*Université Paris I, Panthéon - Sorbonne*

**Première Année Master M.A.E.F.** 2016 – 2017

# Statistiques II

Contrôle continu n°2, avril 2017

*Examen de 1h30. Tout document ou calculatrice est interdit.*

1. **(7 points + 2 points)** On suppose que $M = (M_k)_{k\in\mathbb{Z}}$ est un processus centré stationnaire du second ordre tel qu'il existe $\alpha > 0$ et $\beta \in \mathbf{R}$ vérifiant pour tout $k \in \mathbf{Z}$:

$$\mathbb{E}\big(M_k^2\big) = \alpha, \quad \mathbb{E}\big(M_k M_{k+1}\big) = \beta \quad \text{et} \quad \mathbb{E}\big(M_k M_{k+i}\big) = 0 \quad \text{pour} \quad i \geq 2. \tag{1}$$

(a) Montrer que l'on a nécessairement $|\beta| \leq \alpha$ **(1pt)**.

(b) Rappeler l'expression de la densité spectrale de $M$. En déduire que nécessairement $2|\beta| \leq \alpha$ **(1.5pts)**.

(c) Soit $(\xi_k)_{k\in\mathbf{Z}}$ une suite de v.a.i.i.d. centrées et de variance $\sigma^2 > 0$ et soit $(N_k)_{k\in\mathbf{Z}}$ un processus tel que pour tout $k \in \mathbf{Z}$, $N_k = \xi_k + c\,\xi_{k-1}$ avec $c \in \mathbf{R}$. Quel type de processus est $(N_k)$ **(0.5pts)**? Déterminer l'autocovariance de $(N_k)$ **(1pt)**. En déduire que $(N_k)_{k\in\mathbf{Z}}$ vérifie (1) en précisant $c$ et $\sigma^2$ en fonction de $\alpha$ et $\beta$ **(1pt)**.

(d) Si $M$ est un processus gaussien vérifiant (1), montrer que $M$ est nécessairement un processus MA(1) et donner son écriture à l'aide d'un bruit blanc gaussien $\varepsilon = (\varepsilon_k)_{k\in\mathbb{Z}}$ dont vous préciserez la variance **(2pts)**.

(e) **(Question facultative sur 2 points)** Montrer que si $M$ est la somme d'un processus MA(1) et d'un processus ARCH($p$) stationnaire d'ordre 2, indépendant de ce MA(1), alors $M$ vérifie également la relation (1). Y-a-t-il une contradiction avec la propriété précédente?

*Proof.* (a) On utilise l'inégalité de Cauchy-Schwarz.

(b) On sait que si $\sum_k |r(k)| < \infty$, où $r$ est la covariance de $M$, alors $M$ admet une densité spectrale continue. Ceci est clairement le cas puisque $r(0) = \alpha$, $r(1) = \beta$ et $r(k) = 0$ pour $|k| \geq 2$. On a ainsi $f(\lambda) = \alpha + 2\beta\cos(\lambda)$ pour tout $\lambda \in [-\pi, \pi]$. Mais on sait qu'une densité spectrale est nécessairement positive. Ceci implique donc que $\alpha \pm 2\beta \geq 0$ (pour $\lambda = \pi$ ou 0) donc $2|\beta| \leq \alpha$.

(c) On sait que $(N_k)$ est un processus MA(1) centré et stationnaire. De plus $\mathbb{E}N_k^2 = \mathbb{E}\big((\xi_k + c\,\xi_{k-1})^2\big) = \sigma^2(1 + c^2)$ car $\xi_k$ et $\xi_{k+1}$ sont indépendantes et centrées et $\mathbb{E}N_k N_{k+1} = \mathbb{E}\big((\xi_k + c\,\xi_{k-1})(\xi_{k+1} + c\,\xi_k)\big) = c\sigma^2$. Ainsi $(N_k)$ vérifie (1) lorsque $\alpha = \sigma^2(1 + c^2)$ et $\beta = c\sigma^2$. Donc $\alpha = \sigma^2 + \frac{\beta^2}{\sigma^2}$, soit $\sigma^4 - \alpha\sigma^2 + \beta^2 = 0$; on trouve ainsi que $\sigma^2 = \frac{1}{2}\big(\alpha + \sqrt{\alpha^2 - 4\beta^2}\big)$ (on retrouve ici également la condition $2|\beta| \leq \alpha$) et $c = 0$ si $\beta = 0$, $c = \frac{1}{2\beta}\big(\alpha - \sqrt{\alpha^2 - 4\beta^2}\big)$ si $\beta \neq 0$.

(d) La loi d'un processus gaussien est entièrement donnée par ses moments d'ordre 1 et 2. Or si $M$ est un processus gaussien centré qui vérifie (1) alors $M$ a la même loi que $N$ le processus MA(1) de la question précédente, quand $(\xi_k)$ est gaussien. Donc $M$ est un bien un processus MA(1) qui s'écrit $M_k = \varepsilon_k + c\varepsilon_{k-1}$ avec $c$ donné dans la question précédente, où $(\varepsilon_k)$ est un bruit blanc gaussien de variance $\sigma^2$ donné dans la question précédente.

(e) Il est clair que comme la covariance d'un ARCH($p$) est nulle, alors si on note $\gamma^2$ la variance du processus ARCH($p$) on obtient pour nouvelles équations: $\alpha = (1 + c^2)\sigma^2 + \gamma^2$ et $\beta = c\sigma^2$. Ceci permet de trouver tout un ensemble de solutions possibles. Ainsi $M$ n'est pas seulement un MA(1). Cela n'est pas en contradiction avec la question précédente (qui imposait que $M$ soit seulement un MA(1)) car on sait qu'un ARCH($p$) ne peut pas être un processus gaussien.

□

2. **(7 points)** On considère deux processus stationnaires indépendants $X = (X_k)_{k\in\mathbb{Z}}$ et $Y = (Y_k)_{k\in\mathbb{Z}}$ tels que pour tout $k \in \mathbb{Z}$,

$$X_k = \rho_1 X_{k-1} + u_k, \qquad \text{avec } 0 < |\rho_1| < 1, \tag{2}$$

$$Y_k = \rho_2 Y_{k-1} + v_k, \qquad \text{avec } 0 < |\rho_2| < 1, \tag{3}$$

où $u = (u_k)_{k\in\mathbb{Z}}$ et $v = (v_k)_{k\in\mathbb{Z}}$ sont deux bruits blancs gaussiens indépendants de variances respectives $\sigma_u^2$ et $\sigma_v^2$. Soit enfin $Z = (Z_k)_{k\in\mathbb{Z}}$ tel que

$$Z_k = X_k + Y_k \quad \text{pour tout } k \in \mathbb{Z}.$$

(a) Si $\rho_1 = \rho_2$, quel type de processus est $Z$ **(1pt)**?

(b) Si $\rho_1 \neq \rho_2$, montrer que $Z$ est un processus ARMA$(2,1)$ dont vous donnerez l'écriture (on pourra considérer $Z_k - (\rho_1 + \rho_2)Z_{k-1} + \rho_1\rho_2 Z_{k-2}$, puis s'aider de la question 1.(d)) **(4pts)**.

(c) Quelle relation doivent satisfaire $\rho_1$, $\rho_2$, $\sigma_u^2$ et $\sigma_v^2$ pour que $Z$ soit un AR(2) **(2pts)**?

*Proof.* (a) Si $\rho_1 = \rho_2$ alors $(X_k + Y_k) = \rho_1(X_{k-1} + Y_{k-1}) + (u_k + v_k)$ pour tout $k \in \mathbf{Z}$. Donc si on pose $\varepsilon_k = u_k + v_k$ pour tout $k \in \mathbf{Z}$, alors $(\varepsilon_k)_k$ est un bruit blanc gaussien de variance $\sigma_u^2 + \sigma_v^2$ et $Z_k = \rho_1 Z_{k-1} + \varepsilon_k$ pour tout $k \in \mathbf{Z}$: $(Z_k)_k$ est un processus AR(1) gaussien stationnaire.

(b) On peut écrire que $Z_k = (\rho_1+\rho_2)Z_{k-1} - \rho_1\rho_2 Z_{k-2} + u_k + v_k - \rho_2 u_{k-1} - \rho_1 v_{k-1}$. Posons $M_k = u_k + v_k - \rho_2 u_{k-1} - \rho_1 v_{k-1}$. Alors $(M_k)_k$ est un processus gaussien (filtre linéaire d'un processus gaussien) $\mathbb{E}M_k^2 = \sigma_u^2(1+\rho_2^2)+\sigma_v^2(1+\rho_1^2)$ et $\mathbb{E}M_k M_{k-1} = -\rho_2\sigma_u^2 - \rho_1\sigma_v^2$ et $\mathbb{E}M_k M_{k-j} = 0$ dès que $|j| \geq 2$. Donc d'après la question 1. (c), $(M_k)_k$ est un processus AR(1) gaussien et peut s'écrire sous la forme $M_k = \varepsilon_k + c\varepsilon_{k-1}$ avec (après calculs):

$$\begin{aligned}
\mathbb{E}M_k^2 &= \frac{1}{2}\Big(\sigma_u^2(1+\rho_2^2) + \sigma_v^2(1+\rho_1^2) + \sqrt{\big(\sigma_u^2(1-\rho_2^2) + \sigma_v^2(1-\rho_1^2)\big)^2 + 4\sigma_u^2\sigma_v^2(\rho_1+\rho_2)^2}\ \Big) \\
\text{et } \ c &= -\frac{1}{2(\rho_2\sigma_u^2 + \rho_1\sigma_v^2)}\Big(\sigma_u^2(1+\rho_2^2) + \sigma_v^2(1+\rho_1^2) - \sqrt{\big(\sigma_u^2(1-\rho_2^2) + \sigma_v^2(1-\rho_1^2)\big)^2 + 4\sigma_u^2\sigma_v^2(\rho_1+\rho_2)^2}\ \Big).
\end{aligned}$$

Donc $Z_k - (\rho_1+\rho_2)Z_{k-1} + \rho_1\rho_2 Z_{k-2} = M_k - cM_{k-1}$ où $(M_k)_k$ est un bruit blanc gaussien de variance $\mathbb{E}M_k^2$: $(Z)$ est bien un processus ARMA$(2,1)$.

(c) $(Z_k)_k$ est un processus AR(2) si $\mathbb{E}M_k M_{k-1} = -\rho_2\sigma_u^2 - \rho_1\sigma_v^2 = 0$, donc si $\rho_2\sigma_u^2 = -\rho_1\sigma_v^2$ (on remarque que ceci n'est pas possible si $\rho_1 = \rho_2 \neq 0$, cas qui correspondait à un AR(1)).

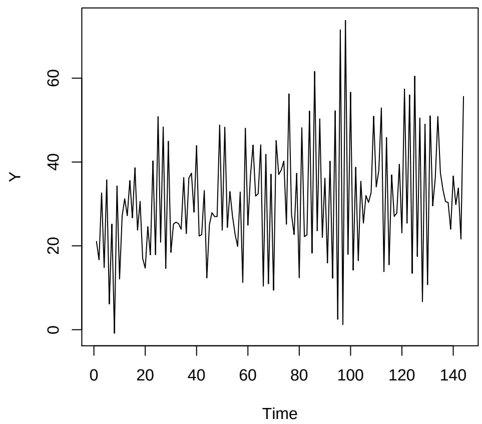
□

3. **(9 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)* **(0.5pts)**? *Qelles sont les tendances et saisonnalités de Y* **(1pt)**? *Quelle est sa variance (théorique)* **(2pts)**?

*Proof.* $X$ est un ARMA$(1,1)$ d'équation $X_n + 0.7X_{n-1} = epsilon_n - 2epsilon_{n-1}$ .
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 . 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, $\mathrm{var}(X_0) = 9\big(1 + (2.7)^2\sum_{k=1}^{\infty} 0.49^{k-1}\big) = 9\big(1 + (2.7)^2/0.51\big) = 9 * 260/17 \simeq 137$. □

(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* **(0.5pts)***? Que représentent les valeurs* 10.6447975*,* 0.593669 *et* 1.5746636 **(1.5pts)***? Que conclure à la lecture de ces résultats* **(0.5pts)***?*

*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.
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 1.5746636 est la valeur de l'écart-type estimé de l'estimateur du coefficient devant la variable $Z5$.
Il apparaît que le modèle n'est pas satisfaisant (en particulier les variables $Z1$, $Z2$ et $Z3$ ne semble pas forcément explicatives). □

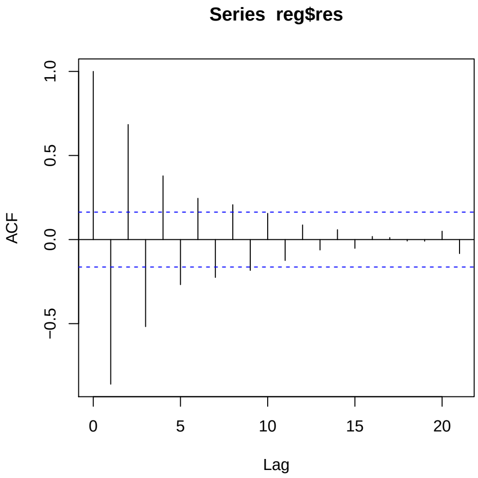
(c) On tape enfin les commandes suivantes:

```
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* **(0.5pts)***? Est-on satisfait du modèle obtenu* **(0.5pts)***? Pouvait-on s'attendre aux valeurs numériques* 3.4173 *et* 5.9084 **(0.5pts)***? Expliquez ce que représente le graphe* **(0.5pts)***. La valeur numérique décrite par la seconde barre est approximativement de* $-0.85$*. Expliquez ce que l'on pouvait espérer comme valeur théorique* **(1pt)***.*

*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.
On est satisfait a priori du modèle choisi, car toutes les $p$-values des tests de student sont inférieures à 0.05.
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.
Le graphe représente les autocorrélations empiriques des résidus de la régression, qui sont des estimations des $epsilon_i$.
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 $1.49\, r(0) + 1.4 r(1) = 45$. 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!. $\square$