STT3010 (Statistique informatique) — Hiver 2024
Devoir 3

Instructions

Date limite de remise: 27 mars 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.

Question 1

On observe des données X_{ij}, $i \in \{1, \dots, k\}$, $j \in \{1, \dots, n_i\}$, groupées en strates indexées par i et où la $i$ème strate contient n_i observations. Considérez le modèle hiérarchique bayésien

$$
\begin{aligned}
X_{ij} \mid \{\theta_1, \dots, \theta_k, \sigma^2\} &\overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, \sigma^2), \quad i \in \{1, \dots, k\}, j \in \{1, \dots, n_i\} \\
\theta_i \mid \{\nu, \tau^2\} &\overset{\text{ind}}{\sim} \mathrm{N}(\nu, \tau^2), \quad i \in \{1, \dots, k\} \\
\sigma^2 &\sim \mathrm{InvGamma}(3, 1) \\
\nu &\sim \mathrm{N}(0, 1) \\
\tau^2 &\sim \mathrm{Exp}(1),
\end{aligned}
$$

où la loi Gamma inverse $\mathrm{InvGamma}(\alpha, \beta)$ a une densité donnée par

$$f(y) \propto y^{-(\alpha+1)} \exp\{-\beta/y\}, \quad y > 0.$$

On peut démontrer que la log densité *a posteriori* conjointe de tous les paramètres est donnée par

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

où K est une constante, $N = \sum_{i=1}^{k} n_i$, et $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$.

Vous trouverez sur Moodle un fichier de type `.RDS` qui contient les données X_{ij} so forme d'une matrice. Commencez par l'importer en utilisant la commande

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

en n'oubliant pas de remplacer "..." par le répertoire dans lequel vous avez enregistré le fichier. Les données sont sous la forme d'une matrice de format $k \times \max_{1 \le i \le k} n_i$ — pour les strates i telles que $n_i < \max_{1 \le i \le k} n_i$, les dernières entrées de la ligne i sont `NA`.

(a) En ignorant la constante K, c'est-à-dire en supposant que $K = 0$, créez une fonction R `lpost` qui calcule $\log f$. Votre fonction devrait prendre en entrée 4 arguments: un vecteur `theta` (qui contient les valeurs $\theta_1, \ldots, \theta_k$), `sig2` (qui correspond à σ^2), `nu` (qui correspond à ν), et `tau2` (qui correspond à τ^2).

Indice: Afin d'alléger la définition de votre fonction `lpost`, je vous suggère de d'abord calculer les quantités k, n_i, N, $\bar{X}_i$ et $\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2$. Pour calculer ces différentes sommes et moyennes, il pourrait vous être utile de savoir que les fonctions `sum`, `mean`, `rowMeans`, etc. acceptent un argument `na.rm`. Si vous spécifiez `na.rm=TRUE`, R ignorera les `NA`.

(b) En utilisant la fonction `optim` en R, calculez le maximum *a posteriori* de $(\theta_1, \ldots, \theta_k, \sigma^2, \nu, \tau^2)$. Pour ce faire, il vous faudra définir une version "vectorielle" de votre fonction `lpost`, qui prend en entrée un vecteur à la place d'arguments individuels:

```
lpost_optim <- function(param) lpost(param[1:k],
param[k+1], param[k+2], param[k+3])
```

Je vous suggère de jeter un coup d'oeil à la documentation sur la fonction `optim` (à l'aide de la commande `help(optim)`) et/ou à mon propre code de la démonstration R sur l'inférence bayésienne.

(c) En utilisant l'approximation de Laplace, c'est-à-dire l'approximation normale de la loi *a posteriori*, calculez des intervalles de crédibilité de niveau 95% pour chacun des paramètres θ_i. Rappelons qu'un tel intervalle pour θ_i est défini comme un intervalle R tel que la probabilité que $\theta_i \in R$ *a posteriori* est de 0.95.

(d) Soit la fonction $h : \mathbb{R}^k \to \mathbb{R}$ définie par

$$h(\theta_1, \ldots, \theta_k) = \frac{\exp\{\sum_{i=1}^{k} \theta_i\}}{1 + \exp\{\sum_{i=1}^{k} \theta_i\}}.$$

Utilisez l'approximation de Laplace exponentielle, telle que vue en cours, pour estimer l'espérance *a posteriori* de $h(\theta_1, \ldots, \theta_k)$.