Examen du 22/03/2024

MAP-STA2 : Séries chronologiques

Yannig Goude [yannig.goude@edf.fr]
Bianca Marin-Moreno [bianca.marin-moreno@edf.fr]

Exercice 1 /2

Soit X un processus possédant une tendance polynomiale d'ordre k:

$$X_t = \sum_{i=0}^{k} a_i t^i + \varepsilon_t$$

ou $a_i \in \mathbb{R}$ et ε_t un bruit blanc faible.

1. Quelle opération doit on réaliser pour stationnariser ce processus (argumenter votre réponse)? /1

   différenciation d'ordre k

2. On considère ensuite le processus :

$$Y_t = X_T + S_t$$

   où S_t est une fonction périodique de période s. Comment peut on stationnariser ce processus? /1

   si s est la période, différencier $y_t - y_{t-s}$

Exercice 2 /4

Soit un processus $(X_t)_{t\in Z}$ moyenne mobile infinie tel que $X_t = \sum_{i=-\infty}^{+\infty} a_i \varepsilon_{t-i}$ ou ε_t est un bruit blanc de variance σ^2. La suite $(a_i)_{i\in Z}$ est une suite absolument sommable.

1. montrer que sa fonction d'autocovariance $\gamma(h)$ est bien définie et donner son expression. /1

$$\gamma(h) = \sigma^2 \sum_{i\in\mathbf{Z}} a_i a_{i-h}$$

   La suite $(a_i)_{i\in Z}$ est une suite absolument sommable donc la série $\gamma(h)$ est absolument sommable.

2. donner la définition de sa densité spectrale, montrer qu'elle existe et que c'est une fonction réelle. /1

   la densité spectrale du processus est la fonction définie par:

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

   cette fonction existe car la série $(\gamma(h)\mathrm{e}^{i\omega h})_{h\in\mathbf{Z}}$ est absolument sommable.

   D'autre part c'est une fonction réelle car γ est paire:

$$f(\omega) = \frac{1}{2\pi}[\gamma(0) + \sum_{h=1}^{+\infty} \gamma(h)(\mathrm{e}^{i\omega h} + \mathrm{e}^{-i\omega h})]$$

$$f(\omega) = \frac{\gamma(0)}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{+\infty} \gamma(h)\cos(\omega h) = \frac{1}{2\pi}\sum_{h=-\infty}^{+\infty} \gamma(h)\cos(\omega h)$$

3. montrer qu'il est équivalent de connaitre sa fonction d'autocovariance ou sa densité spectrale. /2

   on a

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

   et

$$= \int_{-\pi}^{\pi} \frac{1}{2\pi}\Sigma_{j=-\infty}^{+\infty}\gamma(j)\mathrm{e}^{i\omega j}\mathrm{e}^{-i\omega h}d\omega$$

   et

$$= \frac{1}{2\pi}\Sigma_{j=-\infty}^{+\infty}\gamma(j)\int_{-\pi}^{\pi} \mathrm{e}^{-i\omega(h-j)}$$

   or $\int_{-\pi}^{\pi} \mathrm{e}^{-i\omega(h-j)} = 2\pi$ si $j = h$ et 0 sinon. D'ou:

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

Exercice 3 /4

Soit le processus suivant, supposé stationnaire, avec ε_t un bruit blanc de variance σ^2:

$$y_t = \sum_{k=1}^{p} a_k y_{t-k} + \varepsilon_t$$

1. Quel est le nom de ce processus? Donner la formule de sa fonction d'autocovariance $\gamma(h)$ et celle de sa fonction d'autocorrélation $\rho(h)$. /1

C'est un AR(p).

$$\gamma(h) = cov(y_t, y_{t-h})$$

$$\rho(h) = \gamma(h)/\gamma(0)$$

et $\rho(h)$ est définie comme la solution d'une suite récurrente d'ordre p, donc racines non-nulles de l'équation:

$$\rho(h) + \sum_{j=1}^{p} \phi_j \rho(h-j) = 0, \quad h = 1, 2...$$

soit:

$$1 + \sum_{j=1}^{p} \phi_j \lambda^{-j} = 0$$

2. Vérifiez si les deux processus suivants sont stationnaires puis calculer les autocorrélations et autocorrélations partielles d'ordre 1, 2, et 3 pour ces processus. /3

$$y_t = 0.7y_{t-1} - 0.49y_{t-2} + \varepsilon_t$$

$$y_t = 0.8y_{t-1} + \varepsilon_t$$

Il faut vérifier que les racines des polynômes caractéristiques sont de modules >1.

C'est le cas ici et:

Pour l'AR(2):

pour un AR(2) $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$ et l'équation de récurrence vérifiée par la fonction d'autocorrelation $\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}$ nous permet d'obtenir: $\rho_1 = \phi_1 + \phi_2 \rho_1$ et $\rho_2 = \phi_1 \rho_1 + \phi_2$ soit $\rho_1 = \frac{\phi_1}{1-\phi_2}$ et $\rho_2 = \frac{\phi_1^2}{1-\phi_2} + \phi_2$ et $\rho_3 = \phi_1 \rho_2 + \phi_2 \rho_1$.

Cela donne:

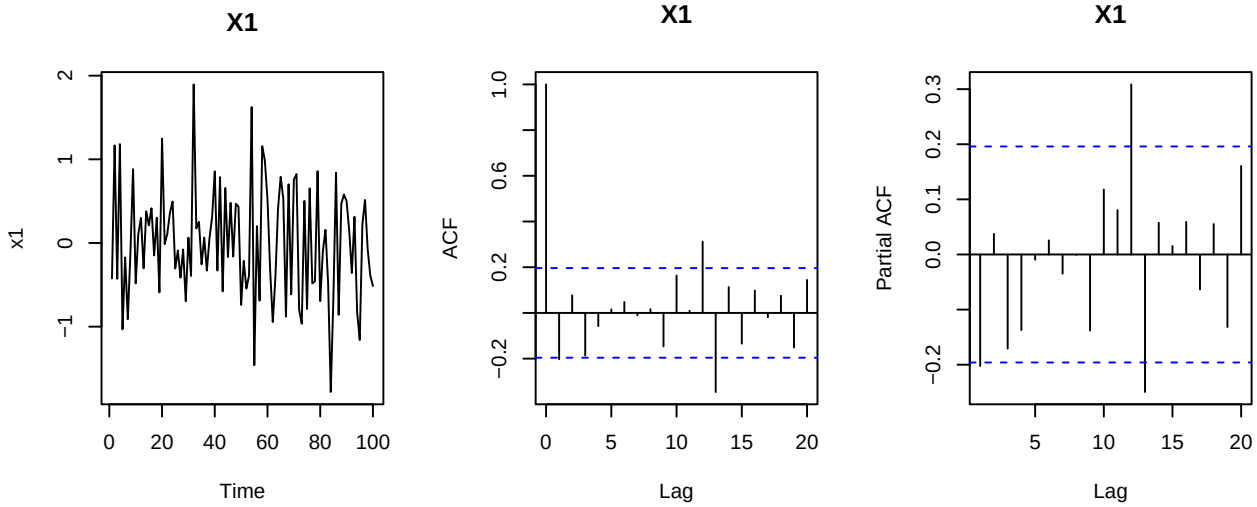
- $\rho(1) = 0.47$, $\rho(2) = -0.17$, $\rho(3) = -0.343$
- $r(1) = \rho(1) = 0.47$, $r(2) = -0.49$, $r(3) = 0$

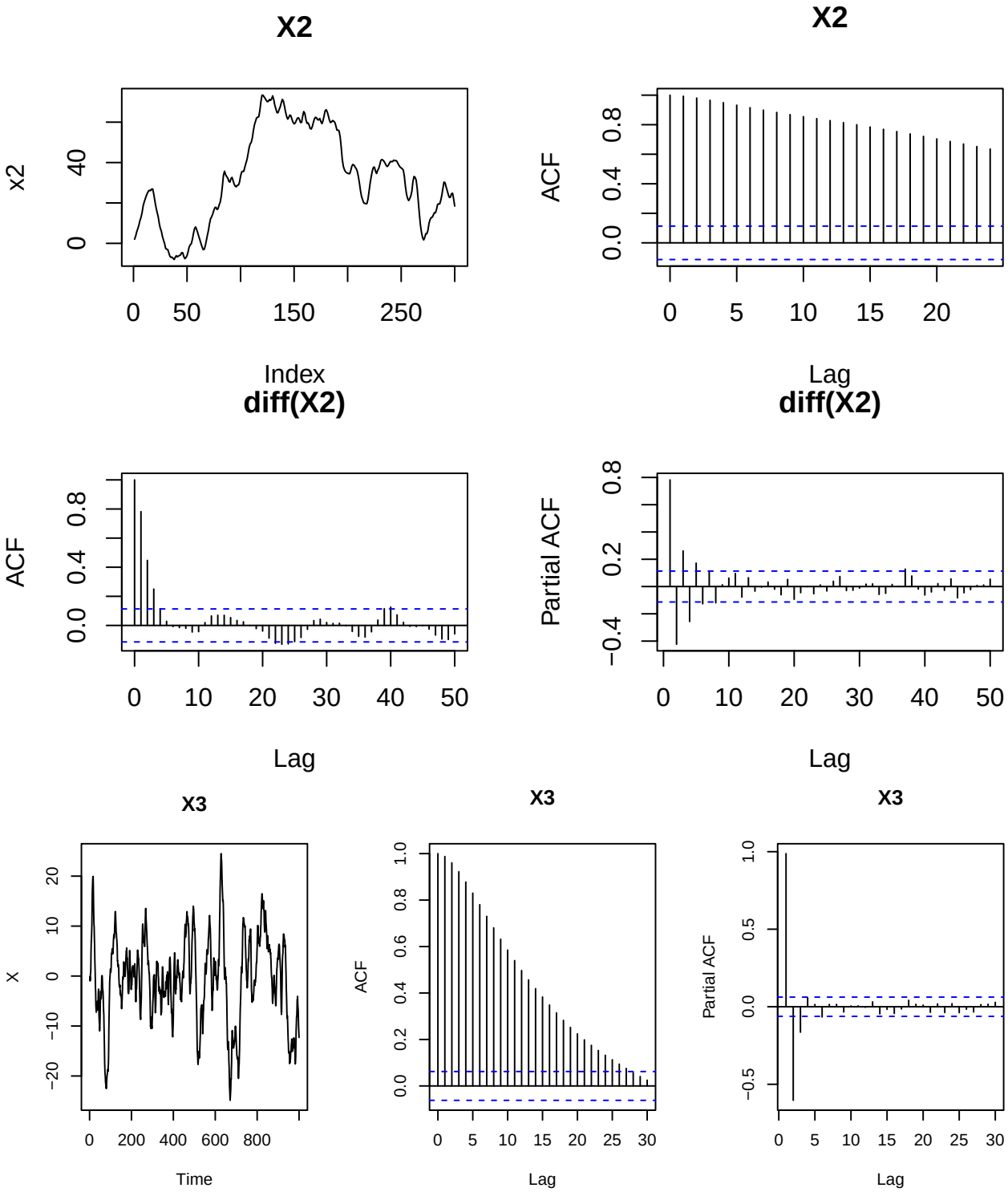
Pour l'AR(1):

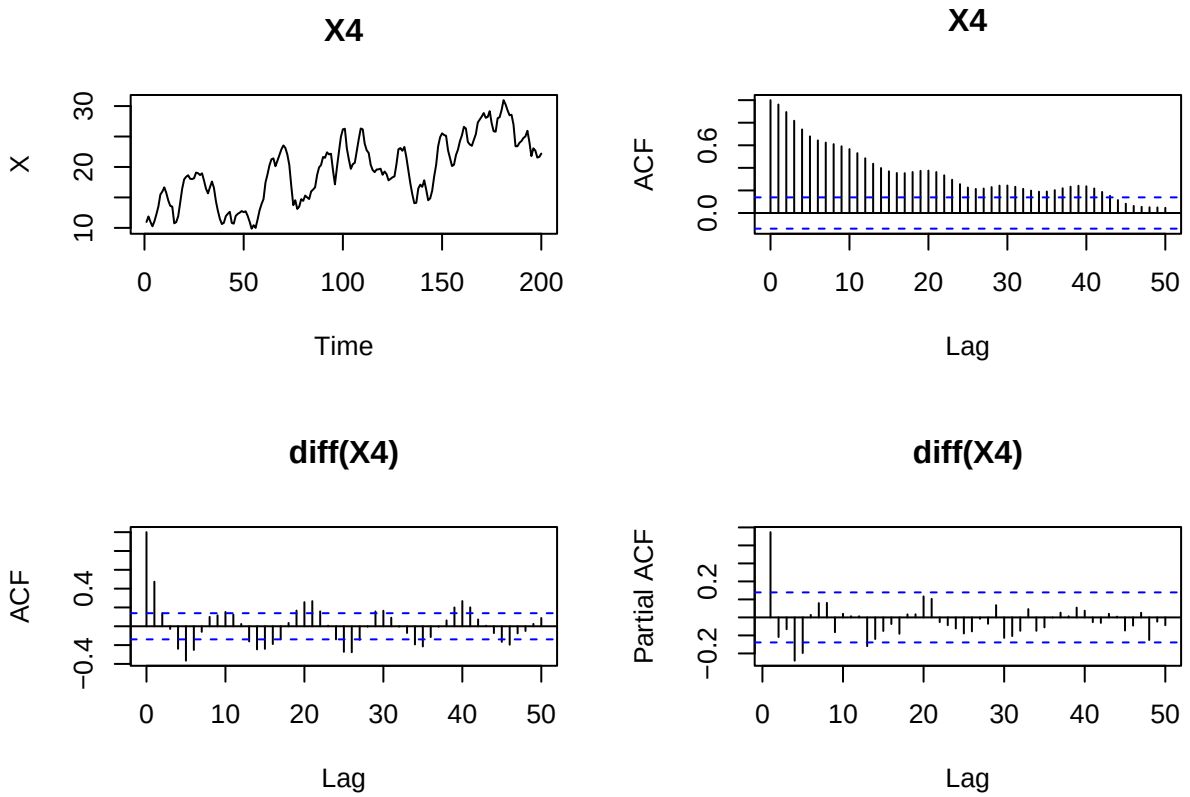
- $\gamma(1) = (0.8\sigma^2)/(1 - 0.8^2) = 2.22\sigma^2$, $\gamma(2) = (\sigma^2 0.8^2)/(1 - 0.8^2) = 1.77\sigma^2$, $\gamma(3) = (\sigma^2 0.8^3)/(1 - 0.8^2) = 1.42\sigma^2$
- $r(1) = 0.8$, $r(2) = r(3) = 0$

Exercice 4 /10

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




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



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

- Série 1: la série ne présente pas de comportement évident de type tendance ou saisonnalité. Sa variance semble également stable dans le temps. 2 choix sont possibles: l'ACF s'annule à partir du lag 13 inclu, nous sommes en présence d'un MA(13). Une modélisation plus parcimonieuse pourrait être proposée, en remarquant que l'ACF est nulle jusqu'à l'ordre 12, ce qui est modélisable par une saisonnalité d'ordre 12, donc un Seasonnal MA d'odre Q=1 pour la partie saisonnière serait à étudier.

- 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. Notre statisticien a donc bien fait de différencier la série. L'ACF du processus différencié à l'ordre 1 décroit exponentiellement vers 0, le processus différencié est donc bien stationnaire. Au vu de la PACF, il présente une composante AR d'ordre max $p = 4$. 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 $p \leq 4$, $q \leq 3$ et $d = 1$.

- Série 3: la série ne présente pas de comportement évident de type tendance ou saisonnalité. Sa variance semble également stable dans le temps. L'ACF décroit rapidement vers 0, on peut donc supposer que c'est un processus stationnaire. La PACF s'annule à partir du lag 4, nous sommes donc en présence d'un AR(3).

- Série 4: 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 pour la partie tendance, sur série de Fourier pour la partie saisonnière. On peut également différencier la série pour la corriger de sa tendance, on constate ici que l'acf de la série différenciée décroit plus rapidement vers 0 et comporte des oscillations caractéristiques d'une saisonnalité de période 10. Une autre approche aurait pu être d'effectuer une analyse spectrale.On peut donc ensuite soit la différencier avec un opérateur de différenciation de lag 10 ou effectuer une régression sur série de Fourier. 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.

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"? Comment est calculée la p-value du test de student de nullité de chacun des coefficients? A quoi correspond AIC? /1

```
##
## Call:
## arima(x = X5, order = c(3, 0, 4), include.mean = F, method = c("ML"))
##
## Coefficients:
##          ar1      ar2     ar3     ma1     ma2     ma3      ma4
##       0.6221  -0.9286  0.2128  0.2669  0.3883  0.4204  -0.2087
## s.e.  0.6855   0.3011  0.6282  0.6804  0.3470  0.3937   0.4591
##
## sigma^2 estimated as 5.381:  log likelihood = -678.86,  aic = 1373.73
```

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

```
##  ar1  ar2  ar3  ma1  ma2  ma3  ma4
## 0.36 0.00 0.73 0.69 0.26 0.29 0.65
```

La sortie s.e. correspond aux écarts types estimés des paramètres du modèle. Les p-values affichées correspondent au test de student de significativité des coefficients du modèle au seuil 0.05.

L'AIC est calculé ainsi: $\text{AIC}(\phi, \theta, \sigma^2) = -2\log(L(\theta, \phi, \sigma^2)) + 2k$ ou L est la vraisemblance du modèle, k le nombre de paramètre du modèle ($p+q+1$ si la constante est inclue, $p+q$ sinon). Ce critère permet d'effectuer un choix de modèle parmi une famille de modèles ARMA, c'est un compromis entre un bon ajustement aux données et la dimension du modèle.

3. Au vu de ces résultats que doit faire notre statisticien? /1

L'hypothèse H_0 du test de student de nullité des coefficients AR(3) n'est pas rejetée au seuil 5%. Le statisticien doit donc proposer un modèle ARMA(2,4) toujours sans la constante (option: include.mean = F) et réeffectuer les tests.

4. 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  35.67892 20 0.01677255
```

à quoi correspond "df"? Quelle est la conclusion de ce test? /2

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:

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

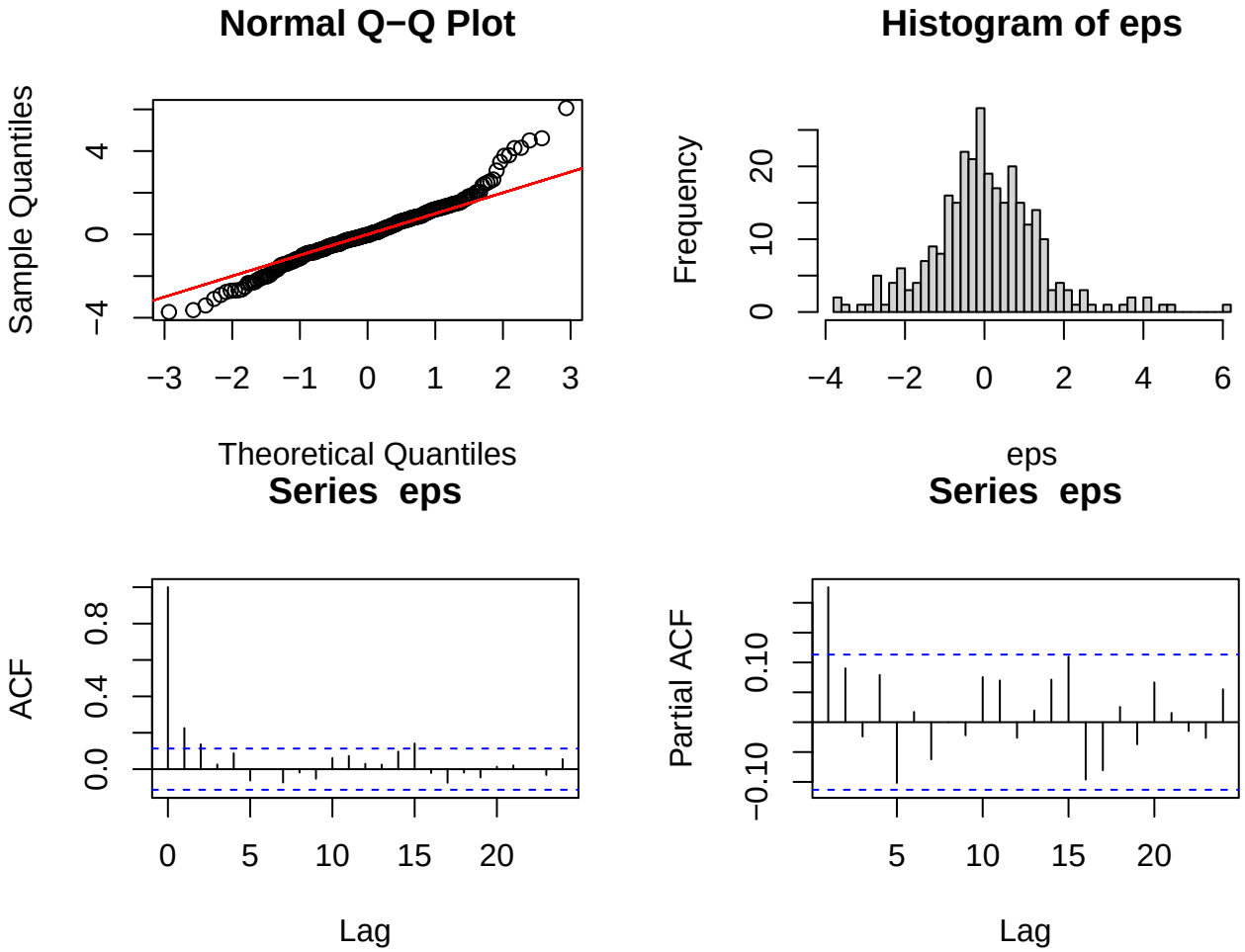
Il se base sur la statistique de Box-Pierce:

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

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

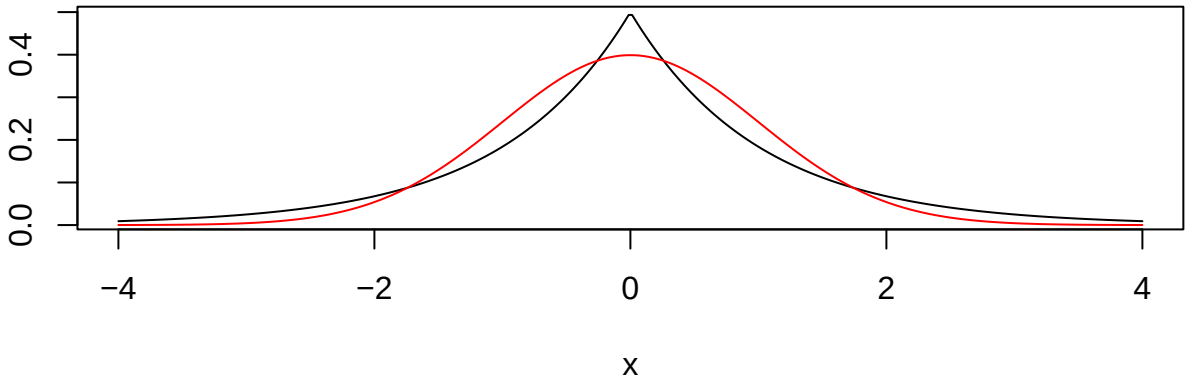
Au seuil 5% l'hypothèse H_0 est rejetée, il n'y a donc présence d'autocorrélations significatives jusqu'à l'ordre 10. La modélisation de notre statisticien est donc à revoir.

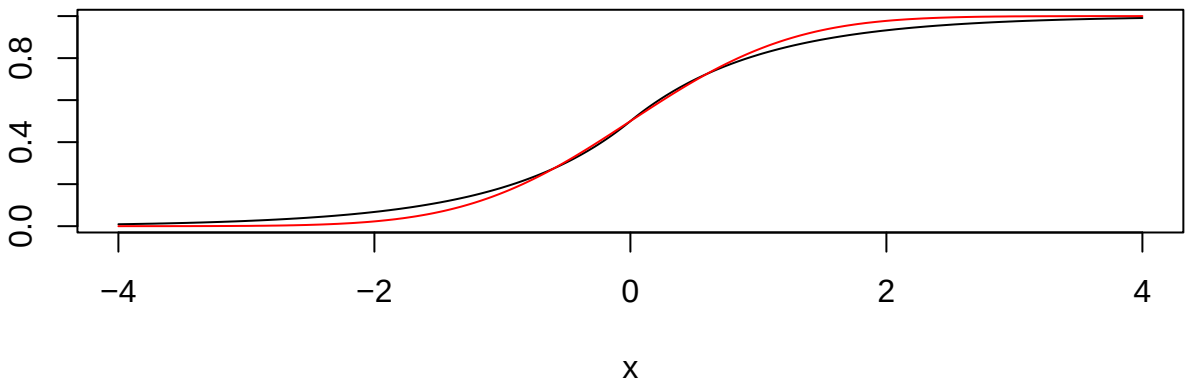
5. Un de ces collègues cherche à diagnostiquer un modèle de prévision qu'il a conçu. Il étudie les résidus d'apprentissage de son modèle (un ARIMA Gaussien). Que doit il dire à son collègue? /2



Au vu du qqplot, l'hypothèse Gaussienne est à remettre en cause. Il y a notamment des problèmes dans les queues de distribution.

L'histogramme montre que la loi des données est plus proche d'une loi de Laplace (en noir ci-dessous la densité (puis la fonction de distribution) d'une Laplace, et en rouge celle d'une Gaussienne).





D'autre part, on constate qu'il substiste une autocorrélation à l'ordre 1 significative dans les résidus. Il faut revoir le modèle et intégrer cette autocorrélation.