Chapter 1
Time Series

Time series are sequences of data points indexed in discrete time. This chapter reviews the Moving Average (MA), Autoregressive (AR), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) time series models. The stationarity properties of time series are considered via their autocovariance graph and unit root testing. Several examples are considered, including the fitting of time series models to financial data in R. We conclude with an application to a pair trading algorithm on a financial market, which relies the Dickey-Fuller stationarity test.

1.1 Autoregressive Moving Average (ARMA)

White noise

A white noise is a sequence $(Z_n)_{n\in\mathbb{Z}}$ of independent, centered and identically distributed random variables with unit variance, *i.e.*

$$\mathbb{E}[Z_n] = 0, \quad \text{and} \quad \mathrm{Cov}(Z_n, Z_m) = \mathbb{1}_{\{n=m\}}, \qquad n, m \in \mathbb{Z}.$$

```
Zn<-rnorm(100,0,1)
Zn
```

Moving Average (MA) Model

Definition 1.1. *In the $MA(q)$ model of order $q \geqslant 1$, the current state X_n of the system is expressed as the linear combination*

$$\begin{aligned} X_n &:= Z_n + \beta_1 Z_{n-1} + \cdots + \beta_q Z_{n-q} \\ &= Z_n + \sum_{k=1}^{q} \beta_k Z_{n-k}, \qquad n \geqslant 0, \end{aligned} \tag{1.1}$$

of the q previous values $Z_{n-1}, \ldots, Z_{n-q}$. Here, $\beta_1, \ldots, \beta_q$ is a sequence of deterministic coefficients.

We consider the "lag operator" or "backward time shift operator" L defined as

$$LZ_n := Z_{n-1}, \qquad n \geqslant 1. \tag{1.2}$$

```
library(zoo)
N=5
Zn<-zoo(rnorm(N,0,1))
Zn
lag(Zn,-1, na.pad = TRUE)
```

Using the lag operator L, we can rewrite (1.1) as

$$\begin{aligned} X_n &= Z_n + \beta_1 LZ_n + \cdots + \beta_q L^q Z_n \\ &= Z_n + \sum_{k=1}^{q} \beta_k L^k Z_n \\ &= Z_n + \psi(L) Z_n, \qquad n \geqslant q, \end{aligned}$$

where

$$\psi(L) = \beta_1 L + \cdots + \beta_q L^q = \sum_{k=1}^{q} \beta_k L^k Z_n$$

is the *moving average operator* given by the function

$$\psi(x) = \beta_1 x + \cdots + \beta_q x^q = \sum_{k=1}^{q} \beta_k x^k.$$

Example: generating MA(2) samples in R

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

```
n=41
ma.sim<-arima.sim(model=list(ma=c(-.7,.1)),n.start=100,n)
x=seq(100,100+n-1)
plot(x,ma.sim,pch=19, ylab="X", xlab="n",main = 'MA(2) samples',col='blue')
lines(x,ma.sim, main = 'MA(2) samples',col='blue')
```

The ARIMA command uses a parameter "n.start", here taken equal to 100, which creates a "burn-in" initial time interval which ensures sufficient randomness in the behavior of X_n.

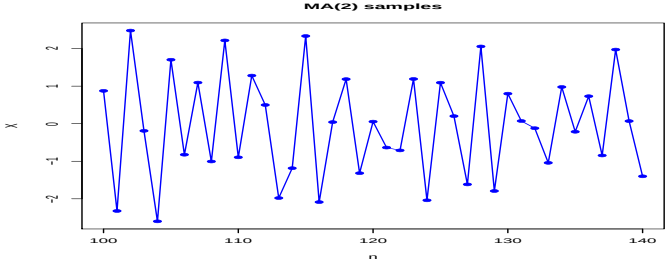


Fig. 1.1: MA(2) samples.

Autoregressive (AR) Model

In the simplest AR(1) model, the current state X_n of the system is expressed as

$$X_n := Z_n + \alpha_1 X_{n-1}, \qquad n \geqslant p, \tag{1.3}$$

Definition 1.2. *In the AR(p) model, $p \geqslant 1$, the state X_n of the system is expressed as the linear feedback combination*

$$\begin{aligned} X_n &:= Z_n + \alpha_1 X_{n-1} + \cdots + \alpha_p X_{n-p} \\ &= Z_n + \sum_{k=1}^{p} \alpha_k X_{n-k}, \qquad n \geqslant p, \end{aligned} \tag{1.4}$$

of the p previous values $X_{n-1}, \ldots, X_{n-p}$ of the process, where $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$ is a sequence of deterministic coefficients.

Using again the lag operator L defined in (1.2) we can rewrite (1.4) as

$$\begin{aligned} X_n &= Z_n + \alpha_1 L X_n + \cdots + \alpha_p L^p X_n \\ &= Z_n + \sum_{k=1}^{p} \alpha_k L^k X_n \\ &= Z_n + \phi(L) X_n, \qquad n \geqslant p, \end{aligned}$$

where

$$\phi(L) = \alpha_1 L + \cdots + \alpha_p L^p = \sum_{k=1}^{p} \alpha_k L^k$$

is the operator given by the function

$$\phi(x) = \alpha_1 x + \cdots + \alpha_p x^p = \sum_{k=1}^{p} \alpha_k x^k.$$

Proposition 1.3. *The equation*

$$X_n := Z_n + \alpha_1 X_{n-1}, \qquad n \geqslant 1, \tag{1.5}$$

defining the AR(1) process $(X_n)_{n\geqslant 0}$ *can be solved recursively in the following cases:*

a) When $|\alpha_1| < 1$, (1.5) *admits the converging* causal *moving average solution*

$$X_n = \sum_{k\geqslant 0} \alpha_1^k Z_{n-k}, \qquad n \geqslant 0, \tag{1.6}$$

with

$$\mathrm{Var}[X_n] = \sum_{k\geqslant 0} |\alpha_1|^{2k} = \frac{1}{1-|\alpha_1|^2}, \qquad n \geqslant 0. \tag{1.7}$$

b) When $|\alpha_1| > 1$, (1.5) *admits the converging* non-causal *moving average solution*

$$X_n = -\sum_{k\geqslant 1} \frac{1}{\alpha_1^k} Z_{n+k}, \qquad n \geqslant 0, \tag{1.8}$$

with

$$\mathrm{Var}[X_n] = \sum_{k\geqslant 1} |\alpha_1|^{-2k} = \frac{1}{|\alpha_1|^2-1}, \qquad n \geqslant 0. \tag{1.9}$$

No such converging solution exists when $|\alpha_1| = 1$.

Proof. *a)* When $|\alpha_1| < 1$ we may write, using backward induction,

$$\begin{aligned}
X_n &= Z_n + \alpha_1 X_{n-1} \\
&= Z_n + \alpha_1 (Z_{n-1} + \alpha_1 X_{n-2}) \\
&= Z_n + \alpha_1 (Z_{n-1} + \alpha_1 (Z_{n-2} + \alpha_1 X_{n-3})) \\
&= Z_n + \alpha_1 (Z_{n-1} + \alpha_1 (Z_{n-2} + \alpha_1 (Z_{n-3} + \alpha_1 X_{n-4}))) \\
&= Z_n + \alpha_1 Z_{n-1} + \alpha_1^2 Z_{n-2} + \alpha_1^3 Z_{n-3} + \alpha_1^4 X_{n-4} \\
&= \cdots \\
&= \sum_{k\geqslant 0} \alpha_1^k Z_{n-k},
\end{aligned}$$

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

which converges when the solution $z = 1/\alpha_1$ of the equation $\phi(z) = \alpha_1 z = 1$ satisfies $|\alpha_1| < 1$, *i.e.* $|z| > 1$.

b) When $|\alpha_1| > 1$, we write

$$X_n = -\alpha_1^{-1} Z_{n+1} + \alpha_1^{-1} X_{n+1}, \qquad n \geqslant 0,$$

which can be solved by forward induction as

$$\begin{aligned} X_n &= -\alpha_1^{-1} Z_{n+1} + \alpha_1^{-1} X_{n+1} \\ &= -\alpha_1^{-1} Z_{n+1} + \alpha_1^{-1}(-\alpha_1^{-1} Z_{n+2} + \alpha_1^{-1} X_{n+2}) \\ &= -\alpha_1^{-1} Z_{n+1} + \alpha_1^{-1}(-\alpha_1^{-1} Z_{n+2} + \alpha_1^{-1}(-\alpha_1^{-1} Z_{n+3} + \alpha_1^{-1} X_{n+3})) \\ &= -\alpha_1^{-1} Z_{n+1} - \alpha_1^{-2} Z_{n+2} - \alpha_1^{-3} Z_{n+3} + \alpha_1^{-4} X_{n+3} \\ &= \cdots \\ &= -\sum_{k \geqslant 1} \frac{1}{\alpha_1^k} Z_{n+k}, \end{aligned}$$

when the solution $z = 1/\alpha_1$ of the equation $\phi(z) = \alpha_1 z = 1$ satisfies $|z| < 1$. □

Example: generating AR(2) samples in R

Consider the AR(2) times series

$$X_n := Z_n + 0.9 \times X_{n-1} - 0.2 \times X_{n-2}. \tag{1.10}$$

```
n=4100
ar.sim<-arima.sim(model=list(ar=c(.9,-.2)),n.start=100,n)
x=seq(100,100+n-1)
plot(x,ar.sim,pch=19, ylab="X", xlab="n", main = 'AR$(2)$ samples',col='blue')
lines(x,ar.sim, main = 'AR$(2)$ samples',col='blue')
```

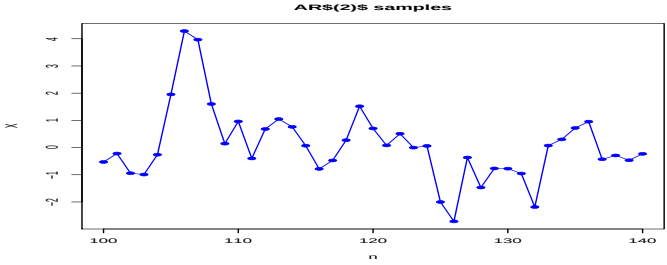


Fig. 1.2: AR(2) samples.

Autoregressive Moving Average (ARMA) Model

Definition 1.4. *In the ARMA(p,q) model with orders $p \geqslant 1$ and $q \geqslant 1$, the current state X_n of the system is expressed as the linear combination*

$$\begin{aligned} X_n &:= Z_n + \alpha_1 X_{n-1} + \cdots + \alpha_p X_{n-p} + \beta_1 Z_{n-1} + \cdots + \beta_p Z_{n-q} \\ &= Z_n + \sum_{k=1}^{p} \alpha_k X_{n-k} + \sum_{k=1}^{q} \beta_k Z_{n-k}, \qquad n \geqslant \max(p,q), \end{aligned} \tag{1.11}$$

of the p previous values $X_{n-1}, \ldots, X_{n-p}$ and $Z_{n-1}, \ldots, Z_{n-p}$.

Using again the lag operator L defined in (1.2) we can rewrite (1.11) as

$$\begin{aligned} X_n &= Z_n + \sum_{k=1}^{p} \alpha_k L^k X_n + \sum_{k=1}^{q} \beta_k L^k Z_n \\ &= Z_n + \phi(L) X_n + \psi(L) Z_n, \qquad n \geqslant 1. \end{aligned}$$

Example: generating ARMA(2) samples in R

```
n=41
arma.sim<-arima.sim(model=list(ar=c(.9,-.2),ma=c(-.7,.1)),n.start=100,n)
x=seq(100,100+n-1)
plot(x,arma.sim,pch=19, ylab="X", xlab="n", main = 'ARMA$(2)$ samples',col='blue')
lines(x,arma.sim, main = 'ARMA$(2)$ samples',col='blue')
```


This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

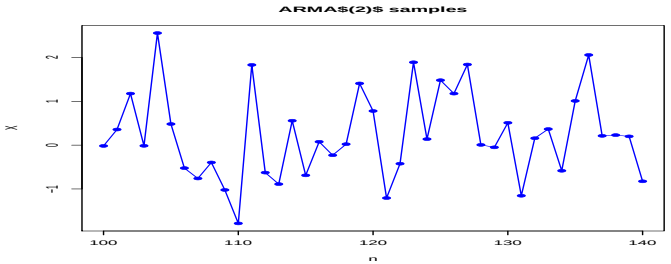


Fig. 1.3: ARMA(2) samples.

1.2 Autoregressive Integrated Moving Average

Consider the difference operator $\nabla := \mathrm{I}_d - L$ defined as

$$\nabla X_n := X_n - X_{n-1}, \qquad n \geqslant 1.$$

The time series $(X_n)_{k\geqslant 1}$ can be recovered by integrating $(\nabla X_k)_{k\geqslant 1}$ using the telescoping identity

$$X_n = X_0 + \sum_{k=1}^{n}(X_k - X_{k-1}) = X_0 + \sum_{k=1}^{n}\nabla X_k, \qquad n \geqslant 1. \tag{1.12}$$

Proposition 1.5. *The iterated operator ∇^d satisfies*

$$\nabla^d X_n = \sum_{k=0}^{d}\binom{d}{k}(-1)^k X_{n-k}, \qquad n \geqslant d \geqslant 1.$$

Proof. This is a consequence of the binomial operator identity

$$\begin{aligned}\nabla^d &= (\mathrm{I}_d - L)^d \\ &= \sum_{k=0}^{d}\binom{d}{k}(\mathrm{I}_d)^{n-k}(-L)^k \\ &= \sum_{k=0}^{d}\binom{d}{k}(-1)^k L^k.\end{aligned}$$

□

For example, we have

$$\nabla^2 X_n = \nabla\nabla X_n$$

$$
\begin{aligned}
&= \nabla(X_n - X_{n-1}) \\
&= \nabla X_n - \nabla X_{n-1} \\
&= X_n - X_{n-1} - (X_{n-1} - X_{n-2}) \\
&= X_n - 2X_{n-1} + X_{n-2},
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla^3 X_n &= \nabla\nabla^2 X_n \\
&= \nabla X_n - 2\nabla X_{n-1} + \nabla X_{n-2} \\
&= X_n - X_{n-1} - 2(X_{n-1} - X_{n-2}) + X_{n-2} - X_{n-3} \\
&= X_n - 3X_{n-1} + 3X_{n-2} - X_{n-3}.
\end{aligned}
$$

Definition 1.6. *In the ARIMA(p,d,q) model, the iterated difference process $(\nabla^d X_n)_{n\geqslant 0}$ is modeled as the ARMA(p,q) time series*

$$
\begin{aligned}
\nabla^d X_n := {} & Z_n + \alpha_1 \nabla^d X_{n-1} + \cdots + \alpha_p \nabla^d X_{n-p} \\
& + \beta_1 Z_{n-1} + \cdots + \beta_p Z_{n-q} \\
= {} & Z_n + \sum_{k=1}^{p} \alpha_k \nabla^d X_{n-k} + \sum_{k=1}^{p} \beta_k X_{n-k},
\end{aligned}
\tag{1.13}
$$

$n \geqslant \max(p+d, q+d)$, based on the p previous values $\nabla^d X_{n-1}, \ldots, \nabla^d X_{n-p}$ and $Z_{n-1}, \ldots, Z_{n-p}$.

Using the backward time shift operator L we can rewrite (1.13) as

$$
\nabla^d X_n = Z_n + \phi(L)\nabla^d X_n + \psi(L) Z_n,
$$

$n \geqslant \max(p+d, q+d)$, where the functions $\phi(z)$ and $\psi(z)$ are given by

$$
\phi(z) = \sum_{k=1}^{p} \alpha_k z^k \quad \text{and} \quad \psi(z) = \sum_{k=1}^{p} \beta_k z^k.
$$

In other words, we have

$$
(\mathrm{I}_d - \phi(L))\nabla^d X_n = Z_n + \psi(L) Z_n,
$$

$n \geqslant \max(p+d, q+d)$. The process $(X_n)_{n\geqslant 0}$ can then be recovered by successive applications of the discrete integration formula (1.12).

Proposition 1.7. *We can recover X_n from ∇X_n as*

$$
X_n = \sum_{k=0}^{d} \binom{d}{k} \nabla^k X_{n+k-d}.
$$

Proof. We apply the binomial operator identity

$$\mathrm{I}_d = (\mathrm{I}_d - L + L)^d = (L + \nabla)^d = \sum_{k=0}^{d} \binom{d}{k} L^{d-k} \nabla^k.$$

□

Alternatively, we can start by recovering $\nabla^{d-1} X_n$ from $\nabla^d X_n$ as

$$\begin{aligned} \nabla^{d-1} X_n &= \nabla^{d-1} X_0 + \sum_{k=1}^{n} \left(\nabla^{d-1} X_k - \nabla^{d-1} X_{k-1} \right) \\ &= \nabla^{d-1} X_0 + \sum_{k=1}^{n} \nabla^d X_k, \end{aligned}$$

followed by

$$\nabla^{d-2} X_n, \nabla^{d-3} X_n, \ldots, \nabla^2 X_n, \nabla X_n, X_n,$$

by induction.

Example: generating ARIMA$(1,2,3)$ samples in R

```
n=41
arima.sim<-arima.sim(
model=list(ar=c(0.5),ma=c(0.5, 0.5, -0.5),order=c(1,2,3)),n.start=100,n)
x=seq(100,100+n+1)
plot(x,arima.sim,pch=19, ylab="X", xlab="n", main = 'ARIMA$(0,2,3)$ samples',col='blue')
lines(x,arima.sim, main = 'ARIMA$(1,2,3)$ samples',col='blue')
```

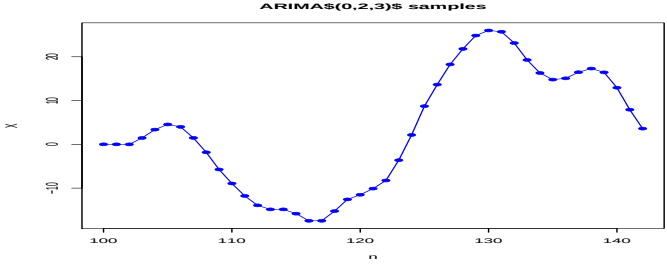


Fig. 1.4: ARIMA$(1,2,3)$ samples.

Note that the ARIMA graph of Figure 1.4 has more potential for prediction than the ARMA graph of Figure 1.3 due to increased dependence on past samples in the considered model, or longer memory.

1.3 Time Series Stationarity

Definition 1.8. *A time series* $(X_n)_{n\geqslant 0}$ *is* strictly stationary *if the equality holds*

$$(X_n, X_{n-1}, \ldots, X_{n-p}) \overset{d}{\simeq} (X_{n+m}, X_{n+m-1}, \ldots, X_{n+m-p}),$$

for all $n \in \mathbb{Z}$ *and* $p \geqslant 1$.

In other words, Definition 1.8 states that the random vectors

$$(X_n, X_{n-1}, \ldots, X_{n-p}) \quad \text{and} \quad (X_{n+m}, X_{n+m-1}, \ldots, X_{n+m-p})$$

have same distribution for all $m \in \mathbb{Z}$ and $p \geqslant 1$.

Example. The MA(q) time series $(X_n)_{n\geqslant 0}$ is *strictly stationary* since

$$\begin{aligned} X_{n+m} &= Z_{n+m} + \beta_1 Z_{n+m-1} + \cdots + \beta_q Z_{n+m-q} \\ &= Z_{n+m} + \sum_{k=1}^{q} \beta_k Z_{n+m-k}, \qquad n \geqslant q, \end{aligned}$$

satisfies the equality

$$\begin{aligned} X_{n+m} &\overset{d}{\simeq} Z_n + \beta_1 Z_{n-1} + \cdots + \beta_q Z_{n-q} \\ &= Z_n + \sum_{k=1}^{q} \beta_k Z_{n-k} \\ &= X_n, \qquad n \geqslant q, \end{aligned}$$

in distribution as $(Z_n)_{n\geqslant 0}$ is an *i.i.d.* sequence.

Definition 1.9. *A time series* $(X_n)_{n\geqslant 0}$ *is* weakly stationary *if*

i) $\mathbb{E}[X_n] = \mathbb{E}[X_0]$, $n \geqslant 0$, *and*

ii) the autocovariance [*]

$$(n, m) \longmapsto \mathrm{Cov}(X_n, X_m)$$

depends only on the absolute difference $|n-m|$, $n, m \geqslant 0$.

We note that the expressions (1.6)-(1.8) correspond to strictly stationary times series. More generally, by representing an AR(q) series as a vector-valued AR(1) series we can obtain the following result, see *e.g.* Theorem 4.4 page 119 of Pourahmadi (2001).

[*] The covariance $\mathrm{Cov}(X, Y)$ is defined as $\mathrm{Cov}(X, Y) := \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

Theorem 1.10. Unit root test. *The equation* (1.4) *written as*

$$X_n := Z_n + \phi(L)X_n = Z_n + \alpha_1 X_{n-1} + \cdots + \alpha_q X_{n-q}$$

with

$$\phi(z) = \alpha_1 z + \cdots + \alpha_q z^q, \qquad z \in \mathbb{C},$$

admits an AR(q) solution $(X_n)_{n\geqslant 0}$ *which is* weakly stationary *if and only if* all *(complex) solutions of the equation* $\phi(z) = 1$ *lie* outside the complex unit circle $\{z \in \mathbb{C} \ : \ |z| = 1\}$ *in the complex plane.*

The autocovariances $\mathrm{Cov}(X_n, X_{n+l})$ and cross-covariances $\mathrm{Cov}(Y_n, X_{n+l})$ of time series with lag parameter $l \in \mathbb{Z}$ can be respectively estimated as follows:

```
ar.acf<-acf(ar.sim,type="covariance",plot=T,col='blue')
ar.ccf<-ccf(ar.sim,ar.sim,type="covariance",plot=T,lwd=2,col='blue')
```

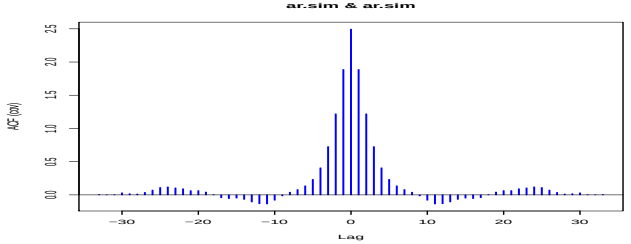


Fig. 1.5: Autocovariances of AR(2) samples.

Examples.

i) In the AR(2) example

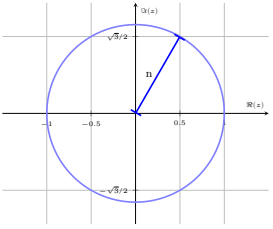
$$X_n := Z_n + 0.9 \times X_{n-1} - 0.2 \times X_{n-2}$$

of Figure 1.2 with $\phi(z) = 0.9z - 0.2z^2$. The solutions $z = 2, 2.5$ of the equation $\phi(z) = 1$ do not belong to the complex unit circle, hence by Theorem 1.10 the time series $(X_n)_{n\geqslant 2}$ is *not* (weakly) stationarity.

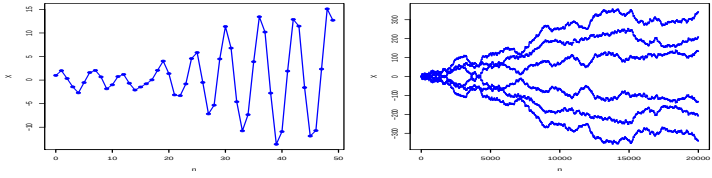
ii) Consider the AR(2) time series

$$X_n := Z_n + X_{n-1} - X_{n-2} = Z_n + \phi(L)X_n, \qquad n \geqslant 2,$$

with $\phi(z) = z - z^2$.



The solutions $z = (1 \pm i\sqrt{3})/4$ of the equation $\phi(z) = 1$ lie on the unit circle, hence by Theorem 1.10 the time series $(X_n)_{n\geqslant 2}$ is not (weakly) stationarity. The next Figure 1.6 presents a simulation of non-stationary time series according to the attached **R code**.



(a) Simulation with 50 time samples.

(b) Simulation with 20000 time samples.

Fig. 1.6: Nonstationarity of AR(2) time series with $a_1 = 1$ and $a_2 = -1$.

In the AR(1) example

$$X_n := Z_n + \alpha_1 X_{n-1}, \qquad n \geqslant 1,$$

of Proposition 1.3 with $\phi(z) = \alpha_1 z$ the unique solution $z = 1/\alpha_1$ of the equation $\phi(z) = 1$ belongs to the complex unit circle if and only if $|\alpha_1| = 1$, hence by Theorem 1.10 the time series $(X_n)_{n\geqslant 2}$ is (weakly) stationarity if and only if $|\alpha_1| = 1$. More precisely, we have

$$X_n := Z_n + \alpha_1 X_{n-1} = Z_n + \phi(L)X_n,$$

with $\phi(z) = \alpha_1 z$. The equation $\phi(z) = \alpha_1 z = 1$ has the unique solution $z = 1/\alpha_1$ which lies outside the unit circle if and only if $\alpha_1 \neq \pm 1$, *i.e.* $|\alpha_1| \neq 1$. In this case we have the causal representations

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

$$\begin{cases} X_n = \sum_{k\geqslant 0} \alpha_1^k Z_{n-k}, & |\alpha_1| < 1, \quad (1.14a) \\ X_n = -\sum_{k\geqslant 1} \frac{1}{\alpha_1^k} Z_{n+k}, & |\alpha_1| > 1, \quad (1.14b) \end{cases}$$

which converge when $|\alpha_1| \neq 1$, with

$$\mathbb{E}\left[X_n^2\right] = \frac{1}{|1-|\alpha_1|^2|}, \qquad n \geqslant 1,$$

see (1.7) and (1.9).

i) In case (1.14a) with $|\alpha_1| < 1$, the solution

$$X_n = \sum_{k\geqslant 0} \alpha_1^k Z_{n+k}, \qquad n \geqslant 0,$$

satisfies

$$\begin{aligned} \mathrm{Cov}(X_n, X_m) &= \mathbb{E}\left[\sum_{k\geqslant 0} \alpha_1^k Z_{n+k} \sum_{l\geqslant 0} \bar{\alpha}_1^l Z_{m+l}\right] \\ &= \alpha^{n-m} \sum_{k\geqslant 0} |\alpha_1|^{2k} \\ &= \frac{\alpha^{n-m}}{1-|\alpha_1|^2}, \qquad n \geqslant m \geqslant 0. \end{aligned}$$

ii) In case (1.14b) with $|\alpha_2| < 1$, the solution

$$X_n = -\sum_{k\geqslant 1} \frac{1}{\alpha_1^k} Z_{n+k}, \qquad n \geqslant 0,$$

satisfies

$$\begin{aligned} \mathrm{Cov}(X_n, X_m) &= \mathbb{E}\left[\sum_{k\geqslant 1} \frac{1}{\alpha_1^k} Z_{n+k} \sum_{l\geqslant 1} \frac{1}{\bar{\alpha}_1^l} Z_{m+l}\right] \\ &= \frac{\alpha^{m-n}}{|\alpha_1|^2-1}, \qquad n \geqslant m \geqslant 0. \end{aligned}$$

Stationarity test

The Dickey-Fuller test allows us to test the null hypothesis H_0, *i.e.* "$|\alpha_1| = 1$", against the alternative stationarity hypothesis "$|\alpha_1| \neq 1$".

a) When $|\alpha_1| \neq 1$ we consider the OLS test statistic

$$\begin{aligned}\widehat{\rho}_n &:= \frac{\sum_{t=1}^n X_{t-1}X_t}{\sum_{t=1}^n X_{t-1}^2}\\ &= \frac{\sum_{t=1}^n X_{t-1}(Z_t+\alpha_1 X_{t-1})}{\sum_{t=1}^n X_{t-1}^2}\\ &= \alpha_1 + \frac{\sum_{t=1}^n X_{t-1}Z_t}{\sum_{t=1}^n X_{t-1}^2}\end{aligned}$$

which minimizes the value of α_1 in the residual

$$\sum_{t=1}^n (X_t - \alpha_1 X_{t-1})^2$$

representing the quadratic distance between $(X_t)_{t=1,2,\ldots,n}$ and $\big(\alpha_1 X_{t-1}\big)_{t=1,2,\ldots,n}$. By (1.7), (1.9) and the Central Limit Theorem we have

$$\begin{aligned}\sqrt{n}(\widehat{\rho}_n - \alpha_1) &\simeq \sqrt{n}\frac{\sum_{t=1}^n X_{t-1}Z_t}{\sum_{t=1}^n X_{t-1}^2}\\ &\simeq \frac{|1-|\alpha_1|^2|}{\sqrt{n}}\sum_{t=1}^n X_{t-1}Z_t\\ &\simeq \sqrt{\frac{|1-|\alpha_1|^2|}{n}}\sum_{t=1}^n Z_t\end{aligned}$$

converges in distribution to $\mathcal{N}(0, 1-|\alpha_1|^2)$ as n tends to infinity, see Chapters 8 and 17 of Hamilton (1994).

b) When $|\alpha_1| = 1$, the test statistic

$$\hat{t}_n := \frac{\sum_{t=1}^{n} X_{t-1} Z_t}{n \sqrt{\sum_{t=1}^{n} X_{t-1}^2 \sum_{k=1}^{n} (X_t - \widehat{\rho}_k X_{k-1})^2 / (n-k)}}$$

has the asymptotic distribution of $\int_0^1 B_s dB_s / \int_0^1 B_s^2 ds$, where $(B_s)_{s\in[0,1]}$ is a standard Brownian motion, see § 17.4 of Hamilton (1989), and it can be used to test the null hypothesis H_0, *i.e.* "$|\alpha_1| = 1$", against the alternative stationarity hypothesis "$|\alpha_1| \neq 1$".

The (Augmented) Dickey-Fuller stationarity test uses an additional lag parameter and is implemented in the 'tseries' R package, as follows:

```
install.packages('tseries')
library('tseries')
adf.test(ar.sim)
```

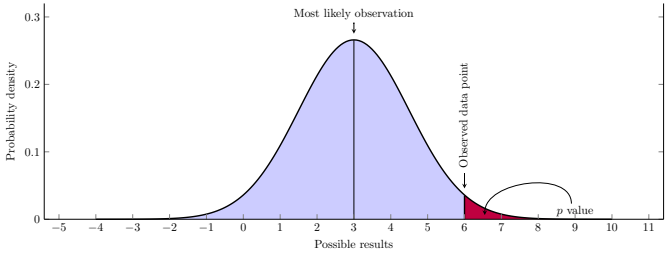


Fig. 1.7: Hypothesis testing.

Its output leads us to reject the nonstationarity (null) hypothesis H_0 at the level 1% for the AR(2) time series (1.10) of Figure 1.2:

```
        Augmented Dickey-Fuller Test
data: ar.sim
Dickey-Fuller = -13.765, Lag order = 16, p-value = 0.01
alternative hypothesis: stationary
Warning message:
In adf.test(ar.sim) : p-value smaller than printed p-value
```

Applying the Augmented Dickey-Fuller Test to the ARIMA time series of Figure 1.4 would not allow us to reject the nonstationarity (null) hypothesis H_0. Other stationarity tests for time series include the KPSS test.

1.4 Fitting Time Series to Financial Data

Data can be fitted to an ARIMA(p, d, q) model in R using the command

Fitting market returns

```
arima(data,order=c(p,d,q))
```

For an example based on market returns, we use:

```
library(quantmod)
getSymbols("1800.HK",from="2013-01-01",to="2014-11-30",src="yahoo")
stock.rtn=diff(log(Ad(`1800.HK`)))
chartSeries(stock.rtn,up.col="blue",theme="white")
n = sum(!is.na(stock.rtn))
```

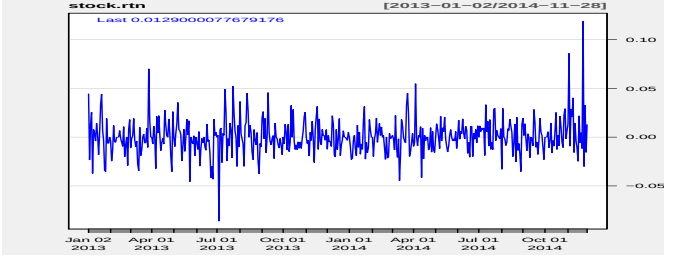


Fig. 1.8: Graph of stock returns.

We fit this data to an ARMA(2,2) time series using the Arima command

```
arima(stock.rtn,order=c(2,0,2))
```

is as follows:

Coefficients:

	ar1	ar2	ma1	ma2
	-1.0593	-0.9048	1.0509	0.9508
s.e.	0.0679	0.0444	0.0474	0.0273

Sample data from this time series can be obtained (up to rescaling) from

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

```
n=498
arima.sim<-arima.sim(model=list(ar=c(-1.0593,-0.9048),ma=c(1.0509,0.9508),order=c(2,0,2)),
        n.start=100,n)
x=seq(100,100+n-1)
myTheme <- chart_theme();myTheme$col$line.col <- "blue"
par(mfrow=c(1,2));chart_Series(stock.rtn,theme=myTheme)
plot(x,arima.sim, type="l",col="blue",ylab="X", xlab="n", main = 'ARIMA$(2,0,2)$
        samples')
```

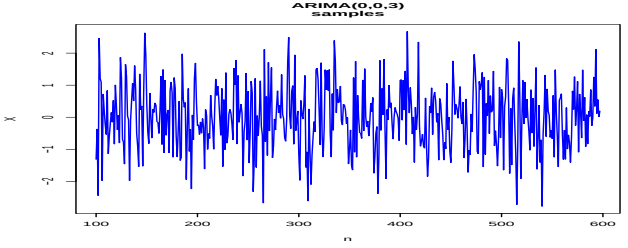


Fig. 1.9: ARIMA$(2,0,2)$ samples.

Next, we fit this data to an MA(3) time series using the command

```
arima(stock.rtn,order=c(0,0,3))
```

yields the output:

Coefficients:

	ma1	ma2	ma3
	0.0029	0.0470	-0.0416
s.e.	0.0452	0.0467	0.0465

```
n=498
arima.sim<-arima.sim(model=list(ma=c(0.0029, 0.0470,
        -0.0416),order=c(0,0,3)),n.start=100,n)
x=seq(100,100+n-1)
myTheme <- chart_theme();myTheme$col$line.col <- "blue"
par(mfrow=c(1,2));chart_Series(stock.rtn,theme=myTheme)
plot(x,arima.sim, type="l",col="blue",ylab="X", xlab="n", main = 'ARIMA$(0,0,3)$
        samples')
```

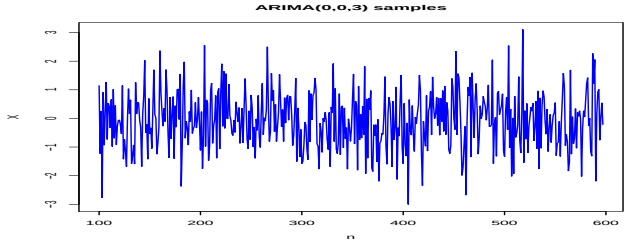


Fig. 1.10: ARIMA$(0,0,3)$ samples.

Fitting market prices

Next, we fit market price data to an ARIMA time series.

```
library(quantmod)
getSymbols("1800.HK",from="2007-01-03",to="2011-12-02",src="yahoo")
stock=Ad(`1800.HK`)
chartSeries(stock,up.col="blue",theme="white")
n = sum(!is.na(stock.rtn))
arima(stock,order=c(2,1,2))
```

Fig. 1.11: Cumulative stock returns.

Coefficients:

	ar1	ar2	ma1	ma2
	-0.3133	-0.9464	0.3535	0.9637
s.e.	0.0213	0.0239	0.0206	0.0172

```
n=1236
arima.sim<-arima.sim(model=list(ar=c(-0.3133,-0.9464),ma=c(0.3535,0.9637),order=c(2,1,2)),
        n.start=100,n)
x=seq(100,100+n)
myTheme <- chart_theme();myTheme$col$line.col <- "blue"
par(mfrow=c(1,2));chart_Series(stock,theme=myTheme)
plot(x,arima.sim, type="l",col="blue",ylab="X", xlab="n", main = 'ARIMA$(2,1,2)$
        samples')
```

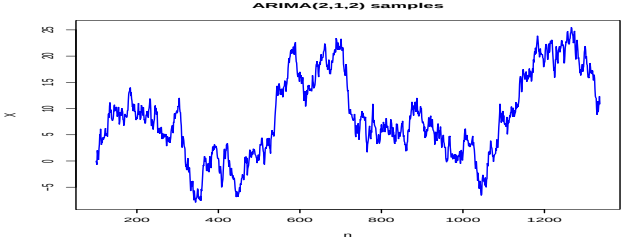


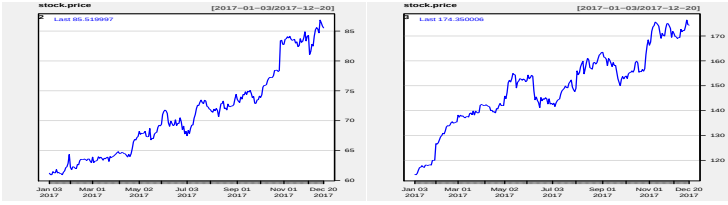
Fig. 1.12: ARIMA$(2, 1, 2)$ samples.

1.5 Application: Pair Trading

Pair trading data

We consider two asset prices that can be traded in pairs.

```
install.packages("quantmod")
library(quantmod)
symbols = c("1800.HK","KO","PEP")
symbols = c("1800.HK","MSFT","AAPL")
getSymbols(symbols, from="2017-01-01", to="2018-01-01")
ClosePrices <- lapply(symbols, function(x) Ad(get(x)))
stock<-do.call(merge, ClosePrices)
stock.price<-stock[rowSums(is.na(stock[ , 1:3])) == 0, ]
price.pair <- stock.price[,2:3]["2017-02-01::"]
chartSeries(stock.price[,2],up.col="blue",theme="white",name = symbols[2])
chartSeries(stock.price[,3],up.col="blue",theme="white",name = symbols[3])
myTheme <- chart_theme()
myTheme$col$line.col <- "blue"
par(mfrow=c(1,3))
chart_Series(stock.price[,2],theme=myTheme,name = symbols[2])
chart_Series(stock.price[,3],theme=myTheme,name = symbols[3])
add_TA(stock.price[,2], col='purple', lw =2, on = 1)
```



(a) MSFT graph. (b) AAPL graph.

Fig. 1.13: MSFT *vs* APPL graphs.

Linear regression

As the two assets may evolve within different price ranges, we use a linear regression to put them both on the scale of the second asset.

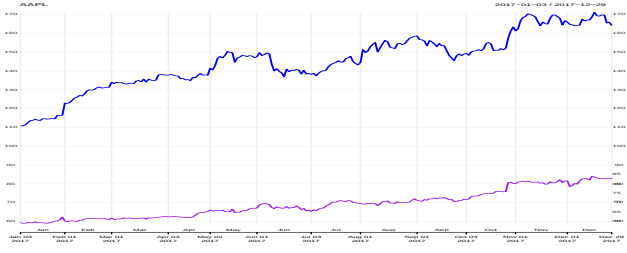


Fig. 1.14: Comparison graph before linear regression.

Letting

$$r_t^{(1)} := \log(S_t^{(1)}) \quad \text{and} \quad r_t^{(2)} := \log(S_t^{(2)}), \qquad t \geqslant 1,$$

by an Ordinary Least Square (OLS) regression using the R command *lm* (linear model), we derive a linear relationship of the form

$$r_t^{(2)} = a + b r_t^{(1)} + X_t, \quad t \geqslant 1, \tag{1.15}$$

between $(r_k^{(1)})_{k\geqslant 1}$ and $(r_k^{(2)})_{k\geqslant 1}$, where X_k is a random remainder term, by minimization of the quadratic residual distance

$$\sum_{t=1}^{n} \left(r_t^{(2)} - a - b r_t^{(1)}\right)^2 \tag{1.16}$$

between $(r_t^{(2)})_{t=1,2,\dots,n}$ and $\left(a+br_t^{(1)}\right)_{t=1,2,\dots,n}$, *i.e.*

$$\begin{cases} \widehat{a} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}\left(r_k^{(2)} - \widehat{b}r_k^{(1)}\right), \\ \\ \text{and} \\ \\ \widehat{b} = \dfrac{\displaystyle\sum_{k=1}^{n} r_k^{(1)} r_k^{(2)} - \frac{1}{n}\sum_{k,l=1}^{n} r_k^{(1)}\tilde{r}_l^{(2)}}{\displaystyle\sum_{k=1}^{n}(r_k^{(1)})^2 - \frac{1}{n}\sum_{k,l=1}^{n} r_k^{(1)} r_l^{(1)}} = \dfrac{\displaystyle\sum_{k=1}^{n}\left(r_k^{(1)} - \frac{1}{n}\sum_{l=1}^{n} r_l^{(1)}\right)\left(r_k^{(2)} - \frac{1}{n}\sum_{l=1}^{n}\tilde{r}_l^{(2)}\right)}{\displaystyle\sum_{k=1}^{n}\left(r_k^{(1)} - \frac{1}{n}\sum_{k=1}^{n} r_k^{(1)}\right)^2}. \end{cases}$$

cf. Exercise 1.5. The coefficient a in (1.15) is called the *premium*, and b is called the *hedge ratio* .

```
reg <- lm(log(price.pair[,2]) ~ log(price.pair[,1]))
hedge.ratio <- as.numeric(reg$coef[2])
premium <- as.numeric(reg$coef[1])
myTheme <- chart_theme()
myTheme$col$line.col <- "blue"
chart_Series(price.pair[,2],name="",theme=myTheme)
pdf("compare2.pdf")
add_TA(exp(premium+hedge.ratio*log(price.pair[,1])), col='purple', lw =2,on = 1)
```

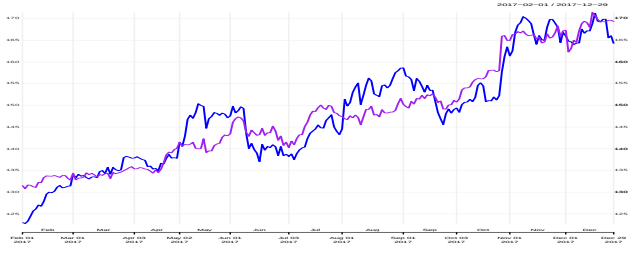


Fig. 1.15: Comparison graph after linear regression.

This allows us to defined the *spread* $(X_t)_{t\geqslant 1}$ via the linear relationship

$$X_t := \log(S_t^{(2)}) - (\text{premium} + \text{hedge.ratio} \times \log(S_t^{(1)})), \qquad \geqslant 0.$$

```
spread <- log(price.pair[,2]) - ( hedge.ratio * log(price.pair[,1]) + premium )
list(spread = spread, hedge.ratio = hedge.ratio, premium = premium)
plot(spread,col='blue', main = "Spread")
```

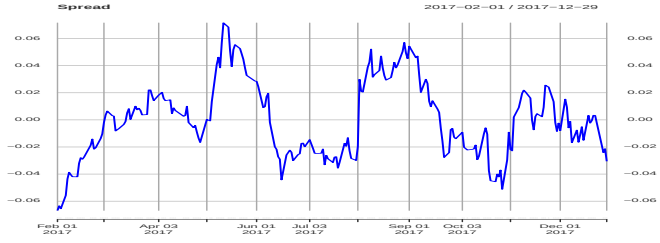


Fig. 1.16: Spread graph.

Next, we model the spread X_t using an AR(1) time series and check for its stationarity, in which case the log-price processes $\log\left(S_t^{(1)}\right)_{t\geqslant 0}$ and $\log\left(S_t^{(2)}\right)_{t\geqslant 0}$ are said to be *cointegrated*.

This will be interpreted as the existence of a statistically significant connection between $\left(S_t^{(1)}\right)_{t\geqslant 0}$ and $\left(S_t^{(2)}\right)_{t\geqslant 0}$, also named *cointegration*. See Engle and Granger (1987) and Chapter 6 of Enders (2009) for more information on *cointegration*.

Dickey-Fuller test

Consider an AR(1) time series $(X_n)_{n\geqslant 0}$ given by

$$X_n := Z_n + \alpha_1 X_{n-1}.$$

The Dickey-Fuller test allows us to test the null hypothesis H_0, *i.e.* "$|\alpha_1| = 1$", against the alternative stationarity hypothesis "$|\alpha_1| \neq 1$".

```
install.packages('tseries')
library('tseries')
adf.test(spread)
```

Its output leads us to reject the nonstationarity (null) hypothesis H_0 at the level 1%:

```
        Augmented Dickey-Fuller Test
```

```
data: spread
Dickey-Fuller = -2.8771, Lag order = 6, p-value = 0.2077
alternative hypothesis: stationary
```

We reject the (nonstationarity) null hypothesis H_0 at the confidence level 5% when the p-value is lower than 0.05. The Phillips-Perron test is another unit root test.

Pair trading

The trading signal is $\{-1, 1\}$-valued and determined by the alternating crossing times of a threshold level by the spread.

```
signal<-spread;threshold <- 0.02
signal[1] = -sign(as.numeric(spread[1]));i=1
while (i<length(spread)){i=i+1;
while (i<length(spread) &&
      sign(as.numeric(spread[i+1])-threshold)==sign(as.numeric(spread[i])-threshold))
      {signal[i]=sign(threshold-as.numeric(spread[i-1]));i=i+1;print(i);}
signal[i]=sign(threshold-as.numeric(spread[i-1]));
threshold=-threshold;print(i);}
signal[i]=sign(threshold-as.numeric(spread[i-1]));
threshold <- abs(threshold)
ratio1=range(spread)[1]/threshold
ratio2=range(spread)[2]/threshold
tblue <- rgb(0, 0, 1, alpha=0.8)
tred <- rgb(1, 0, 0, alpha=0.5)
barplot(spread,col = tblue,lwd = 3, main = "Spread vs (-Signal)");par(new=TRUE);
barplot(signal,offset=(range(spread)[1]+range(spread)[2])/threshold,ylim=c(ratio2,ratio1),
      xpd = FALSE, col=tred,space = 0, border
      ="blue",xaxt="n",yaxt="n",xlab="",ylab="")
```

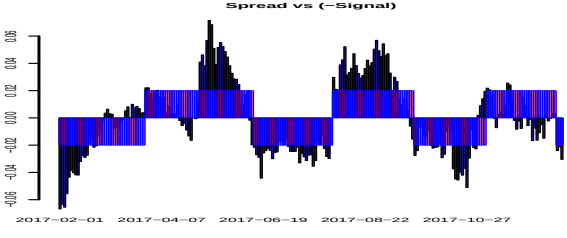


Fig. 1.17: Pair trading signal.

Backtesting

The performance of the pair trading algorithm can be evaluated by the following code.

```
return.pairtrading=(lag(signal)*(price.pair[,2]-lag(price.pair[,2]))-lag(signal)*hedge.ratio
      *(price.pair[,1]-lag(price.pair[,1])))/(hedge.ratio*lag(price.pair[,1])+lag(price.pair[,2]))
return.pairtrading<-return.pairtrading[2:length(return.pairtrading)]
par(mfrow=c(1,2))
plot(return.pairtrading,col='blue', main = "Returns")
plot(100 * cumprod(1 + return.pairtrading),col='blue',main = "Performance")
```

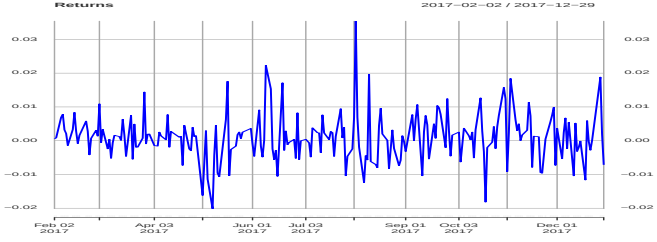


Fig. 1.18: Pair trading returns.

The return of the portfolio is defined as

$$\frac{\xi_t^{(2)}(S_t^{(2)}-S_{t-1}^{(2)})+\xi_t^{(1)}(S_t^{(1)}-S_{t-1}^{(1)})}{\xi_t^{(2)}S_{t-1}^{(2)}+\xi_t^{(1)}S_{t-1}^{(1)}},$$

where

$$\begin{cases}\xi_t^{(2)} := \text{signal}_{t-1}\times\dfrac{1}{1+\text{hedge.ratio}},\\[2ex] \xi_t^{(1)} := (-\text{signal}_{t-1})\times\dfrac{\text{hedge.ratio}}{1+\text{hedge.ratio}}.\end{cases}$$

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

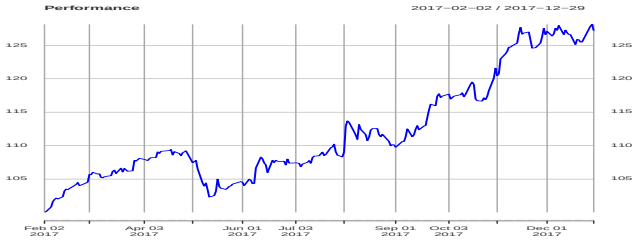


Fig. 1.19: Performance of pair trading.

The next Figure 1.20 provides another example of pair trading backtesting.*

```
> source("pairtrading.R")
Examples of pairs: 005930.KS vs AAPL, 2600.HK vs 1919.HK
Enter Stock 1 (Ex: GOOG):2600.HK
Enter Stock 2 (Ex: AAPL):1919.HK

        Augmented Dickey-Fuller Test
data: spread
Dickey-Fuller = -3.4553, Lag order = 8, p-value = 0.0466
alternative hypothesis: stationary
```

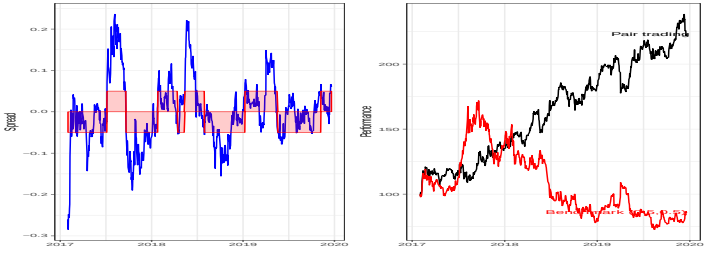


Fig. 1.20: Pair trading algorithm.

See also the R PairTrading package in Takayanagi and Ishikawa (2017).

* Download the corresponding **R code**.

Exercises

Exercise 1.1 Consider a white noise $(Z_n)_{n\geqslant 1}$.

a) Check the weak stationarity of the AR(1) time series $(X_n)_{n\geqslant 1}$ given by

$$X_n = Z_n + X_{n-1}, \qquad n \geqslant 1.$$

b) Check the weak stationarity of the AR(2) time series $(Y_n)_{n\geqslant 1}$ given by

$$X_n = Z_n + 0.75 \times X_{n-1} - 0.125 \times X_{n-2}, \qquad n \geqslant 2.$$

Hint: Consider the roots of $\varphi(z) = 1$ where $\varphi(z)$ is the polynomial defined by $X_n = Z_n + \varphi(L)X_n$ and L is the lag operator $LX_n = X_{n-1}$.

Exercise 1.2 Let $\alpha \in \mathbb{R}$. Consider an *i.i.d.* white noise sequence $(Z_n)_{n\geqslant 0}$ with mean $\mathbb{E}[Z_n] = 0$ and variance $\mathrm{Var}[Z_n] = 1$, $n \geqslant 1$, and the AR(1) time series $(X_n)_{n\geqslant 0}$ given by $X_0 := 0$ and

$$X_n := Z_n + \alpha X_{n-1}, \qquad n \geqslant 1. \tag{1.17}$$

a) Find a recurrence relation for the mean $\mathbb{E}[X_n]$ in the parameter $n \geqslant 0$, and deduce the value of $\mathbb{E}[X_n]$ for all $n \geqslant 0$.
b) For fixed $n \geqslant 1$, find a recurrence relation in the parameter $k \geqslant 0$ for the covariance

$$\mathrm{Cov}(X_{n+k}, X_n) = \mathbb{E}[X_{n+k}X_n] - \mathbb{E}[X_{n+k}]\,\mathbb{E}[X_n], \qquad k \geqslant 0.$$

c) Find a recurrence relation in the parameter $n \geqslant 1$ for the variance

$$\mathrm{Var}[X_n] = \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2 = \mathbb{E}[X_n^2], \qquad n \geqslant 0.$$

d) When is the time series $(X_n)_{n\geqslant 0}$ weakly stationary?

Exercise 1.3 Consider an *i.i.d.* white noise sequence $(Z_n)_{n\geqslant 0}$ with mean $\mathbb{E}[Z_n] = 0$ and variance $\mathrm{Var}[Z_n] = 1$, $n \geqslant 1$, and the AR(3) time series $(X_n)_{n\geqslant 3}$ given by

$$X_n := Z_{n-1} - Z_{n-2} + \alpha Z_{n-3}, \qquad n \geqslant 3.$$

a) Find the autocovariances

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html

$$\begin{cases} \mathrm{Cov}(X_n, X_n) = \mathrm{Var}[X_n], \\ \\ \mathrm{Cov}(X_{n+1}, X_n), \\ \\ \mathrm{Cov}(X_{n+2}, X_n), \\ \\ \mathrm{Cov}(X_{n+k}, X_n), \qquad k \geqslant 3. \end{cases}$$

b) Show that $(X_n)_{n\geqslant 3}$ has same distribution as an MA(q) time series $(Y_n)_{n\geqslant 3}$ of the form

$$Y_n = Z_n + \sum_{k=1}^{q} \beta_k Z_{n-k},$$

whose order q and coefficients $(\beta_k)_{1\leqslant k\leqslant q}$ will be determined.

Exercise 1.4 Consider an AR(1) time series $(X_n)_{n\geqslant 0}$ given by $X_0 = 0$ and

$$X_n := Z_n + \alpha_1 X_{n-1}, \qquad n \geqslant 1,$$

and the difference operator

$$\nabla X_n := X_n - X_{n-1}, \qquad n \geqslant 1,$$

also written $\nabla = I - L$, which can be integrated by the telescoping identity

$$X_n = X_0 + \sum_{k=1}^{n} (X_k - X_{k-1}) = \sum_{k=1}^{n} \nabla X_k, \qquad n \geqslant 1.$$

a) Show that the first order difference process $(\nabla X_n)_{n\geqslant 1} = (X_n - X_{n-1})_{n\geqslant 1}$. forms an ARMA$(1,2)$ time series.

b) Show that the second order difference process

$$(\nabla^2 X_n)_{n\geqslant 2} = (\nabla X_n - \nabla X_{n-1})_{n\geqslant 2}$$

forms an ARMA$(1,3)$ time series.

Exercise 1.5 Consider two sequences $(r_k^{(1)})_{k\geqslant 1}$ and $(r_k^{(2)})_{k\geqslant 1}$ of market returns. We aim at deriving a linear relationship of the form

$$r_k^{(2)} = a + b r_k^{(1)} + X_k, \quad k \in \mathbb{N},$$

between $(r_k^{(1)})_{k\geqslant 1}$ and $(r_k^{(2)})_{k\geqslant 1}$, where X_k is a random remainder term, by minimization of the quadratic residual distance

$$\sum_{k=1}^{n} \left(r_k^{(2)} - a - b r_k^{(1)}\right)^2 \tag{1.18}$$

between $(r_k^{(2)})_{k=1,2,\ldots,n}$ and $\left(a + b r_k^{(1)}\right)_{k=1,2,\ldots,n}$.

a) Compute the partial derivatives of (1.18) with respect to the parameters a and b.
b) By equating the derivatives to zero, find the least square estimates $\widehat{a}$ and $\widehat{b}$ of the parameters a and b based on the sequences $(r_k^{(1)})_{k\geqslant 1}$ and $(r_k^{(2)})_{k\geqslant 1}$.

This version: October 15, 2020
https://personal.ntu.edu.sg/nprivault/indext.html