2 - \frac{1}{\sigma^2} - \frac{\nu^2}{2} - \tau^2 \\ - \frac{1}{2\tau^2} \sum_{i=1}^{k} (\theta_i - \nu)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} X_{ij}^2 - 2n_i \bar{X}_i \theta_i + n_i \theta_i^2 \right],\end{aligned}$$

où K est une constante, $N = \sum_{i=1}^{k} n_i$, et $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$. Dans ce devoir, on vise à utiliser les méthodes MCMC vues en classe afin de générer des observations selon cette densité. Par souci de simplicité, on retire $\mathbf{X}$ de la notation — on suppose donc que $(\theta, \sigma^2, \nu, \tau^2) \in \mathbb{R}^{13}$ est un vecteur aléatoire distribué selon la loi *a posteriori* décrite ci-haut, et on dénote par f (plutôt que $f(\cdot \mid \mathbf{X})$) sa fonction de densité.

En utilisant le même fichier `.RDS` que la dernière fois, commencez par importer les données en utilisant la commande

```
X <- readRDS(".../Devoir3_data.RDS")
```

et en n'oubliant pas de remplacer "..." par le répertoire dans lequel vous avez enregistré le fichier. Comme la dernière fois, je vous suggère de créer les objets k, $n_1, \ldots, n_k$, N et $\bar{X}_1, \ldots, \bar{X}_k$ en utilisant le code suivant.

```
k <- nrow(X)
n <- rowSums(!is.na(X))
N <- sum(n)
Xbar <- rowMeans(X, na.rm=TRUE)
```

Question 1

On peut démontrer que les lois conditionnelles du vecteur aléatoire $(\theta, \sigma^2, \nu, \tau^2)$ sont les suivantes:

$$\theta_i \mid (\theta_{-i}, \sigma^2, \nu, \tau^2) \sim \mathrm{N}\Big(\frac{\nu/\tau^2 + n_i \bar{X}_i/\sigma^2}{1/\tau^2 + n_i/\sigma^2}, \frac{1}{1/\tau^2 + n_i/\sigma^2}\Big),$$

$$\sigma^2 \mid (\theta, \nu, \tau^2) \sim \mathrm{InvGamma}\Big(\frac{N+6}{2}, 1 + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \theta_i)^2\Big),$$

$$\nu \mid (\theta, \sigma^2, \tau^2) \sim \mathrm{N}\Big(\frac{\sum_{i=1}^{k}\theta_i}{\tau^2 + k}, \frac{1}{1 + k/\tau^2}\Big);$$

pour ce qui est de τ^2, sa loi conditionnelle n'est pas d'une famille très connue, mais sa densité conditionnelle est donnée par

$$f(\tau^2 \mid \theta, \sigma^2, \nu) \propto \frac{1}{(\tau^2)^{k/2}} \exp\Big\{ -\tau^2 - \frac{1}{2\tau^2}\sum_{i=1}^{k}(\theta_i - \nu)^2 \Big\}.$$

En utilisant ces informations, utiliser l'algorithme de Gibbs afin de générer une chaîne de longueur $m = 10^6$ dont la distribution stationnaire est celle de $(\theta, \sigma^2, \nu, \tau^2)$.

Indice: Vous devriez être en mesure de générer à partir des lois conditionnelles de θ_i, σ^2 et ν en utilisant les fonctions `rnorm` et `rgamma` — rappelons que si $Y \sim \mathrm{Gamma}(\alpha, \beta)$, alors $1/Y \sim \mathrm{InvGamma}(\alpha, \beta)$. Afin de générer une observation de la loi conditionnelle de τ^2, vous pouvez emprunter le code suivant (pouvez-vous deviner sur quelle méthode il se base?).

```
accept <- FALSE
while(!accept){
y <- 1/rgamma(1, shape=k/2-1, rate=sum((theta-nu)^2)/2)
u <- runif(1)
accept <- (log(u) <= -y)
}
tau2 <- y
```

Question 2

Utilisez d'abord le code suivant afin de créer une implémentation `lf` de la log densité $\log f$.

```
SumSquares <- sum(X^2, na.rm=TRUE)
lf <- function(theta, sig2, nu, tau2){
- (N + 8)/2 * log(sig2) - k/2 * log(tau2) - 1/sig2 - nu^2/2 -
tau2 - sum((theta - nu)^2)/(2*tau2) - SumSquares/(2*sig2) +
sum(n*Xbar*theta)/sig2 - sum(n*theta^2)/(2*sig2)
}
```

Utilisez maintenant l'algorithme de Metropolis–Hastings afin de générer une chaîne de longueur $m = 10^6$ dont la distribution stationnaire est celle de $(\theta, \sigma^2, \nu, \tau^2)$. Comme distribution instrumentale, utilisez une loi normale centrée et de covariance proportionnelle à la matrice identité, c'est-à-dire qu'étant donné l'état précédent de la chaîne $(\theta, \sigma^2, \nu, \tau^2)^{(t-1)}$, le candidat au temps t sera défini par

$$(\theta, \sigma^2, \nu, \tau^2)' \mid (\theta, \sigma^2, \nu, \tau^2)^{(t-1)} \sim \mathrm{N}((\theta, \sigma^2, \nu, \tau^2)^{(t-1)}, \ell^2 I),$$

où I représente la matrice identité dans $\mathbb{R}^{13\times 13}$. Pour commencer, utilisez $\ell^2 = 0.1$, puis essayer différentes valeurs de ℓ^2 jusqu'à obtenir un taux d'acceptation des sauts entre 20% et 25%.