Examen du 10/03/2020, corrigé

MAP-STA2 : Séries chronologiques

Yvenn Amara-Ouali -yvenn.amara-ouali@universite-paris-saclay.fr, Ajmal Loodally - ajmaloodally@hotm

Exercice 1 (/3)

Soit Z_t un processus stationnaire de moyenne nulle et de fonction d'autocovariance γ. Soit m_t une tendance et s_t une composante saisonnière de période p.

Calculer, pour tout entier t, l'espérance et la variance de X_t et sa covariance γ_X dans les cas suivants. Préciser à chaque fois le type de modèle dont il est question.

1. $X_t = m_t + s_t + Z_t$ (/1)
2. $X_t = m_t s_t + Z_t$ (/1)
3. $X_t = m_t s_t(1 + z_t)$ (/1)

Correction:

1. $E(X_t) = m_t + s_t$, $cov(X_t, X_{t+k}) = \gamma(k)$
2. $E(X_t) = m_t s_t$, , $cov(X_t, X_{t+k}) = \gamma(k)$
3. $E(X_t) = m_t s_t$, $cov(X_t, X_{t+k}) = m_t m_{t+k} s_t s_{t+k} \gamma(k)$

Exercice 2 (/2)

Soit X le processus défini par:

$$X_t - \phi X_{t-1} = \varepsilon_t$$

avec $|\phi| < 1$ et ε est un processus aléatoire stationnaire blanc de moyenne nulle et d'écart-type σ. Montrer que la seule solution stationnaire de cette équation est causale. Quel est le nom du processus ainsi défini?

Correction: C'est un processus AR(1). Cf cours. $1 - aL$ est une application de l'ensemble des processus stationnaires dans lui même. La série $\sum_{i=0}^{\infty} a^i$ étant absolument sommable, la série $\sum_{i=0}^{\infty} a^i L^i$ est défnie ($|a| < 1$) et nous avons:

$$(\sum_{i=0}^{\infty} a^i L^i)(1 - aL) = \sum_{i=0}^{\infty} a^i L^i - a^{i+1} L^{i+1} = L^0 = 1$$

$1 - aL$ est donc inversible et son inverse vaut $(1 - aL)^{-1} = \sum_{i=0}^{\infty} a^i L^i$.

Exercice 3 (/4)

1. Soit un processus X_t défini par $X_t = \varepsilon_t + 1.5\varepsilon_{t-1} - \varepsilon_{t-2}$ ou ε_t est un bruit blanc faible de moyenne nulle et de variance σ^2. Quel est le nom de ce processus? Calculer sa fonction d'autocovariance. (/1)

2. Calculer la densité spectrale d'un processus AR(1) et celle d'un processus MA(1) (/1)
3. Soit X_t un processus $MA(\infty)$. Exprimer sa densité spectrale. Montrer que connaitre la fonction d'auto-covariance de ce processus est équivalent à connaitre sa densité spectrale. (/1)
4. Soient X_t et Y_t des processus stationnaires centrés indépendants de densités spectrales respectives f_X et f_Y. Calculer la densité spectrale du processus $Z_t = X_t + Y_t$ en fonction de f_X et f_Y. (/1)

Correction:

1. processus MA(2), $\gamma(0) = \sigma^2(2 + 1.5^2)$, $\gamma(1) = 0$, $\gamma(2) = -\sigma^2$
2. la densité spectrale d'un processus est la fonction définie par:

$$f(\omega) = \frac{1}{2\pi} \sum_{h \in \mathbf{Z}} \gamma(h) \mathrm{e}^{i\omega h}, \quad \forall \omega \in \mathbf{R}$$

   - Soit le processus MA(1): $X_t = \varepsilon_t + \theta \varepsilon_{t-1}$, sa densité spectrale est

$$f(\omega) = \frac{\sigma^2}{2\pi}(1 + \theta^2 + 2\theta \cos(\omega))$$

   - Soit le processus AR(1): $X_t = \varepsilon_t + \phi \varepsilon_{t-1}$, si $|\phi| < 1$ il est causal et de sa décomposition $MA(\infty)$ on en déduit: $\gamma(k) = \sigma^2 \frac{\phi^k}{1-\phi^2}$. Si $|\phi| > 1$, on a $X_t = -\sum_{i=1}^{\infty} \phi^{-i} \varepsilon_{t+i}$, $\gamma(0) = \sigma^2 \frac{\phi^{-2}}{1-\phi^{-2}}$, $\gamma(k) = \sigma^2 \frac{\phi^{-k}}{\phi^2 - 1}$.
3. Connaitre la fonction d'auto-covariance d'un processus est équivalent à connaitre sa densité spectrale et on a:

$$\gamma(h) = \int_{-\pi}^{\pi} f(\omega) \cos(\omega h) d\omega = \int_{-\pi}^{\pi} f(\omega) \mathrm{e}^{i\omega h} d\omega$$

   Pour un processus $MA(\infty)$:

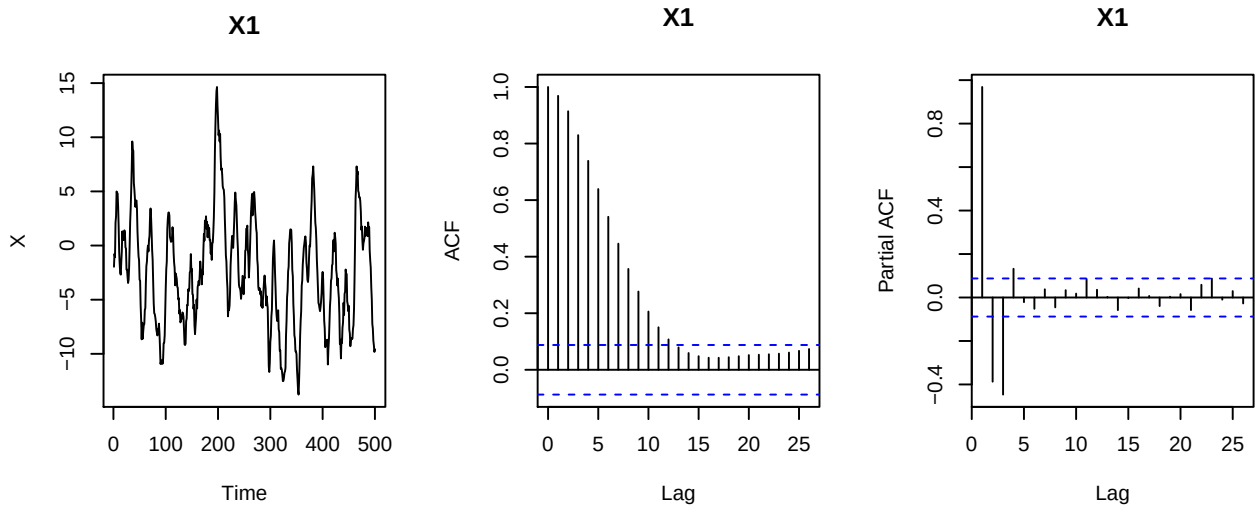
$$f(\omega) = \frac{\sigma^2}{2\pi} |\Theta(\mathrm{e}^{i\omega})|^2$$

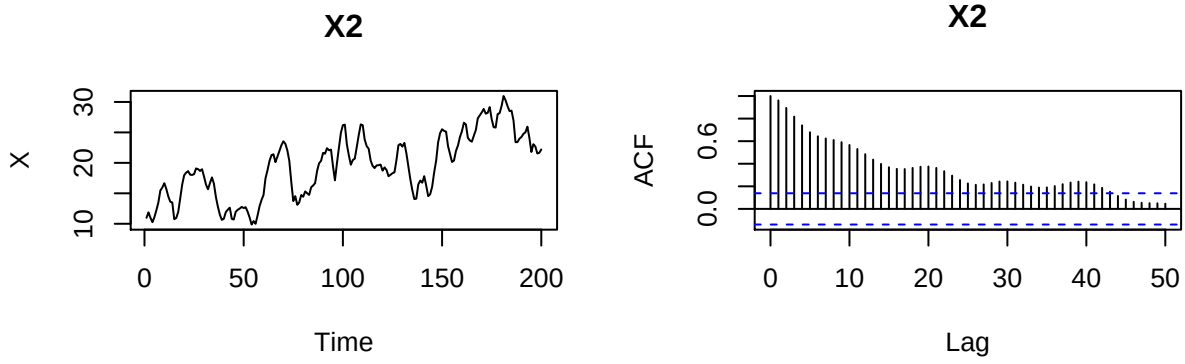
4. $\gamma_Z(k) = E[(X_t + Y_t)(X_{t+k} + Y_{t+k})] = \gamma_X(k) + \gamma_Y(k)$, d'ou

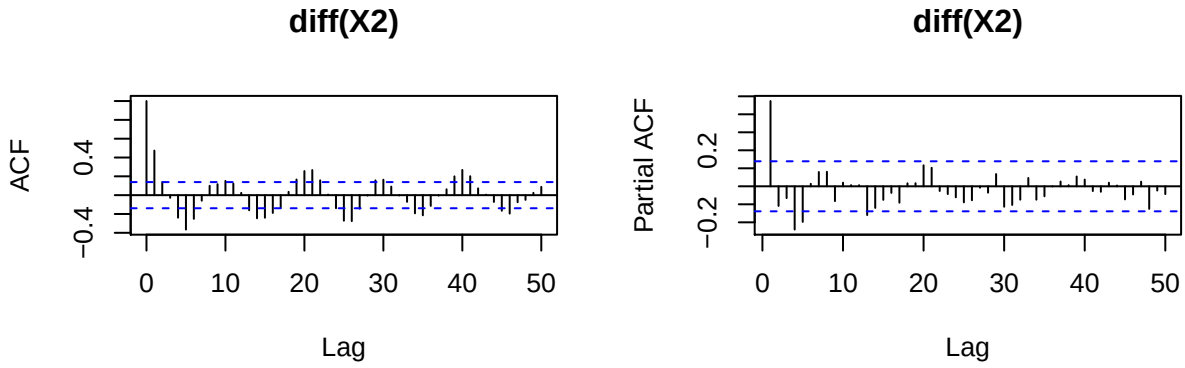
$$f_Z(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} e^{-i\omega k} (\gamma_X(k) + \gamma_Y(k)) = f_X(\omega) + f_Y(\omega)$$

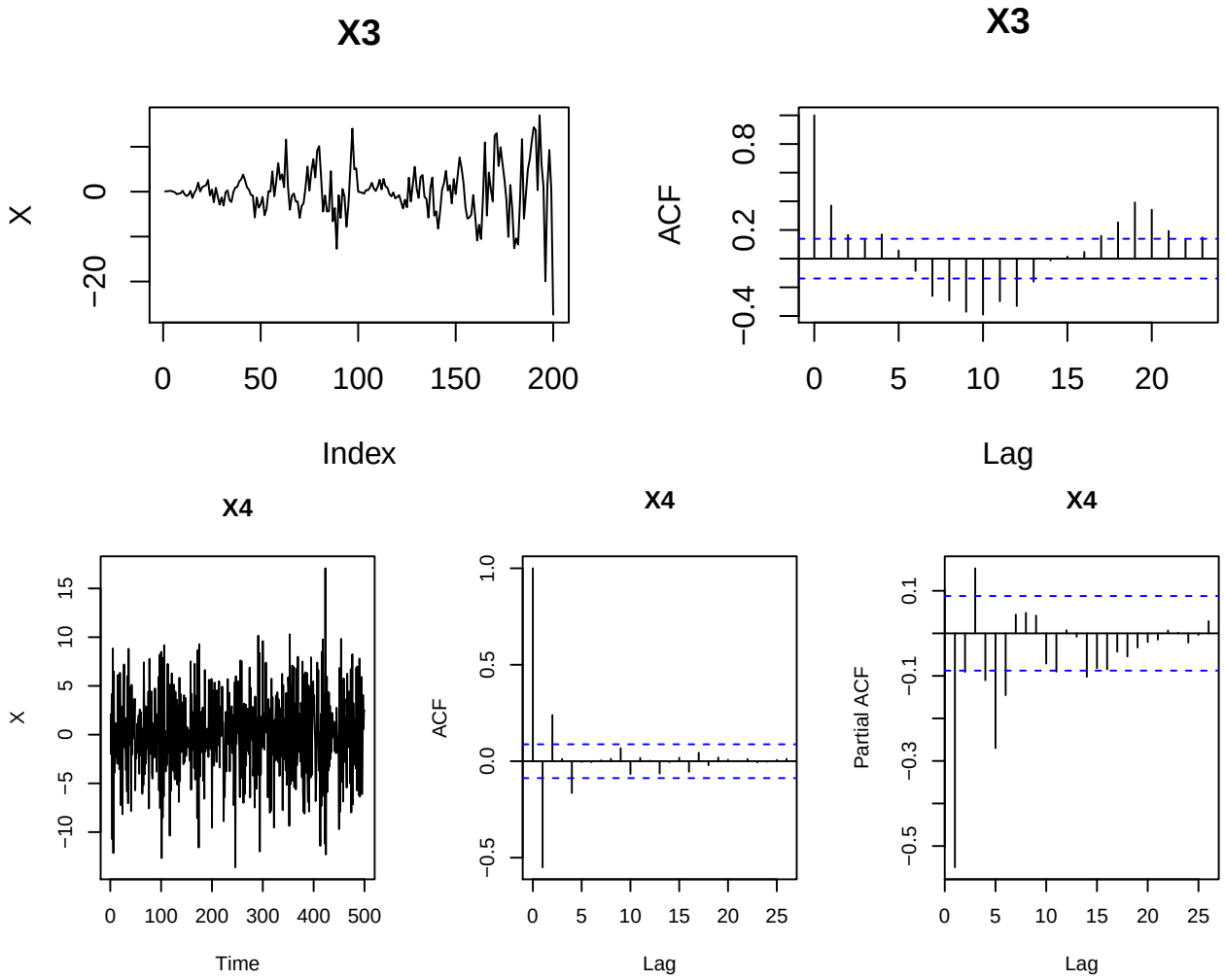
Exercice 4 (/12)

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


X1
X
Time
ACF
Lag
Partial ACF
Lag


X2
X
Time
ACF
Lag


diff(X2)
ACF
Lag
Partial ACF
Lag



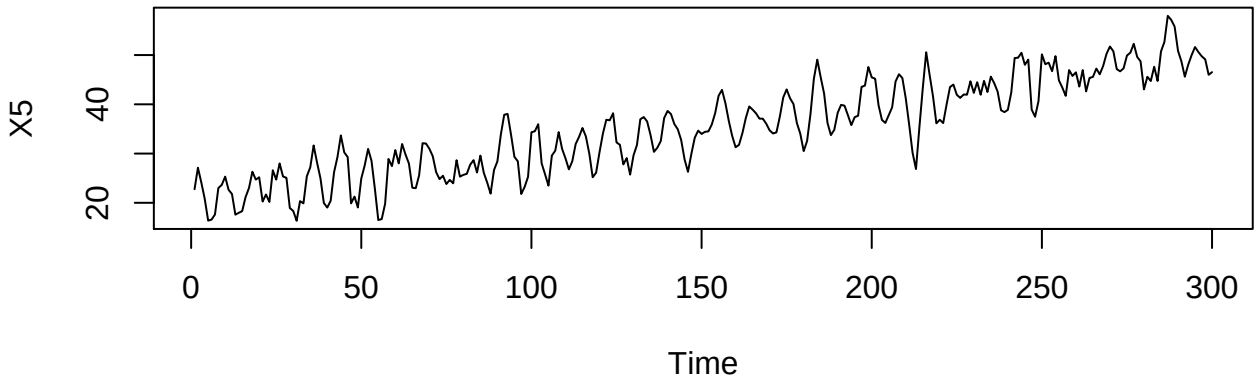
1. Proposer une démarche de modélisation pour chacune de ces séries, justifier. (/4)

- Série 1: 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(3).
- Série 2: au vu de la décroissance lente de l'ACF la série n'est pas stationnaire, cela est également visible sur la réprésentation graphique du processus qui indique la présence d'une tendance additive. En différenciant une fois la série, on élimine la composante de tendance linéaire. Il apparait ainsi dans l'ACF une composante périodique additive de période 10 qu'il faudrait modéliser par une régression sur série de fourier. Le choix du nombre d'harmoniques pourra se faire par une méthode de sélection de variable de type AIC ou selon des performances obtenues sur un échantillon test (sur la série corrigée de la tendance). Sur la série corrigée de la tendance et de la saisonnalité, il faudra représenter l'ACF et la PACF pour rechercher l'existance d'autocorrélation dans la série et ainsi préciser l'ordre d'un modèle ARMA.
- Série 3: le graphique de la série montre clairement l'existance d'une saisonnalité additive de période 10, cela est également visible sur l'ACF. On distingue également une tendance multiplicative par morceaux (augmentation de l'amplitude des oscillations jusqu'à t=100, puis diminution brutale et croissance jusqu'à t=200). Cette tendance est à modéliser en passant au log puis en effectuant une régression spline. Sur la série corrrigée de la tendance la saisonnalité est à modéliser par une régression sur série de fourier. Le choix du nombre d'harmoniques pourra se faire par une méthode de sélection de variable

de type AIC ou selon des performances obtenues sur un échantillon test (sur la série corrigée de la tendance).

- Série 4: la série ne présente pas de comportement évident de type tendance ou saisonnalité. Elle semble homoscédastique. De plus l'acf décroit rapidement vers 0 et on peut donc supposer que c'est un processus stationnaire. D'autre part l'ACF s'anule à partir du rang 5 inclu ce qui indique que le processus peut être modélisé par un MA(4).

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:

```
## pvalue student-test:

##  ar1  ar2  ar3  ar4  ma1  ma2  ma3  ma4  ma5
## 0.00 0.79 0.00 0.62 0.01 0.45 0.00 0.00 0.00
```

2. Préciser et commenter le modèle choisi par notre statisticien. (/1)

Le modèle choisie est un ARIMA (p=4, d=1, q=5) intégrant la constante. 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. De même intégrer la constante semble une bonne idée car l'ordonnée à l'origine n'est pas nulle.

3. A quoi correspond "sigma^2"? Que signifie l'option "method = c("ML")"? A quoi correspondent les p-values affichées? (1.5)

"sigma^2"" est l'estimateur de la variance de l'innovation par maximum de vraisemblance. l'option "ML" signifie que les paramètres du modèle sont obtenus par maximum de vraisemblance gaussienne. Les p-values affichées correspondent au test de student de significativité des coefficients du modèle au seuil 0.05:

$$\frac{|\widehat{\phi}_p|}{\sigma_{\widehat{\phi}_p}} < 1.96$$

4. Il décide de tester un modèle ARIMA(5,1,6) et obtient pour ce modèle une log-vraisemblance de -627.4, cela vous parait-il logique? Quel est l'AIC de ce modèle? (/1.5)

La log vraisemblance du premier modèle est -629.03 pour un modèle comprenant $p+q+1=10$ paramètres. En ajoutant 2 paramètres il est normal que la log vraisemblance augmente (l'erreur quadratique diminue car le modèle ayant plus de paramètres il s'ajuste mieux aux données). L'AIC obtenu par ce nouveau modèle est $2*627.4+2*12=1278.8$ ce qui est supérieur à l'AIC du premier modèle.

5. Au vu de ces résultats que doit faire notre statisticien? (/1)

L'hypothèse de nullité du coefficient AR(4) n'est pas rejettée au seuil 5%. Il faut donc réduire séquentiellement l'ordre du modèle et réeffectuer un test.

6. Il souhaite ensuite valider son modèle, quels autres outils de diagnostique lui proposez vous? (/1.5)

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.

7. Proposer un code R implémentant les démarches proposées pour les question 5. et 6. (/1.5)

Tout d'abord un code R pour diminuer l'ordre p:

```
ordre.init <- c(5, 1, 5)
#calcul du modèle d'ordre p-1
model.opt_pm1<-arima(x=x1,order=ordre.opt-c(1,0,0),method = c("ML"),
                     SSinit = c("Rossignol2011"),optim.method = "BFGS",include.mean = F)
```

Ensuite, on effectue le test de student:

```
##########pvalue du test de student
pvalue<-function(model)
{
  (1-pnorm(abs(model$coef)/sqrt(diag(model$var.coef))))*2
}
pvalue(model.opt_pm1)
```

Si l'hypothèse de nullité est rejetée pour les ordre max $p-1$ et/ou q on réduit encore l'odre du modèle et on procède au test sinon on s'arrête.

Ensuite les diagnostiques se font à l'aide du code:

```
##########################etude des résidus
plot(model.opt$residuals,type='l')

par(mfrow=c(1,2))
acf(model.opt$residuals)
pacf(model.opt$residuals)

par(mfrow=c(1,1))
qqnorm(model.opt$residuals)
hist(model.opt$residuals,breaks=50,freq=F)
x<-seq(min(model.opt$residuals),
       max(model.opt$residuals),length=50)
lines(x,dnorm(x,mean(model.opt$residuals),model.opt$sigma2), col='red')


#######test de box pierce
pvalue_BP<-function(model,K)
{
  rho<-acf(model$residuals,lag.max=K,plot=F)$acf[-1]
  n<-model$nobs
  pval<-(1-pchisq(n*sum(rho^2),df=K-length(model$coef)))
```

```
  return(pval)
}

pvalue_BP(model.opt,K=10)
```

Enfin, pour simuler des prévisions:

```
###exemple de prévision à horizon 10
h <- 10
test <- predict(model.opt, n.ahead=h)
```