TD 4 : Modélisation de séries temporelles de données binaires
Exemple : éruptions de geysers

M2 Statistique et économétrie
Séries temporelles

2010-11

Objectif : L'objectif de ce TP est de découvrir (et de comprendre!) la modélisation des séries temporelles binaires.

1 Chaînes de Markov

Un outil naturel pour modéliser une série temporelle binaire est la chaîne de Markov. Nous allons alors commencer par rappeler les propriétés d'une chaîne de Markov.

Soit $\{S_t, t \in \mathbb{N}\}$ une chaîne de Markov d'ordre un définie sur l'espace $\{1, \cdots, M\}$ et de probabilités initiales $\{\pi_1, \cdots, \pi_M\}$ et de matrice de transition Γ telle que

$$\gamma_{ij} = P(S_t = j | S_{t-1} = i)$$

On suppose que la chaîne de Markov est homogène (c'est à dire que sa matrice de transition ne varie pas dans le temps). Ceci implique que $\{S_t\}$ est stationnaire et admet une unique loi stationnaire que nous noterons $\delta = (\delta_1, \cdots, \delta_M)$.

En notant $v = (1, \cdots, M)$ et V la matrice qui a pour diagonale v et ses autres éléments nuls, montrer que

1. $E(S_t) = \delta v^T$
2. $E(S_t S_{t+k}) = \delta V \Gamma^k v^T$
3. $Cov(S_t, S_{t+k}) = \delta V \Gamma^k v^T - (\delta v^T)^2$

On remarque (ne pas le montrer) que si Γ est inversible, alors $\Gamma^k = U \Omega^k U^{-1}$ et la covariance s'écrit alors aussi sous la forme suivante :

$$\begin{aligned} Cov(S_t, S_{t+k}) &= \Omega^k U^{-1} v^T - ({}^T)^2 \\ &= a \Omega^k b^T - a_1 b_1 \\ &= \sum_{i=1}^{M} a_1 b_i \omega_i^k \end{aligned}$$

avec $a = \delta V U$ et $b^T = U^{-1} v^T$

Montrer que si M=2, alors l'autocorrélation de $\{S_t\}$ est de la forme $\rho_k = \rho_1^k$ (c'est à dire de la même forme que pour un processus AR(1) gaussien).

Montrer que l'estimateur du maximum de vraisemblance de la matrice de transition s'écrit, pour tous i et j dans $\{1, \cdots, M\}$,

$$\hat{\gamma}_{ij} = \frac{\text{Card}(t \in \{2, \cdots, T\}, s_{t-1} = i \text{ et } s_t = j)}{\text{Card}(t \in \{2, \cdots, T\}, s_{t-1} = i)}.$$

2 Application : éruption du geyser "Old Faithful"

Azzilini et Bowman (1990) ont présenté une série temporelle de données d'éruption du geyser Old Faithful dans le parc du Yellowstone aux Etats-Unis (voir photo). Il s'agit de données de durées d'éruption et de temps d'attente entre 2 éruptions.

Ici nous allons nous intéresser à la série des durées d'éruption, mais sous une forme discrétisée. En pratique, on symbolise par 0 les éruptions de courte durée (moins de 3 minutes) et par 1 les séries de longue durée (plus de 3 minutes).

Pour avoir une idée du type de données, on peut, pour commencer, lancer les commandes ci-dessous.

```
library(MASS)
D = faithful
summary(D)
hist(D$eruptions)
d = as.numeric(D$eruptions>=3)
```

Mais dans la suite, nous utiliserons un jeu de données plus long et déjà discrétisé
(`http://www-labsticc.univ-ubs.fr/~monbet/docs/cours/st/faithful.txt`)
(Ref: Azzilini et Bowman (1990), Mc Donald et Zucchini (1997)). On peut, par exemple, représenter la série temporelle de la façon suivante :

```
xx = t(matrix(1:L,L,2)) ; yy = t(matrix(c(0*(1:L),d+1),L,2)) ;
txt = which(d==1)
matplot(xx[,txt],yy[,txt],col="black",lty=1,type="l")
txt = which(d==0)
matlines(xx[,txt],yy[,txt],col="red",lty=1,type="l")
```

2.1 Chaîne de Markov d'ordre un

L'idée la plus naturelle pour modéliser cette série temporelle est de la caractériser par une chaîne de Markov d'ordre un.

1. Estimer les paramètres de la chaîne de Markov.
2. En déduire la loi stationnaire et la matrice d'autocorrélation du modèle (jusqu'à k=8).
3. Comparer ces statistiques avec les statistiques empiriques correspondantes.
4. Commenter.

2.2 Chaîne de Markov d'ordre deux

On a observé dans la question précédente, qu'un modèle de chaîne de Markov d'ordre un ne permet pas de reproduire correctement la fonction d'autocorrélation des données. On propose alors d'ajuster un modèle de chaîne de Markov d'ordre deux.

1. Estimer les paramètres de la chaîne de Markov. Pour simplifier les calculs, on peut se ramener à une chaîne de Markov d'ordre un en regroupant les données par couples. Par exemple, regarder les transitions entre l'état (0,0) et l'état (0,1). L'espace d'état est alors de dimension 4 : $\{(0,0),(0,1),(1,0),(1,1)\}$. Rq : ici l'état $(0,0)$ n'est jamais observé.
2. En déduire la loi stationnaire et la matrice d'autocorrelation du modèle.
3. Comparer ces statistiques avec les statistiques empiriques correspondantes.
4. Commenter.

2.3 Modèle à chaîne de Markov cachée

On a observé dans la question précédente que le modèle à chaîne de Markov d'ordre deux permet d'assez bien reproduire les statistiques observées. Envisageons maintenant un modèle à chaîne de Markov cachée avec deux régimes. Les probabilités d'émission sont binomiales.

1. Estimer les paramètres du modèle par un algorithme forward-backward. On utilise la fonction `baumWelch` du package `HMM`. Pour y avoir accès, il suffit de télécharcger le fichier `http://www-labsticc.univ-ubs.fr/~monbet/doc/cours/ST/hmm.r` et d'exécuter un "source" de ce fichier.

```
hmm = initHMM(c(1,2),c(1,2),
        transProbs=t(matrix(c(0.01,0.99,.8,.2),2,2)),
        emissionProbs=t(matrix(c(0.8,0.2,0.05,95),2,2)))
H = baumWelch(hmm,c(d+1),20,delta=1E-3)
```

2. En déduire la loi stationnaire et la fonction d'autocorrelation du modèle.
3. Comparer ces statistiques avec les statistiques empiriques correspondantes, ainsi qu'avec les statistiques du modèle précédent. On pourra tracer sur un même graphique les fonctions d'autocorrélation des deux modèles avec la fonction d'autocorrélation empirique.
4. Calculer les critères BIC pour les 3 modèles.
5. Commenter l'ensemble des résultats et conclure.