Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2012 – 2013

Statistiques II

Examen final, avril 2013

Examen de 3h00. Tout document ou calculatrice est interdit.

1. **(11 points)** Soit $(\varepsilon_n)_{n\in\mathbf{Z}}$ un bruit blanc gaussien de variance σ_ε^2, où $\sigma_\varepsilon^2>0$ est inconnu et soit $X=(X_n)_{n\in\mathbf{Z}}$ le processus défini par
$$X_n=\alpha X_{n-1}+\varepsilon_n \quad \text{pour} \quad n\in\mathbf{Z},$$
avec $|\alpha|<1$ un réel inconnu.
   (a) Quel processus est X? Est-il centré? stationnaire? gaussien? (justifier)
   (b) Pour $n\in\mathbf{Z}$, déterminer $\mathbb{E}(X_n\,|\,X_{n-1})$ et $\mathbb{E}(X_n^2\,|\,X_{n-1})$. En déduire que $\mathrm{var}(X_n\,|\,X_{n-1})=\sigma_\varepsilon^2$ (on rappelle que $\mathrm{var}(X_n\,|\,X_{n-1})=\mathbb{E}(X_n^2\,|\,X_{n-1})-\big(\mathbb{E}(X_n\,|\,X_{n-1})\big)^2$) et la loi de X_n sachant X_{n-1}.
   (c) On observe une trajectoire $(X_1,\cdots,X_N)$ de X. Montrer que la log-vraisemblance de $(X_2,\cdots,X_N)$ sachant X_1 vaut:
$$LV_{\alpha,\sigma_\varepsilon^2}(X_2,\cdots,X_N)=-\frac{(N-1)}{2}\big(\log(2\pi)+\log(\sigma_\varepsilon^2)\big)-\frac{1}{2\sigma_\varepsilon^2}\sum_{n=2}^{N}(X_n-\alpha X_{n-1})^2.$$
   En déduire les estimateurs $\widehat{\alpha}_N$ et $\widehat{\sigma}_N^2$ de α et σ_ε^2 maximisant cette log-vraisemblance (on montrera que $\widehat{\alpha}_N=\frac{\sum_{n=2}^N X_nX_{n-1}}{\sum_{n=2}^N X_{n-1}^2}$).
   (d) On rappelle que pour un processus ARMA stationnaire causal, $\sqrt{N}\big(\widehat{r}_N(k)-r(k)\big)_{0\le k\le m}\xrightarrow[N\to\infty]{\mathcal{L}}\mathcal{N}_{m+1}\big(0\,,\,\Sigma\big)$, où $\Sigma=\big(4\pi\int_{-\pi}^{\pi}f^2(\lambda)\cos(i\lambda)\cos(j\lambda)d\lambda\big)_{1\le i,j\le m}$, avec r et f l'autocovariance et la densité spectrale du processus et $\widehat{r}_N(k)=\frac{1}{N}\sum_{n=1}^N X_nX_{n+k}$. En déduire un théorème de la limite centrale vérifiée par $\widehat{\alpha}_N$.

Proof. (a) X est un $AR(1)$ centré causal stationnaire car la racine de son polynôme est $1/\alpha$ qui est de module >1 **(1 pt)**.
(b) $\mathbb{E}(X_n\,|\,X_{n-1})=\alpha X_{n-1}+\mathbb{E}(\varepsilon_n\,|\,X_{n-1})$ et comme ε_n est indépendant de X_{n-1}, on a $\mathbb{E}(X_n\,|\,X_{n-1})=\alpha X_{n-1}$ **(1 pt)**.
$\mathbb{E}(X_n^2\,|\,X_{n-1})=\alpha^2X_{n-1}^2+2\alpha\mathbb{E}(\varepsilon_nX_{n-1}\,|\,X_{n-1})+\mathbb{E}(\varepsilon_n^2\,|\,X_{n-1})=\alpha^2X_{n-1}^2+2\alpha X_{n-1}\mathbb{E}(\varepsilon_n)+\mathbb{E}(\varepsilon_n^2)=\alpha^2X_{n-1}^2+\sigma_\varepsilon^2$ toujours du fait que ε_n est indépendant de X_{n-1} et centré **(1 pt)**.
De ces 2 résultats, on en déduit que $\mathrm{var}(X_n\,|\,X_{n-1})=\alpha^2X_{n-1}^2+\sigma_\varepsilon^2-(\alpha X_{n-1})^2=\sigma_\varepsilon^2$ **(0.5 pts)**.
On en déduit que la loi de X_n sachant X_{n-1} est une loi gaussienne (car ε_n est de loi gaussienne) $\mathcal{N}(\alpha X_{n-1},\sigma_\varepsilon^2)$ **(0.5 pts)**.
(c) $f_{(X_2,\cdots,X_N)\,|\,X_1)}(x_2,\cdots,x_N)=f_{(X_N\,|\,X_{N-1})}(x_N)\times f_{(X_{N-1}\,|\,X_{N-2})}(x_{N-1})\times\cdots f_{(X_2\,|\,X_1)}(x_2)$ par itération, donc grâce à la question (b) et la loi de X_n sachant X_{n-1}, on obient que:
$$f_{(X_2,\cdots,X_N)\,|\,X_1)}(X_2,\cdots,X_N)=\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\big(-\frac{1}{2\sigma_\varepsilon^2}(X_N-\alpha X_{N-1})^2\big)\times\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\big(-\frac{1}{2\sigma_\varepsilon^2}(X_{N-1}-\alpha X_{N-2})^2\big)\times\cdots$$
$$\times\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\big(-\frac{1}{2\sigma_\varepsilon^2}(X_2-\alpha X_1)^2\big).$$
En passant au logarithme, on en déduit la log-vraisemblance proposée **(3 pts)**.
En prenant les dérivées partielles et en cherchant le point où le gradient s'annule, on montre bien que $\widehat{\alpha}_N=\frac{\sum_{n=2}^N X_nX_{n-1}}{\sum_{n=2}^N X_{n-1}^2}$ et $\widehat{\sigma}_N^2=\frac{1}{N-1}\sum_{k=2}^N(X_k-\widehat{\alpha}_NX_{k-1})^2$ **(2 pts)** (on vérifie aisément que la matrice hessienne contenant les dérivées secondes est négative).
(d) On a donc $\widehat{\alpha}_N=\widehat{r}_{N-1}(1)/\widehat{r}_{N-1}(0)$ et on sait que la densité spectrale de X est: $f_X(\lambda)=\frac{\sigma_\varepsilon^2}{2\pi}\frac{1}{|1-\alpha e^{i\lambda}|^2}=\frac{\sigma_\varepsilon^2}{2\pi}\frac{1}{1+\alpha^2-2\alpha\cos(\lambda)}$ pour $\lambda\in[-\pi,\pi]$. Avec $m=1$, on a donc un TLC mutlidimensionnel vérifié par $(\widehat{r}_{N-1}(0),\widehat{r}_{N-1}(1))$ et on a $r(0)=\sigma_\varepsilon^2(1-\alpha^2)^{-1}$ et $r(1)=\sigma_\varepsilon^2\alpha(1-\alpha^2)^{-1}$. Il suffit alors d'utiliser la Delta-méthode avec la fonction $g(x_1,x_2)=x_2/x_1$ pris en $(r(1),r(0))$. La matrice jacobienne associée est: $(-r(1)/r(0)^2,1/r(0))=(1-\alpha^2)\sigma_\varepsilon^{-2}(-\alpha,1)$. La matrice Σ vaut $\big(\frac{\sigma_\varepsilon^4}{\pi}\int_{-\pi}^{\pi}\frac{\cos(i\lambda)\cos(j\lambda)}{(1+\alpha^2-2\alpha\cos(\lambda))^2}d\lambda\big)_{0\le i,j\le1}$ et ainsi:
$$\sqrt{N}(\widehat{\alpha}_N-a)\xrightarrow[N\to\infty]{\mathcal{L}}\mathcal{N}\Big(0\,,\,(1-\alpha^2)^2(-\alpha,1)\Big(\frac{\sigma_\varepsilon^4}{\pi}\int_{-\pi}^{\pi}\frac{\cos(i\lambda)\cos(j\lambda)}{(1+\alpha^2-2\alpha\cos(\lambda))^2}d\lambda\Big)_{0\le i,j\le1}\begin{pmatrix}-\alpha\\1\end{pmatrix}\Big)\qquad\textbf{(4 pts)}.$$
□

2. **(11 points)** On considère maintenant $(\xi_n)_{n\in\mathbf{Z}}$ un bruit blanc gaussien de variance 1 admettant un moment d'ordre 4 et soit $Y=(Y_n)_{n\in\mathbf{Z}}$ le processus défini par

$$Y_n = \xi_n \sqrt{a_0 + a_1 Y_{n-1}^2} \quad \text{pour} \quad n \in \mathbf{Z},$$

où a_0 et a_1 sont deux réels strictement positifs inconnus.

(a) Quel processus est Y? Rappeler la condition portant sur a_1 permettant à Y d'être stationnaire d'ordre 2 et causal.

(b) On se place désormais sous cette condition. Calculer alors $\mathbb{E}(Y_0)$ et $\mathrm{cov}(Y_0, Y_k)$ pour tout $k \in \mathbf{Z}$.

(c) Pour $n \in \mathbf{Z}$, déterminer $\mathbb{E}(Y_n \,|\, Y_{n-1})$ et $\mathbb{E}(Y_n^2 \,|\, Y_{n-1})$. En déduire la loi de Y_n sachant Y_{n-1}. Que vaut $\mathrm{cov}(Y_n, Y_{n+k} \,|\, Y_{n+k-1})$ pour $k \in \mathbf{N}$?

(d) Soit le processus $Z=(Z_n)_{n\in\mathbf{Z}}$ tel que $Z_n = Y_n^2 - a_0(1-a_1)^{-1}$ pour $n \in \mathbf{Z}$. Montrer que Z vérifie $Z_n = a_1 Z_{n-1} + u_n$, où $u_n = (\xi_n^2 - 1)\big(a_0 + a_1(Z_{n-1} + a_0(1-a_1)^{-1})\big)$ pour $n \in \mathbf{Z}$. Montrer que (u_n) est un bruit blanc faible.

(e) On observe une trajectoire $(Y_1, \cdots, Y_N)$ de Y. En utilisant 1.(c) et (d), proposer un estimateur de a_1 fonction de $(Y_1, \cdots, Y_N)$ et dont on donnera un théorème de la limite centrale (on admettra que le Théorème obtenu en 1.(d) est valide pour (Z_n)), puis en utilisant la valeur de $\mathrm{var}(Y_0)$ proposer un estimateur convergeant de a_0.

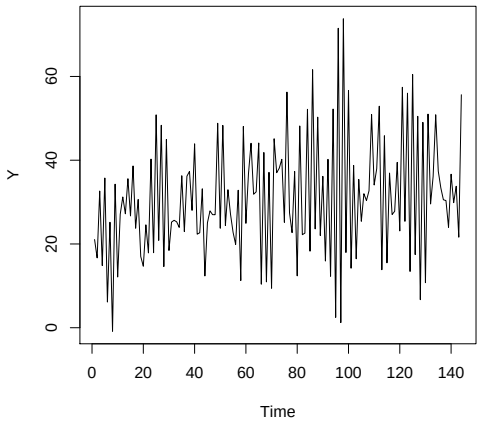
Proof. (a) Y est un $ARCH(1)$. Il est centré causal stationnaire d'ordre 2 lorsque $a_1 \mathbb{E}\xi_0^2 < 1$ soit $a_1 < 1$ **(1 pt)**.
(b) On a clairement (voir le cours) $\mathbb{E}Y_n = 0$ et $\mathrm{cov}(Y_0, Y_k) = 0$ **(1 pt)**.
(c) $\mathbb{E}(Y_n \,|\, Y_{n-1}) = 0$ et et $\mathbb{E}(Y_n^2 \,|\, Y_{n-1}) = (a_0 + a_1 Y_{n-1}^2)\mathbb{E}\xi_0^2 = a_0 + a_1 Y_{n-1}^2$ (voir le cours) **(0.5 pts)**.
La loi de de Y_n sachant Y_{n-1} est donc une loi gaussienne centrée et de variance $a_0 + a_1 Y_{n-1}^2$ **(0.5 pts)**.
Si $k \geq 1$, on a $\mathrm{cov}(Y_n, Y_{n+k} \,|\, Y_{n+k-1}) = \mathbb{E}\big[Y_n Y_{n+k} \,|\, Y_{n+k-1}\big] = \mathbb{E}\xi_{n+k}\mathbb{E}\big[Y_n \sqrt{a_0 + a_1 Y_{n+k-1}}\big] = 0$ du fait de l'indépendance entre ξ_{n+k} et Y_n ou Y_{n+k-1} **(1 pt)**.
(d) On a $\mathbb{E}Z_n = 0$ car comme $\mathbb{E}Y_n^2 = a_0 + a_1 \mathbb{E}Y_{n-1}^2$, il vient que $\mathbb{E}Y_n^2 = a_0(1-a_1)^{-1}$ **(1 pt)**.
Comme $Y_n^2 = Z_n^2 + a_0(1-a_1)^{-1}$ et comme $Y_n^2 = \xi_n^2(a_0 + a_1 Y_{n-1}^2)$, en remplaçant, on arrive à montrer que $Z_n = a_1 Z_{n-1} + u_n$ avec l'expression de u_n voulue **(1 pt)**.
Pour $k \geq 1$, $\mathrm{cov}(u_n, u_{n-k}) = \mathbb{E}\Big[(\xi_n^2 - 1)\big(a_0 + a_1(Z_{n-1} + a_0(1-a_1)^{-1})\big)(\xi_{n-k}^2 - 1)\big(a_0 + a_1(Z_{n-k-1} + a_0(1-a_1)^{-1})\big)\Big] = \mathbb{E}\big[\xi_n^2 - 1_b ig]\mathbb{E}\Big[\big(a_0 + a_1(Z_{n-1} + a_0(1-a_1)^{-1})\big)(\xi_{n-k}^2 - 1)\big(a_0 + a_1(Z_{n-k-1} + a_0(1-a_1)^{-1})\big)\Big] = 0$ car ξ_n est indépendant du reste **(1 pt)**.
(e) En s'aidant de la partie 1., on peut estimer a_1 comme $\widehat{\alpha}_N$ estime α. Ainsi on peut considérerla covariance empirique des Y_n^2, soit $\widehat{a}_1 = \frac{\sum_{n=2}^N (Y_n^2 - \overline{Y_N^2})(Y_{n-1}^2 - \overline{Y_N^2})}{\sum_{n=2}^N (Y_{n-1}^2 - \overline{Y_N^2})^2}$, où $\overline{Y_N^2} = \frac{1}{N}\sum_{k=1}^N Y_k^2$ et le théorème limite est le même sauf que l'on remplace α par a_1 **(1.5 pts)**.
Comme $\mathrm{var}(Y_0) = a_0(1-a_1)^{-1}$, on a donc a_1 qui peut être estimé, et $\mathrm{var}(Y_0)$ peut être estimé par la variance empirique, soit $\widehat{s}_Y^2 = \frac{1}{N}\sum_{n=1}^N (Y_n - \overline{Y_N})^2$. Ainsi $\widehat{a}_0 = \widehat{s}_Y^2(1-\widehat{a}_1)$ est un estimateur de a_0 **(1.5 pts)**.
On a vu que $\widehat{a}_1$ converge vers a_1 (en probabilité par exemple). On sait également que pour un ARCH(p), $\widehat{s}_Y^2$ converge (en probabilité par exemple) vers $\mathrm{var}(Y_0)$. On en déduit donc que $\widehat{a}_0$ converge en probabilité vers a_0 **(1 pt)**.
□

3. **(9.5 points)** Voici des simulations effectuées avec le logiciel R.

(a) On tape d'accord les commandes suivantes:

```
epsilon=3*rnorm(144)
X=0
for (j in c(1:143))
 {X[j+1]=-0.7*X[j]+epsilon[j+1]-2*epsilon[j]}
t=c(1:144)
Y=15+4*log(t)+7*cos(pi*(t-2)/6)+X
ts.plot(Y)
```

Voici le graphe obtenu:



Questions: Quel est le processus simulé par le vecteur X (préciser tous ses paramètres)? Quelles sont les tendances et saisonnalités de Y ? Quelle est sa variance (théorique)?

Proof. X est un ARMA$(1,1)$ d'équation $X_n + 0.7X_{n-1} = epsilon_n - 2epsilon_{n-1}$ **(1 pt)**.
La tendance de Y est $a(t) = 15 + 4\ln(t)$ et sa saisonnalité est $7 * cos(pi * (t-2)/6)$ avec une période de 12 **(0.5 pts)**. La variance théorique de Y est celle de X. On montre facilement que $X = \big(\sum_{n=0}^{\infty}(-0.7)^n B^n\big)(I - 2B)epsilon$ donc $X_n = epsilon_n - 2.7\sum_{k=1}^{\infty}(-0.7)^k epsilon_{n-k}$. En conséquence, $\text{var}(X_0) = 9\big(1+(2.7)^2\sum_{k=1}^{\infty}0.49^{k-1}\big) = 9\big(1+(2.7)^2/0.51\big) = 9 * 260/17 \simeq 137$ **(2 pts)**. □

(b) Voici les commandes tapées ensuite:

```
Z1=t
Z2=t^2
Z3=sqrt(t)
Z4=cos(pi*t/6)
Z5=sin(pi*t/6)
reg=lm(Y~Z1+Z2+Z3+Z4+Z5)
summary(reg)
```

Voici les résultats obtenus:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.6447975 11.1173261   0.957 0.339991
Z1          -0.4077850  0.5735870  -0.711 0.478322
Z2           0.0009389  0.0017558   0.535 0.593669
Z3           5.3872548  5.0063068   1.076 0.283763
Z4           3.4346042  1.5587953   2.203 0.029227 *
Z5           6.0576026  1.5746636   3.847 0.000182 ***

Residual standard error: 13.22 on 138 degrees of freedom
Multiple R-squared: 0.1789,     Adjusted R-squared: 0.1492
```

Questions: Qu'a-t-on fait par ces commandes? Que représentent les valeurs 10.6447975, 0.593669 *et* 0.1492*? Que conclure à la lecture de ces résultats?*

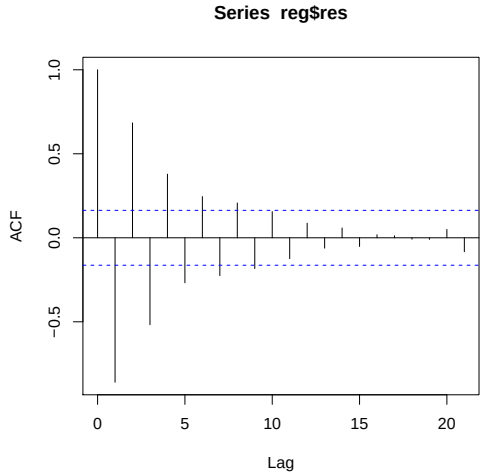
Proof. On estime la tendance et la stationarité de Y par régression. On régresse Y par rapport à des variables diverses pour la tendance ($Z1$ à $Z3$) et par rapport à $Z4$ et $Z5$ qui modélisent la saisonnalité **(0.5 pts)**.
10.6447975 est l'estimation de l'intercept (15 dans le modèle simulé), 0.593669 est la p-value de la variable $Z2$ (testant si le coefficient devant $Z2$ est nul) et 0.1492 est la valeur du R^2-ajusté **(1 pt)**.
Il apparaît que le modèle n'est pas satisfaisant (en particulier les variables $Z1$, $Z2$ et $Z3$ ne semble pas forcément explicatives) **(0.5 pts)**. □

(c) On tape enfin les commandes suivantes:

```
reg=lm(Y~Z1+Z3+Z4+Z5)
summary(reg)
reg=lm(Y~Z3+Z4+Z5)
summary(reg)
acf(reg$res)
```

Voici les résultats obtenus ainsi que le graphe tracé:

```
lm(formula = Y ~ Z1 + Z3 + Z4 + Z5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.2777     6.9492   2.198   0.0296 *
Z1           -0.1099     0.1367  -0.804   0.4225
Z3            2.9408     2.0282   1.450   0.1493
Z4            3.4318     1.5548   2.207   0.0289 *
Z5            5.9595     1.5599   3.820   0.0002 ***

Residual standard error: 13.19 on 139 degrees of freedom
Multiple R-squared: 0.1772,     Adjusted R-squared: 0.1536

lm(formula = Y ~ Z3 + Z4 + Z5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1762     3.3436   6.034 1.35e-08 ***
Z3            1.3403     0.3928   3.412 0.000843 ***
Z4            3.4173     1.5527   2.201 0.029384 *
Z5            5.9084     1.5567   3.796 0.000219 ***

Residual standard error: 13.17 on 140 degrees of freedom
Multiple R-squared: 0.1734,     Adjusted R-squared: 0.1557
```



Questions: Expliquez pourquoi on a répété deux fois la commande lm et quel(s) critère(s) guide(nt) cette répétition? Est-on satisfait du modèle obtenu? Pouvait-on s'attendre aux valeurs numériques 3.4173 *et* 5.9084*? Expliquez ce que représente le graphe. La valeur numérique décrite par la seconde barre est approximativement de* -0.85*. Expliquez ce que l'on pouvait espérer comme valeur théorique.*

Proof. On élimine la variable ayant la plus forte p-value lorsque celle-ci est supérieure à 0.05 et on obtient au final un modèle avec $Z3$ (pour la tendance), $Z4$ et $Z5$ (pour la saisonnalité). On peut aussi avoir choisi le critère du R^2-ajusté que l'on maximise pour chosir le meilleur modèle possible **(0.5 pts**.
On est satisfait a priori du modèle choisi, car toutes les p-values des tests de student sont inférieures à 0.05 **(0.5 pts)**.
3.4173 et 5.9084 sont les estimations des paramètres apparaissant devant $Z4$ et $Z5$. Dans le modèle simulé, la saisonnalité est $7 * cos(pi * (t-2)/6) = 7\big(\frac{1}{2}cos(pi*t/6) + \frac{\sqrt{3}}{2}cos(pi*t/6)\big) \simeq 3.5\,Z4 + 5.5\,Z5$. On voit que les estimations sont bonnes **(1 pt)**.
Le graphe représente les autocorrélations empiriques des résidus de la régression, qui sont des estimations des $epsilon_i$ **(0.5 pts)**.
La valeur numérique décrite par la seconde barre est approximativement de -0.85 est une approximation de la corrélation d'ordre 1 des $epsilon_i$. Or comme $\mathbb{E}(X_n + 0.7X_{n-1})^2 = \mathbb{E}(epsilon_n - 2 * epsilon_{n-1})^2 = 45$, donc $1.49\,r(0) + 1.4r(1) = 45$. On trouve ainsi $r(1) \simeq (45 - 1.49 * 137)/1.4 \simeq -114$. Par suite la corrélation théorique vaut $\rho(1) = r(1)/r(0) \simeq -114/137 \simeq -0.83$.
L'estimation est donc très proche! **(1.5 pts)**. □