7. Analyser les jeux de données envoyés en pièce jointe par l'enseignant. Décrire l'incertitude sur le(s) point(s) de changement pour ces jeux de données (localisation des points des changements et intervalles contenant chacun des points avec une probabilité supérieure à 50%).

TP 3 : Modèle de Poisson et mélanges – A remettre avant le lundi de la semaine 9. On souhaite estimer le nombre moyen d'occurrences d'un phénomène donné, correspondant par exemple au nombre de clics journaliers sur un type de produit spécifique dans un site de vente en ligne. Pour cela, on dispose de n observations entières positives ou nulles, notées $y_1, \ldots, y_n$. Au moins l'une de ces observations est non nulle.

Première partie.

1. On suppose les observations indépendantes et de loi de Poisson de paramètre $\theta > 0$. Déterminer la loi générative des données, y, dans ce modèle.
2. On suppose que la loi a priori est non informative

$$p(\theta) \propto \frac{1}{\theta}, \quad \theta > 0\,.$$

   Déterminer la loi a posteriori du paramètre θ. Quelle est l'espérance de la loi a posteriori ?
3. Rappeler le principe de l'algorithme de Metropolis-Hasting. Ecrire dans un programme en `R` une version de cet algorithme pour simuler la loi a posteriori du paramètre θ. On pourra, par exemple, choisir une loi instrumentale exponentielle.

4. Fournir une estimation ponctuelle du paramètre θ (moyenne et médiane a posteriori). Donner un intervalle de crédibilité à 95% pour ce paramètre.

5. Représenter graphiquement l'histogramme de la loi a posteriori du paramètre θ obtenu suivant l'algorithme de Metropolis-Hasting. Superposer la loi obtenue à la question 2.

6. Déterminer la loi prédictive *a posteriori* d'une nouvelle donnée $\tilde{y}$. Ecrire un algorithme de simulation de cette loi et comparer l'histogramme des résultats simulés aux données. Quelles critiques pouvez-vous faire du modèle proposé pour les données et le paramètre θ.

Seconde partie. Soit $y = y_1, \ldots, y_n$ un échantillon constitué de n données de comptage, entiers positifs ou nuls. On suppose les données indépendantes et provenant de K sources poissonniennes de moyennes inconnues $\theta = (\theta_1, \ldots, \theta_K)$. On définit un vecteur de variables non-observées $z = (z_1, \ldots, z_n)$, $z_i \in \{1, \ldots, K\}$, et un modèle pour (y, z, θ) de la manière suivante

$$p(y_i|z_i, \theta) = (\theta_{z_i})^{y_i} e^{-\theta_{z_i}} / y_i!$$

$$p(z) \propto 1$$

$$p(\theta_k) \propto 1/\theta_k \, .$$

On suppose que les $K+n$ variables $(\theta_k, z_i)_{k,i}$ sont mutuellement indépendantes.

1. Décrire la loi $p(y|z, \theta)$ en séparant les produits faisant intervenir les ensembles d'indices $I_k = \{i : z_i = k\}$, $k = 1, \ldots, K$.

2. Décrire la loi a posteriori du vecteur de variables (z, θ).

3. Soit $i \in \{1, \ldots, n\}$, calculer la probabilité conditionnelle $p(z_i = k|y, \theta)$.

4. Soit n_k le nombre d'éléments dans I_k et $\bar{y}_k$ la moyenne empirique des

données y_i pour $i \in I_k$. On note $\theta_{-k} = (\ldots, \theta_{k-1}, \theta_{k+1}, \ldots)$ le vecteur θ privé de sa coordonnées k. Montrer que

$$p(\theta_k|\theta_{-k}, y, z) \propto \theta_k^{n_k \bar{y}_k - 1} \exp(-n_k \theta_k).$$

5. Décrire l'implantation d'un cycle de l'algorithme d'échantillonnage de Gibbs pour la loi a posteriori en langage `R`. On pourra utiliser la commande `rgamma` pour simuler des réalisations de la loi gamma.

6. À l'issue d'une exécution de l'algorithme précédent, comment peut-on estimer l'espérance de θ_k sachant y ? Comment peut-on estimer les proportions de chacune des composantes du mélange ?

7. On considère à nouveau l'échantillon y donné au début de l'énoncé. Pour cet échantillon, représenter un histogramme et superposer la loi de mélange obtenue par l'algorithme d'échantillonnage de Gibbs pour diverses valeurs de K. Quelle valeur de K vous semble-t-elle la plus appropriée pour modéliser l'échantillon de données ? Proposer une ou plusieurs statistiques pour vérifier le modèle et en calculer les lois prédictives (code `R` à donner). Critiquer la modélisation.