Université de Caen Examen TP de Novembre 2018. Durée 2h Par Faïcel Chamroukhi
UFR des Sciences **Séries temporelles** **M2 Statistique & Analyse Décisionnelle**

Consignes : Vous devez séparer pour chacun des exercices votre programme principal (l'équivalent du main) du reste du code (fonctions etc). Le dépôt de votre travail se fera, sous la forme d'une archive, à déposer sur ecampus (exclusivement), le 20/11/2018 avant 12h30, heure stricte.(Les envois par mail ne seront pas pris en compte).

Le barème (4pts Etape E, 8 pts étape M, 3pts le reste du code) est donné à titre indicatif. Les exercices pour lesquels les codes ne s'exécutent pas, seront notés sur la moitié des points du barème.

On considère une série temporelle $(Y_1, \ldots, Y_n)$, avec $Y_t \in \mathbb{N}$, régie par un processus latent $(Z_1, \ldots, Z_n)$ à K états où Z_t représente l'état latent à l'instant $t\,(t = 1, \ldots, n)$. $Z_t \in \{1, 2, \ldots, K\}$, $K \in \mathbb{N}^\star$.

On suppose que pour chaque état k, la variable observée Y_t est modélisé en fonction du temps par le modèle log-linéaire suivant

$$\ln(\lambda_k(t, \boldsymbol{\beta}_k)) = \boldsymbol{\beta}_k^T \boldsymbol{x}_t \tag{1}$$

où $\boldsymbol{x}_t = (1, t)^\top$ et $Y_t | Z_t = k; \boldsymbol{\beta}_k \sim \mathcal{P}\big(\lambda_k(t; \boldsymbol{\beta}_k)\big)$ de loi Poisson

$$\mathbb{P}(Y_t = y_t | Z_t = k; \lambda_k(t; \boldsymbol{\beta}_k)) = \frac{e^{-\lambda_k(t;\boldsymbol{\beta}_k)}(\lambda_k(t; \boldsymbol{\beta}_k))^{y_t}}{y_t!}, \; \forall y_t \in \mathbb{N}, \tag{2}$$

qu'on notera $\text{Poisson}(y_t; \lambda_k(t, \boldsymbol{\beta}_k))$. La probabilité de l'état k à l'instant t définie par le modèle logistique :

$$\mathbb{P}(Z_t = k; \mathbf{w}) = \frac{e^{\boldsymbol{w}_k^\top \boldsymbol{x}_t}}{1 + \sum_{\ell=1}^{K} e^{\boldsymbol{w}_\ell^\top \boldsymbol{x}_t}}, \tag{3}$$

où $\mathbf{w} = (\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_{K-1})^\top$ et $\boldsymbol{w}_K = \mathbf{0}$. On notera par $\pi_k(t; \mathbf{w})$ cette probabilité.

La loi de la variable observée est dans ce cas définie par :

$$\mathbb{P}(Y_t = y_t; \boldsymbol{\theta}) \;\; = \;\; \sum_{k=1}^{K} \pi_k(t; \mathbf{w}) \text{Poisson}(y_t; \lambda_k(t, \boldsymbol{\beta}_k)) \tag{4}$$

$\boldsymbol{\theta}$ étant le vecteur paramètre inconnu du modèle.

L'objectif est de segmenter la série temporelle sur la base du modèle (4). Pour cela, on estime $\boldsymbol{\theta}$ à partir des données en maximisant la log-vraisemblance $L(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ln \mathbb{P}(Y_t = y_t; \boldsymbol{\theta})$ par l'algorithme EM. On montre que celui-ci consiste à partir d'un modèle initial de paramètre $\boldsymbol{\theta}^{(0)}$ et alterner à chaque itération q entre les deux étapes E- et M- suivantes jusqu'à la convergence :

Étape E : Calculer la fonction $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ définie par :

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = \sum_{t=1}^{n} \sum_{k=1}^{K} \tau_k(t; \boldsymbol{\theta}^{(q)}) \ln[\pi_k(t; \mathbf{w}) \text{Poisson}(y_t; \lambda_k(t, \boldsymbol{\beta}_k))] \tag{5}$$

où $\tau_k(t;\boldsymbol{\theta}^{(q)}) = \mathbb{P}(Z_t = k|y_t;\boldsymbol{\theta}^{(q)})$ est la probabilité a posteriori de l'état k à l'instant t, définie par :

$$\tau_k(t;\boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t;\mathbf{w}^{(q)})\,\text{Poisson}(y_t;\lambda_k(t,\boldsymbol{\beta}_k^{(q)}))}{\mathbb{P}(Y_t = y_t;\boldsymbol{\theta}^{(q)})}, \quad \forall t = 1,\ldots,n \tag{6}$$

Étape M : Mettre à jour les paramètres du modèle en calculant $\boldsymbol{\theta}^{(q+1)}$ définie par :

$$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}). \tag{7}$$

Ce problème de maximization n'admet pas de solution analytique.

1. La maximisation de (5) par rapport aux vecteurs $\boldsymbol{\beta}_k$, peut se faire par l'algorithme Newton-Raphson. A chaque itération m de cet algorithme, la mise à jour est donnée par l'équation suivante :

$$\boldsymbol{\beta}_k^{(m+1)} \;=\; \boldsymbol{\beta}_k^{(m)} + \Big[\sum_{t=1}^{n}\tau_k(t;\boldsymbol{\theta}^{(q)})e^{\boldsymbol{\beta}_k^\top \boldsymbol{x}_t}\boldsymbol{x}_t\boldsymbol{x}_t^\top\Big]^{-1}_{\boldsymbol{\beta}_k=\boldsymbol{\beta}_k^{(m)}} \sum_{t=1}^{n}\tau_k(t;\boldsymbol{\theta}^{(q)})\boldsymbol{x}_t\left(y_t - e^{\boldsymbol{\beta}_k^\top \boldsymbol{x}_t}\right)_{\boldsymbol{\beta}_k=\boldsymbol{\beta}_k^{(m)}} \tag{8}$$

2. On prend $K = 2$. Dans ce cas on montre que la mise à jour itérative du vecteur $\boldsymbol{w}$ à chaque itération s de l'algorithme Newton-Raphson est donnée par :

$$\boldsymbol{w}^{(s+1)} = \boldsymbol{w}^{(s)} + \Big[\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t^\top\pi(t;\boldsymbol{w})(1-\pi(t;\boldsymbol{w}))\Big]^{-1}_{\boldsymbol{w}=\boldsymbol{w}^{(s)}} \sum_{t=1}^{n}\boldsymbol{x}_t\left(\tau_t(\boldsymbol{\theta}^{(q)}) - \pi(t;\boldsymbol{w})\right)_{\boldsymbol{w}=\boldsymbol{w}^{(s)}} \tag{9}$$

   où $\pi(t;\boldsymbol{w}) = \frac{e^{\boldsymbol{w}^\top \boldsymbol{x}_t}}{1+e^{\boldsymbol{w}^\top \boldsymbol{x}_t}}$ est la probabilité de l'un des deux états.

3. Soit $\widehat{\boldsymbol{\theta}}$ le vecteur paramètre estimé par l'algorithme EM. On peut en déduire la séquence des états $(\widehat{Z}_1,\ldots,\widehat{Z}_n)$ où

$$\widehat{Z}_t = \arg\max_k \pi_k(t;\widehat{\boldsymbol{w}}). \tag{10}$$

Travail demandé : Implémenter l'algorithme EM et appliquer le à la série dataPoisson.mat avec $K = 2$. Il est conseillé d'utiliser une formulation vectorielle des mises à jour des paramètres $\boldsymbol{\beta}_k$ et $\boldsymbol{w}$.

— Afficher dans un graphique la série temporelle

— Afficher dans un graphique les probabilités logistiques estimées

— Afficher dans un graphique la log-vraisemblance calculée à chaque itération de l'algorithme EM

— Afficher dans un graphique la série temporelle, et la fonction de régression de Poisson de chaque état k définie par : $\lambda_k(t,\widehat{\boldsymbol{\beta}}_k) = \mathbb{E}[Y_t|Z_t = k;\widehat{\boldsymbol{\beta}}_k] = e^{\widehat{\boldsymbol{\beta}}_k^\top \boldsymbol{x}_t}\ \forall t = 1,\ldots,n$.