# TD 5 : Modèle à chaîne de Markov cachée avec émissions gaussiennes. Modélisation de séries financières

M2 Statistique et économétrie
Séries temporelles

2011-12

---

Ce sujet de TD servira de base à l'examen de fin d'année. Votre préparation doit vous amener à être capable de traiter l'ensemble des questions avec un certain recul. Vous devez notamment comprendre le sens des questions posées ainsi que des commentaires et des codes qui leur sont associés.

---

Dans ce TD on s'intéresse à la modélisation de séries financières, en particulier des séries décrivant l'évolution journalière d'indices tels que le CAC40 ou ses équivalents allemand, suisse et anglais.

## 1 Statistiques descriptives

1. Charger les series temporelles d'évolution des indices d'actions.

```
library(tseries)
data(EuStockMarkets)
summary(EuStockMarkets)
cac = EuStockMarkets[,"CAC"] # France
dax = EuStockMarkets[,"DAX"] # Allemagne
smi = EuStockMarkets[,"SMI"] # Suisse
ftse = EuStockMarkets[,"FTSE"] # R.U.
```

2. Tracer les 4 séries temporelles. Commentez.

3. En général, on ne travaille pas directement avec ces séries temporelles, mais avec les rendements ou le logarithme des rendements (appelés aussi rendements logarithmiques). Si on note $\{x_t\}_{t\in\{1,\cdots,T\}}$ la série temporelle, le logarithme des rendements est défini par

$$y_t = \log\left(\frac{x_t}{x_{t-1}}\right)$$

Tracer les log des rendements pour les 4 séries. Pourquoi fait-on cette transformation? Pouvez vous commentez les graphiques obtenus?

4. Représenter les distributions de chacune des séries (par un histogramme par exemple). Estimer et tracer l'autocorrélation de chacune des séries $\{y_t\}$, $\{y_t^2\}$ et $\{\|y_t\|\}$. Qu'observe t'on?

# 2 Modélisation univariée

Considérons la série des rendements logarithmes de l'indice FTSE. On peut éventuellement travailler avec une série "normalisée" afin de faciliter l'interprétation des paramètres des modèles que l'on va ajuster : $yr_t = y_t/\hat{\sigma}_y$.

## 2.1 Modèle ARMA

1. Quand on étudiela série temporelle $y$ ou $yr$, on ne retire pas de tendance et on ne fait pas de désaisonnalisation. Ce choix vous semble t'il adapté? Justifiez votre réponse.

2. Justifier le fait qu'un modèle ARMA ne serait pas pertinent pour modéliser l'évolution du logrithme des rendements de cet indice.
(Rq : Les modèles usuels pour ce type de séries sont les modèles GARCH.)

## 2.2 Modèle à changement de régime Markovien

Nous proposons de modéliser cette série par un modèle à changements de régime Markovien avec des émissions gaussiennes. Nous allons utiliser le package `RHmm`. Sous linux, on l'installe à l'aide de la commande `install.packages("RHmm")`.

1. Ecrire le modèle et ses propriétés en notant $M$ le nombre de régimes. Lister les paramètres du modèle. Donner la dimension du vecteur de paramètres en fonction de $M$.

2. Ecrire la vraisemblance du modèle.

3. On ajuste ce modèle pour différentes valeurs de $M$ en utilisant un algorithme EM.

```
library(RHmm)
hmm = HMMFit(y,nStates=M)
```

4. Quel est le meilleur modèle au sens du critère BIC? Interprétez les paramètres du modèle : quelles sont les probabilités de changer de régime, quel est le régime le plus stable, quelles sont les caractéristiques des différents régimes, etc.

5. Pour ce modèle, tracer les probabilités de lissage, c'est dire $P(S_t = m|y_1, \cdots, y_T)$. On les obtient à l'aide de l'algorithme forward-backward.

```
fb = forwardBackward(hmm,y)
```

Utiliser ces probabilités de lissage pour inférer les états cachés. Notons `S` la série des états cachées. On peut alors tracer,

```
coul = c(2,1,3,4,5) ;
par(mfrow=c(2,1))
plot(fb$Gamma[,1],type="l",col=coul[1],ylab = "prob. lissage")
for (m in 2:M) {lines(fb$Gamma[,m],col=coul[m])}
for (m in 1:M) {
  tmp = y*NA
  tmp[S==m] = y[S==m]
  if (m==1) {plot(tmp,type="l",col=coul[m],ylim=c(min(y),max(y)))
  }else{lines(tmp,col=coul[m])} }
```

Commentez les graphiques obtenus. Sont-ils cohérents avec l'interprétation que vous avez fait du modèle? Quelles sont les périodes où le marché est volatile?

6. Estimer le temps moyen de séjour dans chacun des régimes.

7. Le modèle qui admet un régime caché ($M' = M + 1$) de plus est-il plus facile à interpréter?

# 3 Validation de modèle

## 3.1 Comparaison des distributions

1. Comparer la distribution empirique des rendements logarithmiques avec la distribution théorique. Représenter les distributions empiriques dans chancun des régimes.
2. Comparer les 4 premiers moments centrés réduits empiriques et théoriques.

## 3.2 Validation par simulation

1. Simuler, en utilisant le modèle, une série temporelle $\{z_t\}$ de même longueur que la série observée.
2. Comparer les fonctions d'autocorrélation empiriques de $\{z_t\}$ (resp. $\{z_t^2\}$, $\{\|z_t\|\}$) et $\{y_t\}$ (resp. $\{y_t^2\}$, $\{\|y_t\|\}$). Commentez : peut-on conclure que l'observation semble être une réalisation possible du modèle? Si oui pourquoi, sinon pourquoi?
   Remarque : vous pouvez recommencer plusieurs fois la simulation pour observer la variabilité des estimations de la fonction d'autocorrélation.

# 4 Modélisation multivariée

Il est facile de généraliser le modèle de la partie précédente pour prendre en compte plusieurs séries temporelles. Les probabilités d'émission sont alors les lois de gauss multivariées.

1. On considère la série bivariée suivante (dont on peut normaliser chaque composante par son écart-type) :

   ```
   y = matrix(cbind(diff(log(ftse)),diff(log(cac))),length(ftse)-1,2)
   ```

   (a) Ajuster un modèle et interpréter les paramètres.

   ```
   mod = HMMFit(y,nStates=M) # choisir M!
   ```

   (b) Inférer la variable cachée et visualiser les distributions marginales empiriques et théoriques de $Y_t$ dans chacune des classes. Pour la distribution empirique, on pourra se contenter de tracer un nuage de points. Pour la distribution théorique, on peut tracer des contours d'équiprobabilités. Par exemple, si $M = 1$, on fait

   ```
   bvdnorm <- function(x1,x2,mean=c(0,0),cov=diag(2)) {
   p = matrix(NA,length(x1),length(x2))
   invcov = solve(cov)
   for (i in 1:length(x1)) {
   for (j in  1:length(x2)) {
   tmp = (matrix(c(x1[i],x2[j]),1,2)- mean)
   p[i,j] = tmp%*%invcov%*%t(tmp) }}
   p = exp(-p/2)/2/pi/det(cov)
   return(p)
   }

   x1 = seq(min(y[,1]),max(y[,1]),length=50)
   x2 = seq(min(y[,2]),max(y[,2]),length=50)
   p.1 = bvdnorm(x1,x2,mod$HMM$distribution$mean[[1]],mod$HMM$distribution$cov[[1]])
   p.2 = bvdnorm(x1,x2,mod$HMM$distribution$mean[[2]],mod$HMM$distribution$cov[[2]])
   par(mfrow=c(1,2))
   zlev = min(max(c(p.1)),max(c(p.2)))
   image(x1,x2,p.1,zlim = c(0,zlev),col = heat.colors(20),xlab="ftse",ylab="CAC40")
    title("Régime 1")
   contour(x1,x2,p.1,add=TRUE)
   points(y[S==1,1],y[S==1,2],pch=".")
   ```

```
image(x1,x2,p.2,zlim = c(0,zlev),col = heat.colors(20),xlab="ftse",ylab="CAC40")
 title("Régime 2")
contour(x1,x2,p.2,add=TRUE)
points(y[S==2,1],y[S==2,2],pch=".")
```

Commentez les graphiques. Faire le lien avec les paramètres estimés.

2. Considerer maintenant la série à 4 variables tenant compte des quatres indices.

   (a) Ajuster un modèle à chaine de Markov cachée.

   (b) Inférer la chaine cachée.

   (c) Comparer la chaine cachée obtenue au résultat d'une classification par une méthode de kmeans. Commenter vos résultats.

   (d) Interpréter les classes dans chacun des deux modèles. Il est intéressant de représenter les densités bivariées de des rendements logarithmiques de FTSE et CAC dans les différentes classes des deux modèles afin de comparer les résultats avec ceux obtenus dans la question précédente.