Séries temporelles

Chapitre 3: Méthodologie de Box et Jenkins

Mohamed Essaied Hamrita

Institut Supèrieur de Mathématiques Appliqués & Informatique - Kairouan

Décembre 2013

M2: Ingénierie Financière

Plan du chapitre

Introduction

La méthode de Box et Jenkins est un outils systématique qui permet :

Introduction

La méthode de Box et Jenkins est un outils systématique qui permet :

- de déterminer le meilleur modèle de type ARMA décrivant le processus stochastique d'une série observée ou d'une transformation stationnaire de celle-ci ;

Introduction

La méthode de Box et Jenkins est un outils systématique qui permet :

- de déterminer le meilleur modèle de type ARMA décrivant le processus stochastique d'une série observée ou d'une transformation stationnaire de celle-ci ;
- d'estimer ce modèle ;

Introduction

La méthode de Box et Jenkins est un outils systématique qui permet :

- de déterminer le meilleur modèle de type ARMA décrivant le processus stochastique d'une série observée ou d'une transformation stationnaire de celle-ci ;
- d'estimer ce modèle ;
- de l'utiliser pour extrapoler les valeurs de la série.

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Étape 1 : Transformation des données afin de stabiliser la variance (log, sqrt,...) et différenciation des données pour les stationariser.

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Étape 1 : Transformation des données afin de stabiliser la variance (log, sqrt,...) et différenciation des données pour les stationariser.

Étape 2 : Visualiser les ACF et les PACF empiriques pour identifier les paramètres p et q appropriés.

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Étape 1 : Transformation des données afin de stabiliser la variance (log, sqrt,...) et différenciation des données pour les stationariser.

Étape 2 : Visualiser les ACF et les PACF empiriques pour identifier les paramètres p et q appropriés.

Étape 3 : Estimation des paramètres du(des) modèle(s) sélectionné(s).

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Étape 1 : Transformation des données afin de stabiliser la variance (log, sqrt,...) et différenciation des données pour les stationariser.

Étape 2 : Visualiser les ACF et les PACF empiriques pour identifier les paramètres p et q appropriés.

Étape 3 : Estimation des paramètres du(des) modèle(s) sélectionné(s).

Étape 4 : Diagnostique et tests adéquation du modèle.

Les étapes de la méthodologie

La méthodologie de Box et Jenkins comporte essentiellement cinq étapes :

Étape 1 : Transformation des données afin de stabiliser la variance (log, sqrt,...) et différenciation des données pour les stationariser.

Étape 2 : Visualiser les ACF et les PACF empiriques pour identifier les paramètres p et q appropriés.

Étape 3 : Estimation des paramètres du(des) modèle(s) sélectionné(s).

Étape 4 : Diagnostique et tests adéquation du modèle.

Étape 5 : Prévision : La dernière étape consiste la prévision des valeurs futures à travers le modèle retenu.

Stationarisation de la série

À travers la représentation graphique de la série, on peut décider si la série en question est stationnaire ou non.
Dans la plus part des cas, une première différenciation rend la série stationnaire.
Il arrive parfois d'appliquer une transformation (log ou racine carré) à la série pour stabiliser la variance.

Exemples

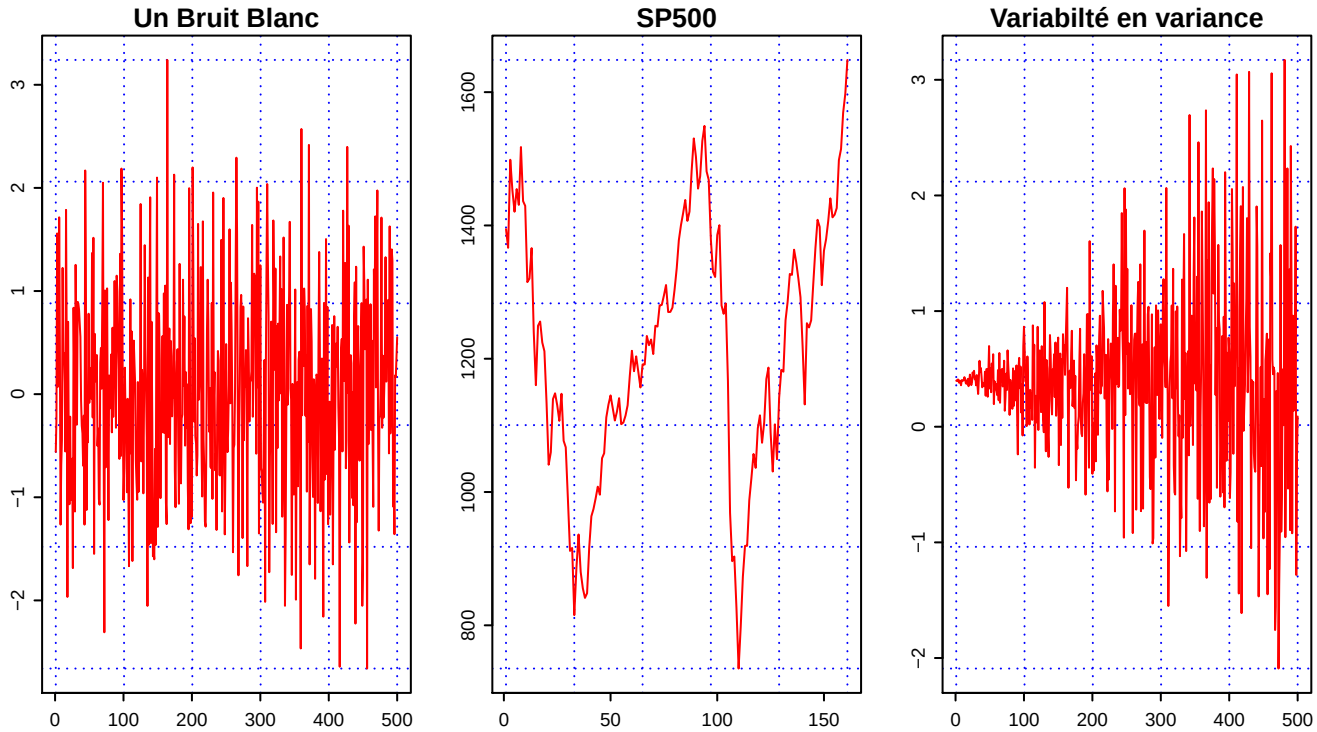


Figure: Exemples des séries temporelles

Exemples

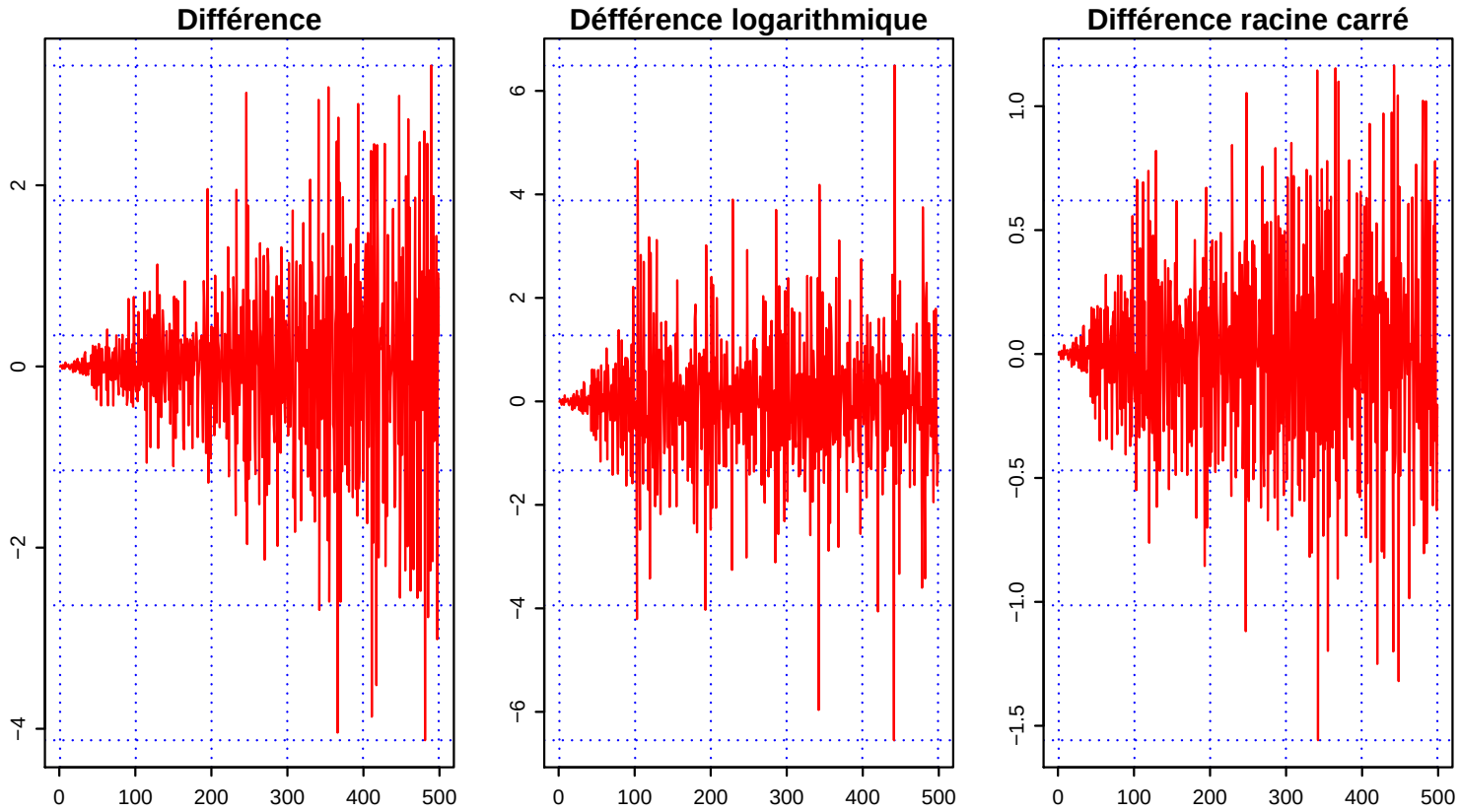


Figure: Stationnarisation de la troisième série

Identification

La méthode d'identification est essentiellement fondée sur l'analyse conjointe des auto-corrélations et des auto-corrélations partielles.
Nous supposons le processus stationnaire. La méthode s'í'appuie sur les résultats suivants :
Pour un processus AR(p) minimal :
les auto-corrélations sont à l'intérieur d'une enveloppe à décroissance géométrique ;
les auto-corrélations partielles sont identiquement nulles au delà de l'ordre p.
Pour un processus MA(q) minimal :
les auto-corrélations sont identiquement nulles au delà de l'ordre q.
les auto-corrélations partielles sont à l'intérieur d'une enveloppe à décroissance géométrique.

Identification

Identification

La fonction d'auto-corrélation empirique vérifie les propriétes suivantes : $E(\widehat{\rho}_k) = -1/n$ et $var(\widehat{\rho}_k) = 1/n$ où n est la taille de la série en question. $\widehat{\rho}_k$ est asymptotiquement normalement distribuées. D'où pour tester :

$$H_0 : \widehat{\rho}_k = 0$$

$$H_1 : \widehat{\rho} \neq 0$$

on compare la statistique $t_{\widehat{\rho}_k} = \widehat{\rho}_k / \widehat{\sigma}(\widehat{\rho}_k)$ à la valeur tabulée de la loi normale.
Si $|t_{\widehat{\rho}_k}| < z_{1-\alpha/2}$: on accepte H_0, c-à-d $\widehat{\rho}_k = 0$, où $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.
Graphiquement, les limites de l'intervalle de confiance des auto-corrélations seront représentées par $-1/n \pm z_{1-\alpha/2}/\sqrt{n}$.

Identification

Identification

De même, la fonction d'auto-corrélation partielle empirique est asymptotiquement normalement distribuée de variance $1/n$. D'où pour tester la nulleté des auto-corrélations partielles, on compare la statistique $t_{\widehat{\Phi}_{kk}} = \widehat{\phi}_{kk} \times \sqrt{n}$ à la valeur tabulée de la loi normale.
Si $|t_{\widehat{\Phi}_{kk}}| < z_{1-\alpha/2}$: on accepte H_0, c-à-d $\widehat{\phi}_{kk} = 0$.

Estimation I

Il existe plusieurs méthodes d'estimation des paramètres (MCO, Yule-Walker, MV).
La méthode la plus repandue, est celle du maximum de vraisemblance (MV). Cette méthode repose sur l'hypothèse de normalité des résidus.($\epsilon \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$). Le logarithme de vraisemblance d'un processus ARMA(p, q)est donnée par :

$$\ln L_n = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma_\epsilon^2) - \frac{1}{2}\ln\left(det(Z'Z)\right) - \frac{S(\phi,\theta)}{2\sigma_\epsilon^2}$$

Z est une matrice de taille $(p+q+n, p+q)$ qui dépends des paramètres ϕ_i $(i = 1, \ldots, p)$ et θ_j $(j = 1, \ldots, q)$.

$$S(\phi,\theta) = \sum_{t=0}^{n} \left(E(\epsilon_t)\right)^2.$$

Estimation II

En maximisant cette fonction, on déduit les estimateurs $\widehat{\phi}_i$, $\widehat{\theta}_j$ et $\widehat{\sigma}^2_\epsilon$.

Remarque :

Pour tester si on inclut le terme constant lors de l'estimation ou non, i.e $H_0 : c = 0$ vs $H_0 : c \neq 0$, on procède comme suit :

Si $|t| = \sqrt{n}\left|\frac{\overline{X}}{\widehat{\sigma}\epsilon}\right| < Z_{1-\alpha/2}$, on accepte H_0.

Validation

Souvent, il n'est pas facile de déterminer un modèle unique qui représente le processus générateur de données, et il n'est pas rare pour estimer plusieurs modèles à l'étape initiale. Le modèle qui est finalement choisi est celui considéré comme le meilleur basé sur un ensemble de critères de contrôle et de diagnostic. Ces critères comprennent :

Validation

Souvent, il n'est pas facile de déterminer un modèle unique qui représente le processus générateur de données, et il n'est pas rare pour estimer plusieurs modèles à l'étape initiale. Le modèle qui est finalement choisi est celui considéré comme le meilleur basé sur un ensemble de critères de contrôle et de diagnostic. Ces critères comprennent :

- Les t-tests de significativité des paramètres estimés.

Validation

Souvent, il n'est pas facile de déterminer un modèle unique qui représente le processus générateur de données, et il n'est pas rare pour estimer plusieurs modèles à l'étape initiale. Le modèle qui est finalement choisi est celui considéré comme le meilleur basé sur un ensemble de critères de contrôle et de diagnostic. Ces critères comprennent :

- Les t-tests de significativité des paramètres estimés.
- Analyse des résidus (normalité, absence d'auto-corrélation, homoscédasticité)

Validation

Souvent, il n'est pas facile de déterminer un modèle unique qui représente le processus générateur de données, et il n'est pas rare pour estimer plusieurs modèles à l'étape initiale. Le modèle qui est finalement choisi est celui considéré comme le meilleur basé sur un ensemble de critères de contrôle et de diagnostic. Ces critères comprennent :

- Les t-tests de significativité des paramètres estimés.
- Analyse des résidus (normalité, absence d'auto-corrélation, homoscédasticité)
- Les critères d'informations.

Les tests de significativité

Il s'agit de tester $H_0 : p' = p - 1$ et $q' = q$ ou $H_0 : p' = p$ et $q' = q - 1$.
En d'autres termes, on test : $H_0 : ARMA(p-1, q)$ vs $H_1 : ARMA(p, q)$.
Il s'agit d'un test de significativité sur les coefficients ϕ_p et θ_q.
On compare les statistiques : $t_{\widehat{\phi}_p} = \widehat{\phi}_p / \widehat{\sigma}(\widehat{\phi}_p)$ et $t_{\widehat{\theta}_q} = \widehat{\theta}_q / \widehat{\sigma}(\widehat{\theta}_q)$ à la valeur tabulée $t_{1-\alpha/2}$.
Si $\left| t_{\widehat{\phi}_p} \right| < t_{1-\alpha/2}$ on accepte H_0 et on retient $ARMA(p, q)$.

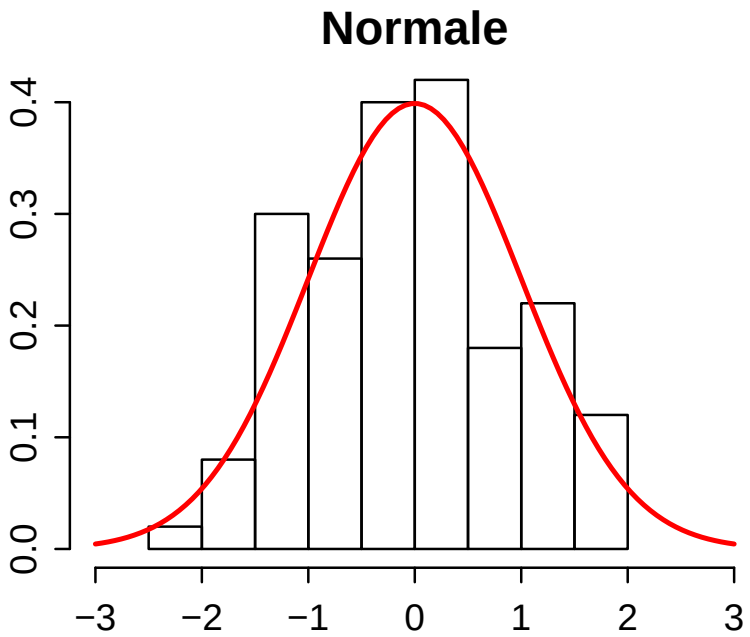
Analyse des résidus

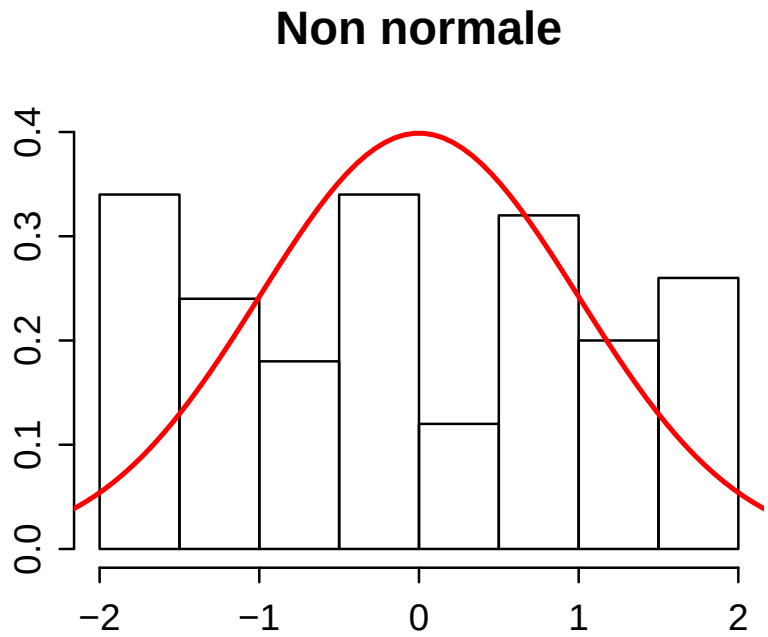
Les résidus issues d'une estimation doivent vérifier quelques propriétés statistiques :

– La normalité.
– Absence d'auto-corrélation.
– Homoscédasticité.

La normalité

La normalité peut être tester graphiquement, soit en représentant l'histogramme des résidus, soit par le graphe quantile-quantile (qq-plot)..





La normalité

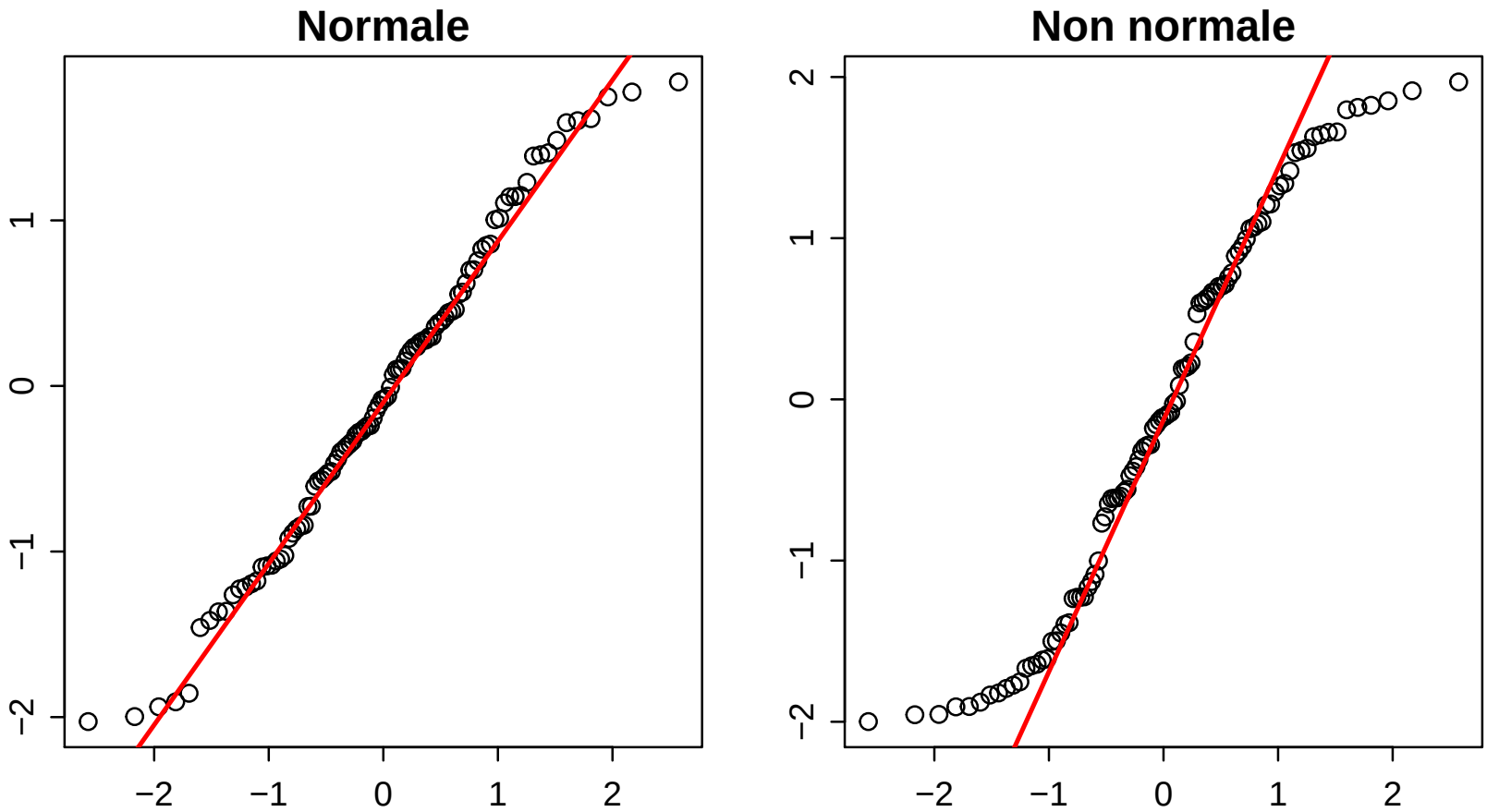


Figure: Graphe quantile-quantile

Il existe plusieurs tests d'absence d'auto-corrélation. Ces tests se regroupent en deux groupe. Un groupe regroupe les tests paramétriques et un groupe constitué des tests non paramétriques.

Il existe plusieurs tests d'absence d'auto-corrélation. Ces tests se regroupent en deux groupe. Un groupe regroupe les tests paramétriques et un groupe constitué des tests non paramétriques.

– Les tests paramétriques les plus connus dans la littérature étant ceux de **Box et Pierce (1970)** et de **Ljung et Box (1978)**.

– Les tests non paramétriques les plus repandus sont : le test de retournement et le test de monotonie.

Les tests paramétriques

Le test de Box et Pierce : Ce test est connu sous le nom du test de portmanteau. Il s'agit de tester :

$$H_0 : \widehat{\rho}_1(\widehat{\epsilon}_t) = \widehat{\rho}_2(\widehat{\epsilon}_t) = \ldots = \widehat{\rho}_k(\widehat{\epsilon}_t) = 0$$

La statistique utilisée est :

$$BP(k) = n \sum_{k=1}^{k} \widehat{\rho}_k^2(\widehat{\epsilon}_t)$$

Sous l'hypothèse nulle, $BP(k) \sim \chi^2(k-q-p)$
Si $BP(k) < \chi^2_{1-\alpha/2}(k-p-q)$, on accepte H_0

Les tests paramétriques

Le test Ljung et Box : Ce test est préféré au test de portmanteau.
La statistique utilisée est :

$$LB(k) = n(n+2)\sum_{k=1}^{k}\frac{\widehat{\rho}_k^2(\widehat{\epsilon}_t)}{n-k}$$

Sous l'hypothèse nulle, $LB(k) \sim \chi^2(k-q-p)$
Si $BP(k) < \chi^2_{1-\alpha/2}(k-p-q)$, on accepte H_0

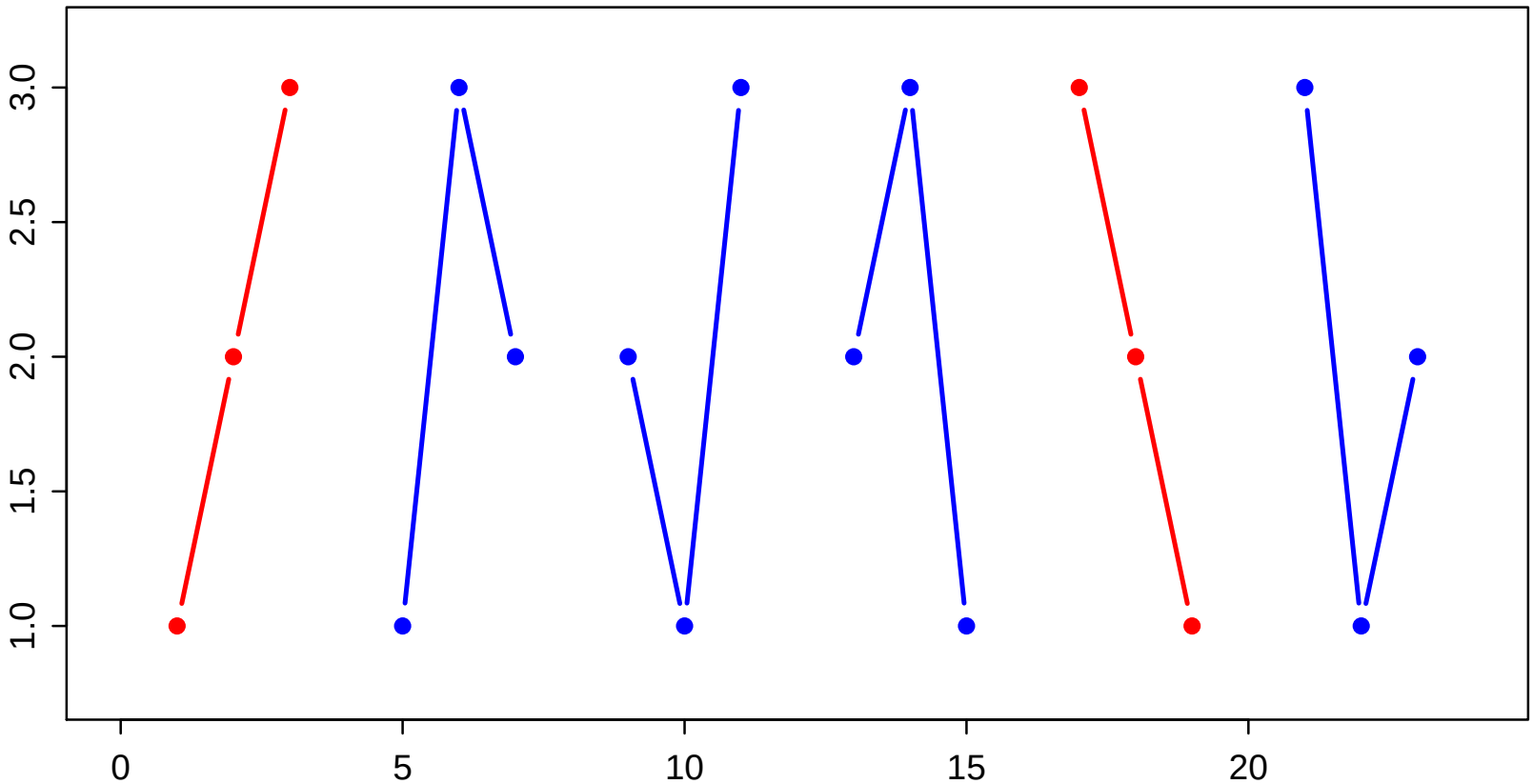
Les tests non paramétriques

Le test de retournement : Il s'agit de tester H_0 : les $\widehat{\epsilon}_t$ sont des v.a iid.
Une série temporelle $\{x_t\}$ possède un point de retournement à l'instant t si l'une des deux situations suivantes est vérifiée :
$x_t > x_{t-1}$ et $x_t > x_{t+1}$ ou $x_t < x_{t-1}$ et $x_t < x_{t+1}$.
Si X_{t-1}, X_t et X_{t+1} sont des v.a iid de distibution continue, les six événements suivants sont également susceptibles :

- $X_{t-1} < X_t < X_{t+1}$ pas de point de retournement.
- $X_{t-1} < X_{t+1} < X_t$ un point de retournement.
- $X_t < X_{t-1} < X_{t+1}$ un point de retournement.
- $X_t < X_{t+1} < X_{t-1}$ un point de retournement.
- $X_{t+1} < X_{t-1} < X_t$ un point de retournement.
- $X_{t+1} < X_t < X_{t-1}$ pas de point de retournement.

Le test de retournement

Graphiquement, les six cas sont représentés comme suit :



Le test de retournement

Soit T le nombre des points de retournement d'une série temporelle $\{x_t\}$ de taille n. Sous H_0, la v.a $T \sim N(\mu_T, \sigma_T^2)$ avec :

$$\mu_T = \frac{2(n-2)}{3} \quad \text{et } \sigma_T^2 = \frac{16n-29)}{90}$$

Si $|T - \mu_T|/\sigma_T < z_{1-\alpha/2}$, on accepte l'hypothèse nulle où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale standard.

Le test de rang

Étant donnée une série temporelle de taille n,on considère l'ensemble P des couples (r, s) tels que : $x_r > x_s$ et $r > s$. Si x_t est iid, alors P est approximativement suit la loi normale de paramètres :

$$\mu_P = \frac{n(n-1)}{4} \quad \text{et } \sigma_P^2 = \frac{n(n-1)(5n+2)}{72}$$

Si $|P - \mu_P| / \sigma_P < z_{1-\alpha/2}$, on accepte l'hypothèse nulle où $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale standard.

Tests d'homoscédasticité

Les tests d'homoscédasticité les plus connus dans la littérature sont au nombre de deux :

– Le test de White (1980).

– Le test de ARCH de Engel (1982).

Test de White

Le test de White est fondé sur l'existence d'une relation entre le carré des résidus et une ou plusieurs variables explicatives en niveau et au carré :

$$\widehat{\epsilon}_t^2 = a_0 + a_1 X_{t-1} + b_1 X_{t-1}^2 + \ldots + a_p X_{t-p} + b_p X_{t-p}^2 + \nu_t, \quad \nu_t \sim BB$$

Si au moins un des coefficients de regression est significatif, on rejette l'hypothèse nulle d'homoscédasticité.
La statistique utilisée est $LM(p) = nR^2$ où n est le nombre d'observation et R^2 est le coefficient de détermination associé à la regression. Sous H_0, cette statistique suit une loi $\chi^2(2p)$.
Si $LM < \chi^2_{1-\alpha/2}(2p)$, on accepte l'hypoyhèse nulle.

Test ARCH d'Engel

Ce test repose sur la regression suivante :

$$\widehat{\epsilon}_t^2 = 0 + \sum_{i=1}^{k} \alpha_i \widehat{\epsilon}_{t-i}^2$$

On utilise la même statistique que précédement, c-à-d
$LM(k) = nR^2 \overset{H_0}{\sim} \chi^2(k)$.
Si $LM(k) < \chi^2_{1-\alpha/2}(k)$, on accepte l'hypoyhèse nulle.

Les critères d'information

Les critères les plus fréquement employés sont :

- Le critère dinformation d'Akaike (1969) :

$$AIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \frac{2(p+q)}{n}$$

Les critères d'information

Les critères les plus fréquement employés sont :

- Le critère dinformation d'Akaike (1969) :

$$AIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \frac{2(p+q)}{n}$$

- Le critère de Schawrz (1978) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + (p+q)\frac{\ln n}{n}$$

Les critères d'information

Les critères les plus fréquement employés sont :

- Le critère dinformation d'Akaike (1969) :

$$AIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \frac{2(p+q)}{n}$$

- Le critère de Schawrz (1978) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + (p+q)\frac{\ln n}{n}$$

- Le critère de Hannan-Quinn (1979) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \alpha(p+q)\frac{\ln n}{n} \quad \text{où } \alpha > 2.$$

Les critères d'information

Les critères les plus fréquement employés sont :

- Le critère dinformation d'Akaike (1969) :

$$AIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \frac{2(p+q)}{n}$$

- Le critère de Schawrz (1978) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + (p+q)\frac{\ln n}{n}$$

- Le critère de Hannan-Quinn (1979) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \alpha(p+q)\frac{\ln n}{n} \quad \text{où } \alpha > 2.$$

Les critères d'information

Les critères les plus fréquement employés sont :

- Le critère dinformation d'Akaike (1969) :

$$AIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \frac{2(p+q)}{n}$$

- Le critère de Schawrz (1978) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + (p+q)\frac{\ln n}{n}$$

- Le critère de Hannan-Quinn (1979) :

$$SIC = \ln\left(\widehat{\sigma}_{\epsilon}^{2}\right) + \alpha(p+q)\frac{\ln n}{n} \quad \text{où } \alpha > 2.$$

Un modèle choisi par critère d'information est un modèle qui minimise un des précédents critères.

Prévision

Considèrons un processus ARMA(p, q) et on note $\widehat{X}_{t+k}$ la prévision de X fait en t à l'horizon k.
$\widehat{X}_{t+k} = E[X_{t+k}|T_t]$, I_t est l'information disponible à l'instant t.
Formule déduite de la forme ARMA :

$$\widehat{X}_{t+k} = \sum_{j=1}^{p+q} \phi_j X_{t+k-j} + \sum_{j=1}^{q} \widehat{\epsilon}_{t+k-j}$$