Examen du 03/04/2025

MAP-STA2 : Séries chronologiques

Yannig Goude [yannig.goude@edf.fr]
Guillaume Principato [guillaume.principato@edf.fr]

Exercice 1 /3

1. Donner la définition du lissage exponentiel simple. De quel problème d'optimisation l'estimateur obtenu est-il la solution (à justifier)? /1

   Cf le cours:

   soit une série temporelle y_t. On appelle lissage exponentiel simple de paramètre $\alpha \in [0,1]$ de cette série le processus $\widehat{y}_t$ définie ainsi: $\widehat{y}_{t+1/t} = \alpha y_t + (1-\alpha)\widehat{y}_{t/t-1}$, on a donc $\widehat{y}_{t+1/t} = \sum_{i=0}^{t-1} \alpha(1-\alpha)^i y_{t-i}$

   la prévision de l'instant $t+1$ est donc une somme pondérée des valeurs passées de la série, les poids décroissants exponentiellement dans le passé. Le prédicteur obtenu par lissage exponentiel simple peut être vu comme celui qui minimise un problème de moindres carrés pondérés:

   En effet, si on note:

   $$\widehat{a} = \text{argmin}_a \sum_{i=0}^{t-1} (1-\alpha)^i (y_{t-i} - a)^2$$

   on a $\widehat{a} = \sum_{i=0}^{t-1}(1-\alpha)^i y_{t-i} / \sum_{i=0}^{t-1}(1-\alpha)^i$, soit

   $\widehat{a} = \frac{\alpha}{1-(1-\alpha)^t} \sum_{i=0}^{t-1}(1-\alpha)^i y_{t-i}$

   ce qui, si t est suffisamment grand tend vers l'estimateur de lissage exponentiel simple définie précédemment.Si on défnie y_t comme $(0, ..., 0, y_1, ..., y_t)$ ie valant 0 pour les indices temporelles inférieurs à 1 alors

   $$\widehat{y}_{t+1} = \text{argmin}_a \sum_{i=0}^{\infty} (1-\alpha)^i (y_{t-i} - a)^2$$

2. Donner la définition du lissage exponentiel double. De quel problème d'optimisation l'estimateur obtenu est-il la solution, (à justifier)? /1

   Soit une série temporelle y_t. On appelle lissage exponentiel double (ou de Holt) de paramètre $\alpha \in [0,1]$ de cette série le processus $\widehat{y}_t$ définie ainsi:

   $$\widehat{y}_{t+h/t} = l_t + b_t h$$

   avec:

   $$\begin{cases} l_t = l_{t-1} + b_{t-1} + (1-(1-\alpha)^2)(y_t - \widehat{y}_{t/t-1}) \\ b_t = b_{t-1} + \alpha^2 (y_t - \widehat{y}_{t/t-1}) \end{cases}$$

l_t et b_t minimise à chaque instant:

$$(\widehat{l}, \widehat{b}) = \text{argmin}_{l,b} \sum_{i=0}^{\infty} (1-\alpha)^i (y_{t-i} - (l + bi))^2$$

3. Quel type de lissage exponentiel doit-on utiliser pour modéliser une série temporelle possédant une tendance et une saisonnalité? /1

   Le lissage de holt-winters saisonnier: soit une série temporelle y_t. On appelle lissage exponentiel double de Holt-Winters saisonnier de paramètres $\alpha \in [0,1]$, $\beta \in [0,1]$, $\delta \in [0,1]$ de cette série le processus $\widehat{y}_t$ définie ainsi:

$$\widehat{y}_{t+h} = l_t + hb_t + s_t$$

   avec

$$\begin{cases} l_t = \alpha(y_t - s_{t-T}) + (1-\alpha)(l_{t-1} + b_{t-1}) \\ b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ s_t = \delta(y_t - l_t) + (1-\delta)s_{t-T} \end{cases}$$

Exercice 2 /4

1. Donner la définition de la stationnarité forte et faible. Expliquer en pratique comment on vérifie cette dernière. /1

   soit un processus aléatoire $(Y_t)_{t\in\mathbf{Z}}$, il est dit stationnaire au sens fort (ou strictement) si pour toute fonction f mesurable $f(Y_1, Y_2, ..., Y_t)$ et $f(Y_{1+h}, Y_{2+h}, ..., Y_{t+h})$ ont la même loi.

   soit un processus aléatoire $(Y_t)_{t\in\mathbf{Z}}$ tel que $E(Y_t^2) < \infty$, il est dit stationnaire au sens faible (ou d'ordre 2) si son espérance est constante et ses auto-covariances sont stables dans le temps ie:

$$\forall t \quad E(Y_t) = \mu$$

$$\forall t \quad , \forall h \quad \text{cov}(Y_t, Y_{t+h}) = \gamma(h)$$

   En pratique, pour apprécier la stationnarité d'un processus, on commence d'abord par vérifier que sa moyenne et sa variance sont constantes dans le temps. Ensuite on regarde si l'ACF décroit rapidement vers 0.

2. Donner 3 exemples de processus stationnaires et justifier. /1.5

   - un bruit blanc ε_t vérifiant $E(\varepsilon_t) = \mu$ et $\text{var}(\varepsilon_t) = \sigma^2$

   - Le processus gaussien $(Y_t)_{t\in\mathbf{Z}}$ tel que $E(Y_t) = \mu$ et $\text{cov}(Y_t, Y_{t+h}) = \alpha^{|h|}$ $(|\alpha| < 1)$ est faiblement stationnaire. Tout processus gaussien stationnaire faible est stationnaire fort.

   - le processus moyenne mobile $X_t = \varepsilon_t + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + ... + a_q\varepsilon_{t-q}$

3. Soit $(X_t)_{t\in\mathbf{Z}}$ un processus stationnaire d'espérance μ et de fonction d'autocovariance $\gamma(h)$, $(a_i)_{i\in\mathbf{Z}}$ une suite de nombres réels absolument sommable $\sum_{i\in\mathbf{Z}} |a_i| < \infty$, montrer que $Y_t = \sum_{i\in\mathbf{Z}} a_i X_{t-i}$ est un processus stationnaire. /1.5

   tout d 'abord, on remarque que

$$\sum_{i\in\mathbf{Z}} ||a_i X_{t-i}|| = \sum_{i\in\mathbf{Z}} |a_i| ||X_{t-i}|| = \sqrt{\mathbf{Var}(X) + E(X)^2} \sum_{i\in\mathbf{Z}} |a_i| < \infty$$

$\sum_{i\in\mathbf{Z}} a_i X_{t-i}$ est donc bien convergente au sens de $\mathbf{L}^2$ et Y_t est de carré intégrable.

on vérifie également que $E(Y_t)$ ne dépend pas de t:

$$E(\sum_{i\in\mathbf{Z}} a_i X_{t-i}) = \sum_{i\in\mathbf{Z}} a_i E(X_{t-i}) = \mu \sum_{i\in\mathbf{Z}} a_i$$

ainsi que sa fonction d'autocovariance:

$$\mathrm{cov}(Y_t, Y_{t+h}) = \mathrm{cov}(\sum_{i\in\mathrm{Z}} a_i X_{t-i}, \sum_{j\in\mathrm{Z}} a_j X_{t+h-j})$$
$$\mathrm{cov}(Y_t, Y_{t+h}) = \sum_{i\in\mathrm{Z}} \sum_{j\in\mathrm{Z}} a_i a_j \mathrm{cov}(X_{t-i}, X_{t+h-j})$$

Exercice 3 /3

Soit X un AR(1) stationnaire défini ainsi:

$$\forall t \in Z, X_t = \phi X_{t-1} + \varepsilon_t$$

avec ε_t un bruit blanc centré de variance σ^2. et $|\phi| < 1$.

Soit ensuite le processus Y:

$$\forall t \in Z, Y_t = X_t - X_{t-1}$$

1. Montrer que Y est un processus stationnaire centré. Donner l'équation de récurrence vérifiée par Y et le nom exact de ce processus? /1.5

   $Y_t = \phi X_{t-1} - X_{t-1}$ $Y_t = \phi(X_{t-1} - X_{t-2}) - \varepsilon_{t-1} + \varepsilon_t = \phi Y_{t-1} - \varepsilon_{t-1} + \varepsilon_t$

   $(1 - \phi L)Y_t = (1 - L)\varepsilon_t$

   On reconnait un ARMA(1,1).

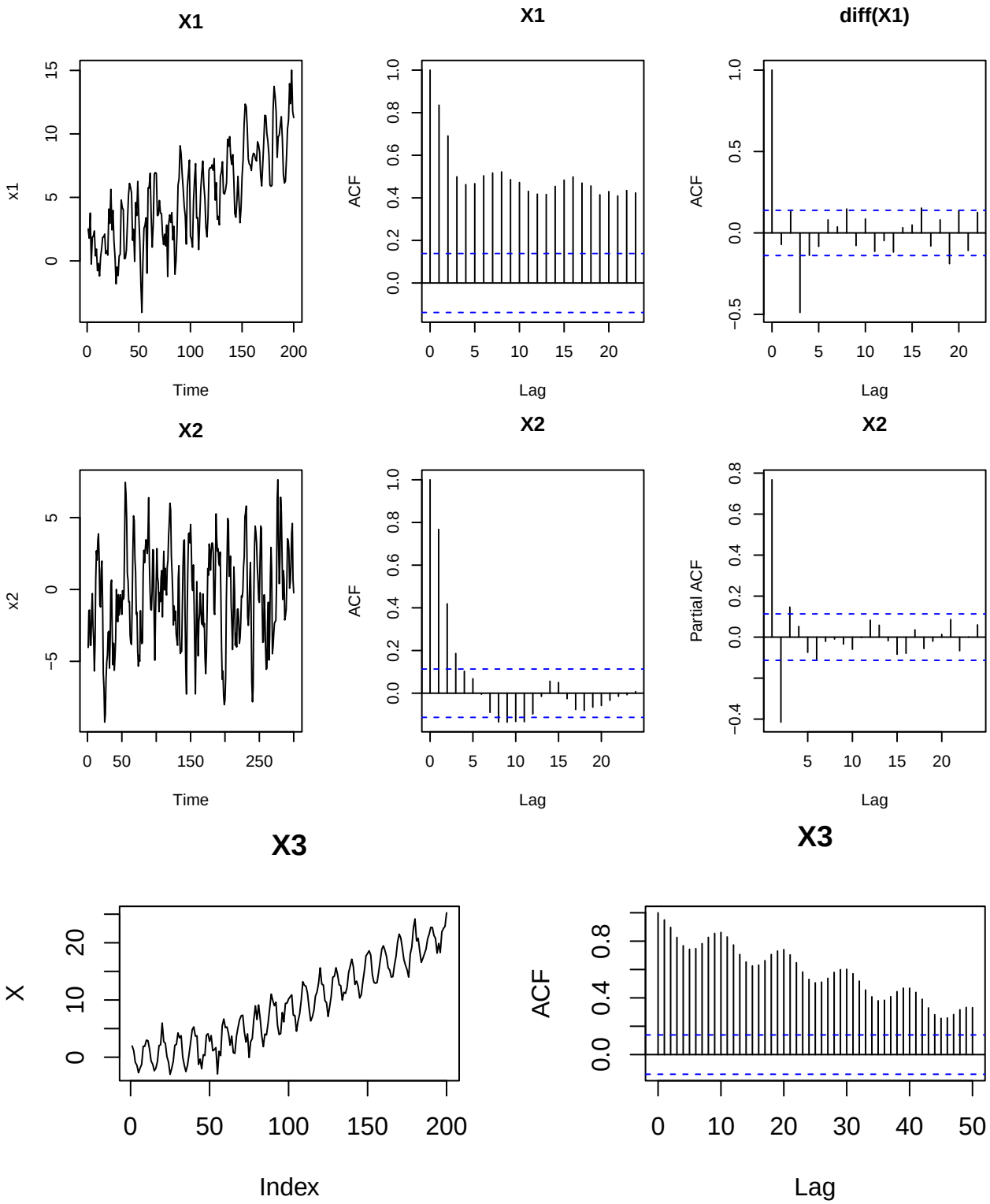
2. Calculer la variance de Y_t. /1.5

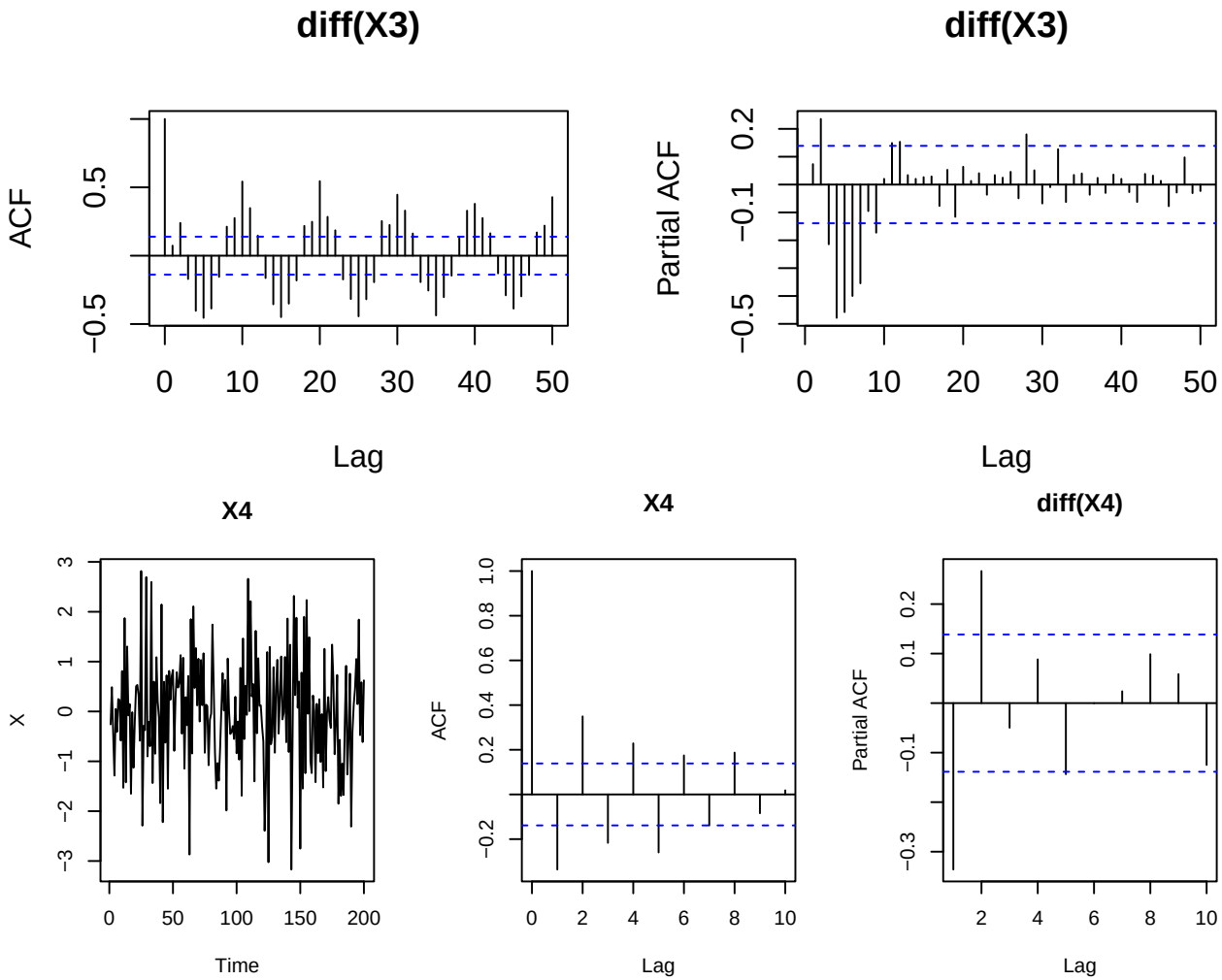
   $Y_t = (1 - L)((1 - \phi L)^{-1})\varepsilon_t$ $Y_t = \sum_{i=0}^{+\infty} \phi^i \varepsilon_{t-i} - \sum_{i=1}^{+\infty} \phi^{i-1} \varepsilon_{t-i}$ $Y_t = \varepsilon_t + \sum_{i=1}^{+\infty} \phi^i (1 - \frac{1}{\phi})\varepsilon_{t-i}$

   $Var(Y_t) = \sigma^2 + \frac{\phi^2}{1-\phi^2}(1 - \frac{1}{\phi})^2$

Exercice 4 /10

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


X1
x1
Time
X1
ACF
Lag
diff(X1)
ACF
Lag
X2
x2
Time
X2
ACF
Lag
X2
Partial ACF
Lag
X3
X
Index
X3
ACF
Lag



1. Proposer une démarche de modélisation pour chacune de ces séries, justifier. /2

- Série 1: la série présente une composante de tendance linéaire clairement visible sur le graphe de gauche. La variance du bruit semble stable dans le temps, il s'agit donc d'une tendance additive. Cette non-stationnarité est confirmée par l'ACF qui ne décroit pas vers 0. L'ACF de la série différenciée s'annule à partir du rang 3, la série différenciée est stationnaire et pourrait être modélisée par un MA(3). Celà qui confirmel'hypothèse de tendance linéaire additive. Outre la différentiation nous aurions pu corriger X1 de sa tendance en effectuant un ajustement de la tendance par régression linéaire puis en travaillant sur les résidus.

- Série 2: au vu du graphique de gauche la série ne présente pas de comportement évident de type tendance ou saisonnalité. La décroissance rapide de l'ACF confirme que la série est a priori stationnaire. Au vu de la PACF, la série présente une composante AR d'ordre max $p = 3$. Au vu de l'ACF une composante MA d'ordre max $q = 3$. Nous suggérons donc une recherche via l'AIC du meilleurs modèle ARIMA avec $p \leq 3$, $q \leq 3$.

- Série 3: la série présente une tendance et une saisonnalité. Elle semble homoscédastique. L'acf décroit lentement vers 0 ce qui confirme la non-stationnarité. Il faut donc mettre en oeuvre une procédure de correction de la tendance et de la saisonnalité. Cela peut se faire par des méthodes de régressions: linéaire par morceau (splines) pour la partie tendance, sur série de Fourier pour la partie saisonnière. L'ACF de la série différenciée nous permet d'estimer la période de la composante saisonnière qui est de 10. Après avoir modéliser la tendance et la saisonnalité il faudra étudier la dépendance temporelle des résidus (ACF, PACF) et éventuellement proposer un modèle ARMA correspondant.

- Série 4: la série ne présente ni tendance ni saisonnalité. Elle semble homoscédastique. L'acf décroit rapidement vers 0 ce qui confirme la stationnarité. Au vu de la PACF, la série présente une composante AR d'ordre 2 ou 5. Au vu de l'ACF une composante MA d'ordre max $q = 6$. Nous suggérons donc une recherche via l'AIC du meilleur modèle ARIMA avec $p \leq 3$, $q \leq 3$.

2. Notre statisticien s'intéresse ensuite à une autre série $X5_t$. Il propose de la modéliser par un ARMA et cherche ensuite à valider son modèle. Il obtient les sorties suivantes, à quoi correspond la ligne "s.e"? Quelle est l'instruction (en R) qui lui a permis d'obtenir cette sortie? /1.5

```
##
## Call:
## arima(x = X5, order = c(2, 0, 5), include.mean = F, method = c("ML"))
##
## Coefficients:
##          ar1      ar2     ma1     ma2      ma3     ma4      ma5
##       0.5778  -0.7899  0.8302  0.6687  -0.0654  0.1970  -0.1158
## s.e.  0.0678   0.0543  0.0879  0.1283   0.1127  0.0992   0.0761
##
## sigma^2 estimated as 8.85:  log likelihood = -755.99,  aic = 1527.97
```

```
## pvalue student-test:
```

```
##  ar1  ar2  ma1  ma2  ma3  ma4  ma5
## 0.00 0.00 0.00 0.00 0.56 0.05 0.13
```

```r
model1<-arima(x=X5,order=c(2,0,5),method = c("ML"),include.mean = F)
model1



pvalue<-function(model)
{
  round((1-pnorm(abs(model$coef)/sqrt(diag(model$var.coef))))*2,digits=2)

}
 pvalue(model1)
```

3. Au vu de ces résultats que doit faire notre statisticien? /1.5

L'hypothèse H_0 du test de student de nullité des coefficients MA(5) n'est pas rejetée au seuil 5%. Le statisticien doit donc proposer un modèle ARMA(2,4) sans la constante (option: include.mean = F) et réeffectuer les tests.

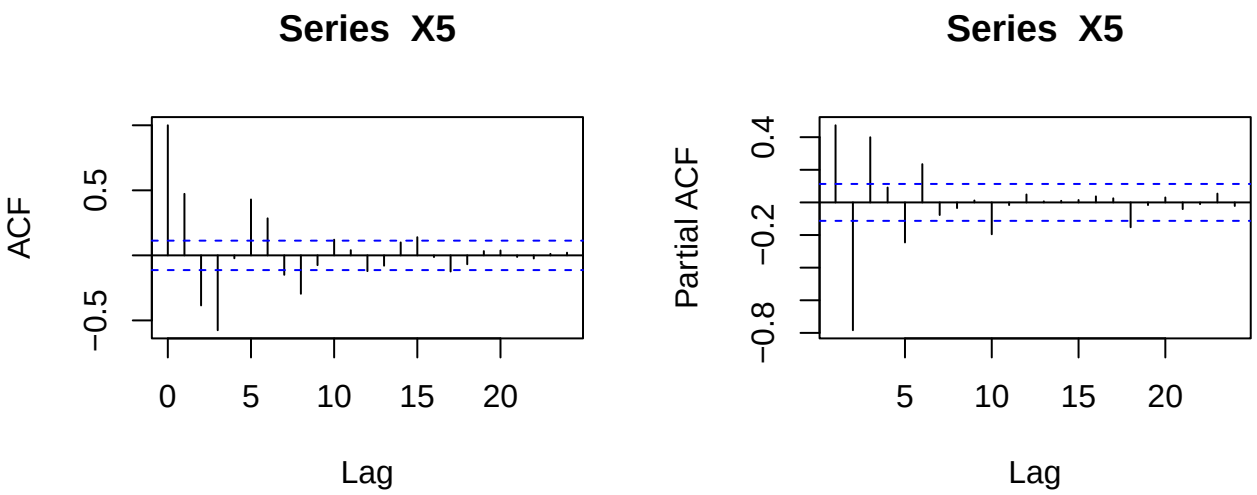
4. Il trace ensuite l'ACF et le PACF de $X5$. Au vu de ces graphiques, il décide d'effectuer une procédure de sélection de modèle via le critère AIC. Expliquer comment est calculé ce critère et quel est le but de cette approche. /1.5

En pratique, lorsque l'on doit ajuster un modèle AR, MA ou ARMA à des données réelles la première question qui se pose est celle du choix des ordres p et q du modèle ARMA (on considère que les AR et MA sont un cas particulier d'ARMA avec respectivement $q = 0$ et $p = 0$). Pour choisir ces ordres, nous regardons ensuite quand s'annule l'ACF (partie MA) et la PACF (partie AR), cela nous donne des ordres pmax et qmax plausibles. Ensuite nous cherchons en minimisant l'AIC à déterminer les ordres optimaux pour la prévision.

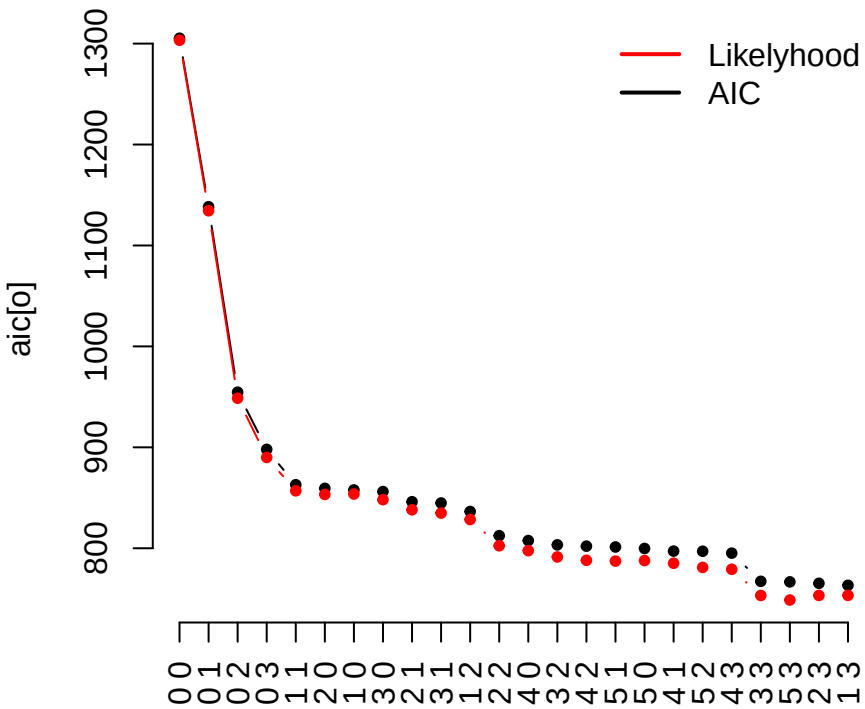
l'AIC (Akaike Information Criterion), se calcule ainsi:

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

où L est la vraisemblance, k est le nombre de paramètres dans le modèle donc $p+q$ ou $p+q+1$ si la constante est dans le modèle.



Commenter le graphique ci-dessous? Quelles conclusions en tirer?/1



Sur ce graphique l'axe des abscisses représente les ordres des modèles rangés par AIC décroissant. On constate que la vraisemblance décroit en fonction de la complexité du modèle ce qui est attendu et correspond à un phénomène de surapprentissage potentiel. L'ARMA(1,3) obtient l'AIC minimal dans un groupe de modèle possédant une partie MA d'ordre 3 et très proches en terme d'AIC.

5. Il s'intéresse ensuite à une série $X6_t$ et effectue un test de Box-Pierce d'ordre 20 sur $X6_t$. Il obtient le résultat suivant:

```
##  lags statistic df    p-value
##    20  35.71442 20 0.01661317
```

à quoi correspond "df"? Quelle est la conclusion de ce test?/1.5

Le test de Box-Pierce 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:

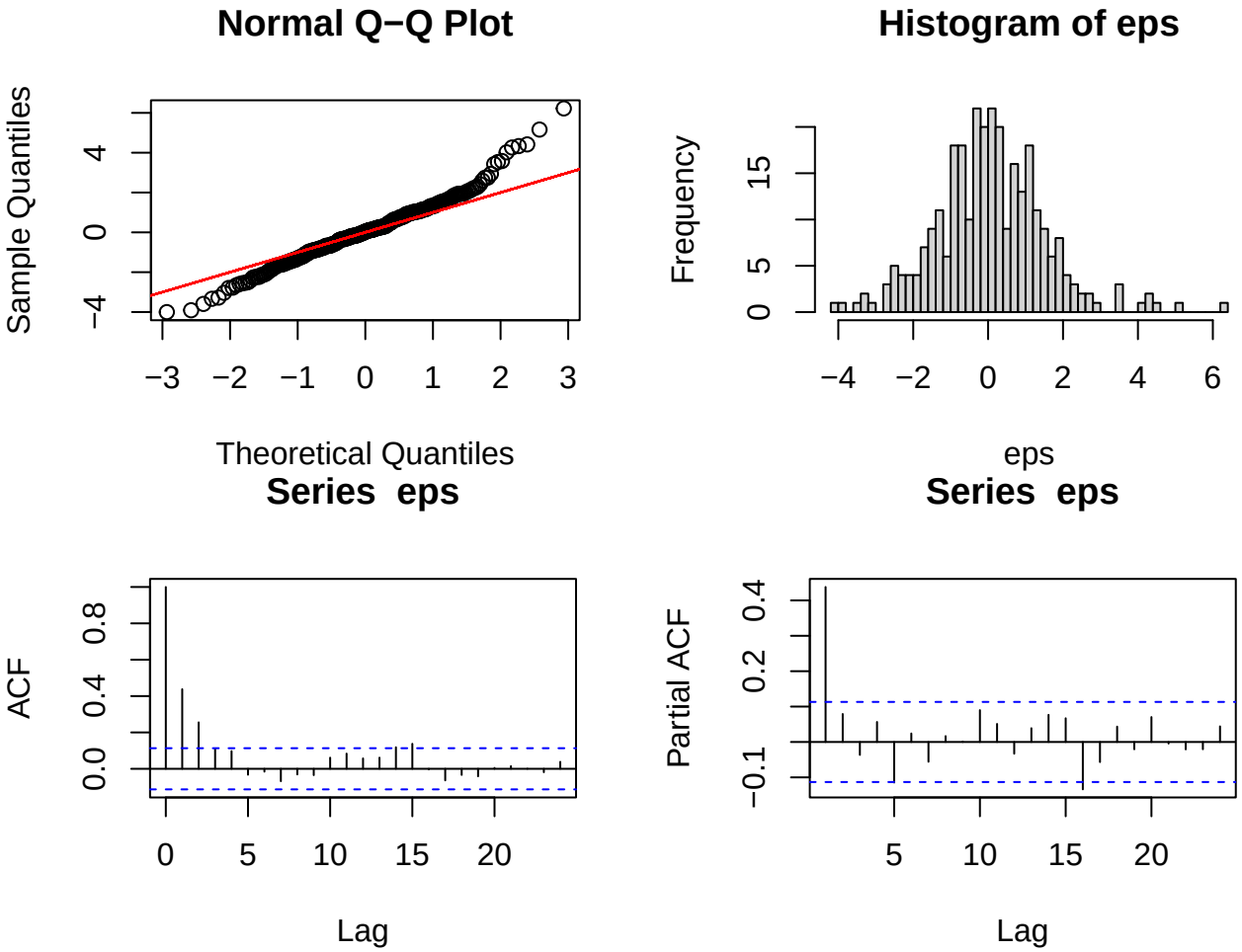
$$H_0(h) : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(h) = 0$$

Il se base sur la statistique de Box-Pierce:

$$Q_{BP}(h) = n \sum_{j=1}^{h} \widehat{\rho}_\varepsilon(j)^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(h-k)$$

lags correspond donc à h l'ordre du test, et df le nombre de degrés de libertés de la χ^2 associée.

Au seuil 5% l'hypothèse H_0 est rejetée. La modélisation de notre statisticien est donc à revoir.

5. Un de ces collègues cherche à diagnostiquer un modèle de prévision. Il étudie les résidus d'apprentissage de son modèle (un ARIMA Gaussien). Que doit il dire à son collègue? /2

L'ACF n'est pas celle d'un bruit blanc, la PACF indique qu'il s'agit probalement d'un AR(1). Il faut revoir l'ordre du modèle ARIMA considéré. le qqplot montre des écarts importants entre les quantiles observés et ceux d'une Gaussienne pour les quantiles faibles et élevés. L'hypothèse Gaussienne est probablement à revoir.