*Université Paris I, Panthéon - Sorbonne*

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

# Statistiques II

## Correction du contrôle continu n°2, avril 2014

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

1. **(Sur 10 points)** Soit $(\delta_n)_{n\in\mathbf{N}}$ une chaîne de Markov prenant ses valeurs dans $\{0,1\}$, telle que pour tout $n\in\mathbf{N}$, $\mathbb{P}(\delta_{n+1}=1 \mid \delta_n=1)=\mathbb{P}(\delta_{n+1}=0 \mid \delta_n=0)=\theta$ avec $\theta\in[0,1]$.
   (a) Donner la matrice de transition de $(\delta_n)_{n\in\mathbf{N}}$ **(0.5pts)**. Est-ce une chaîne homogène **(0.5pts)**?
   (b) Déterminer les valeurs de $\theta$ pour lesquelles $(\delta_n)$ est irréductible, et celles pour lesquelles elle ne l'est pas **(1pt)**.
   (c) Déterminer les mesures invariantes de $(\delta_n)$ suivant $\theta$ **(1pt)**.
   (d) On suppose que la loi de $\delta_0$ est $\mathbb{P}(\delta_0=1)=p_0$ avec $p_0\in[0,1]$. Montrer que $\mathbb{P}(\delta_n=0)=\frac{1}{2}\left(1+(2\theta-1)^n(2p_0-1)\right)$ pour tout $n\in\mathbf{N}$ **(2pts)**. En déduire l'espérance et la variance de $\delta_n$ **(1pt)**. A quelles conditions sur $p_0$ et $\theta$ la suite $(\delta_n)$ est-elle stationnaire **(1pt)**?
   (e) Suivant les valeurs de $p_0$ et $\theta$, déterminer la loi limite de $\delta_n$ quand $n\to\infty$ **(1.5pts)**.
   (f) Si $\theta=p_0=1/2$, montrer que $(\delta_n)_{n\in\mathbf{N}}$ est une suite de variables aléatoires indépendantes identiquement distribuées suivant une loi de Bernoulli dont on précisera le paramètre **(1.5pts)**.

*Proof.* (a) La matrice est $Q=\begin{pmatrix}\theta & 1-\theta\\ 1-\theta & \theta\end{pmatrix}$. C'est une chaîne homogène car la matrice ne dépend pas de $n$.

(b) Si $\theta=1$ les états ne communiquent plus, les 2 états sont absorbants, la chaîne n'est pas irréductible. Pour $\theta\in[0,1[$, elle est bien irréductible.

(c) Il faut résoudre $(p,1-p)Q=(p,1-p)$, ce qui revient à écrire $p(2-2\theta)=1-\theta$ et $(1-\theta)(2p-1)=0$. Donc si $\theta\neq 1$, on trouve une unique mesure invariante $(1/2,1/2)$. Si $\theta=1$, toutes les mesures sont invariantes.

(d) On sait que $\left(\mathbb{P}(\delta_n=0),\mathbb{P}(\delta_n=1)\right)=(p_0,1-p_0)Q^n$. Pour calculer $Q^n$ il suffit de la diagonaliser. Ses valeurs propres sont 1 et $2\theta-1$. Ainsi $Q=\frac{1}{2}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}1&0\\0&2\theta-1\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}$, d'où $Q=\frac{1}{2}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}1&0\\0&(2\theta-1)^n\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}$ et en effectuant le calcul de $(p_0,1-p_0)Q^n$, on trouve $\mathbb{P}(\delta_n=0)=\frac{1}{2}\left(1+(2\theta-1)^n(2p_0-1)\right)$.
Comme $\delta_n$ est une variable de Bernoulli, on en déduit que $\mathbb{E}(\delta_n)=\mathbb{P}(\delta_n=1)=\frac{1}{2}\left(1-(2\theta-1)^n(2p_0-1)\right)$, et $\mathrm{var}(\delta_n)=\mathbb{P}(\delta_n=1)(1-\mathbb{P}(\delta_n=1))=\frac{1}{4}\left(1-(2\theta-1)^{2n}(2p_0-1)^2\right)$.
Il faut déjà que $\mathbb{P}(\delta_n=0)$ ne dépendent pas de $n$ pour avoir la stationnarité: il faut donc $\theta=1$ ou $p_0=1/2$. Si $\theta=1$, on doit nécessairement avoir $p_0=1/2$ car on doit avoir $\mathbb{P}(\delta_n=0)=\mathbb{P}(\delta_0)$. Conclusion: nécessairement on doit avoir $p_0=1/2$ pour avoir la nécessité.

(e) Si $\theta=1$, la loi limite est $(p_0,1-p_0)$, qui est aussi une mesure invariante. Si $\theta=0$, il n'y a pas de loi limite, sauf si $p_0=1/2$ auquel cas la loi limite est la loi de Bernoulli de paramètre $1/2$. Enfin, si $\theta\in]0,1[$, la loi limite est la loi de Bernoulli de paramètre $1/2$ quel que soit $p_0$.

(f) Il est clair d'après ce qui précède que si la loi de Bernoulli de paramètre $1/2$ est la loi de $\delta_n$ pour tout $n$. Il reste à montrer l'indépendance quand $\theta=1/2$. Alors d'après la stationnarité, pour tout $i$, $k\in\mathbf{N}$, $\mathbb{P}(X_i=x_i\cap X_{i+k}=x_j)=(Q^k)_{ij}\mathbb{P}(X_i=x_i)$. Comme $Q^k=Q=\frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}$ et $\mathbb{P}(X_i=x_i)=1/2$ pour $x_i=0$ ou $x_i=1$, on obtient $\mathbb{P}(X_i=x_i\cap X_{i+k}=x_j)=1/4=\mathbb{P}(X_i=x_i)\times\mathbb{P}(X_{i+k}=x_j)$: il y a bien indépendance.

□

2. **(Sur 10 points)** Soit $\varepsilon=(\varepsilon_t)_{t\in\mathbf{Z}}$ un bruit blanc fort centré de variance $\sigma^2>0$, indépendant de $(\delta_n)_{n\in\mathbf{N}}$. On définit le processus $X=(X_n)_{n\in\mathbf{N}}$ tel que

$$X_n=\delta_n(\varepsilon_n-a\varepsilon_{n-2})+(1-\delta_n)\varepsilon_n \qquad \text{pour} \quad n\in\mathbf{N}.$$

avec $a\in\mathbf{R}$, où $(\delta_n)$ est la chaîne de Markov précèdente.

(a) Lorsque $\theta = 1$ et $p_0 = 1$, montrer que pour tout $n \in \mathbf{N}$, $X_n = \varepsilon_n - a\varepsilon_{n-2}$ **(0.5pts)**. Quel processus est alors $(X_n)$ **(0.5pts)**? Déterminer son espérance **(0.5pts)** et sa fonction d'autocovariance **(1pt)**.

(b) On considère maintenant le cas general $\theta \in ]0, 1/2[\cup]1/2, 1[$ avec $p_0 = 1/2$ ($(\delta_n)$ est alors stationnaire). Montrer que $(X_n)$ est stationnaire **(2pts)**. Déterminer l'espérance **(0.5pts)** et l'autocovariance de $(X_n)$ **(1.5pts)**. Montrer que $\mathbb{E}\big[X_n^2 X_{n+3}^2\big] \neq \mathbb{E}\big[X_n^2\big]\mathbb{E}\big[X_{n+3}^2\big]$ **(2.5pts)**. En déduire que $(X_n)$ n'est pas un processus MA(2) **(1pt)**.

*Proof.* (a) Dans ce cas $\delta_n = 1$ pour tout $n$, d'où $X_n = \varepsilon_n - a\varepsilon_{n-2}$.
$(X_n)$ est alors un processus MA(2). Son espérance est 0 et sa fonction d'autocovariance $r(\cdot)$ vaut $r(k) = 0$ si $|k| \geq 3$, $r(0) = \sigma^2(1 + a^2)$, $r(1) = r(-1) = 0$ et $r(2) = r(-2) = -a\sigma^2$.

(b) Soit $n \in \mathbf{N}^*$, $(i_1, \cdots, i_n) \in \mathbf{N}^n$, $c \in \mathbf{N}^*$. Alors du fait des stationnarités de $(\varepsilon_k)$ et $(\delta_k)$, on a $(\varepsilon_{i_1}, \varepsilon_{i_1-2}, \cdots, \varepsilon_{i_n}, \varepsilon_{i_n-2}) \xrightarrow{\mathcal{L}} (\varepsilon_{i_1+c}, \varepsilon_{i_1-2+c}, \cdots, \varepsilon_{i_n+c}, \varepsilon_{i_n-2+c})$ et $(\delta_{i_1}, \cdots, \delta_{i_n}) \xrightarrow{\mathcal{L}} (\delta_{i_1+c}, \cdots, \delta i_n + c)$. Comme les 2 processus sont indépendants entre eux, il est clair que:
$(\varepsilon_{i_1}, \varepsilon_{i_1-2}, \delta_{i_1}, \cdots, \varepsilon_{i_n}, \varepsilon_{i_n-2}, \delta_{i_n}) \xrightarrow{\mathcal{L}} (\varepsilon_{i_1+c}, \varepsilon_{i_1-2+c}, \delta_{i_1+c}, \cdots, \varepsilon_{i_n+c}, \varepsilon_{i_n-2+c}, \delta i_n + c)$. On peut alors appliquer une fonction $g : (x_1, \cdots, x_{3n}) \mapsto (x_1 - ax_2x_3, \cdots, x_{3n-2} - ax_{3n-1}x_{3n}) \in \mathbf{R}^n$, qui est mesurable, des deux côtés de l'égalité en loi et on obtient ainsi que $(X_{i_1}, \cdots, X_{i_n}) \xrightarrow{\mathcal{L}} (X_{i_1+c}, \cdots, X_{i_n+c})$: le processus est bien stationnaire.
On a $\mathbb{E}X_n = \mathbb{E}\varepsilon - a\mathbb{E}(\delta_n\varepsilon_{n-2}) = 0 - a\mathbb{E}(\delta_n)\mathbb{E}(\varepsilon_{n-2}) = 0$ grâce à l'indépendance.
En utilisant l'indépendance entre les $(\varepsilon_i)$ et des $\varepsilon_i$ avec les $\delta_j$, et les fait que les $\varepsilon_i$ sont centrées, on obtient $r_X(0) = \mathbb{E}((\varepsilon_n - a\delta_n\varepsilon_{n-2})^2) = \sigma^2 + a^2\sigma^2\mathbb{E}\delta_n^2 = \sigma^2(1 + a^2/2)$, $r_X(1) = \mathbb{E}(X_nX_{n+1}) = 0$, $r_X(2) = \mathbb{E}((\varepsilon_n - a\delta_n\varepsilon_{n-2})(\varepsilon_{n+2} - a\delta_{n+2}\varepsilon_n) = -a\sigma^2\mathbb{E}\delta_{n+2} = -a\sigma^2/2$ et $r_X(k) = 0$ pour $|k| \geq 3$.
On a $\mathbb{E}\big[X_n^2X_{n+3}^2\big] = \mathbb{E}\big[(\varepsilon_n - a\delta_n\varepsilon_{n-2})^2(\varepsilon_{n+3} - a\delta_{n+3}\varepsilon_{n+1})^2\big] = \sigma^4 + a^2\mathbb{E}(\delta_{n+3}^2)\sigma^4 + a^2\mathbb{E}(\delta_n^2)\sigma^4 + a^4\mathbb{E}(\delta_n^2\delta_{n+3}^2)\sigma^4$. Mais $\mathbb{E}(\delta_i^2) = 1/2$ pour tout $i$, et $\mathbb{E}(\delta_n^2\delta_{n+3}^2) = \mathbb{P}(\delta_n = 1 \cap \delta_{n+3} = 1) = \frac{1}{2}(Q^3)_{22} = \frac{1}{4}(1 + (2\theta - 1)^3)$. D'où $\mathbb{E}\big[X_n^2X_{n+3}^2\big] = \sigma^4 + a^2\sigma^4 + \frac{1}{4}(1 + (2\theta - 1)^3)a^4\sigma^4$.
Par ailleurs, $\mathbb{E}\big[X_n^2\big]\mathbb{E}\big[X_{n+3}^2\big] = \sigma^4(1 + a^2 + a^4/4)$. On a donc bien $\mathbb{E}\big[X_n^2X_{n+3}^2\big] \neq \mathbb{E}\big[X_n^2\big]\mathbb{E}\big[X_{n+3}^2\big]$.
De ceci on déduit que $X_n$ et $X_{n+3}$ ne sont pas indépendantes, donc $(X_n)$ ne peut pas être un processus MA(2).

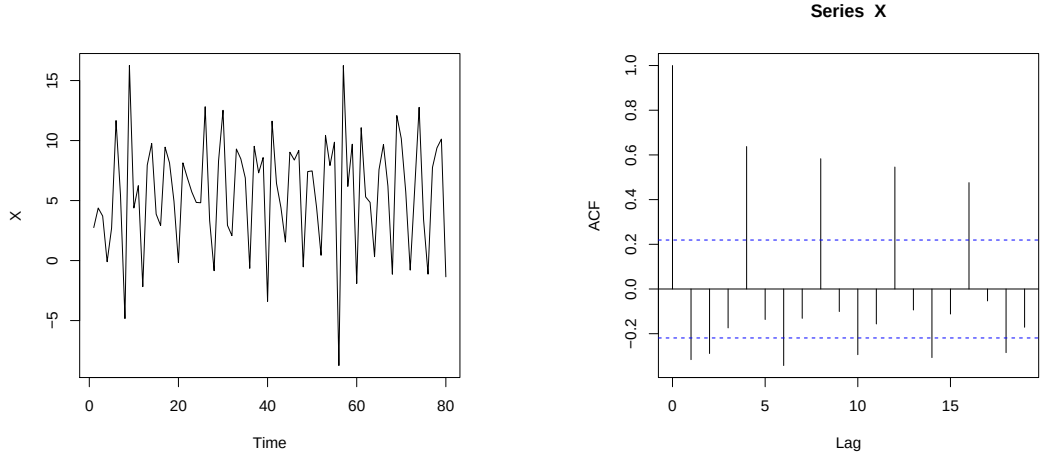
□

3. **(Sur 8 points)** Voici des simulations effectuées avec le logiciel R.

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

```
eps=2*rnorm(81)
X=6-6/c(1:80)+rep(c(3,4,0,-7),20)+eps[2:81]-0.8*eps[1:80]
ts.plot(X)
acf(X)
```

Voici les deux graphes obtenus:




*Questions: Quel est le processus simulé par le vecteur X (écrire le processus formellement en détaillant ses éventuelles tendance et saisonnalité)* **(1pt)***? Que peut-on déduire de la commande* `ACF` **(0.5pts)***?*

*Proof.* On a $X_n = 6 - 6/n + s(n) + \varepsilon_n - 0.8\varepsilon_{n-1}$ où $s(n) = 3\mathbb{1}_{n=1[4]} + 4\mathbb{1}_{n=2[4]} - 7\mathbb{1}_{n=4[4]}$ est la saisonnalité de période 4, la tendance étant $a(n) = 6 - 6/n$, le bruit étant un MA(1).
On ne peut rien déduire de l'ACF car $X$ n'est pas stationnaire. On s'aperçoit cependant que la période de la saisonnalité est apparente (un pic toutes les 4 unités). □

(b) Voici les commandes tapées ensuite:

```
Z1=c(1:80)
Z2=rep(c(1,0,0,-1),20)
Z3=rep(c(0,1,0,-1),20)
Z4=rep(c(0,0,1,-1),20)
reg=lm(X~Z1+Z2+Z3+Z4)
summary(reg)
```

Voici les résultats obtenus:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.88262    0.64595   7.559 8.21e-11 ***
Z1           0.01901    0.01386   1.372    0.174
Z2           3.29274    0.55407   5.943 8.22e-08 ***
Z3           2.85634    0.55372   5.158 1.97e-06 ***
Z4           0.31241    0.55372   0.564    0.574

Residual standard error: 2.859 on 75 degrees of freedom
Multiple R-squared: 0.6656,     Adjusted R-squared: 0.6478
F-statistic: 37.32 on 4 and 75 DF,  p-value: < 2.2e-16
```

*Questions: Qu'a-t-on fait par ces commandes* **(1pt)***? Expliquer pourquoi la* $p-$*value associée à* $Z1$ *est importante* **(0.5pts)**.

*Proof.* On a effectué une régression de $(X_1,\cdots,X_{80})$ par $Z1$, $Z2$, $Z3$ et $Z4$, $Z1$ décrit une tendance linéaire, $Z2$, $Z3$ et $Z4$ étant présentes pour la saisonnalité.
Le modèle simulé n'ayant pas une tendance linéaire, il n'est pas surprenant que le coefficient devant cette tendance linéaire soit considéré comme nul. □

(c) Voici les commandes tapées ensuite:

```
Z11=1/c(1:80)
reg=lm(X~Z11+Z2+Z3+Z4)
summary(reg)
```

Voici les résultats obtenus:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0998     0.3401  17.936  < 2e-16 ***
Z11          -7.2057     2.3984  -3.004  0.00362 **
Z2            3.4665     0.5339   6.493 8.15e-09 ***
Z3            2.8463     0.5296   5.374 8.35e-07 ***
Z4            0.2434     0.5303   0.459  0.64751

Residual standard error: 2.735 on 75 degrees of freedom
Multiple R-squared: 0.694,      Adjusted R-squared: 0.6777
F-statistic: 42.53 on 4 and 75 DF,  p-value: < 2.2e-16
```

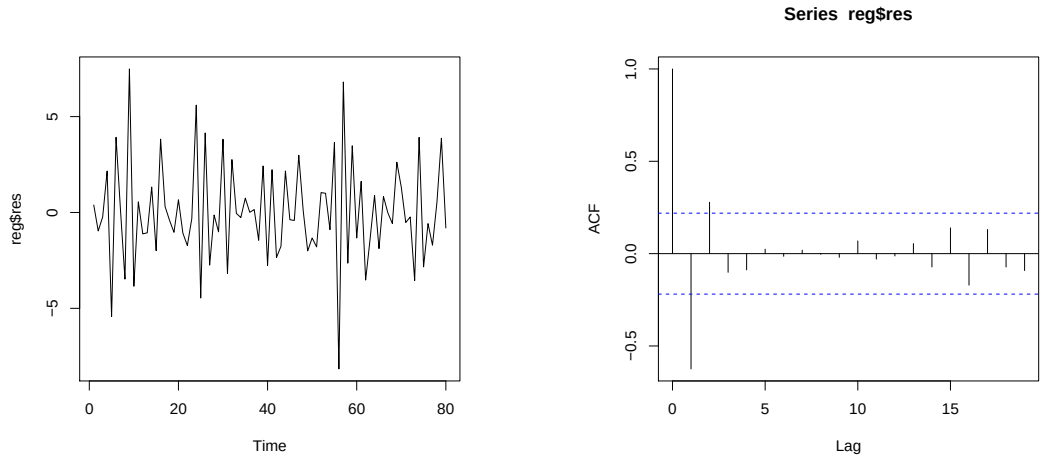
*Questions: Qu'a-t-on fait par ces commandes par rapport à celles qui précèdent et a-t-on gagné quelque chose* **(0.5pts)***? Que représentent formellement les valeurs* $6.0998$, $-7.2057$, $3.4665$ **(0.5pts)***? S'attendait-on approximativement à ces valeurs et pourquoi* **(0.5pts)***? Quelle sont la tendance et la saisonnalité estimées* **(0.5pts)***? Que peut-on conclure quant au modèle utilisé pour expliquer* $(X_t)$ **(0.5pts)***?*

*Proof.* On a effectué une régression de $(X_1,\cdots,X_{80})$ par $Z11$, $Z2$, $Z3$ et $Z4$, $Z11$ étant la variable $1/i$ comme dans le modèle simulé. C'est donc le bon modèle que l'on utilise et ainsi on a amélioré le coefficient $R^2$.
Ces valeurs sont les coefficients estimés devant les différentes variables.
Les valeurs théoriques que l'on espère retrouver par la régression sont celles du modèle simulé, donc 6, $-6$ et 3, les valeurs estimées sont donc assez proches.
La tendance estimée est $\widehat{a}(i)=6.0998-7.2057/i$, la saisonnalité estimée est $\widehat{s}(i)=3.4665\mathbb{1}_{i=1[4]}+2.8463\mathbb{1}_{i=2[4]}+0.2434\mathbb{1}_{n=3[4]}-6.5562\mathbb{1}_{n=4[4]}$.
Le test de Fisher indique une bonne adéquation du modèle. Seul bémol apparent la p-value associée à $Z4$ est élevé mais c'est finalement normal puisque le vrai coefficient est nul! □

(d) On tape enfin les commandes suivantes:

```
ts.plot(reg$res)
acf(reg$res)
```

Voici les deux graphes obtenus:




*Questions: Que conclure de l'ACF? Les valeurs numériques des deux premières barres sont* 1 *et* $-0.64$. *Pouvait-on s'attendre à cela numériquement* **(1pt)***?*

*Proof.* On retrouve à peu près la structure d'un MA(1) qui est le bruit simulé. En effet, théoriquement on aurait du obtenir $\rho(0) = 1$ et $\rho(1) = -0.8/(1+0.8^2) \simeq -0.48$. Les valeurs numériques que l'on trouve sont assez cohérentes avec ces valeurs. □

(e) *Questions: Donner le code pour simuler une trajectoire de taille* 100 *du processus* $ARMA(2,1)$*:* $X_t - 0.3X_{t-1} - 0.1X_{t-2} = \varepsilon_t + 2\varepsilon_{t-1}$*, avec un bruit* $(\varepsilon_t)$ *qui suit une loi normale de variance* 9 *(on prendra soin de vérifier auparavant que l'on a bien un ARMA causal...)* **(1.5pts)**

*Proof.* On a bien un ARMA causal car le polynôme caractéristique est $P(X) = 1 - 0.3X - 0.1X^2 = (1 - 0.5X)(1 + 0.2X)$, donc les racines sont 2 et $-5$: à l'extérieur du disque unité.
Les commandes peuvent être alors:

```
eps=3*rnorm(102)
Z=0; Z[1]=0; Z[2]=0;
X=0;
for (k in c(3:102))
{Z[k]=0.3*Z[k-1]-0.1*Z[k-2]+eps[k]+2*eps[k-1]
X[k-2]=Z[k]}
```

□