Examen du 15/01/2019

MAP-STA2 : Séries chronologiques

Yannig Goude - yannig.goude@edf.fr; Ajmal Loodally - ajmalloodally@hotmail.com;

Exercice 1 (7 points)

1. énoncer le théorême de Wold. En quoi est il important pour la modélisation de séries chronologiques? (/1)

voir cours

2. donner 3 exemples de processus stationnaires. (/1.5)

- $y_t = ay_{t-1} + \varepsilon_t$ avec $|a| < 1$ et ε_t un bruit blanc
- ε_t un bruit blanc
- Z_t une suite de variables aléatoires iid de loi normale de moyenne 0 et de variance σ^2. a, b des constantes, $X_t = a + Z_t + cZ_{t-2}$ est stationnaire

3. donner 3 exemples de processus non-stationnaires. (/1.5)

- $z_t = at^2 + bt + \varepsilon_t$ est non-stationnaire, en effet $E(y_t) = at^2 + bt$ dépend du temps
- $z_t = t * \varepsilon_t$ est non-stationnaire, en effet $V(y_t) = t^2\sigma^2$ dépend du temps
- Z_t une suite de variables aléatoires iid de loi normale de moyenne 0 et de variance σ^2, $X_t = Z_t \cos(ct) + Z_{t-1}\sin(ct)$, non stationnaire, $E(X_t) = 0$, $V(X_t) = \sigma^2$ mais $\gamma(k) = \sigma^2 \sin(ct)\cos(c(t-k))$ dépend de t

4. qu'est ce que la fonction l'autocorrélation d'un processus? Proposer un estimateur empirique et un exemple de code R réalisant cette estimation. (/1.5)

voir cours, exemple de code R:

```
####fonction calculant l'autocorr?lation d'ordre h
autoCorr<-function(x,h)
{
  x.lag<-lag.xts(x,k=h,na.pad=T)
  return(cor(x.lag,x,use="pairwise.complete.obs"))
}


autoCorr2<-function(x,h)
{
  x.lag<-c(x[1:h],head(x,length(x)-h))
  return(cor(x.lag,x))
}

a1<-sapply(c(1:336),autoCorr,x)
a2<-sapply(c(1:336),autoCorr2,x)
```

5. qu'est ce que la fonction l'autocorrélation partiel d'un processus? Proposer un estimateur empirique et exemple de code R réalisant cette estimation. (/1.5)

voir cours, exemple de code R:

```
PartialAutoCorr<-function(x,h)
{
  x.lag<-lapply(c(1:h),lag.xts,x=x,na.pad=T)
```

```
  x.lag<-matrix(unlist(x.lag),ncol=length(x.lag))
  reg<-lm(x~x.lag-1)
  return(tail(reg$coef,1))
}

PartialAutoCorr(conso,h=1)
pa1<-sapply(c(1:50),PartialAutoCorr,x)
```

Exercice 2 (8 points)

Soit U_n un bruit blanc centré de variance σ^2 et soit Y_n le processus défini par

$$Y_n = U_n + 2U_{n-1}, n \in Z$$

1. déterminer la densité spectrale de Y_n. (/1)

$f(x) = \frac{\sigma^2}{2\pi}|1 + 2e^{ix}| = \frac{\sigma^2}{2\pi}(5 + 4\cos(x))$

2. déterminer la fonction d'autocovariance de Y_n. (/1) on note $a = 2$, $\theta = -1/a$,

- $\gamma(0) = \sigma^2(1 + a^2)$
- $\gamma(1) = a\sigma^2$
- $\gamma(k) = 0, k \geq 2$

3. calculer la variance τ^2 de l'erreur de prédiction à horizon 1 conditionnellement au passé. (/1)

$Var(Y_n - E(Y_n/Y_{n-1}, Y_{n-2}, ...)) = Var(2U_{n-1}) = 4\sigma^2$

4. on pose $V_n = \sum_{j=0}^{\infty}(-2)^{-j}Y_{n-j}$. Exprimer V_n en fonction de $U_k, k \leq n$. (/1.5)

on note $a = 2$, $\theta = -1/a$,

$V_n = \sum_{j=0}^{\infty} \theta^j(U_{n-j} + aU_{n-j-1})$

$= U_n + \sum_{j=1}^{\infty}(\theta^j + a\theta^{j-1})U_{n-j}$

$= U_n + (1 + a/\theta)\sum_{j=1}^{\infty} \theta^j U_{n-j}$

et on a donc: $V_n = U_n - 3\sum_{j=1}^{\infty}(-1/2)^j U_{n-j}$

5. calculer la variance et la fonction d'autocovariance de V_n. (/1.5)

En utilisant l'écriture précédente, on a :

$Var(V_n) = \sigma^2(1 + (1 - a^2)^2\sum_{j=1}^{\infty}(-1/a)^{2j}) = \sigma^2 a^2$

$cov(V_n, V_{n+h}) = cov(U_n - 3\sum_{j=1}^{\infty} \theta^j U_{n-j}, U_{n+h} - 3\sum_{j=1}^{\infty} \theta^j U_{n+h-j}) = \sigma^2(9\theta^h + \frac{\theta^2}{1-\theta^2}) - 3\theta^h\sigma^2$ soit comme $\theta = -1/2$, $cov(V_n, V_{n+h}) = 0$

6. exprimer Y_n en fonction de V_n et V_{n-1}. (/1)

$Y_n = V_n + \frac{1}{2}V_{n-1}$

7. soit $\widehat{Y}_{n+1}$ la prédiction de Y_{n+1} conditionnellement au passé $Y_k, k \leq n$. Donner l'erreur de prédiction et le risque quadratique moyen de prévision à un pas. (/1)

V_n est un bruit blanc (cf question 5) et constitue l'innovation de Y_n, l'erreur de prédiction est donc V_n et son risque quadratique $\sigma^2 a^2$.

Exercice 3 (2 points)

Soit le processus y_t suivant, supposé stationnaire, avec ε_t un bruit blanc de variance σ^2:

$$y_t = \sum_{k=1}^{p} a_k y_{t-k} + \varepsilon_t$$

1. quel est le nom de ce processus? (/0.5)

AR(p)

2. à quelle(s) condition(s) ce processus est il stationnaire? (/0.5)

Si les racines du polynôme caractéristique sont de module différent de 1 et écriture MB infinie sur le passé si >1.

3. on fait l'hypothèse que $p = 1$. A quelle(s) condition(s) ce processus est il stationnaire? (/0.5)

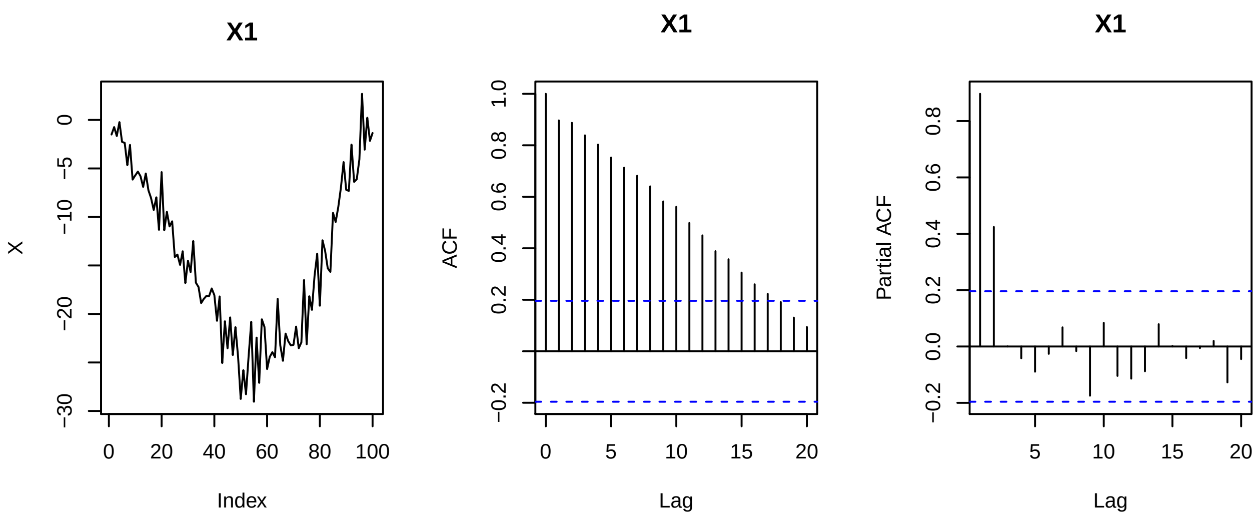
$|a_1| < 1$

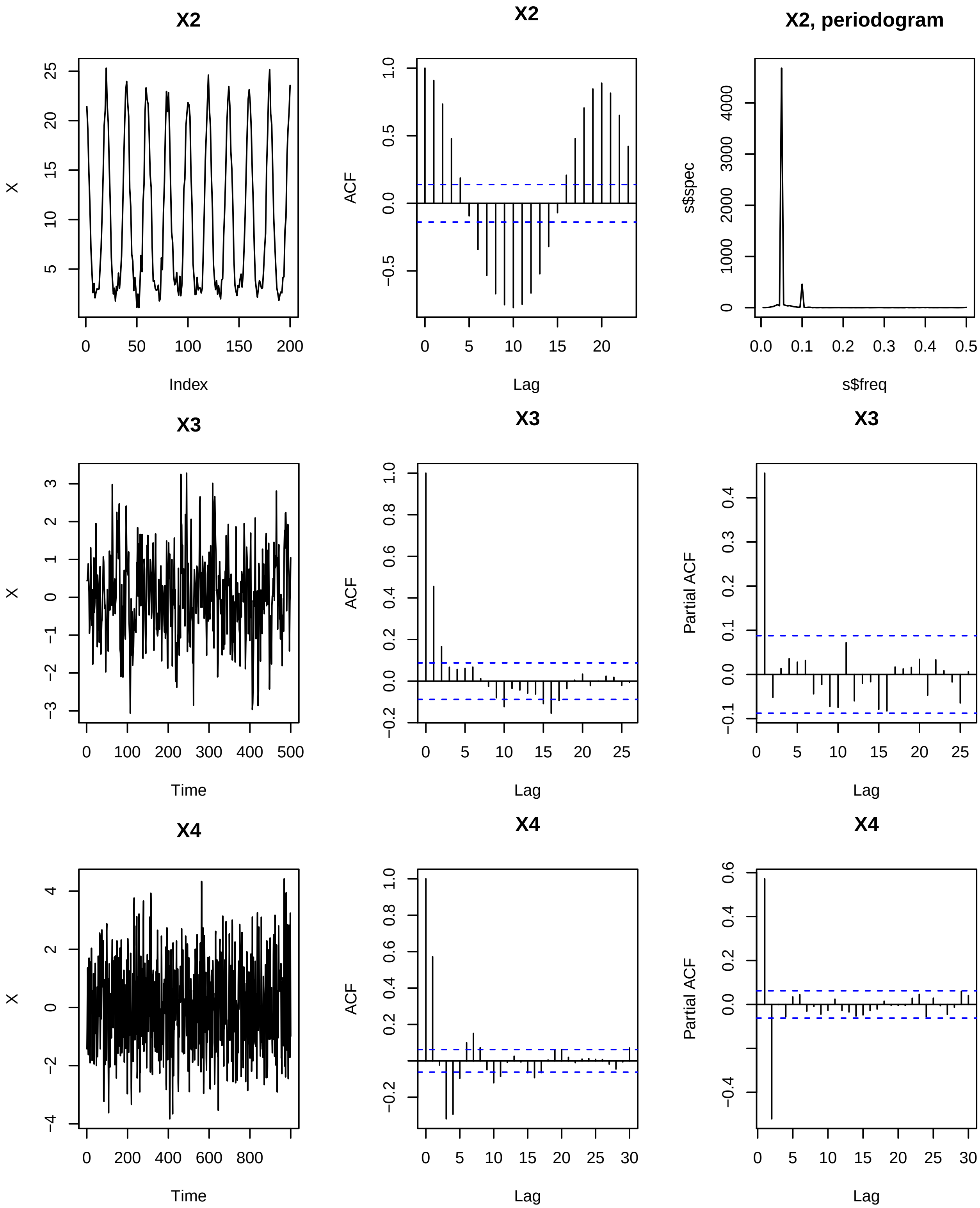
4. dans ce cas, exprimer l'écriture moyenne mobile infinie de y_t. (0.5)

$$y_t = (\sum_{i=0}^{\infty} a^i L^i)\varepsilon_t$$

Exercice 4 (9 points)

Un statisticien étudie un jeux de données composé de 4 séries temporelles pour lesquelles il a représenté/calculé les statistiques suivantes:





1. Proposer une démarche de modélisation pour chacune de ces séries, justifier. (/4, 1 point par série)

- série 1:

La série présente une tendance additive linéaire par morceau. Il s'agit donc d'un processus non stationnaire. Il faut donc modéliser cette tendance par régression sur base de spline ou loess, soustraire cette tendance puis étudie la série résiduelle (ACF, PACF). L'étude de l'ACF ou PACF sur la série brute n'apporte rien.

- série 2:

La série présente une composante périodique visible sur le périodogramme et/ou l'autocorrélogramme. Il s'agit donc d'un processus non stationnaire puisque son espérance dépend du temps. De plus cette composante semble additive au vu du graphique 1. On peut identifier sur le périodogramme les harmoniques de Fourier pertinentes et modéliser cette composante périodique par régression sur la base de Fourier associée. Après avoir soustrait cette composante à la série originale il faut ensuite analysé la stationnarité du processus obtenu.

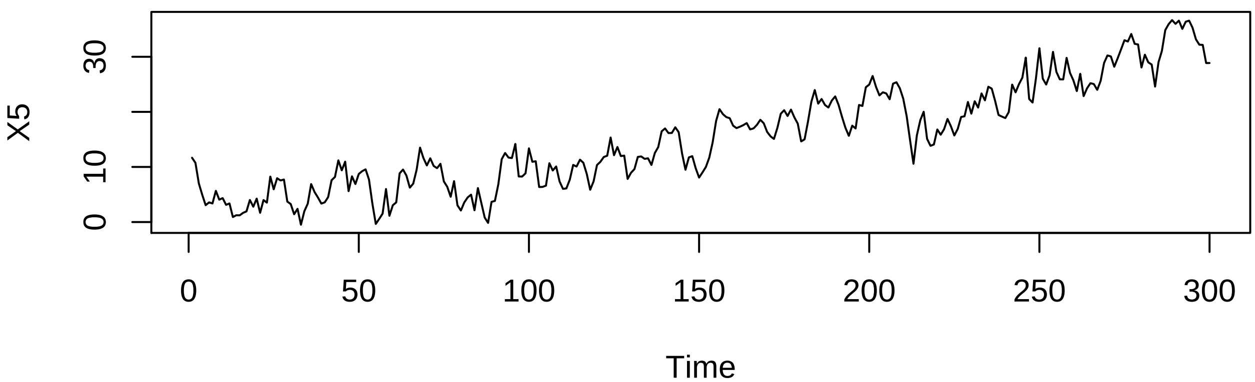
- série 3:

La série ne présente pas de comportement évident de type tendance ou saisonnalité. Sa variance semble également stable dans le temps. De plus l'acf décroit rapidement vers 0 (décroissance exponentiel) et on peut donc supposer que c'est un processus stationnaire. D'autre part la pacf indique que le processus peut être modélisé par un AR(1).

- série 4:

La série ne présente pas de comportement évident de type tendance ou saisonnalité. Sa variance semble également stable dans le temps. De plus l'acf décroit rapidement vers 0 en valeur absolué (décroissance exponentiel) et on peut donc supposer que c'est un processus stationnaire. Plus précisément, l'analyse de la pacf indique que le processus peut être modélisé par un AR(2).

Notre statisticien s'intéresse ensuite à une autre série X_t^5 représentée ci-dessous.



Il propose et cherche ensuite à valider un modèle. Il obtient les sorties suivantes:

```
##
## Call:
## arima(x = X5, order = c(3, 1, 5), include.mean = F, method = c("ML"))
##
## Coefficients:
##          ar1     ar2      ar3      ma1      ma2      ma3     ma4      ma5
##       0.4211  0.1058  -0.0110  -0.3375  -0.1663  -0.5026  0.4278  -0.2350
## s.e.  0.3656  0.2426   0.1545   0.3613   0.2370   0.1593  0.2237   0.1854
##
## sigma^2 estimated as 3.671:  log likelihood = -619.71,  aic = 1257.42

## pvalue student-test:

##  ar1  ar2  ar3  ma1  ma2  ma3  ma4  ma5
## 0.25 0.66 0.94 0.35 0.48 0.00 0.06 0.20
```

2. Préciser le modèle choisi par notre statisticien. A-t-il raison de choisir un ordre de différentiation de 1? /1

Le modèle choisie est un ARIMA (p=3, d=1, q=5). L'ordre de différenciation de 1 est une bonne idée car cela permet d'éliminer la composante de tendance linéaire visible sur la trajectoir de la série X_5.

3. Expliquer comment sont estimés les coefficients de ce modèle? /0.5

Les coefficients sont estimés par maximum de vraisemble gaussienne. Si ε_t est gaussien, le vecteur des observations $(X_1, ..., X_n)$ est gaussien et on peut alors calculer sa vraisemblance puis la minimiser en les paramètres du modèle.

4. Quel critère de choix de modèle a-t-il choisi? Expliquer comment celui-ci a été obtenu. /1

Le critère de choix est visiblement l'AIC, Akaike Information Critérion. Il correspond à:

$$\text{AIC}(\phi, \theta, \sigma^2) = -2\log(L(\theta, \phi, \sigma^2)) + 2k$$

où L est la vraisemblance, k est le nombre de paramètres dans le modèle. Ici $p + q + 1$ car la constante est dans le modèle. En effet, l'instruction include.mean=T est ignoré dans le cas ou d>0 (cf aide la fonction arima, argument include.mean).

5. A quoi correspond la ligne "s.e"? A quoi correspondent les p-values affichées? /0.5

s.e. correspond à l'écart type estimé des coefficients du modèle, diminuer d'une unité l'ordre de l'AR(p) ou du MA(q) revient à tester la significativité du coefficient ϕ_p (resp. θ_q) ce qui peut être fait par un test de student.

6. Au vu de ces résultats que doit faire notre statisticien? /1

Les tests de student pour les coefficients AR(3) et MA(5) ne rejette pas au seuil 5% la nullité de ces coefficients. Il faut donc réduire séquentiellement l'ordre du modèle et réeffectué un test.

7. Il souhaite ensuite valider son modèle, quels autres outils de diagnostique lui proposez vous? /1

plusieurs pistes sont à regarder pour valider son modèle:

- analyse des résidus:

test de Box-Pierce, ce test permet de tester l'hypothèse que les résidus d'une série X_t suivant une modélisation ARMA(p,q) sont un bruit blanc ie, pour une série X_t et ses résidus associés $\widehat{\varepsilon}_t = \widehat{\Theta}(L)^{-1}\widehat{\Phi}(L)(1 - B)^d X_t$ de fonction d'autocorrélation $\rho_\varepsilon(h)$ et son estimateur empirique asssocié:*

$$H_0(h) : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(h) = 0$$

qqplot et tests d'adéquation à une loi gaussienne

- validation par simulation d'une prévision sur un échantillon test ou par validation croisée, éventuellement dans ce cas comparer avec plusieurs modèles candidats.