

MÉTHODE DE ZONAGE PAR LISSAGE SPATIAL MODÈLE DE BOSKOV ET VERRALL

*L'objectif de ce TP est de construire un zonier par lissage spatial bayésien. Nous travaillerons à partir de la base de données **freMTPL** de la librairie **CASdatasets**. On suivra les différentes étapes décrites en cours afin de construire ce zonier.*

Partie introductive : Echantillonneur de Gibbs.

Rappel : l'échantillonneur de Gibbs permet de retrouver une densité multivariée par simulations MCMC (Monte Carlo Markov Chain), notamment lorsque cette densité est d'une forme inconnue ou non simulable facilement. Nous allons prendre ici un cas simple à titre d'illustration : une gaussienne bivariee.

Supposons $X \sim \mathcal{N}(0, 1)$, et $Y|X \sim \mathcal{N}(\rho X, \sqrt{1 - \rho^2})$. On cherche la loi de (X, Y) . Autrement dit, la densité conditionnelle est connue ici, mais nous chercherons à la retrouver par la technique de l'échantillonneur. Par analogie avec le zonier, on ne connaît pas la loi explicite de (N, U, V, τ, Λ) , mais on connaît les lois univariées de $\tau|(N, U, V, \Lambda)$, $\lambda|(N, U, V, \tau)$, $U|(N, V, \tau, \Lambda)$ et $V|(N, U, \tau, \Lambda)$ (comme pour $X|Y$ ci-dessous).

- (1) Programmer une fonction `R`, avec pour paramètres n (nombre de simulations) et ρ (coefficient de corrélation), qui vous permet de simuler X et $Y|X$. Stocker ces résultats dans un vecteur bivarié (X, Y) .
- (2) Simuler 10 000 réalisations de ce vecteur grâce à votre fonction, avec un coefficient de corrélation ρ égal à 0,98. Afficher sous forme graphique ces couples de points, par exemple en faisant un histogramme pour X et un histogramme pour Y .
- (3) Programmer maintenant une fonction `gibbs` avec les mêmes paramètres (n et ρ) qui vous permet de créer des couples réalisés de (X, Y) et de stocker chaque itération dans une matrice. Pour cela, remarquez que :

$$f(x|y) \sim f(y|x) f(x) \sim \mathcal{N}(\rho x, \sqrt{1 - \rho^2}) \mathcal{N}(0, 1) \sim \mathcal{N}(\rho y, \sqrt{1 - \rho^2}),$$

et appliquer le raisonnement suivant :

- on se donne une valeur initiale pour X , notée x_0 ;
 - on trouve y_0 en la simulant à partir de x_0 , ce qui donne le couple (x_0, y_0) .
 - on cherche maintenant (x_1, y_1) qui dépend de (x_0, y_0) , à partir de la loi de $X_1|Y_0, Y_1|X_1, \dots$
 - on répète ces étapes un grand nombre de fois.
- (4) Simuler 10 000 réalisations du vecteur (X, Y) avec votre fonction, et comparer les graphiques obtenus avec les graphiques précédents. Avez-vous réussi (les distributions convergent-elles) ?

Implémentation du zonier : modèle de Boskov-Verrall.

- (1) Les questions relatives à la construction des données et à la modélisation de la fréquence hors variable géographique sont identiques au précédent TP. Les réaliser.

- (2) Nous allons maintenant intégrer l'effet géographique via la covariable **Région**. Pour commencer, créer une matrice d'adjacence des régions.
- (3) On peut maintenant commencer à préparer les données en vue du lissage spatial :
 - extraire les prévisions du nombre de sinistres par région à partir du modèle retenu à la question 1.
 - extraire les nombres de sinistres observés par région.
 - extraire l'exposition globale par région.
 - calculer le nombre de voisins par région.
- (4) Mise en place du lissage spatial : usage de l'algorithme de Metropolis-Hastings, généralisation de Gibbs au cas où les distributions conditionnelles ne sont pas de forme connue et nécessitent donc d'utiliser l'Adaptive Rejection Sampling (avec 1000 simulations). Vous aurez besoin ici de la librairie **Runuran**.
 - Programmer les densités a posteriori des paramètres du problème (cf compléments ci-dessous. Les 2 premières densités ne sont en réalité pas à programmer, cf remarque ci-dessous).
 - programmer la mise à jour des paramètres grâce à la méthode Monte Carlo Markov Chain vue en cours. Pour cela, vous utiliserez notamment :
 - la fonction **urchisq** pour la simulation des lois des variances de U, V ;
 - les fonctions **ur** et **ars.new** pour la mise en place nécessaire de l'algorithme *Adaptive Rejection Sampling* lors de la simulation de U et V (conseil : pour ne pas faire exploser les temps de calcul, fixer les bornes suivantes pour la fonction **ars.new** : lower = -10 et upper = 10.
 - accéder aux densités a posteriori empiriques simulées.
 - en déduire les résultats finaux.

Compléments au TP.

On donne ci-après les distributions conditionnelles nécessaires à la simulation MCMC :

$$\begin{aligned}
 f(\tau | u, v, \lambda, n) &\sim \tau^{-r/2} \exp \left(-\frac{1}{2\tau} \left[\xi + \sum_{i \sim j} (u_i - u_j)^2 \right] \right) \\
 f(\lambda | u, v, \tau, n) &\sim \lambda^{-r/2} \exp \left(-\frac{1}{2\lambda} \left[\xi + \sum_{i=1}^r v_i^2 \right] \right) \\
 f(u_i | u_{-i}, v, \tau, \lambda, n) &\sim \exp \left(-c_i e^{u_i + v_i} + u_i n_i - \frac{nb_i}{2\tau} [u_i - \bar{u}_i]^2 \right) \\
 f(v_i | u, v_{-i}, \tau, \lambda, n) &\sim \exp \left(-c_i e^{u_i + v_i} + v_i n_i - \frac{1}{2\lambda} v_i^2 \right)
 \end{aligned}$$

Avec

- nb_i le nombre de voisins de la région i (régions limitrophes),
- \bar{u}_i : moyenne des effets spatiaux des voisins de i à i ,
- c_i : prévision du modèle GLM sans la covariable géographique.
- n_i : le nombre de sinistres observés pour la région i .

Remarque.

Les paramètres de variance suivent en fait une loi du χ^2 , dont les paramètres seront à déterminer. Pour la simulation du paramètre Λ par exemple, on utilisera l'expression

$$\frac{\xi + \sum_i^r v_i^2}{urchisq(...)}$$