STAT 720

TIME SERIES ANALYSIS

Spring 2015

Lecture Notes

Dewei Wang

Department of Statistics

University of South Carolina

Contents

1 Introduction

1.1 Some examples

Question: What is a time series?
Answer: It is a random sequence $\{X_t\}$ recorded in a time ordered fashion.
Question: What are its applications?
Answer: Everywhere when data are observed in a time ordered fashion. For example:

- Economics: daily stock market quotations or monthly unemployment rates.
- Social sciences: population series, such as birthrates or school enrollments.
- Epidemiology: the number of influenza cases observed over some time period.
- Medicine: blood pressure measurements traced over time for evaluating drugs.
- Global warming?

Example 1.1. (Johnson & Johnson Quarterly Earnings) Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson.

- 84 quarters (21 years) measured from the 1st quarter of 1960 to the last quarter of 1980.

```
require(astsa)
par(mar=c(4,4,2,.5))
plot(jj, type="o", ylab="Quarterly Earnings per Share",col="blue")
```

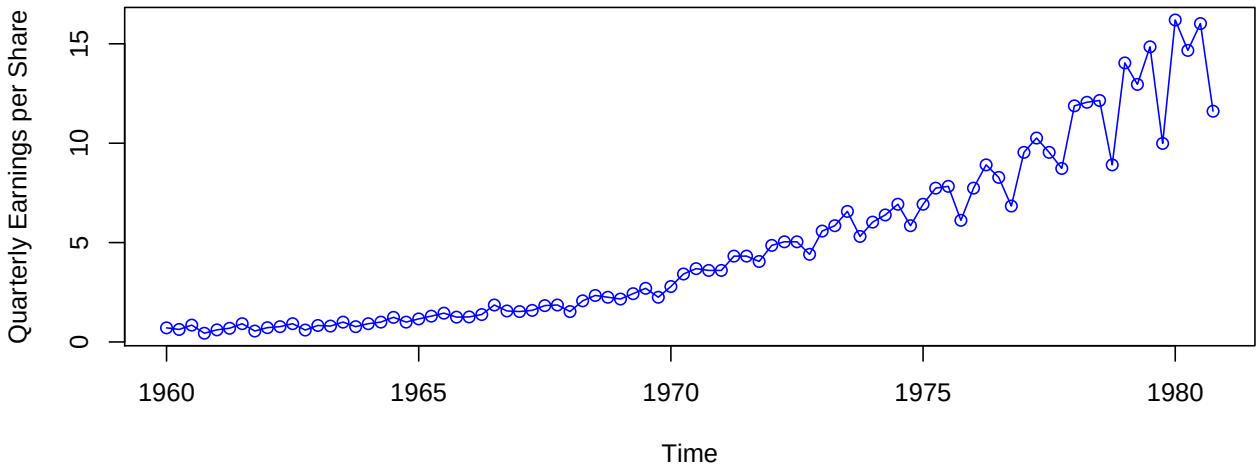


Figure 1.1: Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV

Example 1.2. (Global Warming) Figure 1.2 shows the global mean land-ocean temperature index from 1880 to 2009 with the base period 1951-1980.

```
require(astsa)
par(mar=c(4,4,2,.5))
plot(gtemp, type="o", ylab="Global Temperature Deviations",col="blue")
```

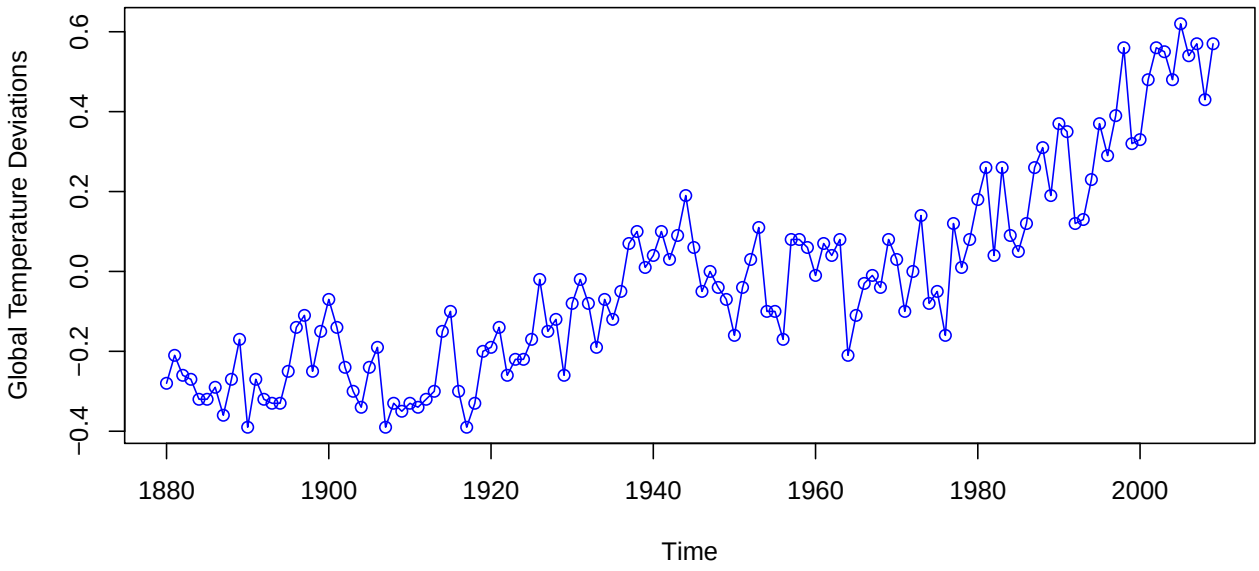


Figure 1.2: Yearly average global temperature deviations (1880-2009) in degrees centigrade.

Example 1.3. (Speech Data) Figure 1.3 shows a small .1 second (1000 point) sample of recorded speech for the phrase *aaa*$\cdots$*hhh*.

```
require(astsa)
par(mar=c(4,4,2,.5))
plot(speech,col="blue")
```

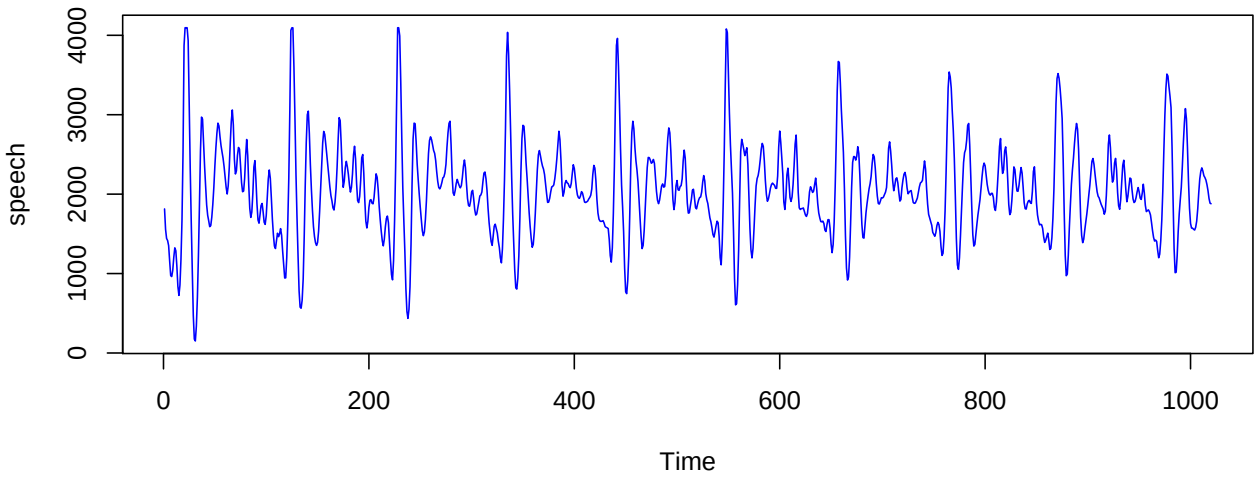


Figure 1.3: Speech recording of the syllable *aaa*$\cdots$*hhh* sampled at 10,000 points per second with $n = 1020$ points

Computer recognition of speech: use spectral analysis to produce a signature of this phrase and then compare it with signatures of various library syllables to look for a match.

1.2 Time Series Statistical Models

A **time series model** specifies the joint distribution of the sequence $\{X_t\}$ of random variables; e.g.,

$$P(X_1 \leq x_1, \ldots, X_t \leq x_t) \text{ for all } t \text{ and } x_1, \ldots, x_t.$$

where $\{X_1, X_2, \ldots\}$ is a stochastic process, and $\{x_1, x_2, \ldots\}$ is a single realization. Through this course, we will mostly restrict our attention to **the first- and second-order properties** only: $E(X_t), \text{Cov}(X_{t_1}, X_{t_2})$

Typically, a time series model can be described as

$$X_t = m_t + s_t + Y_t, \tag{1.1}$$

where

$$\begin{aligned} m_t &: \text{trend component;} \\ s_t &: \text{seasonal component;} \\ Y_t &: \text{Zero-mean error.} \end{aligned}$$

The following are some zero-mean models:

Example 1.4. (**iid noise**) The simplest time series model is the one with no trend or seasonal component, and the observations X_ts are simply independent and identically distribution random variables with zero mean. Such a sequence of random variable $\{X_t\}$ is referred to as **iid noise**. Mathematically, for any t and $x_1, \ldots, x_t$,

$$P(X_1 \leq x_1, \ldots, X_t \leq x_t) = \prod_t P(X_t \leq x_t) = \prod_t F(x_t),$$

where $F(\cdot)$ is the cdf of each X_t. Further $\text{E}(X_t) = 0$ for all t. We denote such sequence as $X_t \sim \text{IID}(0, \sigma^2)$. **IID noise is not interesting for forecasting since** $X_t \mid X_1, \ldots, X_{t-1} = X_t$**.**

Example 1.5. (A binary {discrete} process, see Figure 1.4) As an example of iid noise, a binary process $\{X_t\}$is a sequence of iid random variables X_ts with

$$P(X_t = 1) = 0.5, \quad P(X_t = -1) = 0.5.$$

Example 1.6. (A continues process: Gaussian noise, see Figure 1.4) $\{X_t\}$ is a sequence of iid normal random variables with zero mean and σ^2 variance; i.e.,

$$X_t \sim \text{N}(0, \sigma^2) \text{ iid}$$

Example 1.7. (**Random walk**) The random walt $\{S_t, t = 0, 1, 2, \dots\}$ (starting at zero, $S_0 = 0$) is obtained by cumulatively summing (or "integrating") random variables; i.e., $S_0 = 0$ and

$$S_t = X_1 + \cdots + X_t, \quad \text{for } t = 1, 2, \dots,$$

where $\{X_t\}$ is iid noise (see Figure 1.4) with zero mean and σ^2 variance. Note that by differencing, we can recover X_t; i.e.,

$$\nabla S_t = S_t - S_{t-1} = X_t.$$

Further, we have

$$\mathrm{E}(S_t) = \mathrm{E}\left(\sum_t X_t\right) = \sum_t \mathrm{E}(X_t) = \sum_t 0 = 0; \quad \mathrm{Var}(S_t) = \mathrm{Var}\left(\sum_t X_t\right) = \sum_t \mathrm{Var}(X_t) = t\sigma^2.$$

```
set.seed(100); par(mfrow=c(2,2)); par(mar=c(4,4,2,.5))
t=seq(1,60,by=1); Xt1=rbinom(length(t),1,.5)*2-1
plot(t,Xt1,type="o",col="blue",xlab="t",ylab=expression(X[t]))
t=seq(1,60,by=1); Xt2=rnorm(length(t),0,1)
plot(t,Xt2,type="o",col="blue",xlab="t",ylab=expression(X[t]))
plot(c(0,t),c(0,cumsum(Xt1)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
plot(c(0,t),c(0,cumsum(Xt2)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
```

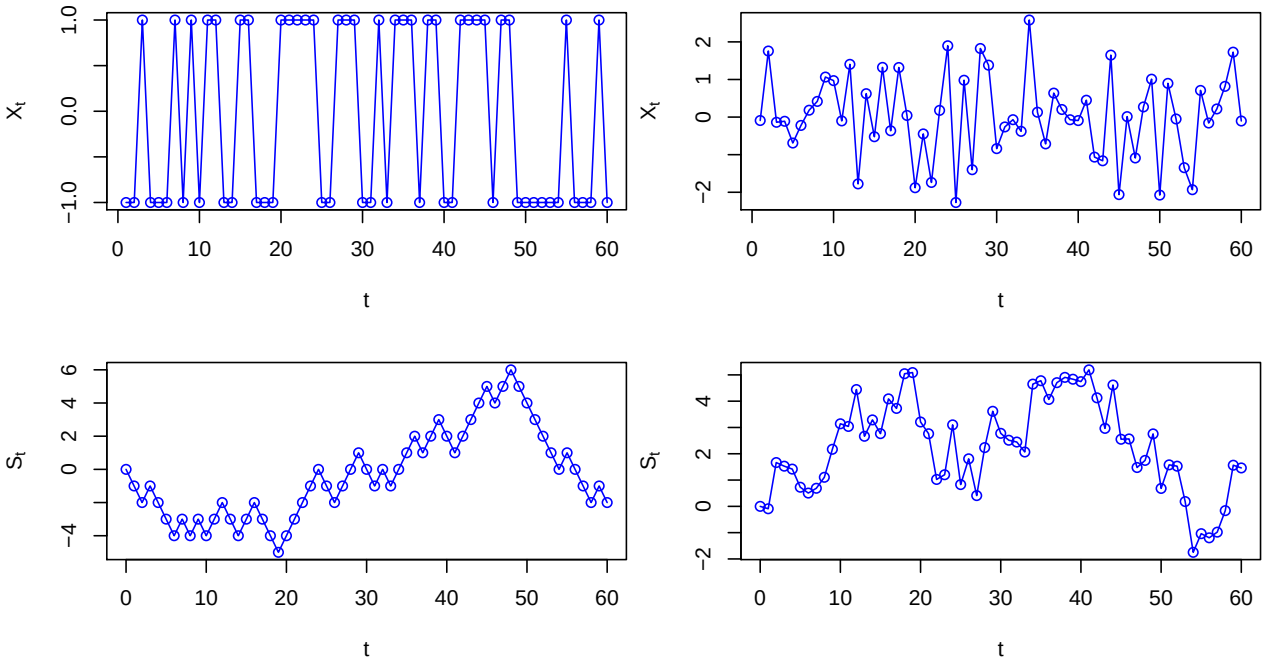


Figure 1.4: Top: One realization of a binary process (left) and a Gaussian noise (right). Bottom: the corresponding random walk

Example 1.8. (**white noise**) We say $\{X_t\}$ is a white noise; i.e., $X_t \sim \text{WN}(0, \sigma^2)$, if

$\{X_t\}$ is uncorrelated, i.e., $\text{Cov}(X_{t_1}, X_{t_2}) = 0$ for any t_1 and t_2, with $\text{E}X_t = 0$, $\text{Var}X_t = \sigma^2$.

Note that every $\text{IID}(0, \sigma^2)$ sequence is $\text{WN}(0, \sigma^2)$ but not conversely.

Example 1.9. (An example of white noise but not IID noise) Define $X_t = Z_t$ when t is odd, $X_t = \sqrt{3}Z_{t-1}^2 - 2/\sqrt{3}$ when t is even, where $\{Z_t, t = 1, 3, \dots\}$ is an iid sequence from distribution with pmt $f_Z(-1) = 1/3, f_Z(0) = 1/3, f_Z(1) = 1/3$. It can be seen that $E(X_t) = 0$, $\text{Var}(X_t) = 2/3$ for all t, $\text{Cov}(X_{t_1}, X_{t_2}) = 0$ for all t_1 and t_2, since

$$\text{Cov}(Z_t, \sqrt{3}Z_{t-1}^2 - 2/\sqrt{3}) = \sqrt{3}\text{Cov}(Z_t, Z_t^2) = 0.$$

However, $\{X_t\}$ is not an iid sequence. Since when Z_{2k} is determined fully by Z_{2k-1}.

$$Z_{2k-1} = 0 \Rightarrow Z_{2k} = -2/\sqrt{3},$$
$$Z_{2k-1} = \pm 1 \Rightarrow Z_{2k} = \sqrt{3} - 2/\sqrt{3}.$$

A realization of this white noise can be seen from Figure 1.5.

```
set.seed(100); par(mfrow=c(1,2)); par(mar=c(4,4,2,.5))
t=seq(1,100,by=1); res=c(-1,0,1)
Zt=sample(res,length(t)/2,replace=TRUE); Xt=c()
for(i in 1:length(Zt)){
Xt=c(Xt,c(Zt[i], sqrt(3)*Zt[i]^2-2/sqrt(3)))}
plot(t,Xt,type="o",col="blue",xlab="t",ylab=expression(X[t]))
plot(c(0,t),c(0,cumsum(Xt)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
```

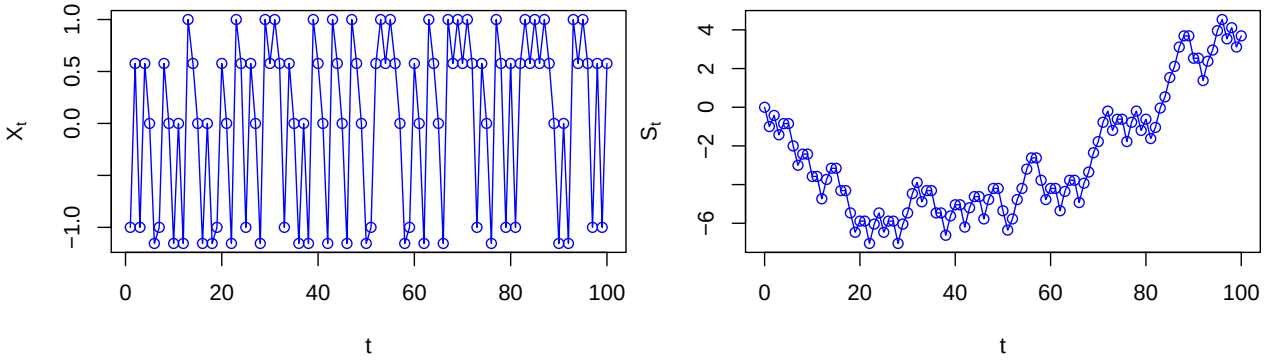


Figure 1.5: One realization of Example 1.9

If the stochastic behavior of all time series could be explained in terms of the white noise model, classical statistical methods would suffice. Two ways of introducing serial correlation and more smoothness into time series models are given in Examples 1.10 and 1.11.

Example 1.10. (Moving Averages Smoother) This is an essentially nonparametric method for trend estimation. It takes averages of observations around t; i.e., it smooths the series. For example, let

$$X_t = \frac{1}{3}(W_{t-1} + W_t + W_{t+1}), \tag{1.2}$$

which is a three-point moving average of the white noise series W_t. See Figure 1.9 for a realization. Inspecting the series shows a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster oscillations are taken out.

```
set.seed(100); w = rnorm(500,0,1) # 500 N(0,1) variates
v = filter(w, sides=2, rep(1/3,3)) # moving average
par(mfrow=c(2,1)); par(mar=c(4,4,2,.5))
plot.ts(w, main="white noise",col="blue")
plot.ts(v, main="moving average",col="blue")
```

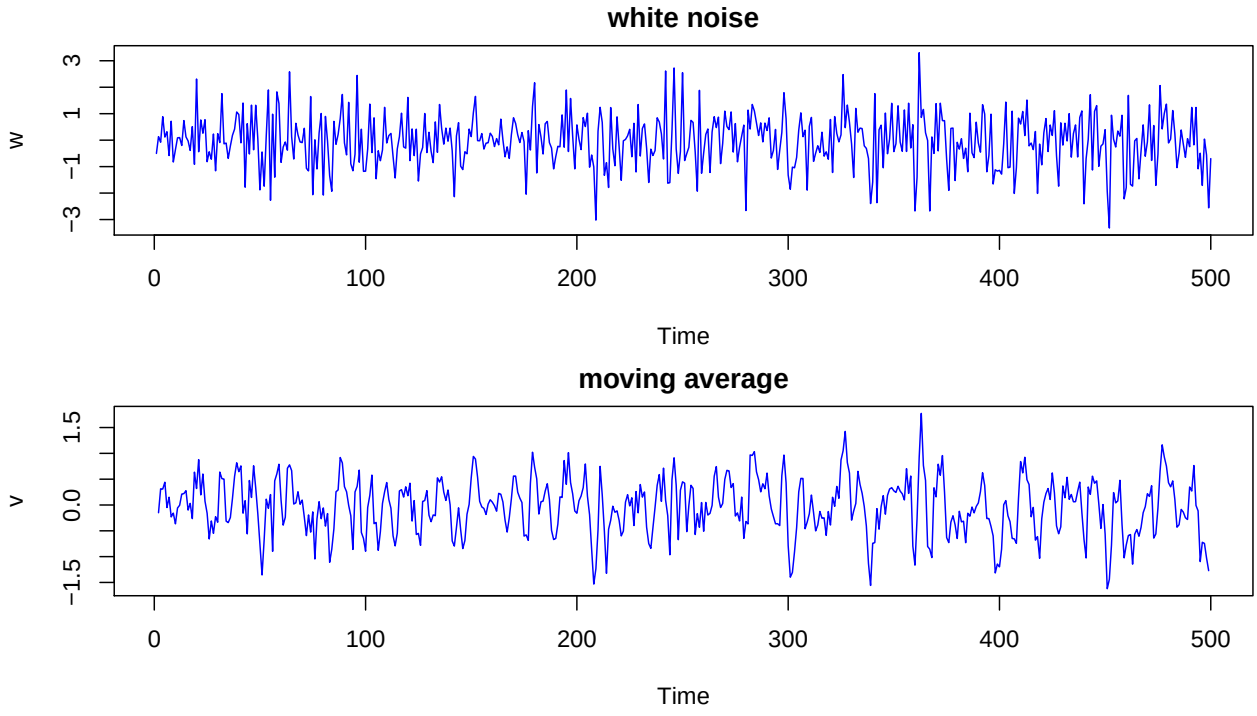


Figure 1.6: Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom).

Example 1.11. AR(1) model (Autoregression of order 1): Let

$$X_t = 0.6X_{t-1} + W_t \tag{1.3}$$

where W_t is a white noise series. It represents a regression or prediction of the current value X_t of a time series as a function of the past two values of the series.

```
set.seed(100); par(mar=c(4,4,2,.5))
w = rnorm(550,0,1) # 50 extra to avoid startup problems
x = filter(w, filter=c(.6), method="recursive")[-(1:50)]
plot.ts(x, main="autoregression",col="blue",ylab=expression(X[t]))
```

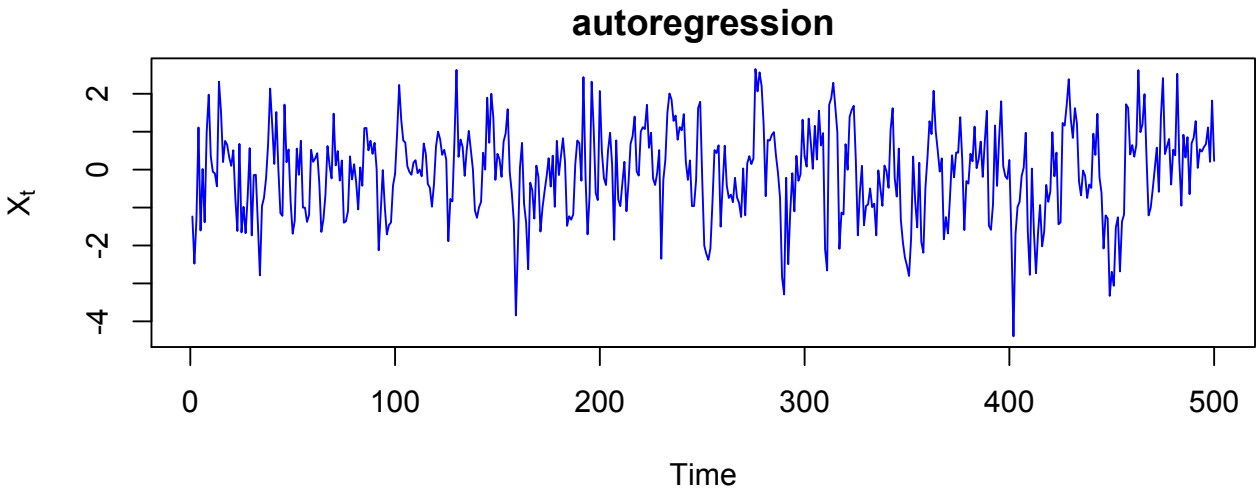


Figure 1.7: A realization of autoregression model (1.3)

Example 1.12. (Random Walk with Drift) Let

$$X_t = \delta + X_{t-1} + W_t \tag{1.4}$$

for $t = 1, 2, \ldots$ with $X_0 = 0$, where W_t is WN$(0, \sigma^2)$. The constant δ is called the drift, and when $\delta = 0$, we have X_t being simply a random walk (see Example 1.7, and see Figure 1.8 for a realization). X_t can also be rewritten as

$$X_t = \delta t + \sum_{j=1}^{t} W_j.$$

```
set.seed(150); w = rnorm(200,0,1); x = cumsum(w);
wd = w +.2; xd = cumsum(wd); par(mar=c(4,4,2,.5))
plot.ts(xd, ylim=c(-5,45), main="random walk",col="blue")
lines(x); lines(.2*(1:200), lty="dashed",col="blue")
```

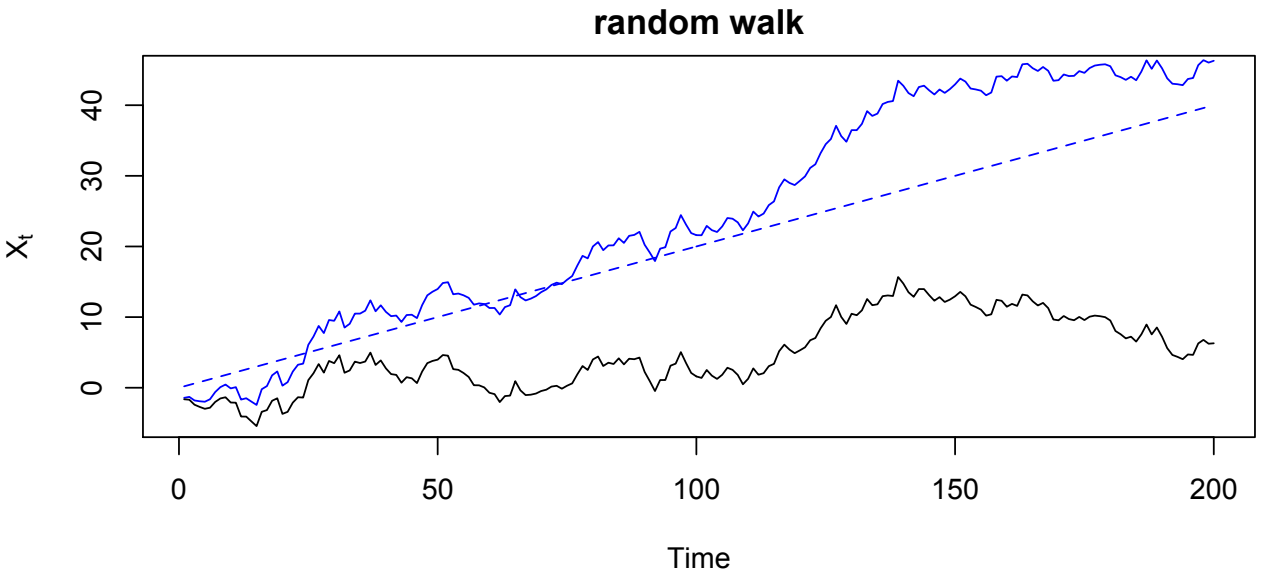


Figure 1.8: Random walk, $\sigma = 1$, with drift $\delta = 0.2$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and a straight line with slope .2 (dashed line).

Example 1.13. (Signal in Noise) Consider the model

$$X_t = 2\cos(2\pi t/50 + 0.6\pi) + W_t \tag{1.5}$$

for $t = 1, 2, \ldots$, where the first term is regarded as the signal, and $W_t \sim \text{WN}(0, \sigma^2)$. Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise. Note that, for any sinusoidal waveform,

$$A\cos(2\pi\omega t + \phi) \tag{1.6}$$

where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift.

```
set.seed(100); cs = 2*cos(2*pi*1:500/50 + .6*pi); w = rnorm(500,0,1)
par(mfrow=c(3,1), mar=c(3,2,2,1), cex.main=1.5)
plot.ts(cs, main=expression(2*cos(2*pi*t/50+.6*pi)),col="blue")
plot.ts(cs+w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,1)),col="blue")
plot.ts(cs+5*w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,25)),col="blue")
```

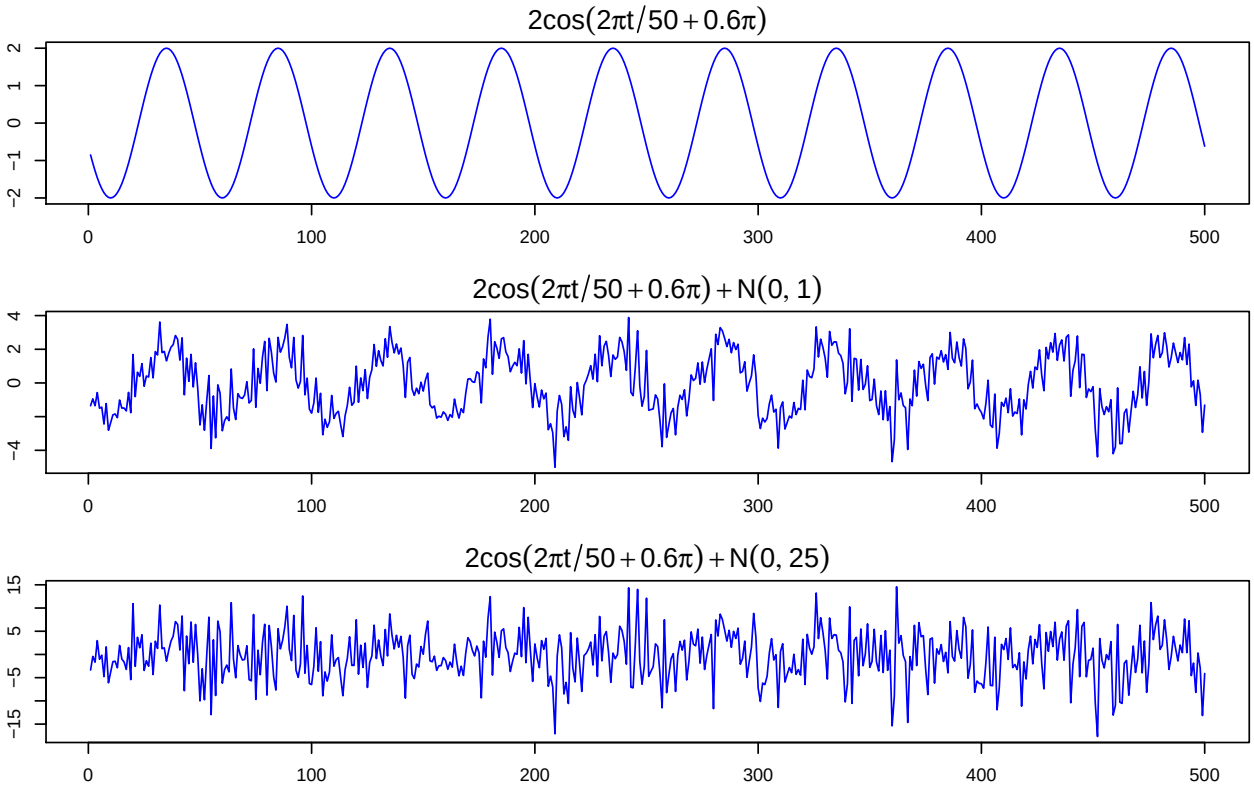


Figure 1.9: Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma = 1$ (middle panel) and $\sigma = 5$ (bottom panel).

2 Stationary Processes

2.1 Measure of Dependence

Denote the mean function of $\{X_t\}$ as

$$\mu_X(t) = \mathrm{E}(X_t),$$

provided it exists. And the **autocovariance function** of $\{X_t\}$ is

$$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{E}[\{X_s - \mu_X(s)\}\{X_t - \mu_X(t)\}]$$

Preliminary results of covariance and correlation: for any random variables X, Y and Z,

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \quad \text{and} \quad \mathrm{Corr}(X,Y) = \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

1. $-1 \le \rho_{XY} \le 1$ for any X and Y
2. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
3. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
4. $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$
5. $\mathrm{Cov}(a+X,Y) = \mathrm{Cov}(X,Y)$
6. If X and Y are independent, $\mathrm{Cov}(X,Y) = 0$
7. $\mathrm{Cov}(X,Y) = 0$ does not imply X and Y are independent
8. $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
9. $\mathrm{Cov}(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{Cov}(X_i, Y_j)$

Verify 1–9 as a **HW** problem.

The time series $\{X_t\}$ is **(weakly) stationary** if

1. $\mu_X(t)$ is independent of t;
2. $\gamma_X(t+h,h)$ is independent of t for each h.

We say $\{X_t\}$ is **strictly (or strongly) stationary** if

$$(X_{t_1}, \ldots, X_{t_k}) \text{ and } (X_{t_1+h}, \ldots, X_{t_k+h}) \text{ have the same joint distributions}$$

for all $k = 1, 2, \ldots$, $h = 0, \pm 1, \pm 2, \ldots$, and time points $t_1, \ldots, t_k$. This is a very strong condition.

Theorem 2.1. Basic properties of a strictly stationary time series $\{X_t\}$:

1. X_ts are from the same distribution.
2. $(X_t, X_{t+h}) =^d (X_1, X_{1+h})$ for all integers t and h.
3. $\{X_t\}$ is weakly stationary if $\mathrm{E}(X_t^2) < \infty$ for all t.
4. Weak stationary does not imply strict stationary.
5. An iid sequence is strictly stationary.

Proof. The proof is quite straightforward and thus left as a **HW** problem. □

Example 2.1. ($q-$dependent strictly stationary time series:) One of the simplest ways to construct a time series $\{X_t\}$ that is strictly stationary is to "filter" an iid sequence. Let $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$, define

$$X_t = g(Z_t, Z_{t-1}, \ldots, Z_{t-q})$$

for some real-valued function g. Then $\{X_t\}$ is strictly stationary and also $q-$dependent; i.e., X_s and X_t are independent whenever $|t - s| > q$.

A process, $\{X_t\}$ is said to be a **Gaussian process** if the n dimensional vector $\boldsymbol{X} = (X_{t_1}, \ldots, X_{t_n})$, for every collection of time points $t_1, \ldots, t_n$, and every positive integer n, have a multivariate normal distribution.

Lemma 2.1. For Gaussian processes, weakly stationary is equivalent to strictly stationary.

Proof. It suffices to show that every weakly stationary Gaussian process $\{X_t\}$ is strictly stationary. Suppose it is not, then there must exists $(t_1, t_2)^{\mathrm{T}}$ and $(t_1 + h, t_2 + h)^{\mathrm{T}}$ such that $(X_{t_1}, X_{t_2})^{\mathrm{T}}$ and $(X_{t_1+h}, X_{t_2+h})^{\mathrm{T}}$ have different distributions, which contradicts the assumption of weakly stationary. □

In this following, unless indicated specifically, ***stationary* always refers to *weakly stationary***. Note, when $\{X_t\}$ is stationary, $r_X(t+h, h)$ can be written as $\gamma_X(h)$ for simplicity since $\gamma_X(t+h, h)$ does not depend on t for any given h.

Let $\{X_t\}$ be a stationary time series. Its mean is $\mu_X = \mu_X(t)$. Its **autocovariance function (ACVF)** of $\{X_t\}$ at lag h is

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t).$$

Its **autocorrelation function (ACF)** of $\{X_t\}$ at lag h is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Corr}(X_{t+h}, X_t)$$

Theorem 2.2. Basic properties of $\gamma_X(\cdot)$:

1. $\gamma_X(0) \geq 0$;
2. $|\gamma_X(h)| \leq \gamma(0)$ for all h;
3. $\gamma_X(h) = \gamma_X(-h)$ for all h;
4. γ_X is nonnegative definite; i.e., a real valued function K defined on the integers is **nonnegative definite** if and only if
$$\sum_{i,j=1}^{n} a_i K(i-j) a_j \geq 0$$
for all positive integers n and real vectors $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$.

Proof. The first one is trivial since $\gamma_X(0) = \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t) \geq 0$ for all t. The second is based on the Cauchy-Schwarz inequality:

$$|\gamma_X(h)| = |\mathrm{Cov}(X_{t+h}, X_t)| \leq \sqrt{\mathrm{Var}(X_{t+h})}\sqrt{\mathrm{Var}(X_t)} = \gamma_X(0).$$

The third one is established by observing that

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_t, X_{t+h}) = \gamma_X(-h).$$

The last statement can be verified by

$$0 \leq \mathrm{Var}(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{X}_n) = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{\Gamma}_n \boldsymbol{a} = \sum_{i,j=1}^{n} a_i \gamma_X(i-j) a_j$$

where $\boldsymbol{X}_n = (X_n, \ldots, X_1)^{\mathrm{T}}$ and

$$\boldsymbol{\Gamma}_n = \mathrm{Var}(\boldsymbol{X}_n) = \begin{pmatrix} \mathrm{Cov}(X_n, X_n) & \mathrm{Cov}(X_n, X_{n-1}) & \cdots & \mathrm{Cov}(X_n, X_2) & \mathrm{Cov}(X_n, X_1) \\ \mathrm{Cov}(X_{n-1}, X_n) & \mathrm{Cov}(X_{n-1}, X_{n-1}) & \cdots & \mathrm{Cov}(X_{n-1}, X_2) & \mathrm{Cov}(X_{n-1}, X_1) \\ & & \vdots & & \\ \mathrm{Cov}(X_2, X_n) & \mathrm{Cov}(X_2, X_{n-1}) & \cdots & \mathrm{Cov}(X_2, X_2) & \mathrm{Cov}(X_2, X_1) \\ \mathrm{Cov}(X_1, X_n) & \mathrm{Cov}(X_1, X_{n-1}) & \cdots & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_1) \end{pmatrix}$$
$$= \begin{pmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(n-2) & \gamma_X(n-1) \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(n-3) & \gamma_X(n-2) \\ & & \vdots & & \\ \gamma_X(n-2) & \gamma_X(n-3) & \cdots & \gamma_X(0) & \gamma_X(1) \\ \gamma_X(n-1) & \gamma_X(n-2) & \cdots & \gamma_X(1) & \gamma_X(0) \end{pmatrix}$$

□

Remark 2.1. An autocorrelation function $\rho(\cdot)$ has all the properties of an autocovariance function and satisfies the additional condition $\rho(0) = 1$.

Theorem 2.3. A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and non-negative definite.

Proof. We only need prove that for any even and non-negative definite $K(\cdot)$, we can find a stationary process $\{X_t\}$ such that $\gamma_X(h) = K(h)$ for any integer h. It is quite trivial to choose $\{X_t\}$ to be a Gaussian process such that $\mathrm{Cov}(X_i, X_j) = K(i-j)$ for any i and j. □

2.1.1 Examples

Example 2.2. Consider

$$\{X_t = A\cos(\theta t) + B\sin(\theta t)\}$$

where A and B are two uncorrelated random variables with zero means and unit variances with $\theta \in [-\pi, \pi]$. Then

$$\mu_X(t) = 0$$

$$\begin{aligned}
\gamma_X(t+h, t) &= \mathrm{E}(X_{t+h}X_t) \\
&= \mathrm{E}[\{A\cos(\theta t + \theta h) + B\sin(\theta t + \theta h)\}\{A\cos(\theta t) + B\sin(\theta t)\}] \\
&= \cos(\theta t + \theta h)\cos(\theta t) + \sin(\theta t + \theta h)\sin(\theta t) \\
&= \cos(\theta t + \theta h - \theta t) = \cos(\theta h)
\end{aligned}$$

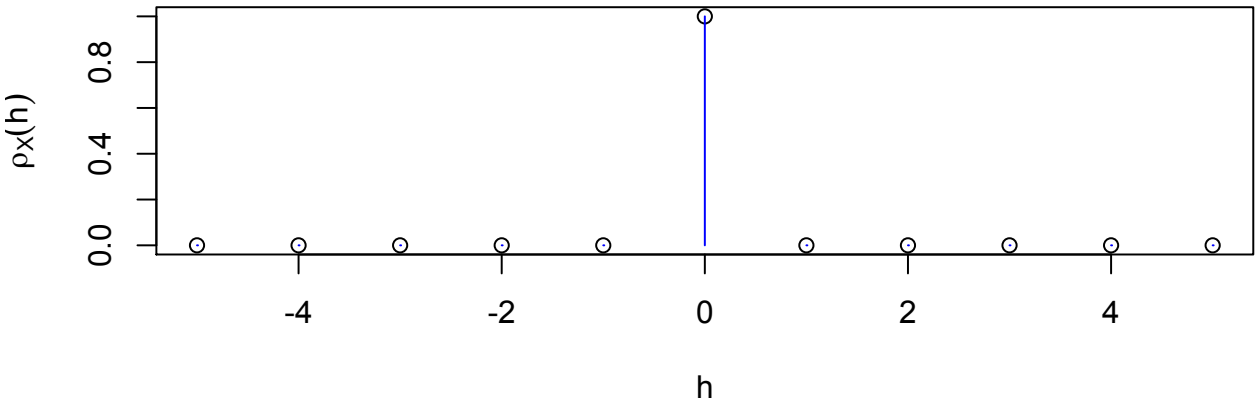
which is free of h. Thus $\{X_t\}$ is a stationary process. Further

$$\rho_X(h) = \cos(\theta h)$$

Example 2.3. For white noise $\{W_t\} \sim \text{WN}(0, \sigma^2)$, we have

$$\mu_W = 0, \quad \gamma_W(h) = \begin{cases} \sigma^2 & \text{if } h = 0; \\ 0 & \text{otherwise,} \end{cases}, \quad \rho_W(h) = \begin{cases} 1 & \text{if } h = 0; \\ 0 & \text{otherwise,} \end{cases}$$

```
rho=function(h,theta){I(h==0)*1}
h=seq(-5,5,1); s=1:length(h); y=rho(h,.6)
plot(h,y,xlab="h",ylab=expression(rho[X](h)))
segments(h[s],y[s],h[s],0,col="blue")
```



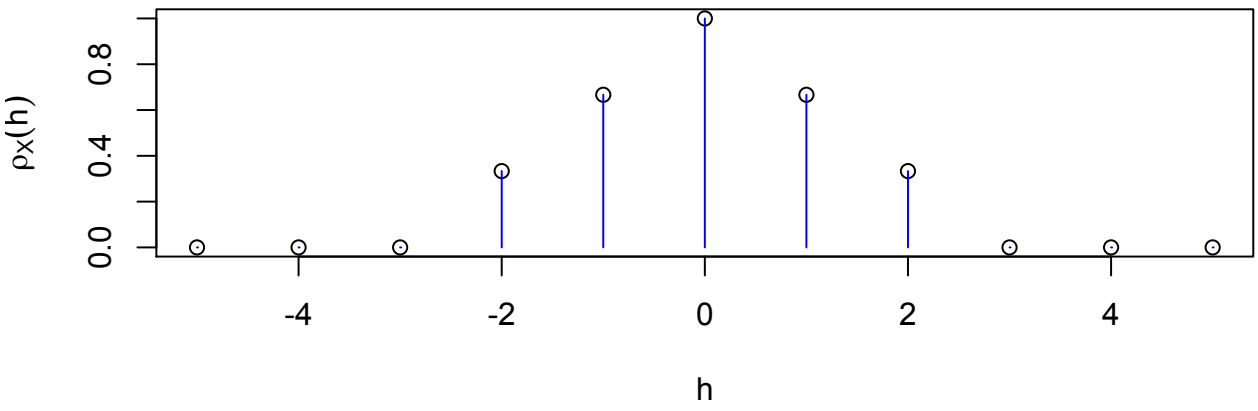
Example 2.4. (Mean Function of a three-point Moving Average Smoother). See Example 1.10, we have $X_t = 3^{-1}(W_{t-1} + W_t + W_{t+1})$, where $\{W_t\} \sim \text{WN}(0, \sigma^2)$. Then

$$\begin{aligned}\mu_X(t) =& \text{E}(X_t) = \frac{1}{3}[\text{E}(W_{t-1}) + \text{E}(W_t) + \text{E}(W_{t+1})] = 0,\\ \gamma_X(t+h,t) =& \frac{3}{9}\sigma^2 I(h=0) + \frac{2}{9}\sigma^2 I(|h|=1) + \frac{1}{9}\sigma^2 I(|h|=2)\end{aligned}$$

does not depend on t for any h. Thus, $\{X_t\}$ is stationary. Further

$$\rho_X(h) = I(h=0) + \frac{2}{3}I(|h|=1) + \frac{1}{3}I(|h|=2).$$

```
rho=function(h,theta){I(h==0)+2/3*I(abs(h)==1)+1/3*I(abs(h)==2)};
h=seq(-5,5,1); s=1:length(h); y=rho(h,.6);
plot(h,y,xlab="h",ylab=expression(rho[X](h))); segments(h[s],y[s],h[s],0,col="blue")
```



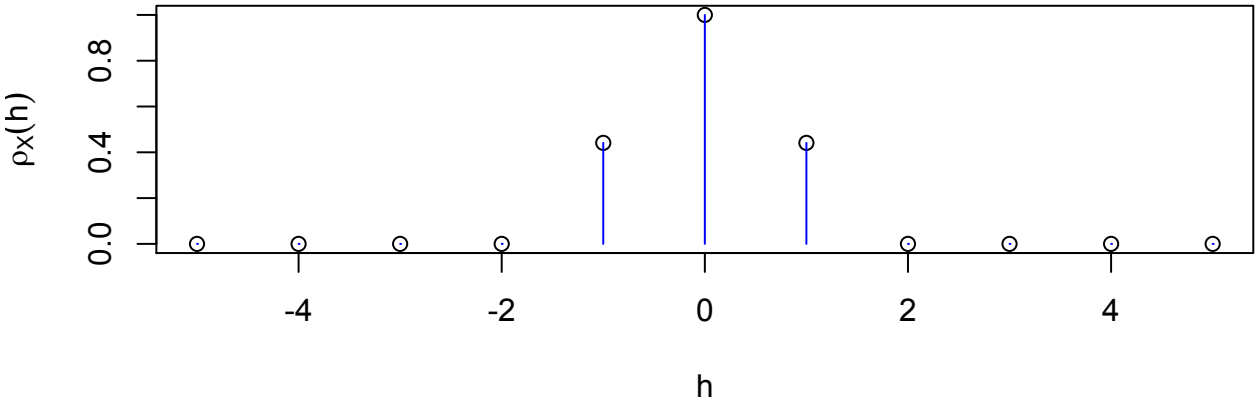
Example 2.5. MA(1) process (First-order moving average):

$$X_t = W_t + \theta W_{t-1}, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$ and θ is a constant. Then

$$\begin{aligned}
\mu_X(t) =& 0 \\
\gamma_X(h) =& \sigma^2(1+\theta^2) I(h=0) + \theta\sigma^2 I(|h|=1) \\
\rho_X(h) =& I(h=0) + \frac{\theta}{1+\theta^2} I(|h|=1).
\end{aligned}$$

```
rho=function(h,theta){I(h==0)+theta/(1+theta^2)*I(abs(h)==1)}
h=seq(-5,5,1); s=1:length(h); y=rho(h,.6)
plot(h,y,xlab="h",ylab=expression(rho[X](h))); segments(h[s],y[s],h[s],0,col="blue")
```



Example 2.6. AR(1) model (Autoregression of order 1). Consider the following model:

$$X_t = \phi X_{t-1} + W_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$ and W_t is uncorrelated with X_s for $s < t$. Assume that $\{X_t\}$ is stationary and $0 < |\phi| < 1$, we have

$$\mu_X = \phi \mu_X \Rightarrow \mu_X = 0$$

Further for $h > 0$

$$\begin{aligned}\gamma_X(h) =& \text{E}(X_t X_{t-h}) = \text{E}(\phi X_{t-1} X_{t-h} + W_t X_{t-h}) \\ =& \phi \text{E}(X_{t-1} X_{t-h}) + 0 = \phi \text{Cov}(X_{t-1} X_{t-h}) \\ =& \phi \gamma_X(h-1) = \cdots = \phi^h \gamma_X(0).\end{aligned}$$

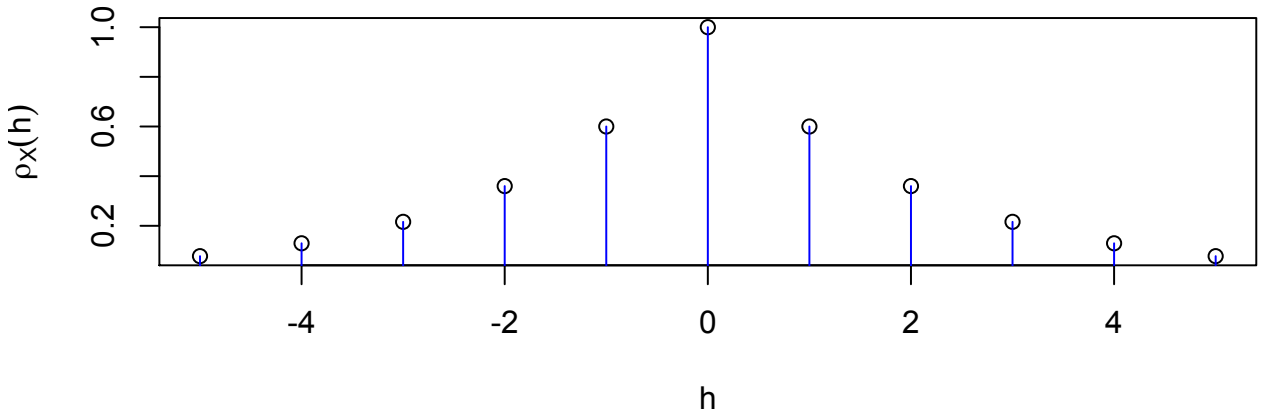
And

$$\gamma_X(0) = \text{Cov}(\phi X_{t-1} + W_t, \phi X_{t-1} + W_t) = \phi^2 \gamma_X(0) + \sigma^2 \Rightarrow \gamma_X(0) = \frac{\sigma^2}{1 - \phi^2}.$$

Further, we have $\gamma_X(h) = \gamma_X(-h)$, and

$$\rho_X(h) = \phi^{|h|}.$$

```
rho=function(h,phi){phi^(abs(h))}
h=seq(-5,5,1); s=1:length(h); y=rho(h,.6)
plot(h,y,xlab="h",ylab=expression(rho[X](h))); segments(h[s],y[s],h[s],0,col="blue")
```



Example 2.7. (Mean Function of a Random Walk with Drift). See Example 1.12, we have

$$X_t = \delta t + \sum_{j=1}^{t} W_j, \quad t = 1, 2, \ldots,$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$Then

$$\mu_X(t) = \text{E}(X_t) = \delta t.$$

Obviously, when δ is not zero, $\{X_t\}$ is not stationary, since its mean is not a constant. Further, if $\delta = 0$,

$$\begin{aligned}\gamma_X(t+h, t) =& \text{Cov}\left\{\sum_{j=1}^{t+h} W_j, \sum_{j=1}^{t} W_j\right\} \\ =& \min\{t+h, t\}\sigma^2\end{aligned}$$

is, again, not free of t. Thus $\{X_t\}$ is not stationary for any δ.

Example 2.8. The MA(q) Process: $\{X_t\}$ is a **moving-average process of order** q if

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$ and $\theta_1, \ldots, \theta_q$ are constants. We have

$$\begin{aligned}\mu_X(t) =& 0 \\ \gamma_X(h) =& \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} I(|h| \leq q).\end{aligned}$$

Proposition 2.1. If $\{X_t\}$ is a stationary $q-$correlated time series (i.e., $\text{Cov}(X_s, X_t) = 0$ whenever $|s-t| > q$) with mean 0, then it can be represented as an MA(q) process.

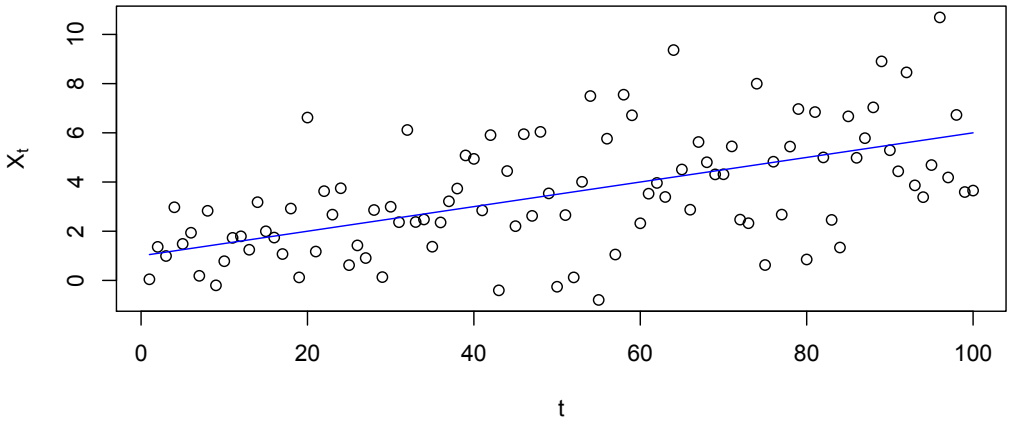
Proof. See Proposition 3.2.1 on page 89 of Brockwell and Davis (2009, Time Series Theory and Methods). □

2.1.2 Identify Non-Stationary Time Series

After learning all the above stationary time series, one question would naturally arise is that, what kind of time series is not stationary? Plotting the data would always be helpful to identify the stationarity of your time series data.

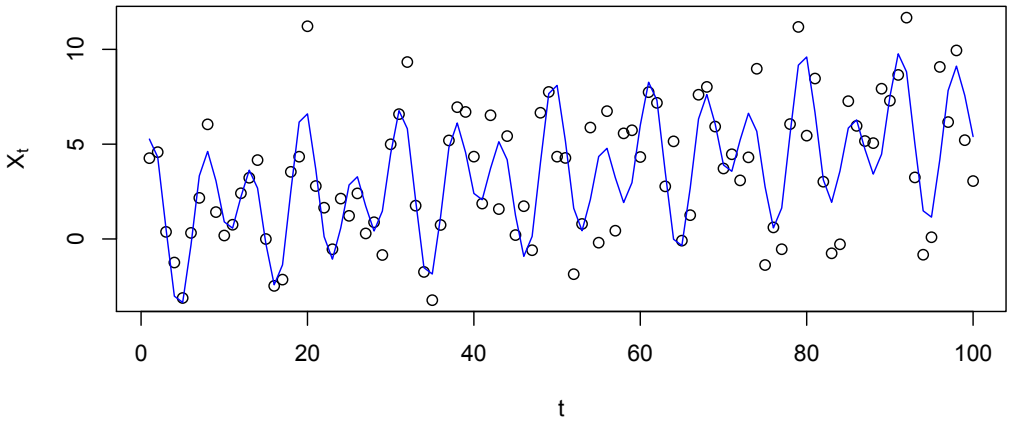
- Any time series with non-constant trend is not stationary. For example, if $X_t = m_t + Y_t$ with trend m_t and zero-mean error Y_t. Then $\mu_X(t) = m_t$ is not a constant. For example, the following figure plots a realization of $X_t = 1 + 0.5t + Y_t$, where $\{Y_t\} \sim \mathrm{N}(0, 1)$ iid.

```
set.seed(100); par(mar=c(4,4,2,.5))
t=seq(1,100,1); Tt=1+.05*t; Xt=Tt+rnorm(length(t),0,2)
plot(t,Xt,xlab="t",ylab=expression(X[t])); lines(t,Tt,col="blue")
```



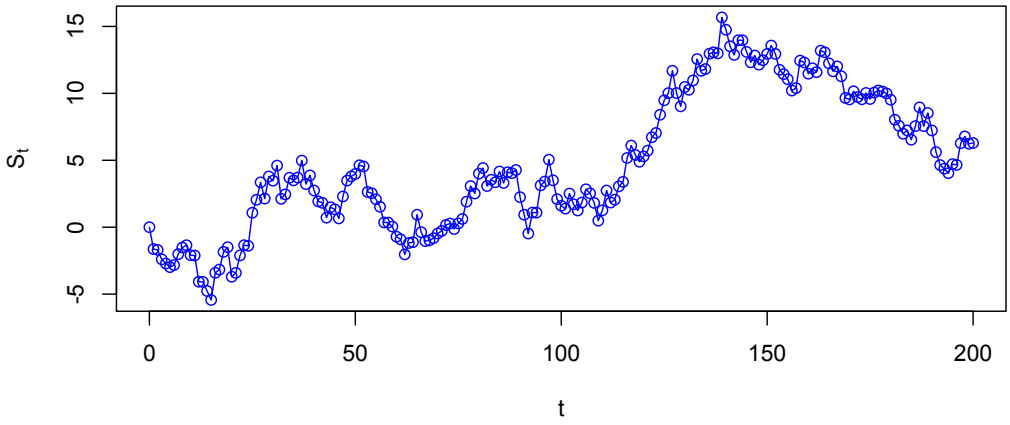
- Any time series with seasonal trend is not stationary. For example, if $X_t = s_t + Y_t$ with seasonal trend s_t and zero-mean error Y_t. Then $\mu_X(t) = s_t$ is not a constant. For example, the following figure plots a realization of $X_t = 1 + 0.5t + 2\cos(\pi t/5) + 3\sin(\pi t/3) + W_t$, where $\{Y_t\} \sim \text{N}(0, 1)$ iid.

```
set.seed(100); par(mar=c(4,4,2,.5)); t=seq(1,100,1); Tt=1+.05*t;
St=2*cos(pi*t/5)+3*sin(pi*t/3); Xt=Tt+St+rnorm(length(t),0,2)
plot(t,Xt,xlab="t",ylab=expression(X[t])); lines(t,Tt+St,col="blue")
```



- Any time series with non-constant variance is not stationary. For example, random walk $\{S_t = \sum_{j=1}^{t} X_t\}$ with X_t being iid $\text{N}(0, 1)$.

```
set.seed(150); par(mar=c(4,4,2,.5)); t=seq(1,200,by=1); Xt1=rnorm(length(t),0,1)
plot(c(0,t),c(0,cumsum(Xt1)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
```



Another way you may have already figured out by yourself of identifying stationarity is based on the shape of the autocorrelation function (ACF). However, in applications, you can never know the true ACF. Thus, a sample version of it could be useful. In the following, we produce the estimators of μ_X, ACVF, and ACF. Later, we will introduce the asymptotic properties of these estimators.

For observations $x_1, \ldots, x_n$ of a time series, the **sample mean** is

$$\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t.$$

The **sample auto covariance function** is

$$\widehat{\gamma}_X(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}), \quad \text{for } -n < h < n.$$

This is like the sample covariance of $(x_1, x_{h+1}), \ldots, (x_{n-h}, x_n)$, except that

- we normalize it by n instead of $n - h$,
- we subtract the full sample mean.

This setting ensures that the sample covariance matrix $\widehat{\mathbf{\Gamma}}_n = [\widehat{\gamma}_X(i-j)]_{i,j=1}^n$ is nonnegative definite.
The **sample autocorrelation function (sample ACF)** is

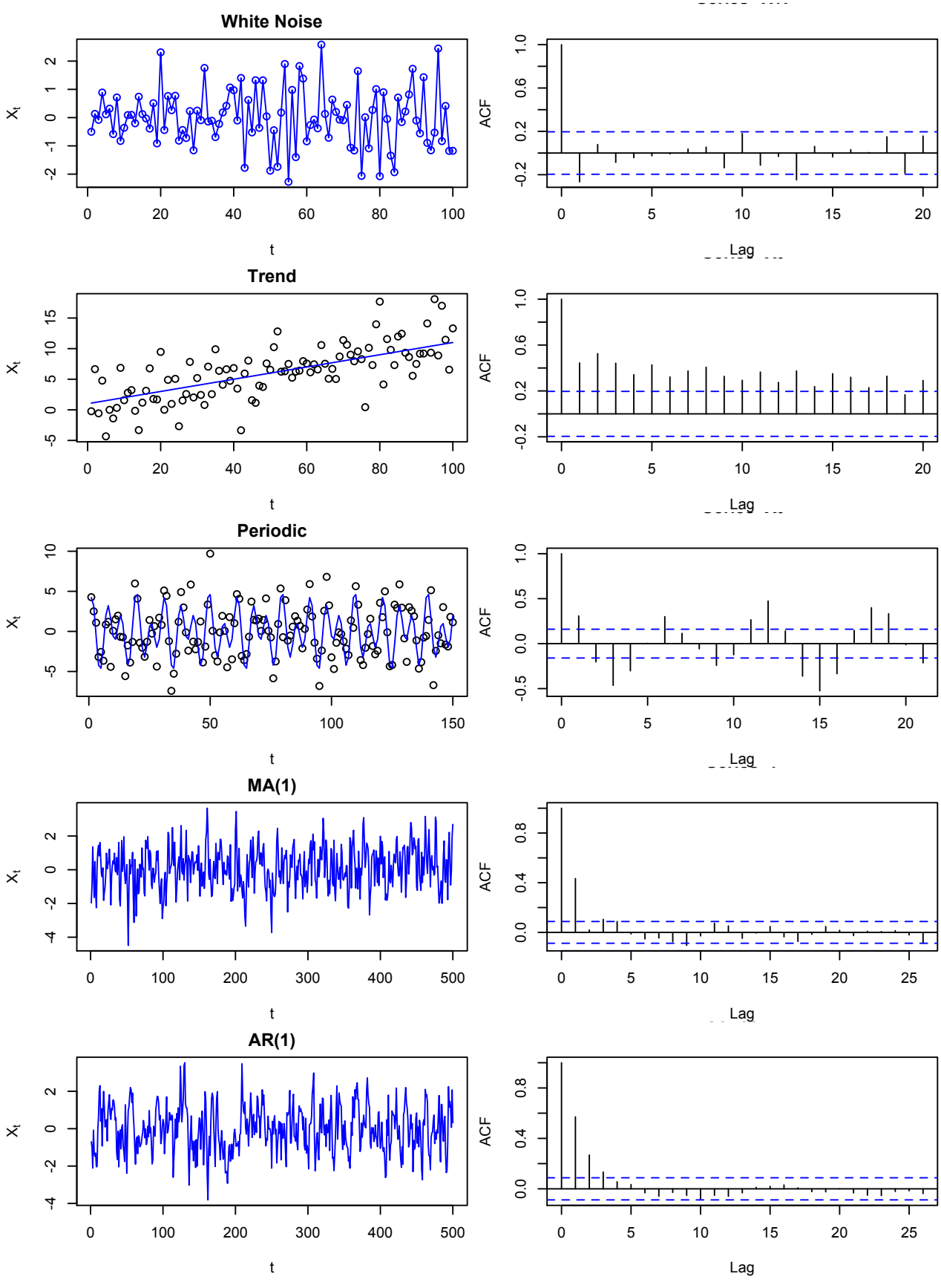
$$\widehat{\rho}_X(h) = \frac{\widehat{\gamma}_X(h)}{\widehat{\gamma}_X(0)}, \quad \text{for } -n < h < n.$$

The sample ACF can help us recognize many non-white (even non-stationary) time series.

Some guidelines:

Time series:	**Sample ACF**
White noise	Zero for $\|h\| > 0$
Trend	Slow decay
Periodic	Periodic
MA(q)	Zero for $\|h\| > q$
AR(1)	Decays to zero exponentially

```
set.seed(100);
par(mfrow=c(5,2))
par(mar=c(4,4,2,.5))
     #White Noise
WN=rnorm(100,0,1);
plot(1:n,WN,type="o",col="blue",main="White Noise",ylab=expression(X[t]),xlab="t");
acf(WN)
     #Trend
t=seq(1,100,1);
Tt=1+.1*t;
Xt=Tt+rnorm(length(t),0,4)
plot(t,Xt,xlab="t",ylab=expression(X[t]),main="Trend")
lines(t,Tt,col="blue")
acf(Xt)
     #Periodic
t=seq(1,150,1)
St=2*cos(pi*t/5)+3*sin(pi*t/3)
Xt=St+rnorm(length(t),0,2)
plot(t,Xt,xlab="t",ylab=expression(X[t]), main="Periodic")
lines(t,St,col="blue")
acf(Xt)
     #MA(1)
w = rnorm(550,0,1)
v = filter(w, sides=1, c(1,.6))[-(1:50)]
plot.ts(v, main="MA(1)",col="blue",ylab=expression(X[t]),xlab="t")
acf(v)
     #AR(1)
w = rnorm(550,0,1)
x = filter(w, filter=c(.6), method="recursive")[-(1:50)]
plot.ts(x, main="AR(1)",col="blue",ylab=expression(X[t]),xlab="t")
acf(x)
```


White Noise
Trend
Periodic
MA(1)
AR(1)
X_t
t
ACF
Lag

The typical procedure of time series modeling can be described as

1. Plot the time series (look for trends, seasonal components, step changes, outliers).
2. Transform data so that residuals are **stationary**.
   (a) Estimate and subtract m_t, s_t.
   (b) Differencing
   (c) Nonlinear transformations ($\log$, $\sqrt{\cdot}$)
3. Fit model to residuals.

Now, we introduce the difference operator ∇ and the backshift operator B.

- Define the lag-1 **difference operator**: (think 'first derivative')

$$\nabla X_t = \frac{X_t - X_{t-1}}{t-(t-1)} = X_t - X_{t-1} = (1-B)X_t$$

  where B is the **backshift** operator, $BX_t = X_{t-1}$.

- Define the lag-s difference operator,

$$\nabla_s X_t = X_t - X_{t-s} = (1-B^s)X_t,$$

  where B^s is the backshift operator applied s times, $B^s X_t = B(B^{s-1}X_t)$ and $B^1 X_t = BX_t$.

Note that

- If $X_t = \beta_0 + \beta_1 t + Y_t$, then
$$\nabla X_t = \beta_1 + \nabla Y_t.$$
- if $X_t = \sum_{i=0}^{k} \beta_i t^i + Y_t$, then
$$\nabla^k X_t = k!\beta_k + \nabla^k Y_t,$$
  where $\nabla^k X_t = \nabla(\nabla^{k-1}X_t)$ and $\nabla^1 X_t = \nabla X_t$.
- if $X_t = m_t + s_t + Y_t$ and s_t has period s (i.e., $s_t = s_{t-s}$ for all t), then
$$\nabla_s X_t = m_t - m_{t-s} + \nabla_s Y_t.$$

2.2 Linear Processes

Every second-order stationary process is either a linear process or can be transformed to a linear process by subtracting a deterministic component, which will be discussed later.

The time series $\{X_t\}$ is a **linear process** if it has the representation

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j},$$

for all t, where $\{W_t\} \sim \text{WN}(0, \sigma^2)$, μ is a constant, and $\{\psi_j\}$ is a sequence of constants with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. if we define $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$, then the linear process $X_t = \mu + \psi(B)W_t$.

Note that the condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ ensures that X_t is meaningful; i.e., $|X_t| < \infty$ almost surely. Since $\text{E}|W_t| \leq \sigma$ for all t and

$$\begin{aligned} P(|X_t| \geq \alpha) \leq & \frac{1}{\alpha}\text{E}|X_t| \leq \frac{1}{\alpha}\left(|\mu| + \sum_{j=-\infty}^{\infty} |\psi_j|\text{E}|W_{t-j}|\right) \\ \leq & \frac{1}{\alpha}\left(|\mu| + \sigma \sum_{j=-\infty}^{\infty} |\psi_j|\right) \to 0 \quad \text{as } \alpha \to 0. \end{aligned}$$

Before proceeding, we provide a brief introduction of several types of convergence in statistics. We say a sequence of random variables X_n converges in mean square to a random variable X (denoted by $X_n \xrightarrow{L^2} X$) if

$$E(X_n - X)^2 \to 0 \quad \text{as } n \to \infty.$$

More generally, we have convergence in r-th mean, denoted by $X_n \xrightarrow{L^r} X$, if

$$E(|X_n - X|^r) \to 0 \quad \text{as } n \to \infty.$$

Also, we say that X_n converges in probability to X, denoted by $X_n \xrightarrow{p} X$, if

$$P\{|X_n - X| > a\} \to 0, \quad \text{for all } a > 0, \text{ as } n \to \infty.$$

X_n converges in distribution to X, denoted by $X_n \xrightarrow{p} X$, if

$$F_n(X) \to F(X) \quad \text{as } n \to \infty,$$

at the continuity points of $F(\cdot)$. The last one is convergence almost surely denoted by $X_n \xrightarrow{a.s.} X$ (which will not be used in this course). The relationship between these convergences is, for $r > 2$,

$$\xrightarrow{L^r} \Rightarrow \xrightarrow{L^2} \Rightarrow \xrightarrow{p} \Rightarrow \xrightarrow{d}.$$

This course mainly focuses on convergence in mean square. One easy way to prove this convergence is through the use of the following theorem:

Theorem 2.4. (**Riesz-Fisher Theorem**, Cauchy criterion.) X_n converges in mean square if and only if

$$\lim_{m,n\to\infty} \mathrm{E}(X_m - X_n)^2 = 0.$$

Example 2.9. Linear process $X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$, then if $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, we have $\sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$ converges in mean square.

Proof. Defining $S_n = \sum_{j=-n}^{n} \psi_j W_{t-j}$, we have

$$\begin{aligned} E(S_m - S_n)^2 =& E\left(\sum_{m\le j\le n} \psi_j W_{t-j}\right)^2 \\ =& \sum_{m\le j\le n} \psi_j^2\sigma^2 \le \sigma^2 \left(\sum_{m\le j\le n} |\psi_j|\right)^2 \to 0 \quad \text{as } m,n\to\infty. \end{aligned}$$

□

Lemma 2.2. Linear process $\{X_t\}$ defined above is stationary with

$$\mu_X = \mu \tag{2.1}$$

$$\gamma_X(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j. \tag{2.2}$$

Proof. Equation (2.1) is trivial. For (2.2), we have

$$\begin{aligned} \gamma_X(h) =& \mathrm{E}\left[\left(\sum_{j=-\infty}^{\infty} \psi_j W_{t+h-j}\right)\left(\sum_{j=-\infty}^{\infty} \psi_j W_{t-j}\right)\right] \\ =& \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j\psi_k \mathrm{E}(W_{t+h-j}W_{t-k}) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j\psi_k\gamma_W(h-j+k) \\ =& \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j\psi_k I(k=j-h)\sigma^2 = \sigma^2\sum_{j=-\infty}^{\infty} \psi_j\psi_{j-h} = \sigma^2\sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j. \end{aligned}$$

□

Proposition 2.2. Let $\{Y_t\}$ be a stationary time series with mean 0 and covariance function γ_Y. If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then the time series

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(B)Y_t$$

is stationary with mean 0 and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h - j + k)$$

It can be easily seen that white noise, MA(1), AR(1), MA(q) and MA(∞) are all special examples of linear processes.

- White noise: choose μ, and $\psi_j = I(j = 0)$, we have $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} = \mu + W_t \sim \text{WN}(\mu, \sigma^2)$.
- MA(1): choose $\mu = 0$, $\psi_j = I(j = 0) + \theta I(j = 1)$, we have $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} = W_t + \theta W_{t-1}$.
- AR(1): choose $\mu = 0$, $\psi_j = \phi^j I(j \geq 0)$, we have $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} = \sum_{j=0}^{\infty} \phi^j W_{t-j} = W_t + \phi \sum_{j=0}^{\infty} \phi^j W_{t-1-j} = W_t + \phi X_{t-1}$
- MA(q): choose $\mu = 0$, $\psi_j = I(j = 0) + \sum_{k=1}^{q} \theta_k I(j = k)$, we have $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}$.
- MA(∞): choose $\mu = 0$, $\psi_j = \sum_{k=0}^{\infty} \theta_k I(j = k)$, we have $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} = \sum_{j=0}^{\infty} \psi_j W_{t-j}$.

2.3 AR(1) and AR(p) Processes

2.3.1 AR(1) process

In this section, we provide closer investigation on the AR(1) process which has been briefly introduced in Example 2.6. An AR(1) process was defined in Example 2.6 as a stationary solution $\{X_t\}$ of the equations

$$X_t - \phi X_{t-1} = W_t, \quad \text{for all } t, \tag{2.3}$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$, and Z_t is uncorrelated with X_s for $s < t$.

- When $\phi = 0$, it is so trivial that $X_t = W_t$.
- When $0 < |\phi| < 1$, we now show that such a solution exists and is the unique stationary solution of (2.3). In the above, we have already shown that

$$X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}, \tag{2.4}$$

This can be found easily through the aid of $\phi(B) = 1 - \phi B$ and $\pi(B) = \sum_{j=0}^{\infty} \phi^j B^j$. We have

$$\begin{aligned} & X_t - \phi X_{t-1} = W_t \\ \Rightarrow\ & \pi(B)(X_t - \phi X_{t-1}) = \pi(B) W_t \\ \Rightarrow\ & \pi(B)\phi(B) X_t = \pi(B) W_t \\ \Rightarrow\ & X_t = \pi(B) W_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}. \end{aligned}$$

The last step is due to

$$\pi(B)\phi(B) = (1 - \phi B) \sum_{j=0}^{\infty} \phi^j B^j = \sum_{j=0}^{\infty} \phi^j B^j - \sum_{j=1}^{\infty} \phi^j B^j = 1$$

which is similarly to the summation of geometric series:

$$\sum_{j=0}^{\infty} \phi^j B^j = \sum_{j=0}^{\infty} (\phi B)^j = \frac{1}{1 - \phi B}.$$

It can be easily seen that it is stationary with mean 0 and ACVF $\gamma_X(h) = \sigma^2 \phi^h / (1 - \phi^2)$, which are the same as in Example 2.6. Further, we show this solution is unique. Suppose $\{Y_t\}$ is another stationary solution, then by iterating, we have

$$\begin{aligned} Y_t =& \phi Y_{t-1} + W_t \\ =& W_t + \phi W_{t-1} + \phi^2 Y_{t-2} \\ =& \cdots \\ =& W_t + \phi W_{t-1} + \cdots + \phi^k W_{t-k} + \phi^{k+1} Y_{t-k-1} \end{aligned}$$

Then

$$\mathrm{E}\left(Y_t - \sum_{j=0}^{k} \phi^j W_{t-j} \right)^2 = \phi^{2k+2} \mathrm{E}(Y_{t-k-1}^2) \to 0 \quad \text{as } k \to \infty.$$

This implies that Y_t is equal to the mean square limit $\sum_{j=0}^{\infty} \phi^j W_{t-j}$ and hence the uniqueness is proved.

- When $|\phi| > 1$, the series defined in (2.4) does not converge. However, we can rewrite (2.3) in the form

$$X_t = -\phi^{-1} W_{t+1} + \phi^{-1} X_{t+1}.$$

By iterating, we have

$$X_t = -\phi^{-1} W_{t+1} - \cdots - \phi^{-k-1} W_{t+k+1} + \phi^{-k-1} X_{t+k+1},$$

which shows that

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} W_{t+j}$$

is the unique stationary solution of (2.3). However, this is very hard to interpret, since X_t is defined to be correlated with *future* values of Z_s. Another way to look at this case is to define a new sequence

$$\begin{aligned} W_t^* =& X_t - \frac{1}{\phi} X_{t-1} = (\phi - \phi^{-1}) X_{t-1} + W_t = -(\phi - \phi^{-1}) \sum_{j=1}^{\infty} \phi^{-j} W_{t-1+j} + W_t \\ =& \frac{1}{\phi^2} W_t - (1 - \phi^{-2}) \sum_{j=1}^{\infty} \phi^{-j} W_{t+j} \end{aligned}$$

Standard arguments yields that (left as a **HW** problem)

$$\begin{aligned} \mathrm{E}(W_t^*) =& 0 \\ \gamma_{W^*}(h) =& \frac{\sigma^2}{\phi^2} I(h = 0); \end{aligned}$$

i.e., $\{W_t^*\}$ is a new white noise sequence with mean 0 and variance σ^2/ϕ^2, then we have a new AR(1) model

$$X_t = \phi^* X_{t-1} + W_t^*$$

with $|\phi^*| = 1/|\phi| < 1$. Thus, we can rewrite the unique stationary solution as

$$X_t = \sum_{j=0}^{\infty} \phi^{*j} W_{t-j}^*$$

which now does not depend on future values. Thus, for an AR(1) model, people typically assumes that $|\phi| < 1$.

- When $|\phi| = 1$. If there is a stationary solution to (2.3), check

$$\mathrm{Cov}(X_{t-1}, W_t) = \mathrm{Cov}(X_{t-1}, X_t - \phi X_{t-1}) = \gamma_X(1) - \phi \gamma_X(0) = 0$$

This holds if and only if $X_t = \phi X_{t-1} + b$ for some constant b. Then $\{W_t = b\}$ is a constant process. Since $\{W_t\}$ is a white noise, then b has to be zero. Now we have

$$X_t = \phi X_{t-1}.$$

When $\phi = -1$, X_t has to be all zeros. When $\phi = 1$, then X_t are all constants. So if we quire $\sigma > 0$, then there is no stationary solution; if more broadly, we allow $\sigma = 0$, i.e., $\{W_t = 0\}$, then when $\phi = -1$, there is a stationary solution which is $X_t = 0$; when $\phi = 1$, there is also a stationary solution that $X_t = \mu_X$. In the following, we require $\sigma > 0$.

Remark 2.2. This example introduced a very important terminology: **causality**. We say that $\{X_t\}$ is a **causal** function of $\{W_t\}$, or more concisely that $\{X_t\}$ is a **causal autoregressive process**, if X_t has a representation in terms of $\{W_s, s \le t\}$; i.e., the current status only relates to the past events, not the future.

A linear process $\{X_t\}$ is **causal** (strictly, **a causal function of** $\{W_t\}$), if there is a

$$\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdots$$

with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

$$X_t = \psi(B) W_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}.$$

- When $|\phi| < 1$, AR(1) process $\{X_t\}$ is a causal function of $\{W_t\}$.
- When $|\phi| > 1$, AR (1) process is not causal.

Proposition 2.3. AR(1) process $\phi(B)X_t = W_t$ with $\phi(B) = 1 - \phi B$ is causal if and only if $|\phi| < 1$ or the root z_1 of the polynomial $\phi(z) = 1 - \phi z$ satisfies $|z_1| > 1$.

2.3.2 AR(p) process

An **AR(p) process** $\{X_t\}$ is a stationary process that satisfies

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$. Equivalently, $\phi(B)X_t = W_t$ where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$.

Recall that, for $p = 1$, $\phi(B) = 1 - \phi_1 B$, and for this AR(1) model, X_t is stationary only if $|\phi_1| \neq 1$. This is equivalent to that for any $z \in \mathbb{R}$ such that $\phi(z) = 1 - \phi z$ satisfis $|z| \neq 1$, or

$$\text{for any } z \in \mathbb{C} \text{ such that } \phi(z) = 1 - \phi z \text{ satisfis } |z| \neq 1.$$

Now for the AR(p) model, similarly, we should have

$$\text{for any } z \in \mathbb{C} \text{ such that } \phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \text{ satisfis } |z| \neq 1.$$

Theorem 2.5. A (unique) stationary solution to $\phi(B)X_t = W_t$ exists if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0 \;\Rightarrow\; |z| \neq 1$$

Further, this AR(p) process is causal if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0 \;\Rightarrow\; |z| > 1 \tag{2.5}$$

When (2.5) is satisfied, based on causality, we can write

$$X_t = \psi(B)W_t$$

where $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdots$ for some ψ_js satisfying $\sum_{j=0}^{\infty} |\psi_j| < \infty$. The question is then, how to calculate ψ_js? One way is to matching the coefficients.

$$\begin{aligned}
&\phi(B)X_t = W_t \quad \text{and} \quad X_t = \psi(B)W_t \\
\Rightarrow\ & 1 = \psi(B)\phi(B) \\
\Leftrightarrow\ & 1 = (\psi_0 + \psi_1 B + \psi_2 B^2 + \cdots)(1 - \phi_1 B - \cdots - \phi_p B^p) \\
\Leftrightarrow\ & 1 = \psi_0, \\
& 0 = \psi_1 - \phi_1\psi_0, \\
& 0 = \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 \\
& \vdots \\
\Leftrightarrow\ & 1 = \psi_0, \quad 0 = \psi_j \quad (j < 0), \quad 0 = \phi(B)\psi_j \quad (j > 0).
\end{aligned}$$

2.4 MA(1) and MA(q) Processes

Now, we look at the MA(1) process defined in Example 2.5. An MA(1) process $\{X_t\}$ is defined as

$$X_t = W_t + \theta W_{t-1}$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$. Obviously, $\{X_t\}$ is a causal function of $\{W_t\}$. But more importantly is about another terminology: **invertibility**. Just as causality means that X_t is expressible in terms of $\{W_s, s \le t\}$, **invertibility** means that W_t is expressible in terms of $\{X_s, s \le t\}$.

A linear process $\{X_t\}$ is **invertible** (strictly, **a invertible function of** $\{W_t\}$), if there is a

$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \cdots$$

with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ such that

$$W_t = \pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

Obviously, AR(1) process is invertible. Back to the MA(1) process:

- When $|\theta| < 1$, we have

$$(1 + \theta B)^{-1} = \sum_{j=0}^{\infty} (-\theta)^j B^j$$

Thus

$$\begin{aligned} & X_t = W_t + \theta W_{t-1} = (1+\theta B) W_t \\ \Rightarrow & (1+\theta B)^{-1} X_t = W_t \\ \Leftrightarrow & W_t = \sum_{j=0}^{\infty} (-\theta)^j X_{t-j}. \end{aligned}$$

We have $\{X_t\}$ as a invertible function of $\{W_t\}$.

- when $|\theta| > 1$, the sum $\sum_{j=0}^{\infty}(-\theta)^j X_{t-j}$ diverges, but we can write

$$\begin{aligned} W_t = & -\theta^{-1} W_{t+1} + \theta^{-1} X_{t+1} \\ = & \theta^{-1} X_{t+1} - \theta^{-2} X_{t+2} + \theta^{-2} W_{t+2} \\ = & \cdots = -\sum_{j=1}^{\infty} (-\theta)^{-j} X_{t+j}. \end{aligned}$$

  Now, MA(1) is not invertible.

- When $\theta = 1$, we have $X_t = W_t + W_{t-1}$. If we have $W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$, then

$$X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} + \sum_{j=0}^{\infty} \pi_j X_{t-1-j} = \sum_{j=0}^{\infty} \pi_j X_{t-j} + \sum_{j=1}^{\infty} \pi_{j-1} X_{t-j} = \pi_0 X_t + \sum_{j=1}^{\infty} (\pi_j + \pi_{j-1}) X_{t-j}.$$

  Thus, we have have $\pi_j + \pi_{j-1} = 0$ and $\pi_0 = 1$, which means $\pi_j = (-1)^j$. Then

$$\sum_{j=0}^{\infty} |\pi_j| < \infty$$

  is not possible. So MA(1) is not invertible when $\theta = 1$; similarly when $\theta = -1$. One may notice that, similarly as the case of $|\phi| = 1$ in the AR(1) model, if we allow $\sigma = 0$, then we have $X_t = 0$ and $W_t = 0 = X_t$ so invertible. But this is a nonsense case. So in the following, we require $\sigma > 0$.

  Sec 4.4 in Brockwell and Davis (2009, Time Series Theory and Methods) defines invertibility in a more general way that is if we can express W_t as $W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$. It does not require that $\sum_{j=0}^{\infty} |\pi_j| < \infty$. With this more general meaning invertibility, we have X_t is invertible when $|\theta| = 1$. In the remaining context, we will keep our more realistic restriction of $\sum_{j=0}^{\infty} |\pi_j| < \infty$.

Proposition 2.4. MA(1) process $X_t = \theta(B) W_t$ where $\theta(B) = 1 + \theta B$ is not invertible if and only if $|\theta| \geq 1$ or the root z_1 of the polynomial $\theta(z) = 1 + \theta z$ satisfies $|z_1| \leq 1$.

Theorem 2.6. The MA(q) process $X_t = \pi(B)W_t$ where

$$\pi(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

is not invertible if and only if

$$\pi(z) = 1 + \theta_1 z + \cdots + \theta_q z^q = 0 \;\Rightarrow\; |z| \leq 1.$$

Based on invertibility, we can write

$$W_t = \pi(B)X_t$$

where $\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \cdots$ for some π_js satisfying $\sum_{j=0}^{\infty} |\pi_j| < \infty$. The question is then, how to calculate π_js? One way is to matching the coefficients.

$$\begin{aligned}
& X_t = \theta(B)W_t \quad \text{and} \quad W_t = \pi(B)X_t \\
\Rightarrow\; & 1 = \pi(B)\theta(B) \\
\Leftrightarrow\; & 1 = (\pi_0 + \pi_1 B + \pi_2 B^2 + \cdots)(1 + \theta_1 B + \cdots + \theta_p B^p) \\
\Leftrightarrow\; & 1 = \pi_0, \\
& 0 = \pi_1 + \theta_1 \pi_0, \\
& 0 = \pi_2 + \theta_1 \pi_1 + \theta_2 \pi_0 \\
& \vdots \\
\Leftrightarrow\; & 1 = \pi_0, \quad 0 = \pi_j \quad (j < 0), \quad 0 = \theta(B)\pi_j \quad (j > 0).
\end{aligned}$$

2.5 ARMA(1,1) Processes

In this subsection we introduce, through an example, some of the key properties of an important class of linear processes known as **ARMA** (autoregressive moving average) processes. This example is the ARMA(1,1) processes. Higher-order ARMA processes will be discussed later.

The time series $\{X_t\}$ is an **ARMA(1,1) process** if it is stationary and satisfies (for every t)

$$X_t - \phi X_{t-1} = W_t + \theta W_{t-1}, \tag{2.6}$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$, $\sigma > 0$, and $\phi + \theta \neq 0$.

Let us find the expression of $\{X_t\}$ in terms of $\{W_t\}$:

- When $|\phi| = 0$, we have the trivial solution $X_t = W_t + \theta W_{t-1}$.
- When $0 < |\phi| < 1$, we have meaning full definition of $\sum_{j=0}^{\infty} \phi^j B^j$. Then applying it to both sides of (2.6) provides that

$$\begin{aligned}\sum_{j=0}^{\infty} \phi^j B^j (1 - \phi B) X_t = X_t &= \left(\sum_{j=0}^{\infty} \phi^j B^j \right) (1 + \theta B) W_t \\ &= \left(\sum_{j=0}^{\infty} \phi^j B^j + \theta \sum_{j=0}^{\infty} \phi^j B^{j+1} \right) W_t \\ &= W_t + (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} W_{t-j}. \end{aligned} \tag{2.7}$$

This is one MA(∞) process, and of course stationary. For the uniqueness, suppose we have another stationary solution Y_t, then we have

$$\begin{aligned} Y_t &= W_t + \theta W_{t-1} + \phi Y_{t-1} \\ &= W_t + (\theta + \phi) W_{t-1} + \theta\phi W_{t-2} + \phi^2 Y_{t-2} \\ &= \cdots = W_t + (\theta + \phi) W_{t-1} + (\theta + \phi)\phi W_{t-2} + \cdots + (\theta + \phi)\phi^{k-1} W_{t-k} + \phi^{k+1} Y_{t-k-1} \end{aligned}$$

Then

$$\text{E}\left(Y_t - W_t - (\phi + \theta) \sum_{j=1}^{k} \phi^{j-1} W_{t-j} \right)^2 = \phi^{2k+2} \text{E}(Y_{t-k-1}^2) \to 0 \quad \text{as } k \to \infty.$$

Hence, solution (2.7) is the unique stationary solution of (2.6) providing that $|\phi| < 1$.

- When $|\phi| > 1$, we have

$$\begin{aligned} X_t =& -\theta\phi^{-1}W_t - \phi^{-1}W_{t+1} + \phi^{-1}X_{t+1} \\ =& -\theta\phi^{-1}W_t - (\theta+\phi)\phi^{-2}W_{t+1} - \phi^{-2}W_{t+2} + \phi^{-2}X_{t+2} \\ =& \cdots = -\theta\phi^{-1}W_t - (\theta+\phi)\sum_{j=1}^{k}\phi^{-j-1}W_{t+j} - \phi^{-k-1}W_{t+k+1} + \phi^{-k-1}X_{t+k+1} \end{aligned}$$

  Then

$$\begin{aligned} &\mathrm{E}\left(X_t - \left\{-\theta\phi^{-1}W_t - (\theta+\phi)\sum_{j=1}^{k}\phi^{-j-1}W_{t+j}\right\}\right)^2 \\ =&\phi^{-2k-2}\mathrm{E}(W_{t+k+1} + X_{t+k+1})^2 \to 0 \quad \text{as } k \to \infty. \end{aligned}$$

  Thus, we have a unique stationary solution of (2.6) when $|\phi| > 1$ as

$$X_t = -\theta\phi^{-1}W_t - (\theta+\phi)\sum_{j=1}^{\infty}\phi^{-j-1}W_{t+j}. \tag{2.8}$$

  Again, this solution depends on future values of W_t.

- When $|\phi| = 1$, there is no stationary solution of (2.6) (left as a **HW** problem). Thus, no stationary ARMA(1,1) process when $|\phi| = 1$.

Summary:

- A stationary solution of the ARMA(1,1) equations exists if and only if $|\phi| \neq 1$.
- If $|\phi| < 1$, then the unique stationary solution is given by (2.7). In this case, we say that $\{X_t\}$ is **causal** or a causal function of $\{W_t\}$, since X_t can be expressed in terms of the current and past values $\{W_s, s \leq t\}$.
- If $|\phi| > 1$, then the unique stationary solution is given by (2.8). In this case, we say that $\{X_t\}$ is **noncausal** since X_t is then a function of current and future values $\{W_s, s \geq t\}$.

For invertibility, we have, by switching the roles of X_t and W_t, and the roles of ϕ and θ,

- If $|\theta| < 1$, then ARMA(1,1) process is invertible as

$$W_t = X_t - (\phi+\theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{t-j}.$$

- If $|\theta| > 1$, then ARMA(1,1) process is noninvertible as

$$W_t = -\phi\theta^{-1}X_t + (\theta+\phi)\sum_{j=1}^{\infty}(-\theta)^{-j-1}W_{t+j}.$$

- If $|\theta| = 1$, the ARMA(1,1) process is invertible under the more general definition of invertibility same as in the MA(1) process. Without this more general setting, we say the ARMA(1,1) process is noninvertible when $|\theta| = 1$.

Like the argument in last subsection of AR(1) model, if the ARMA(1,1) process $\{X_t\}$ is noncausal and noninvertible; i.e., $|\phi| > 1$ and $|\theta| > 1$, then we define

$$\tilde{\phi}(B) = 1 - \phi^{-1}B \quad \text{and} \quad \tilde{\theta}(B) = 1 + \theta^{-1}B$$

and let

$$W_t^* = \tilde{\theta}^{-1}(B)\tilde{\phi}(B)X_t$$

Once verifying that

$$\{W_t^*\} \sim \text{WN}(0, \sigma_*^2) \quad \text{and} \quad \tilde{\phi}(B)X_t = \tilde{\theta}(B)W_t^*, \tag{2.9}$$

we have $\{X_t\}$ being a causal and invertible ARMA(1,1) process relative to the white noise sequence $\{W_t^*\}$. Threre, from a second-order point of view, nothing is lost by restricting attention to causal and invertible ARMA(1,1) models. This statement is also true for higher-ordered ARMA models.

Now, we show (2.9). It is easy to see that $\tilde{\theta}(B)W_t^* = \tilde{\theta}(B)\tilde{\theta}^{-1}(B)\tilde{\phi}(B)X_t = \tilde{\phi}(B)X_t$. It suffices to show $\{W_t^*\}$ is a white noise. This is left as a **HW** problem.

2.6 Properties of $\overline{X}_n$, $\widehat{\gamma}_X(h)$ and $\widehat{\rho}_X(h)$

2.6.1 For $\overline{X}_n$

Recall that, for observations $x_1, \ldots, x_n$ of a time series, the **sample mean** is

$$\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

The **sample auto covariance function** is

$$\widehat{\gamma}_X(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x}), \quad \text{for } -n < h < n.$$

The **sample autocorrelation function (sample ACF)** is

$$\widehat{\rho}_X(h) = \frac{\widehat{\gamma}_X(h)}{\widehat{\gamma}_X(0)}.$$

Estimation of μ_X: The moment estimator of the mean μ_X of a stationary process $\{X_t\}$ is the sample mean

$$\overline{X}_n = n^{-1}\sum_{t=1}^{n} X_t. \tag{2.10}$$

Obviously, it is unbiased; i.e., $\mathrm{E}(\overline{X}_n) = \mu_X$. Its mean squared error is

$$\begin{aligned}
\mathrm{Var}(\overline{X}_n) =& \mathrm{E}(\overline{X}_n - \mu_X)^2 \\
=& n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) \\
=& n^{-2}\sum_{i-j=-n}^{n}(n - |i-j|)\gamma_X(i-j) = n^{-1}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_X(h) \\
=& \underbrace{\frac{\gamma_X(0)}{n}}_{\text{is } \mathrm{Var}(\overline{X}_n) \text{ when } \{X_t\} \text{ are iid}} + \frac{2}{n}\sum_{h=1}^{n-1}\left(1 - \frac{|h|}{n}\right)\gamma_X(h).
\end{aligned}$$

- Depending on the nature of the correlation structure, the standard error of $\overline{X}_n$ may be smaller or larger than the white noise case.
  - Consider $X_t = \mu + W_t - 0.8W_{t-1}$, where $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$, then

$$\mathrm{Var}(\overline{X}_n) = \frac{\gamma_X(0)}{n} + \frac{2}{n}\sum_{h=1}^{n-1}\left(1 - \frac{|h|}{n}\right)\gamma_X(h) = \frac{1.64\sigma^2}{n} - \frac{1.6(n-1)\sigma^2}{n^2} < \frac{1.64\sigma^2}{n}.$$

– And if $X_t = \mu + W_t + 0.8W_{t-1}$, where $\{W_t\} \sim \text{WN}(0, \sigma^2)$, then

$$\text{Var}(\overline{X}_n) = \frac{\gamma_X(0)}{n} + \frac{2}{n}\sum_{h=1}^{n-1}\left(1 - \frac{|h|}{n}\right)\gamma_X(h) = \frac{1.64\sigma^2}{n} + \frac{1.6(n-1)\sigma^2}{n^2} > \frac{1.64\sigma^2}{n}.$$

- If $\gamma_X(h) \to 0$ as $h \to \infty$, we have

$$|\text{Var}(\overline{X}_n)| \leq \frac{\gamma_X(0)}{n} + 2\frac{\sum_{h=1}^{n}|\gamma_X(h)|}{n} \to 0 \quad \text{as } n \to \infty.$$

  Thus, $\overline{X}_n$ converges in mean square to μ.

- If $\sum_{h=-\infty}^{\infty}|\gamma_X(h)| < \infty$, then

$$n\text{Var}(\overline{X}_n) = \sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_X(h) = \gamma_X(0) + 2\frac{\sum_{h=1}^{n}(n-h)\gamma_X(h)}{n} = \gamma_X(0) + 2\frac{\sum_{h=1}^{n-1}\sum_{i=1}^{h}\gamma_X(i)}{n}$$
$$\to \gamma_X(0) + 2\sum_{i=1}^{\infty}\gamma_X(i) = \sum_{h=-\infty}^{\infty}\gamma_X(h) = \gamma_X(0)\sum_{h=-\infty}^{\infty}\rho_X(h).$$

  One interpretation could be that, instead of $\text{Var}(\overline{X}_n) \approx \gamma_X(0)/n$, we have $\text{Var}(\overline{X}_n) \approx \gamma_X(0)/(n/\tau)$ with $\tau = \sum_{h=-\infty}^{\infty}\rho_X(h)$.

The effect of the correlation is a reduction of sample size from n to n/τ.

Example 2.10. For linear processes, i.e., if $X_t = \mu + \sum_{j=-\infty}^{\infty}\psi_j W_{t-j}$ with $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$, then

$$\begin{aligned}\sum_{h=-\infty}^{\infty}|\gamma_X(h)| &= \sum_{h=-\infty}^{\infty}|\sigma^2\sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h}| \\ &\leq \sum_{h=-\infty}^{\infty}\sigma^2\sum_{j=-\infty}^{\infty}|\psi_j|\cdot|\psi_{j+h}| \\ &= \sigma^2\sum_{j=-\infty}^{\infty}|\psi_j|\sum_{h=-\infty}^{\infty}|\psi_{j+h}| \\ &= \sigma^2\left(\sum_{j=-\infty}^{\infty}|\psi_j|\right)^2 < \infty\end{aligned}$$

To make inference about μ_X (e.g., is $\mu_X = 0$?), using the sample mean $\overline{X}_n$, it is necessary to know the asymptotic distribution of $\overline{X}_n$:

If $\{X_t\}$ is Gaussian stationary time series, then, for any n,

$$\sqrt{n}(\overline{X}_n - \mu_X) \sim \text{N}\left(0, \sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_X(h)\right).$$

Then one can obtain exact confidence intervals of estimating μ_X, or approximated confidence intervals if it is necessary to estimate $\gamma_X(\cdot)$.

For the linear process, $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$ with $\{W_t\} \sim \text{IID}(0, \sigma^2)$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$, then

$$\sqrt{n}(\overline{X}_n - \mu_X) \sim \text{AN}(0, \nu), \tag{2.11}$$

where $\nu = \sum_{h=-\infty}^{\infty} \gamma_X(h) = \sigma^2(\sum_{j=-\infty}^{\infty} \psi_j)^2$.

The proof of (2.11) can be found in Page 238 of Brockwell and Davis (2009, Time Series Theory and Methods). Very roughly, recall

$$\gamma_X(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h},$$

then

$$\begin{aligned}
\lim_{n\to\infty} n\text{Var}(\overline{X}_n) &= \lim_{n\to\infty} \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right) \gamma_X(h) \\
&= \lim_{n\to\infty} \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right) \psi_{j+h} \\
&= \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2
\end{aligned}$$

The above results for the linear process, also hold for ARMA models. Naturally,

$$(\overline{X}_n - 1.96\sqrt{\nu/n}, \overline{X}_n + 1.96\sqrt{\nu/n}).$$

is an approximated 95% confidence interval for μ_X.

Since ν is typically unknown, naturally, we have an approximated 95% confidence interval of μ_X as

$$(\overline{X}_n - 1.96\sqrt{\hat{\nu}/n}, \overline{X}_n + 1.96\sqrt{\hat{\nu}/n}),$$

once we can obtain an estimator $\hat{\nu}$ of $\nu = \sum_{h=-\infty}^{\infty} \gamma_X(h)$.

- One intuitive way is to use $\hat{\nu} = \sum_{h=-\infty}^{\infty} \widehat{\gamma}_X(h)$. However, based on finite sample $\{X_1, \ldots, X_n\}$, it is impossible to obtain a reasonable estimator of $\gamma_X(h)$ for $h \geq n$. Then, why not use $\hat{\nu} = \sum_{h=-(n-1)}^{n-1} \widehat{\gamma}_X(h)$. Vary sadly and interestingly, this $\hat{\nu}$ is always zero. A compromising estimator $\hat{\nu}$ is then
$$\hat{\nu} = \sum_{h=-[\sqrt{n}]}^{[\sqrt{n}]} \left(1 - \frac{|h|}{[\sqrt{n}]}\right) \widehat{\gamma}_X(h)$$
- If we known the model of the time series, i.e., we have explicit formula of $\gamma_X(h)$. For example,

say we have an AR(1) $\{X_t\}$ with mean μ_X satisfies

$$X_t - \mu_X = \phi(X_{t-1} - \mu_X) + W_t,$$

we have $\gamma_X(h) = \phi^{|h|}\sigma^2/(1-\phi^2)$ and consequently, $\nu = \sigma^2/(1-\phi)^2$. Then we have

$$\widehat{\nu} = \frac{\widehat{\sigma}^2}{(1-\widehat{\phi})^2}$$

2.6.2 For $\gamma_X(h)$ and $\rho_X(h)$

Estimators of $\gamma_X(h)$ and $\rho_X(h)$ is defined by

$$\widehat{\gamma}_X(h) = n^{-1} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \overline{X}_n)(X_t - \overline{X}_n), \tag{2.12}$$

$$\widehat{\rho}_X(h) = \frac{\widehat{\gamma}_X(h)}{\widehat{\gamma}_X(0)}. \tag{2.13}$$

First let us check that $\hat{\nu} = \sum_{h=-(n-1)}^{n-1} \widehat{\gamma}_X(h)$ is always zero.

$$\begin{aligned}
\hat{\nu} = & \sum_{h=-(n-1)}^{n-1} n^{-1} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \overline{X}_n)(X_t - \overline{X}_n) \\
= & n^{-1} \sum_{t=1}^{n} (X_t - \overline{X}_n)^2 + 2n^{-1} \sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (X_{t+h} - \overline{X}_n)(X_t - \overline{X}_n) \\
= & n^{-1} \sum_{t=1}^{n} (X_t^2 - 2X_t\overline{X}_n + \overline{X}_n^2) + 2n^{-1} \sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (X_{t+h}X_t - X_t\overline{X}_n - X_{t+h}\overline{X}_n + \overline{X}_n^2) \\
= & n^{-1} \sum_{t=1}^{n} X_t^2 - \overline{X}_n^2 + 2n^{-1} \sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (X_{t+h}X_t - X_t\overline{X}_n - X_{t+h}\overline{X}_n + \overline{X}_n^2) \\
= & n^{-1} \sum_{t=1}^{n} X_t^2 - n\overline{X}_n^2 + 2n^{-1} \sum_{h=1}^{n-1} \sum_{t=1}^{n-h} X_{t+h}X_t = 0.
\end{aligned}$$

To check the bias of $\widehat{\gamma}_X(h)$, let us look at the case when $h = 0$. We have

$$\widehat{\gamma}_X(0) = n^{-1} \sum_{t=1}^{n} (X_t - \overline{X}_n)^2.$$

Even in iid case, this is an biased estimator (sample variance is biased which has $(n-1)^{-1}$ instead of n^{-1}). Expression for $\mathrm{E}\{\widehat{\gamma}_X(h)\}$ is messy (try your best to derive it as a **HW** problem). Let's consider instead

$$\overline{\gamma}_X(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \mu_X)(X_t - \mu_X).$$

It can be seen that

$$E\{\overline{\gamma}_X(h)\} = \frac{n-|h|}{n} \gamma_X(h) \neq \gamma_X(h);$$

i.e., biased. Rather than using $\overline{\gamma}_X(h)$, you might seem more natural to consider

$$\widetilde{\gamma}_X(h) = \frac{1}{n-|h|} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \mu_X)(X_t - \mu_X),$$

since now we have $\mathrm{E}\{\widetilde{\gamma}_X(h)\} = \gamma_X(h)$; an unbiased estimator. Now replacing μ_X with $\overline{X}_n$, we have two estimators:

$$\widehat{\gamma}_X(h) \quad \text{and} \quad \frac{n}{n-|h|}\widehat{\gamma}_X(h)$$

respectively, called *biased* and *unbiased* ACVF estimators (even though latter one is actually biased in general!). Generally speaking, both $\widehat{\gamma}_X(h)$ and $\widehat{\rho}_X(h)$ are biased even if the factor n^{-1} is replaced by $(n-h)^{-1}$. Nevertheless, under general assumptions they are nearly unbiased for large sample size (conduct a simulation study to see the bias of both estimators as a **HW** problem). Now, let us talk about the reason of why we like $\widehat{\gamma}_X(h)$, and why I think this is very brilliant.

Lemma 2.3. For any sequence $x_1, \ldots, x_n$, the sample ACVF $\widehat{\gamma}_X$ satisfies:

1. $\widehat{\gamma}_X(h) = \widehat{\gamma}_X(-h)$
2. $\widehat{\gamma}_X$ is nonnegaitve definite, and hence
3. $\widehat{\gamma}_X(0) \geq 0$ and $|\widehat{\gamma}_X(h)| \leq \widehat{\gamma}_X(0)$

Proof. The first one is trivial. It suffices to prove the second property which is equivalent to show that for each $k \geq 1$ the k-dimensional sample covariance matrix

$$\widehat{\boldsymbol{\Gamma}}_k = \begin{pmatrix} \widehat{\gamma}_X(0) & \widehat{\gamma}_X(1) & \cdots & \widehat{\gamma}_X(k-1) \\ \widehat{\gamma}_X(1) & \widehat{\gamma}_X(0) & \cdots & \widehat{\gamma}_X(k-2) \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{\gamma}_X(k-1) & \widehat{\gamma}_X(k-2) & \cdots & \widehat{\gamma}_X(0) \end{pmatrix}$$

is nonnegative definite. To see that, we have, for $k \geq n$,

$$\widehat{\boldsymbol{\Gamma}}_k = n^{-1}\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}},$$

where

$$\boldsymbol{M} = \begin{pmatrix} 0 & \cdots & 0 & 0 & Y_1 & Y_2 & \cdots & Y_k \\ 0 & \cdots & 0 & Y_1 & Y_2 & \cdots & Y_k & 0 \\ \vdots & & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \cdots & Y_k & 0 & \cdots & 0 \end{pmatrix}$$

is a $k \times 2k$ matrix with $Y_i = X_i - \overline{X}_n$ for i=$1, \ldots, n$ and $Y_i = 0$ for $i = n+1, \ldots, k$. Note that, if $\widehat{\boldsymbol{\Gamma}}_m$ is nonnegative definite, then all $\widehat{\boldsymbol{\Gamma}}_k$s are nonnegative definite for all $k < m$. □

The nonnegative definite property is not always true if n^{-1} is replaced by $(n-h)^{-1}$. Further, when $h \geq n$ or for h slightly smaller than$n$, there is no way to reliably estimate $\gamma_X(h)$ and $\rho_X(h)$ since the information around there are too little. Box and Jenkins (1976) suggest that useful estimates of correlation $\rho_X(h)$ can only be made if n is roughly 50 or more and $h \leq n/4$.

It will be important to be able to recognize when sample autocorrelations are significantly different from zero so that we can select the correct model to fit our data. In order to draw such statistical inference, we need the following asymptotic joint distribution.

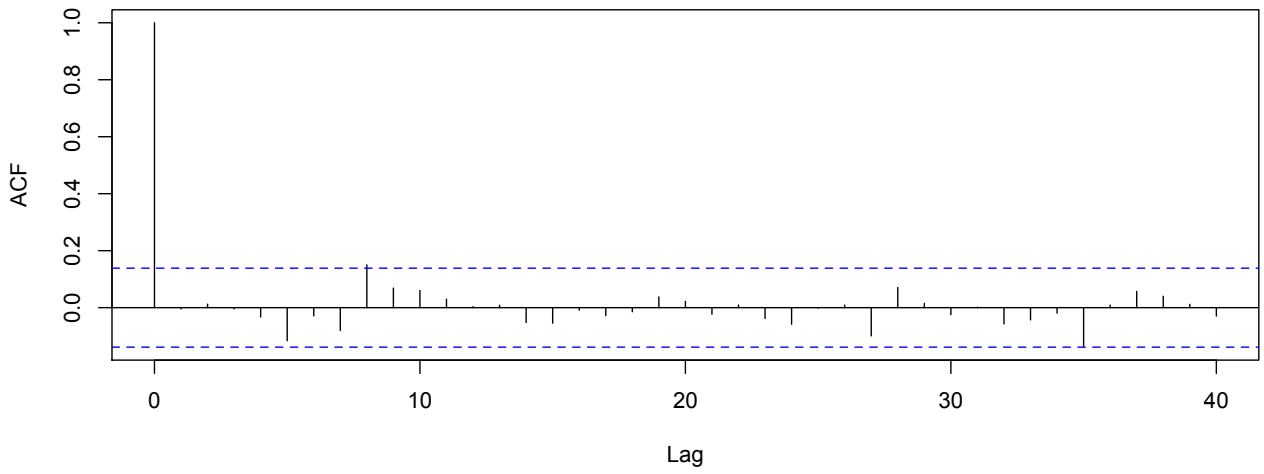
Theorem 2.7. For an IID process $\{W_t\}$, if $\mathrm{E}(W_t^4) < \infty$, we have

$$\widehat{\boldsymbol{\rho}}_W(h) = \begin{pmatrix} \widehat{\rho}_W(1) \\ \vdots \\ \widehat{\rho}_W(h) \end{pmatrix} \sim \mathrm{AN}(0, n^{-1}\boldsymbol{I}_h). \tag{2.14}$$

where $\boldsymbol{I}_h$ is a $h \times h$ identity matrix.

Remark 2.3. For $\{W_t\} \sim \mathrm{IID}(0, \sigma^2)$, then $\rho_W(l) = 0$ for $l \neq 0$. From Theorem 2.7, we have, for large n, $\widehat{\rho}_W(1), \ldots, \widehat{\rho}_W(h)$ is approximately independent and identically distributed normal random variables form $\mathrm{N}(0, n^{-1})$. If we plot the sample autocorrelation function $\widehat{\rho}_W(k)$ as a function of k, approximately 0.95 of the sample autocorrelations should lie between the bounds $\pm 1.96\sqrt{n}$. This can be used as a check that the observations truly are from an IID process. In Figure 2.3, we have plotted the sample autocorrelation $\widehat{\rho}_W(k)$ for $k = 1, \ldots, 40$ for a sample of 200 iid $\mathrm{N}(0, 1)$.

```
set.seed(150); Wt=rnorm(200);par(mar=c(4,4,1.5,.5));acf(Wt,lag.max=40)
```



It can be seen that all but one of the autocorrelations lie between the bounds $\pm 1.96\sqrt{n}$, and

Remark 2.4. This theorem yields several procedures of testing

$$H_0: \text{ iid} \quad \text{vs} \quad H_a: \text{ not iid.}$$

Method 1: Based on the values of sample ACF: If for one h, $\widehat{\rho}_X(h) \pm z_{\alpha/2}/\sqrt{n}$ does not contain zero, reject H_0.

Method 2: The portmanteau test I: Instead of checking $\widehat{\rho}_X(h)$ for each h, it is also possible to consider the single statistics

$$Q = n\sum_{j=1}^{h}\widehat{\rho}_X^2(j).$$

Under H_0, $Q \sim \chi_h^2$. Thus, rejection region is $Q > \chi_h^2(1-\alpha)$.

Method 3: The portmanteau test II (Ljung and Box, 1978).

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\widehat{\rho}_X^2(j)/(n-j)$$

which is better approximated by χ_h^2, thus the same rejection region.

Method 4: The portmanteau test III: if wanted to test residuals $\{R_t\}$ rather than a time series $\{X_t\}$, then

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\widehat{\rho}_R^2(j)/(n-j)$$

which is better approximated by χ_{h-p}^2 instead, where p is the number of parameters estimated in forming $\{R_t\}$.

Method 5: Turning point test

Method 6: Difference-Sign test

Method 7: Rank test

Design a simulation study to compare these testing procedures (as a **HW** problem).

1. Learn them by yourself. At least you should know when is okay to use which test, how to calculate the test statistic and when to reject the null.

2. Set a reasonable sample size and generate an iid sequence. Apply each of these method to test for IID. If rejected, count by 1, if not, count it by zero.

3. Repeat step 2 for 1000 times. Record how many of them lead to the conclusion of rejection (it should be around the value of α)

4. Then, start making your model be more and more Non-IID, for example, you can general $X_t - \phi X_t = W_t$, in the beginning set $\phi = 0$ then you have IID. Then set $\phi = seq(0.02, 1, by = 0.02)$. Each time, you do 1000 replications to obtain a rejection rate (as the power).

5. Plot the power curve to see which methods is the most powerful.

6. Summarize your simulation result and turn in with your homework.

Theorem 2.8. If $\{X_t\}$ is the stationary process,

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

where $\{W_t\} \sim \text{IID}(0, \sigma^2)$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\text{E}W_t^4 < \infty$ (or $\sum_{j=-\infty}^{\infty} \psi_j^2 |j| < \infty$), then for each h, we have

$$\widehat{\boldsymbol{\rho}}_X(h) = \begin{pmatrix} \widehat{\rho}_X(1) \\ \vdots \\ \widehat{\rho}_X(h) \end{pmatrix} \sim \text{AN} \left\{ \boldsymbol{\rho}_X(h) = \begin{pmatrix} \rho_X(1) \\ \vdots \\ \rho_X(h) \end{pmatrix}, n^{-1} \boldsymbol{\Omega} \right\}. \tag{2.15}$$

where $\boldsymbol{\Omega} = [\omega_{ij}]_{i,j=1}^h$ is the covariance matrix whose (i,j)-element is given by Bartlett's formula,

$$\begin{aligned} \omega_{ij} = \sum_{k=-\infty}^{\infty} \Big\{ & \rho_X(k+i)\rho_X(k+j) + \rho_X(k-i)\rho_X(k+j) + 2\rho_X(i)\rho_X(j)\rho_X^2(k) \\ & - 2\rho_X(i)\rho_X(k)\rho_X(k+j) - 2\rho_X(j)\rho_X(k)\rho_X(k+i) \Big\}. \end{aligned}$$

Remark 2.5. Simple algebra shows that

$$\begin{aligned} \omega_{ij} = \sum_{k=1}^{\infty} & \Big\{ \rho_X(k+i) + \rho_X(k-i) - 2\rho_X(i)\rho_X(k) \Big\} \\ & \times \Big\{ \rho_X(k+j) + \rho_X(k-j) - 2\rho_X(j)\rho_X(k) \Big\}, \end{aligned}$$

which is a more convenient form of ω_{ij} for computational purposes. This formula also shows that the asymptotic distribution of $\sqrt{n}\{\widehat{\boldsymbol{\rho}}_X(h) - \boldsymbol{\rho}_X(h)\}$ is the same as the random vector $(Y_1, \ldots, Y_h)^{\text{T}}$ where

$$Y_i = \sum_{k=1}^{\infty} \Big\{ \rho_X(k+i) + \rho_X(k-i) - 2\rho_X(i)\rho_X(k) \Big\} Z_k$$

with $Z_1, Z_2, \ldots$ being iid $\text{N}(0,1)$.

Example 2.11. MA(q): if

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\{W_t\} \sim \text{IID}(0, \sigma^2)$, then from Bartlett's formula, we have

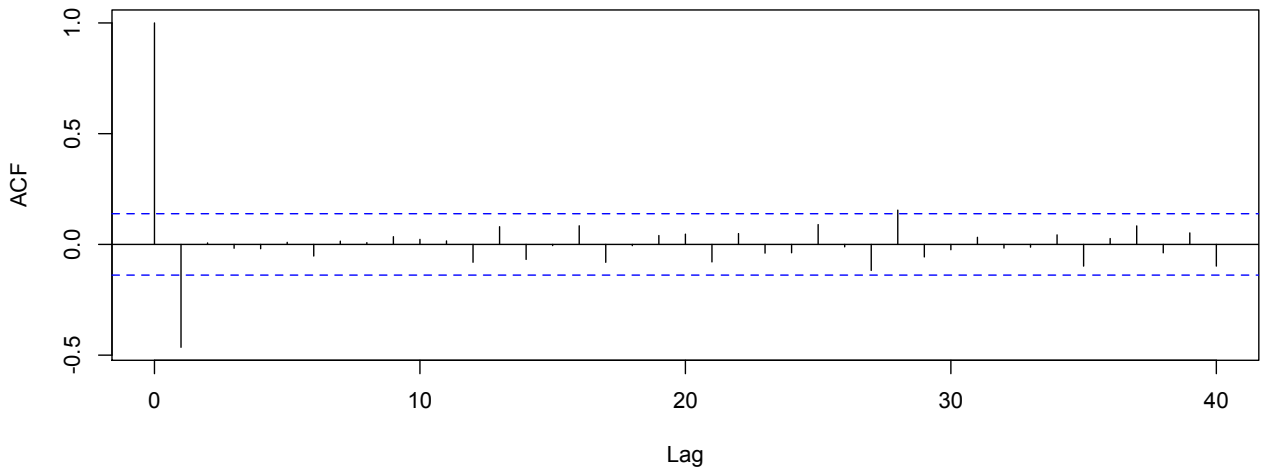
$$\omega_{ii} = 1 + 2\rho_X^2(1) + 2\rho_X^2(2) + \cdots + 2\rho_X^2(q), \quad i > q,$$

as the variance of the asymptotic distribution of $\sqrt{n}\widehat{\rho}_X(i)$ as $n \to \infty$. For MA(1), we have $\rho_X(1) = \theta/(1+\theta^2), \rho_X(h) = 0, |h| > 1$. Then

$$\begin{aligned} \omega_{11} &= 1 - 3\rho_X^2(1) + 4\rho_X^4(1) \\ \omega_{ii} &= 1 + 2\rho_X^2(1), \quad i > q, \end{aligned}$$

Let $\theta = 0.8$. In Figure 2.11 we have plotted the $\widehat{\rho}_X(k)$ for $k = 0, \ldots, 40$, for 200 observations, where $\{W_t\}$ are iid $N(0,1)$. It is found that $\widehat{\rho}_X(1) = -0.465$ and $\rho_X(1) = -0.4878$. Obviously $\widehat{\rho}_X(1)$ is less than $-1.96/\sqrt{n} = -0.1379$. Thus, we would reject the hypothesis that the data are iid. Further, for $h = 2, \ldots, 40$, we have $|\widehat{\rho}_X(h)| \leq 1.96/\sqrt{n}$ which strongly suggests that the data are from a model in which observations are uncorrelated past lag 1. In addition, we have $\rho_X(1) = -0.4878$ is inside the 95% confidence interval $\widehat{\rho}_X(1) \pm 1.96n^{-1/2}\{1 - 3\widehat{\rho}_X^2(1) + 4\widehat{\rho}_X^4(1)\}^{1/2} = (-0.3633, -0.5667)$; i.e., it further supports the compatibility of the data with the model $X_t = W_t - 0.8W_{t-1}$.

```
set.seed(150); Wt=rnorm(250);Xt=filter(Wt,sides=1,c(1,-.8))[-(1:50)]
par(mar=c(4,4,1.5,.5));acf(Xt,lag.max=40)
```



3 Autoregressive Moving Average (ARMA) Processes

3.1 Definition

An **ARMA(p, q) process** $\{X_t\}$ is a stationary process that satisfies

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}$$

which also can be written as

$$\phi(B)X_t = \theta(B)W_t$$

where

$$\begin{aligned}\phi(B) =& 1 - \phi_1 B - \cdots - \phi_p B^p,\\ \theta(B) =& 1 + \theta_1 B + \cdots + \theta_q B^q,\end{aligned}$$

and $\{W_t\} \sim \text{WN}(0, \sigma^2)$. We say $\{X_t\}$ is an ARMA(p, q) process with mean μ_X if $\{X_t - \mu_X\}$ is an ARMA(p, q) process.

Remark 3.1. For an ARMA(p, q) process $\{X_t\}$, we always insist that $\phi_p, \theta_q \neq 0$ and that the polynomials

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad \text{and} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

have no common factors. This implies it is not a lower order ARMA model. For example, consider a white noise process W_t, we can write $X_t = W_t$ or

$$(1 - 2B + B^2)X_t = (1 - 2B + B^2)W_t.$$

It is presented as an ARMA(2,2) model, but essentially it is white noise.

Remark 3.2. ARMA processes can accurately approximate many stationary processes:

- AR(p)=ARMA$(p, 0)$: $\theta(B) = 1$.
- MA(q)=ARMA$(0, q)$: $\phi(B) = 1$.

Further, for any stationary process with ACVF γ, and any $k > 0$, there exists an ARMA process $\{X_t\}$ for which

$$\gamma_X(h) = \gamma(h), \quad h = 0, 1, \ldots, k.$$

3.2 Causality and Invertibility

Recall the definition of *causal* and *invertible*. Let $\{X_t\}$ be an ARMA(p, q) process defined by equations $\phi(B)X_t = \theta(B)W_t$

- $\{X_t\}$ is said to be *causal* (or more specifically to be a causal function of $\{W_t\}$) if there exits a sequence of constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}, \quad t = 0, \pm 1, \ldots. \tag{3.1}$$

- $\{X_t\}$ is said to be *invertible* (or more specifically to be an invertible function of $\{W_t\}$) if there exits a sequence of constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t = 0, \pm 1, \ldots. \tag{3.2}$$

- Neither causality nor invertibility is a property of $\{X_t\}$ alone, but of the relationship between $\{X_t\}$ and $\{W_t\}$.

Theorem 3.1. Let $\{X_t\}$ be an ARMA(p, q) process. Then $\{X_t\}$ is causal if and only if

$$\phi(z) \neq 0 \text{ for all } |z| \leq 1.$$

The coefficients $\{\psi_j\}$ in (3.1) are determined by the relation

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z), \quad |z| \leq 1.$$

Proof. First, we assume that $\phi(z) \neq 0$ if $|z| \leq 1$. Since we have

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = \phi_p (z - z_1) \cdots (z - z_p),$$

then $|z_i| > 1$ for $i = 1, \ldots, p$. For each i,

$$\frac{1}{z - z_i} = -\frac{1}{1 - z/z_i} = -\sum_{k=1}^{\infty} (z/z_i)^k, \qquad \text{for } |z| < |z_i|.$$

This implies that there exists $\epsilon > 0$ such that $1/\phi(z)$ has a power series expansion,

$$1/\phi(z) = \sum_{j=0}^{\infty} \zeta_j z^j \doteq \zeta(z), \quad |z| < 1 + \epsilon \leq \min_i |z_i|.$$

Consequently, $\zeta_j (1 + \epsilon/2)^j \to 0$ as $j \to \infty$ so that there exists $K > 0$ such that

$$|\zeta_j| < K(1 + \epsilon/2)^{-j}, \quad \forall j = 0, 1, 2, \ldots.$$

In particular we have $\sum_{j=0}^{\infty} |\zeta_j| < \infty$ and $\zeta(z)\phi(z) = 1$ for $|z| \leq 1$. There, we can apply the operator

$\zeta(B)$ to both sides of the equation $\phi(B)X_t = \theta(B)W_t$ to obtain

$$X_t = \zeta(B)\theta(B)W_t.$$

Thus we have the desired representation

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$$

where the sequencd $\{\psi_j\}$ is determined by $\theta(z)/\phi(z)$.

Now, assume that $\{X_t\}$ is causal; i.e., $X_t = \psi(B)W_t$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Then

$$\theta(B)W_t = \phi(B)X_t = \phi(B)\psi(B)W_t.$$

If we let $\eta(z) = \phi(z)\psi(z) = \sum_{j=0}^{\infty} \eta_j z^j$, $|z| \leq 1$, we can rewrite this equation as

$$\sum_{j=0}^{q} \theta_j W_{t-j} = \sum_{j=0}^{\infty} \eta_j W_{t-j},$$

and taking inner products of each side with W_{t-k}, we obtain $\eta_k = \theta_k$, $k = 0, \ldots, q$ and $\eta_k = 0, k > q$. Hence

$$\theta(z) = \eta(z) = \phi(z)\psi(z), \quad |z| \leq 1.$$

Since $\theta(z)$ and $\phi(z)$ have no common zeros and since $|\psi(z)| < \infty$ for $|z| \leq 1$, we conclude that $\phi(z)$ cannot be zero for $|z| \leq 1$. □

Remark 3.3. If $\phi(z) = 0$ for some $|z| = 1$, then there is no stationary solution of the ARMA equations $\phi(B)X_t = \theta(B)W_t$.

Theorem 3.2. Let $\{X_t\}$ be an ARMA(p, q) process. Then $\{X_t\}$ is invertible if and only if

$$\theta(z) \neq 0 \text{ for all } |z| \leq 1.$$

The coefficients $\{\pi_j\}$ in (3.2) are determined by the relation

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z), \quad |z| \leq 1.$$

Proof. First assume that $\theta(z) \neq 0$ if $|z| \leq 1$. By the same argument as in the proof of the previous theorem, $1/\theta(z)$ has a power series expansion

$$1/\theta(z) = \sum_{j=0}^{\infty} \eta_j z^j \doteq \eta(z), \quad |z| < 1 + \epsilon.$$

for some $\epsilon > 0$ and $\sum_{j=0}^{\infty} |\eta_j| < \infty$. Then applying $\eta(B)$ to both sides of the ARMA equations, we have

$$\eta(B)\phi(B)X_t = \eta(B)\theta(B)W_t = W_t.$$

Thus, we have the desired representation

$$W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

where the sequence $\{\pi_j\}$ is determined by $\phi(z)/\theta(z)$.

Conversely, if $\{X_t\}$ is invertible then $W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \pi(B)X_t$ for some $\sum_{j=0}^{\infty} |\pi_j| < \infty$, then

$$\phi(B)W_t = \phi(B)\pi(B)X_t = \pi(B)\phi(B)X_t = \pi(B)\theta(B)W_t$$

which leads to

$$\sum_{j=0}^{q} \phi_j W_{t-j} = \sum_{j=0}^{\infty} \zeta_j W_{t-j},$$

where $\zeta(z) = \pi(z)\theta(z) = \sum_{j=0}^{\infty} \zeta_j z^j, |z| \leq 1$. Taking inner products of each side with W_{t-k}, we obtain $\zeta_k = \phi_k$, $k = 0, \ldots, q$ and $\zeta_k = 0, k > q$. Hence

$$\phi(z) = \zeta(z) = \pi(z)\theta(z), \quad |z| \leq 1.$$

Since $\theta(z)$ and $\phi(z)$ have no common zeros and since $|\pi(z)| < \infty$ for $|z| \leq 1$, we conclude that $\theta(z)$ cannot be zero for $|z| \leq 1$. □

Remark 3.4. If $\{X_t\}$ is a stationary solution of the ARMA equations and if $\phi(z)\theta(z) \neq 0$ for $|z| \leq 1$, then

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$$

and

$$W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

where $\sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$ and $\sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z)$, $|z| \leq 1$.

Remark 3.5. Let $\{X_t\}$ be the ARMA process solving the equations $\phi(B)X_t = \theta(B)W_t$, where

$$\phi(z) \neq 0 \text{ and } \theta(z) \neq 0 \text{ for all } |z| = 1.$$

Then there exists polynomials $\widetilde{\phi}(z)$ and $\widetilde{\theta}(z)$, nonzero for $|z| \leq 1$, of degree p and q respectively, and a new white noise sequence $\{W_t^*\}$ such that $\{X_t\}$ satisfies the causal invertible equations

$$\widetilde{\phi}(B)X_t = \widetilde{\theta}(B)W_t^*.$$

Remark 3.6. Uniqueness: If $\phi(z) \neq 0$ for all $|z| = 1$, then the ARMA equations $\phi(B)X_t = \theta(B)W_t$ have the **unique stationary solution**

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

where ψ_j comes from $\theta(z)/\phi(z)$.

3.3 Computing the ACVF of an ARMA(p, q) Process

We now provide three methods for computing the ACVF of an ARMA process. The second one is the most convenient for obtaining a solution in closed form, and the third one is the most convenient for obtaining numerical values.

3.3.1 First Method

Since the causal ARMA(p, q) process $\phi(B)X_t = \theta(B)W_t$ has representation

$$X_t = \psi(B)W_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$$

where

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z), \quad |z| \leq 1.$$

The ACVF of $\{X_t\}$ is then

$$\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

To determine the coefficients ψ_j, herein we use the method of matching coefficients:

$$(1 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \psi_4 z^4 + \cdots)(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = (1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q)$$

which yields the following difference equations for ψ_k:

$$\begin{aligned} \psi_1 - \phi_1 &= \theta_1 \\ \psi_2 - \phi_2 - \psi_1\phi_1 &= \theta_2 \\ \psi_3 - \phi_3 - \psi_2\phi_1 - \psi_1\phi_2 &= \theta_3 \\ &\cdots \end{aligned} \tag{3.3}$$

By defining $\theta_0 = 1$, $\theta_j = 0$ for $j > q$ and $\phi_j = 0$ for $j > p$, we have the results summarized as

$$\psi_j - \sum_{0<k\leq j} \phi_k \psi_{j-k} = \theta_j, \quad 0 \leq j < \max\{p, q+1\} \tag{3.4}$$

$$\psi_j - \sum_{0<k\leq p} \phi_k \psi_{j-k} = 0, \quad j \geq \max\{p, q+1\} \tag{3.5}$$

The general solution of (3.5) can be written down as

$$\psi_n = \sum_{i=1}^{k} \sum_{j=0}^{r_i - 1} \alpha_{ij} n^j \zeta_i^{-n}, \quad n \geq \max(p, q+1) - p,$$

where $\zeta_i, i = 1, \ldots, k$ are the distinct zeros of $\phi(z)$ and r_i is the multiplicity of ζ_i. The p constants α_{ij}s and the coefficients ψ_j, $0 \le j < \max(p, q+1) - p$, are then determined uniquely by the $\max(p, q+1)$ boundary conditions (3.4).

Example 3.1. Consider the ARMA process $X_t - X_{t-1} + 0.25X_{t-2} = W_t + W_{t-1}$. We have $\phi(z) = 1 - z - (-0.25)z^2$ and $\theta(z) = 1 + z$. The root of $\phi(z)$ is 2 ($|2| > 1$) with multiplicity 2 and the root of $\theta(z) = -1$ ($|-1| = 1$) with multiplicity 1. So $\{X_t\}$ is causal but not invertible. To find the ACVF of $\{X_t\}$, we have

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 = \phi_1 + \theta_1 &= 1 + 1 = 2, \\ \psi_j - \psi_{j-1} + 0.25\psi_{j-2} &= 0, \quad j \ge 2. \end{aligned}$$

Transforming the last equation to $\psi_j - 0.5\psi_{j-1} = 0.5(\psi_{j-1} - 0.5\psi_{j-2})$, we see a geometric series with $\psi_1 - 0.5\psi_0 = 1.5$. Thus,

$$\psi_j - 0.5\psi_{j-1} = 3 \times 2^{-j}.$$

Then

$$\psi_j = (1 + 3j)2^{-j}, \quad j = 0, 1, 2, \ldots.$$

Now, we use the general solution; i.e., $r_i - 1 = 2 - 1 = 1$, $\zeta_i = 2$,

$$\psi_n = \sum_{j=0}^{1} \alpha_{ij} n^j 2^{-n}, \quad n \ge \max(p = 2, q + 1 = 1 + 1) - p = 0.$$

The constants α_{10} and α_{11} are found from the boundary conditions $\psi_0 = 1$ and $\psi_1 = 2$ to be

$$\alpha_{10} = 1 \quad \text{and} \quad \alpha_{11} = 3.$$

Then

$$\psi_j = (1 + 3j)2^{-j}, \quad j = 0, 1, 2, \ldots.$$

Thus

$$\begin{aligned} \gamma_X(h) =& \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} \\ =& \sigma^2 \sum_{j=0}^{\infty} (1 + 3j)(1 + 3j + 3h)2^{-2j-h} \\ =& \sigma^2 2^{-h}(32/3 + 8h). \end{aligned}$$

3.3.2 Second Method

The second method is based on the difference equations for $\gamma_X(k)$, $k = 0, 1, 2, \ldots$, which are obtained by multiplying each side of

$$\phi(B)X_t = \theta(B)W_t$$

by X_{t-k} and taking expectations, namely,

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = \sigma^2 \sum_{k \le j \le q} \theta_j\psi_{j-k}, \quad 0 \le k < \max(p, q+1), \tag{3.6}$$

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = 0, \quad k \ge \max(p, q+1). \tag{3.7}$$

The right-hand sides of these equations come from the representation $X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$.

The general solution of (3.6) has the form

$$\gamma_X(h) = \sum_{i=1}^{k} \sum_{j=0}^{r_i-1} \beta_{ij} h^j \zeta_i^{-h}, \quad h \ge \max(p, q+1) - p, \tag{3.8}$$

where the p constants β_{ij} and the covariances $\gamma_X(j), 0 \le j < \max(p, q+1) - p$, are uniquely determined from the boundary conditions (3.6) after computing $\psi_0, \psi_1, \ldots, \psi_q$ from (3.3).

Example 3.2. Consider Example 3.1. We have (3.7) as

$$\gamma(k) - \gamma(k-1) + 0.25\gamma(k-2) = 0, \quad k \ge 2,$$

with general solution

$$\gamma_X(n) = \sum_{j=0}^{1} \beta_{ij} h^j 2^{-h}, \quad h \ge 0.$$

The boundary conditions (3.6) are

$$\begin{aligned} \gamma(0) - \gamma(1) + 0.25\gamma(2) &= \sigma^2(\psi_0 + \psi_1), \\ \gamma(1) - \gamma(0) + 0.25\gamma(1) &= \sigma^2\psi_0, \end{aligned}$$

where $\psi_0 = 1$ and $\psi_1 = 2$. Using the general solution, we have

$$3\beta_{10} - 2\beta_{11} = 16\sigma^2, -3\beta_{10} + 5\beta_{11} = 8\sigma^2,$$

which results in $\beta_{11} = 8\sigma^2$ and $\beta_{10} = 32\sigma^2/3$. Finally, we have Then

$$\gamma_X(h) = \sigma^2 2^{-h}(32/3 + 8h).$$

3.3.3 Third Method

The numerical determination of the autocovariance function $\gamma_X(h)$ from equations (3.6) and (3.7) can be carried out readily by first finding $\gamma_X(0), \ldots, \gamma_X(p)$ from the equations with $k = 0, 1, \ldots, p$, and then using the subsequent equations to determine $\gamma_X(p+1), \gamma_X(p+2), \ldots$ recursively.

Example 3.3. Consider Example 3.1. We have

$$\begin{aligned}\gamma(2) - \gamma(1) + 0.25\gamma(2) =& 0,\\ \gamma(0) - \gamma(1) + 0.25\gamma(2) =& \sigma^2(\psi_0 + \psi_1),\\ \gamma(1) - \gamma(0) + 0.25\gamma(1) =& \sigma^2\psi_0,\end{aligned}$$

providing $\gamma_X(0) = 32\sigma^2/3$, $\gamma_X(1) = 28\sigma^2/3$ and $\gamma_X(0) = 20\sigma^2/3$. Then the higher lag autocovariances can now easily be found recursively from the equations

$$\gamma_X(k) = \gamma_X(k-1) - 0.25\gamma_X(k-2), \quad k = 3, 4, \ldots$$

Example 3.4. Now, we consider the causal AR(2) process,

$$(1 - \zeta_1^{-1}B)(1 - \zeta_2^{-1}B)X_t = W_t, \quad |\zeta_1|, |\zeta_2| > 1, \zeta_1 \neq \zeta_2.$$

Then,

$$\begin{aligned}\phi_1 =& \zeta_1^{-1} + \zeta_2^{-1},\\ \phi_2 =& -\zeta_1^{-1}\zeta_2^{-1}.\end{aligned}$$

Based on (3.8), we have

$$\gamma_X(h) = \sum_{i=1}^{2} \beta_{i1}\zeta_i^{-h}, \quad h \geq 0.$$

Boundary conditions provide

$$\begin{aligned}\gamma_X(0) - \phi_1\gamma_X(1) - \phi_2\gamma_X(2) &= \sigma^2\\ \gamma_X(1) - \phi_1\gamma_X(0) - \phi_2\gamma_X(1) &= 0\end{aligned}$$

Tedious calculation yields that

$$\gamma_X(h) = \frac{\sigma^2\zeta_1^2\zeta_2^2}{(\zeta_1\zeta_2 - 1)(\zeta_2 - \zeta_1)}\left\{\frac{\zeta_1^{1-h}}{(\zeta_1^2 - 1)} - \frac{\zeta_2^{1-h}}{(\zeta_2^2 - 1)}\right\}, \quad h \geq 0.$$

```
rho=function(h,z1,z2){rho0=z1/(z1^2-1)-z2/(z2^2-1)
res=(z1^(1-h)/(z1^2-1)-z2^(1-h)/(z2^2-1))/rho0
return(res)}
```

```
par(mfrow=c(3,1));par(mar=c(4,4,2,.5));h=seq(0,20,1)
plot(h,rho(h,2,5),type="o",xlab="Lag",
ylab=expression(rho[X](h)), ylim=c(-1,1),col="blue")
segments(-1,0,21,0,lty=2)
plot(h,rho(h,-10/9,2),type="o",xlab="Lag",
ylab=expression(rho[X](h)), ylim=c(-1,1),col="blue")
segments(-1,0,21,0,lty=2)
plot(h,rho(h,10/9,2),type="o",xlab="Lag",
ylab=expression(rho[X](h)), ylim=c(-1,1),col="blue")
segments(-1,0,21,0,lty=2)
```

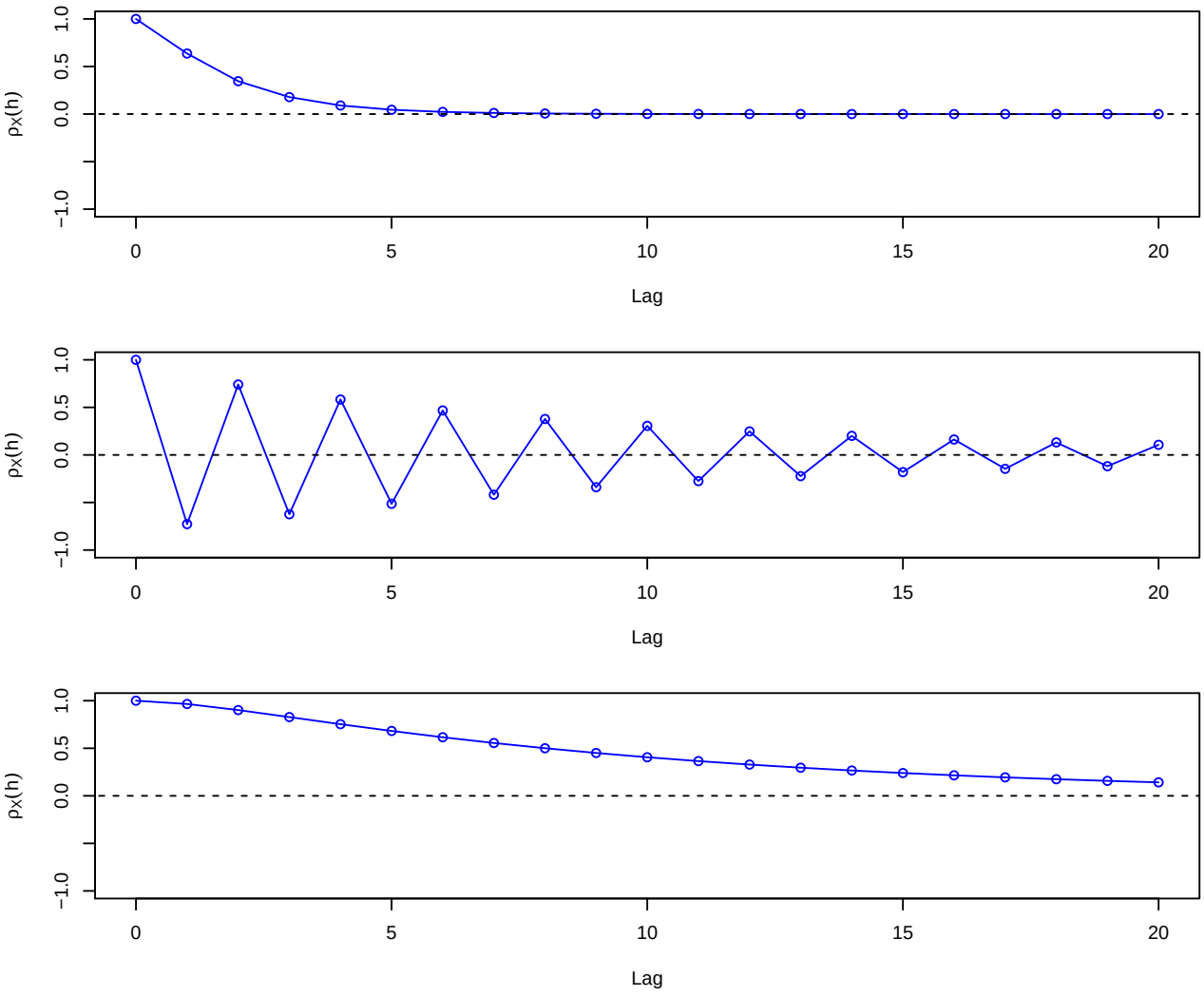


Figure 3.1: $\rho_X(h)$ for AR(2) with from top to bottom: $(\zeta_1, \zeta_2) = (2, 5)$,$(\zeta_1, \zeta_2) = (-10/9, 2)$, and $(\zeta_1, \zeta_2) = (10/9, 2)$

4 The Spectral Representation of a Stationary Process

In class, we did a brief introduction on complex valued numbers and complex valued random variables.

4.1 Complex-Valued Stationary Time Series

The process $\{X_t\}$ is a complex-valued stationary process if $\mathrm{E}|X_t|^2 < \infty$, $\mathrm{E}X_t$ is independent of t and $\mathrm{E}(X_{t+h}\bar{X}_t)$ is independent of t.The autocovariance function $\gamma_X(\cdot)$ is

$$\gamma_X(h) = \mathrm{E}(X_{t+h}\bar{X}_t) - \mathrm{E}(X_{t+j})\mathrm{E}(\bar{X}_t).$$

Similarly as the real-values stationary process, we have the properties of $\gamma_X(h)$:

Theorem 4.1. Basic properties of $\gamma_X(\cdot)$:

1. $\gamma_X(0) \geq 0$;
2. $|\gamma_X(h)| \leq \gamma(0)$ for all h;
3. $\gamma_X(h) = \overline{\gamma_X(-h)}$ for all h;
4. γ_X is Hermitian and nonnegative definite; i.e., a (possible complex) valued function κ defined on the integers is **Hermitian and nonnegative definite** if and only if $K(n) = \overline{K(-n)}$ and

$$\sum_{i,j=1}^{n} a_i K(i-j)\bar{a}_j \geq 0$$

for all positive integers n and real vectors $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{C}^n$.

4.2 The Spectral Distribution of a Linear Combination of Sinusoids

Consider the following simple complex-valued process,

$$X_t = \sum_{j=1}^{n} A(\lambda_j) e^{it\lambda_j} = \sum_{j=1}^{n} A(\lambda_j)\{\cos(\lambda_j t) + i\sin(\lambda_j t)\},$$

noting that $e^{ix} = \cos(x) + i\sin(x)$, in which $-\pi < \lambda_1 < \lambda_2 < \cdots < \lambda_n = \pi$ and $A(\lambda_1), \ldots, A(\lambda_n)$ are uncorrelated complex-valued random coefficients (possible zero) such that

$$\mathrm{E}\{A(\lambda_j)\} = 0, \quad j = 1, \ldots, n,$$

and

$$E\{A(\lambda_j)\overline{A(\lambda_j)}\} = \sigma_j^2, \quad j = 1, \ldots, n.$$

To check its stationarity, we have

$$\mathrm{E}(X_t) = 0$$

and

$$\begin{aligned}
\mathrm{E}(X_{t+h}\bar{X}_t) &= \mathrm{E}\left\{\sum_{j=1}^{n} A(\lambda_j) e^{i(t+h)\lambda_j} \times \sum_{j=1}^{n} \overline{A(\lambda_j)} e^{-it\lambda_j}\right\} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n} \mathrm{E}\left\{A(\lambda_j)\overline{A(\lambda_i)}\right\} e^{i(t+h)\lambda_j} e^{-it\lambda_i} \\
&= \sum_{j=1}^{n} \sigma_j^2 e^{ih\lambda_j}.
\end{aligned}$$

Thus, we have a complex-valued stationary process $\{X_t\}$ with autocovariance function

$$\begin{aligned}
\gamma_X(h) &= \sum_{j=1}^{n} \sigma_j^2 e^{ih\lambda_j} \\
&= \int_{(-\pi,\pi]} e^{ih\nu} dF(\nu),
\end{aligned}$$

where

$$F(\lambda) = \sum_{j:\lambda_j \le \lambda} \sigma_j^2.$$

The function F is known as the *spectral distribution function* of $\{X_t\}$.

Theorem 4.2. (Herglotz). A complex-valued function $\gamma_X(\cdot)$ defined on the integers is non-negative definite if and only if

$$\gamma_X(h) = \int_{(-\pi,\pi]} e^{ih\nu} dF(\nu), \quad \forall h = 0, \pm 1, \pm 2, \ldots,$$

where $F(\cdot)$ is a right-continuous, non-decreasing, bounded function on $[-\pi, \pi]$ and $F(-\pi) = 0$.

- The function F is called the **spectral distribution function** of γ_X and
- if $F(\lambda) = \int_{-\pi}^{\lambda} f(\nu) d\nu$, $-\pi \leq \lambda \leq \pi$, then f is called a **spectral density** of $\gamma_X(\cdot)$.

Corollary 4.1. A complex-valued function $\gamma_X(\cdot)$ defined on the integers is the ACVF of a stationary process $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$ if and only if either

(i) $\gamma_X(h) = \int_{(-\pi,\pi]} e^{ih\nu} dF(\nu)$ for all $h = 0, \pm 1, \pm 2, \ldots$, where F is a right-continuous, non-decreasing, bounded function on $[-\pi, \pi]$ and $F(-\pi) = 0$, or

(ii) $\sum_{i,j=1}^{n} a_i \gamma_X(i-j) \bar{a}_j \geq 0$ for all positive integers n and for all $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{C}^n$.

Corollary 4.2. An absolutely summable complex-valued function $\gamma(\cdot)$ defined on the integers is the autocovariance function of a stationary process if and only if

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \gamma(n) \geq 0, \quad \forall \lambda \in [-\pi, \pi],$$

in which case $f(\cdot)$ is the spectral density of $\gamma(\cdot)$.

Corollary (4.2) provides a way to calculate the spectral density of $\gamma_X(\cdot)$ of a stationary process $\{X_t\}$.

Example 4.1. For white noise $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$, we have $\gamma_W(h) = \sigma^2 I(h = 0)$. Its spectral density is then

$$f_W(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \gamma_X(n) = \frac{1}{2\pi} e^{-i0\lambda} \sigma^2 = \frac{\sigma^2}{2\pi}.$$

Example 4.2. Now let us calculate the spectral density of

- MA(1): $X_t = W_t + \theta W_{t-1}$
- AR(1): $X_t - \phi X_{t-1} = W_t$
- Is $f_X(t)$ always real-valued?

4.3 Spectral Densities and ARMA Processes

Theorem 4.3. If $\{Y_t\}$ is any zero-mean, possibly complex-valued stationary process with spectral distribution function $F_Y(\cdot)$, and $\{X_t\}$ is the process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \quad \text{where} \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty, \tag{4.1}$$

then $\{X_t\}$ is stationary with spectral distribution function

$$F_X(\lambda) = \int_{(-\pi,\lambda]} \left| \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\nu} \right|^2 dF_y(\nu), \quad -\pi \le \lambda \le \pi.$$

Proof. Similar argument of the proof of Proposition 2.2 provides that $\{X_\}$ is stationary with mean zero and ACVF

$$\gamma_X(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \bar{\psi}_k \gamma_Y(h-j+k), \quad h = 0, \pm 1, \pm 2, \ldots.$$

Using the spectral representation of $\gamma_Y(\cdot)$ we can write

$$\begin{aligned}
\gamma_X(h) &= \sum_{j,k=-\infty}^{\infty} \psi_j \bar{\psi}_k \int_{(-\pi,\pi]} e^{i(h-j+k)\nu} dF_Y(\nu) \\
&= \int_{(-\pi,\pi]} \left(\sum_{j=-\infty}^{\infty} \psi_j e^{-ij\nu} \right) \left(\sum_{k=-\infty}^{\infty} \bar{\psi}_k e^{ik\nu} \right) e^{ih\nu} dF_Y(\nu) \\
&= \int_{(-\pi,\pi]} e^{ih\nu} \left| \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\nu} \right|^2 dF_Y(\nu),
\end{aligned}$$

which completes the proof. □

Remark 4.1. If $\{Y_t\}$ has a spectral density $f_Y(\cdot)$ and if $\{X_t\}$ is defined by (4.1), then $\{X_t\}$ also has a spectral density $f_X(\cdot)$ given by

$$f_X(\lambda) = |\psi(e^{-i\lambda})|^2 f_Y(\lambda)$$

where $\psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$.

Theorem 4.4. (Spectral Density of an ARMA(p,q) Process). Let $\{X_t\}$ be an ARMA(p,q) process (not necessarily causal or invertible) satisfying

$$\phi(B)X_t = \theta(B)W_t, \quad \{W_t\} \sim \text{WN}(0, \sigma^2),$$

where $\phi(z) = 1 - \phi_1 z - \cdots \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ have no common zeroes and $\phi(z)$ has no zeroes on the unit circle. Then $\{X_t\}$ has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}, \quad -\pi \leq \lambda \leq \pi.$$

Proof. The assumption that $\phi(z)$ has no zeroes on the unit circle guarantees that X_t can be written as

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Based on Example 4.1, $\{W_t\}$ has spectral density $\sigma^2/(2\pi)$, then Theorem 4.3 implies that $\{X_t\}$ has a spectral density. Setting $U_t = \phi(B)X_t = \theta(B)W_t$ and applying Theorem 4.3, we obtain

$$f_U(\lambda) = |\phi(e^{-i\nu})|^2 f_X(\lambda) = |\theta(e^{-i\lambda})|^2 f_W(t)$$

Since $\phi(e^{-i\nu}) \neq 0$ for all $\lambda \in [-\pi, \pi]$ we can divide the above equation by $|\phi(e^{-i\nu})|^2$ to finish the proof. □

4.4 Causality, Invertibility and the Spectral Density of ARMA(p,q)

Consider the ARMA(p,q) process $\{X_t\}$ satisfying $\phi(B)X_t = \theta(B)W_t$, where $\phi(z)\theta(z) \neq 0$ for $|z| = 1$. Factorizing the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ we can rewrite the defining equations in the form,

$$\prod_{j=1}^{p}(1 - a_j^{-1}B)X_t = \prod_{j=1}^{q}(1 - b_j^{-1}B)W_t,$$

where

$$|a_j| > 1, 1 \leq j \leq r, \quad |a_j| < 1, r < j \leq p,$$

and

$$|b_j| > 1, 1 \leq j \leq s, \quad |b_j| < 1, s < j \leq q.$$

By Theorem 4.4, $\{X_t\}$ has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \frac{\prod_{j=1}^{q} |1 - b_j^{-1}e^{-i\lambda}|^2}{\prod_{j=1}^{p} |1 - a_j^{-1}e^{-i\lambda}|^2}.$$

Now define

$$\widetilde{\phi}(B) = \prod_{1\le j\le r}(1-a_j^{-1}B)\prod_{r<j\le p}(1-\bar{a}_jB) \tag{4.2}$$

and

$$\widetilde{\theta}(B) = \prod_{1\le j\le s}(1-b_j^{-1}B)\prod_{s<j\le q}(1-\bar{b}_jB).$$

Then we have $\{X_t\}$ is also the ARMA process defined by

$$\widetilde{\phi}(B)X_t = \widetilde{\theta}(B)\widetilde{W}_t.$$

where

$$\widetilde{W}_t = \frac{\prod_{r<j\le p}(1-\bar{a}_jB)\prod_{s<j\le q}(1-b_j^{-1}B)}{\prod_{r<j\le p}(1-a_j^{-1}B)\prod_{s<j\le q}(1-\bar{b}_jB)}W_t$$

Based on Theorem 4.4 again, $\{W_t^*\}$ has spectral density

$$f_{\widetilde{W}}(\lambda) = \frac{|\prod_{r<j\le p}(1-\bar{a}_je^{-i\lambda})\prod_{s<j\le q}(1-b_j^{-1}e^{-i\lambda})|^2}{|\prod_{r<j\le p}(1-a_j^{-1}e^{-i\lambda})\prod_{s<j\le q}(1-\bar{b}_je^{-i\lambda})|^2}\cdot\frac{\sigma^2}{2\pi}.$$

Since

$$|1-\bar{b}_je^{-i\lambda}| = |1-b_je^{i\lambda}| = |e^{i\lambda}|\cdot|b_j-e^{-i\lambda}| = |b_j|\cdot|1-b_j^{-1}e^{-i\lambda}|,$$

we can rewrite $f_{\widetilde{W}}(\lambda)$ as

$$f_{\widetilde{W}}(\lambda) = \frac{\prod_{r<j\le p}|a_j|^2}{\prod_{s<j\le q}|b_j|^2}\cdot\frac{\sigma^2}{2\pi}.$$

Thus

$$\{\widetilde{W}_t\} \sim \mathrm{WN}\left(0, \sigma^2\left\{\prod_{r<j\le p}|a_j|\right\}^2\left\{\prod_{s<j\le q}|b_j|\right\}^{-2}\right).$$

Noting that both $\widetilde{\phi}(z)$ and $\widetilde{\theta}(z)$ has no root in $|z|\le 1$. Thus, $\{X_t\}$ has the causal invertible representation

$$\widetilde{\phi}(B)X_t = \widetilde{\theta}(B)\widetilde{W}_t.$$

Example 4.3. The ARMA process

$$X_t - 2X_{t-1} = W_t + 4W_{t-1}, \quad \{W_t\} \sim \mathrm{WN}(0,\sigma^2),$$

is neither casual nor invertible. Introducing $\widetilde{\phi}(z) = 1-0.5z$ and $\widetilde{\theta}(z) = 1+0.25z$, we see that $\{X_t\}$ has the causal invertible representation

$$X_t - 0.5X_{t-1} = \widetilde{W}_t + 0.25\widetilde{W}_{t-1}, \quad \{\widetilde{W}_t\} \sim \mathrm{WN}(0,4\sigma^2).$$

5 Prediction of Stationary Processes

In this section, we consider to predict the value X_{n+h} for $h > 0$ of a stationary time series with known mean μ_X and ACVF γ_X in terms of the values $\{X_n, \ldots, X_1\}$. The prediction is constructed as a linear combination of $1, X_n, \ldots, X_1$ by minimizing the mean squared error (called *the optimal linear predictor*); i.e., we have the predictor as

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) = a_0 + a_1 X_n + \cdots + a_1 X_1,$$

where $\boldsymbol{a} = (a_0, \ldots, a_n)^{\mathrm{T}}$ minimizes

$$\mathbb{S}(\boldsymbol{a}) = \mathrm{E}(X_{n+h} - a_0 - a_1 X_n - \cdots - a_1 X_1)^2. \tag{5.1}$$

5.1 Predict X_{n+h} by X_n

We start with predicting X_{n+h} by X_n as $\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) = a_0 + a_1 X_n$. In this case, we have

$$\begin{aligned}
\mathbb{S}(\boldsymbol{a}) &= \mathrm{E}(X_{n+h} - a_0 - a_1 X_n)^2 \\
&= \mathrm{E}(X_{n+h}^2 + a_0^2 + a_1^2 X_n^2 - 2a_0 X_{n+h} - 2a_1 X_n X_{n+h} + 2a_0 a_1 X_n) \\
&= a_0^2 + (a_1^2 + 1)\{\gamma_X(0) + \mu_X^2\} + (2a_0 a_1 - 2a_0)\mu_X - 2a_1\{\gamma_X(h) + \mu_X^2\}.
\end{aligned}$$

Taking partial derivative of $\mathbb{S}(\boldsymbol{a})$ and setting to zero yields

$$\begin{aligned}
\frac{\partial \mathbb{S}(\boldsymbol{a})}{\partial a_0} &= 2a_0 + 2a_1\mu_X - 2\mu_X = 0 \\
\frac{\partial \mathbb{S}(\boldsymbol{a})}{\partial a_1} &= 2a_0\mu_X + 2a_1\{\gamma_X(0) + \mu_X^2\} - 2\{\gamma_X(h) + \mu_X^2\} = 0.
\end{aligned}$$

Solving this provides

$$a_1 = \rho_X(h) \quad \text{and} \quad a_0 = \mu_X\{1 - \rho_X(h)\}.$$

Finally, $\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) = \mu_X + \rho_X(h)\{X_n - \mu\}$ and $\mathrm{E}[\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) - X_n\}^2] = \gamma_X(0)\{1 - \rho_X^2(h)\}$.

- If $|\rho_X(h)| \to 1$, $\mathrm{E}[\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) - X_n\}^2] \to 0$ (accuracy improves)
- If $\rho_X(h) = \pm 1$, $\mathrm{E}[\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) - X_n\}^2] = 0$ (linearity)
- If $\rho_X(h) = 0$, $\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) = \mu_X$, and $\mathrm{E}[\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) - X_n\}^2] = \gamma_X(0)$ (uncorrelated)

If $\{X_t\}$ is Gaussian stationary, the joint distribution of (X_n, X_{n+h}) is then

$$\mathrm{N}\left\{ \begin{pmatrix} \mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \gamma_X(0) & \rho_X(h)\gamma_X(0) \\ \rho_X(h)\gamma_X(0) & \gamma_X(0) \end{pmatrix} \right\},$$

and the conditional distribution of X_{n+h} given X_n is

$$\mathrm{N}\left[\mu_X + \rho_X(h)(X_n - \mu_X), \gamma_X(0)\{1 - \rho_X^2(h)\}\right].$$

Thus

$$E(X_{n+h} \mid X_n) = \mu_X + \rho_X(h)(X_n - \mu_X).$$

Generally speaking, suppose we have a target Y and a set of predictor variables s $\boldsymbol{X}$. The optimal (least square) predictor of Y given $\boldsymbol{X}$ is $E(Y \mid \boldsymbol{X})$:

$$\begin{aligned}\min_f \mathrm{E}\{Y - f(\boldsymbol{X})^2\} =& \min_f \mathrm{E}[\{Y - f(\boldsymbol{X})^2\} \mid \boldsymbol{X}] \\ =& \mathrm{E}\left[\mathrm{E}\{Y - E(Y \mid \boldsymbol{X})\}^2 \mid \boldsymbol{X}\right].\end{aligned}$$

Thus the optimal predictor of Y given $\boldsymbol{X}$ is $E(Y \mid \boldsymbol{X})$.

- If $\{X_t\}$ is stationary, $\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) = \mu_X + \rho_X(h)\{X_n - \mu\}$ is the **optimal linear predictor**.
- If $\{X_t\}$ is also Gaussian, $\overline{\mathbb{P}}(X_{n+h} \mid X_n, 1) = \mu_X + \rho_X(h)\{X_n - \mu\}$ is the **optimal predictor**.
- This holds for longer histories, $\{X_n, X_{n-1}, \ldots, X_1\}$.

5.2 Predict X_{n+h} by $\{X_n, \ldots, X_1, 1\}$

To find $\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)$, we minimize function (5.1) to find the values of $\boldsymbol{a} = (a_0, a_1, \ldots, a_n)^{\mathrm{T}}$. Taking partial derivative and setting to zero, we have a system of equations

$$\frac{\partial \mathbb{S}(\boldsymbol{a})}{\partial a_j} = 0, \quad j = 0, \ldots, n,$$

which is

$$\begin{aligned}\mathrm{E}\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\right) &= 0 \\ \mathrm{E}\left\{\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\right) X_{n+1-j}\right\} &= 0.\end{aligned}$$

It can be seen that we have $\boldsymbol{a}_n = (a_1, \ldots, a_n)^{\mathrm{T}}$ is a solution of

$$\boldsymbol{\Gamma}_n \boldsymbol{a}_n = \boldsymbol{\gamma}_n(h) \tag{5.2}$$

and

$$a_0 = \mu_X \left(1 - \sum_{i=1}^{n} a_i\right).$$

where

$$\boldsymbol{\Gamma}_n = [\gamma_X(i-j)]_{i,j=1}^n, \quad \text{and } \boldsymbol{\gamma}_n(h) = (\gamma_X(h), \ldots, \gamma_X(h+n-1))^{\mathrm{T}}.$$

Hence, we have

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) = \mu_X + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu_X) \tag{5.3}$$

and

$$E\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) - X_{n+h}\}^2 = \gamma_X(0) - \boldsymbol{a}_n^{\mathrm{T}} \boldsymbol{\gamma}_n(h).$$

Now, we show the uniqueness of $\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)$ (left as a **HW** problem). Hint: suppose there are two different set of $\boldsymbol{a}$s: $\{a_{j1}, j = 0, \ldots, n\}$ and $\{a_{j2}, j = 0, \ldots, n\}$ such that

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) = a_{01} + a_{11} X_n + \cdots + a_{n1} X_1 = a_{02} + a_{12} X_n + \cdots + a_{n2} X_1.$$

Denote

$$Z = a_{01} - a_{02} + \sum_{j=1}^{n} (a_{j1} - a_{j2}) X_{n+1-j}.$$

Show $\mathrm{E}(Z^2) = 0$ which implies $Z = 0$.

Proposition 5.1. For a stationary process, if $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$, then the covariance matrix $\boldsymbol{\Gamma}_n = [\gamma_X(i-j)]_{i,j=1}^n$ is positive definite for every n.

Remark 5.1. When $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$, the uniqueness can be seen directly from Proposition **??**; i.e., in this case, $\boldsymbol{\Gamma}_n = [\gamma_X(i-j)]_{i,j=1}^n$ is non-singular for every n, thus (5.2) has a unique solution $\boldsymbol{a}_n = \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n(h)$. Further if $\mu_X = 0$, we have

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) = \sum_{i=1}^{n} \phi_{ni} X_{n+1-i}$$

and

$$E\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) - X_{n+h}\}^2 = \gamma_X(0) - \boldsymbol{\gamma}_n(h)^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n(h).$$

Properties of $\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1)$

1. $\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) = \mu_X + \sum_{i=1}^n a_i (X_{n+1-i} - \mu_X)$, where $\boldsymbol{a}_n = (a_1, \ldots, a_n)^{\mathrm{T}}$ satisfies $\boldsymbol{\Gamma}_n \boldsymbol{a}_n = \boldsymbol{\gamma}_n(h)$.
2. $\mathrm{E}\{\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1) - X_{n+h}\}^2 = \gamma_X(0) - \boldsymbol{a}_n^{\mathrm{T}} \boldsymbol{\gamma}_n(h)$.
3. $\mathrm{E}\{X_{n+h} - \overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1)\} = 0$.
4. $\mathrm{E}[\{X_{n+h} - \overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1, 1)\} X_j] = 0$, for $j = 1, \ldots, n$.

Remark 5.2. Notice that properties 3 and 4 can be interpreted easily by viewing $\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)$ as the projection of X_{n+h} on the linear subspace formed by $\{X_n, \ldots, X_1, 1\}$. This comes from the projection mapping theory of Hilbert spaces.

Hilbert space is a complete **inner product space**. An inner produce space is a vector space with inner product $< a, b >$:

- $< a, b >=< b, a >$
- $< \alpha_1 a_1 + \alpha_2 a_2, b >= \alpha_1 < a_1, b > +\alpha_2 < a_2, b >$
- $< a, a >= 0 \Leftrightarrow a = 0$
- Norm of a is $\|a\| =< a, a >$.

Note that **complete** means that every Cauchy sequence in the space has its limit in the space. Examples of Hilbert spaces include

1. $\mathbb{R}^n$ with $< \boldsymbol{x}, \boldsymbol{y} >= \sum x_i y_i$
2. $\mathcal{H} = \{X : \mathrm{E}X^2 < \infty\}$ with $< X, Y >= \mathrm{E}(XY)$

Hilbert space $\mathcal{H}$ is the one of interest in this course

Theorem 5.1. (The Projection Theorem). If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, then

(i) there is a unique element $\hat{x} \in \mathcal{M}$ such that

$$\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|,$$

and

(ii) $\hat{x} \in \mathcal{M}$ and $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x})$ is orthogonal to $\mathcal{M}$.

And we write $\overline{\mathbb{P}}(x \mid \mathcal{M})$ as the projection of x onto $\mathcal{M}$.

Proposition 5.2. (Properties of Projection Mappings). Let $\mathcal{H}$ be a Hilbert space and let $\overline{\mathbb{P}}(\cdot \mid \mathcal{M})$ denote the projection mapping onto a closed subspace $\mathcal{M}$. Then

(i) $\overline{\mathbb{P}}(\alpha x + \beta y \mid \mathcal{M}) = \alpha \overline{\mathbb{P}}(x \mid \mathcal{M}) + \beta \overline{\mathbb{P}}(x \mid \mathcal{M})$,

(ii) $\|x\|^2 = \|\overline{\mathbb{P}}(x \mid \mathcal{M}) + \{x - \overline{\mathbb{P}}(x \mid \mathcal{M})\}\|^2 = \|\overline{\mathbb{P}}(x \mid \mathcal{M})\|^2 + \|x - \overline{\mathbb{P}}(x \mid \mathcal{M})\|^2$,

(iii) each $x \in \mathcal{H}$ as a unique representation as a sum of an element of $\mathcal{M}$ and an element that is orthogonal to $\mathcal{M}$, i.e.,

$$x = \overline{\mathbb{P}}(x \mid \mathcal{M}) + \{x - \overline{\mathbb{P}}(x \mid \mathcal{M})\},$$

(iv) $\overline{\mathbb{P}}(x_n \mid \mathcal{M}) \to \overline{\mathbb{P}}(x \mid \mathcal{M})$ if $\|x_n - x\| \to 0$,

(v) $x \in \mathcal{M}$ if and only if $\overline{\mathbb{P}}(x \mid \mathcal{M}) = x$,

(vi) x is orthogonal to $\mathcal{M}$ if and only if $\overline{\mathbb{P}}(x \mid \mathcal{M}) = 0$,

(vii) $\mathcal{M}_1 \subset \mathcal{M}_2$ if and only if $\overline{\mathbb{P}}\{\overline{\mathbb{P}}(x \mid \mathcal{M}_2) \mid \mathcal{M}_1\} = \overline{\mathbb{P}}(x \mid \mathcal{M}_1)$ for all $x \in \mathcal{H}$.

5.3 General Case

Suppose now that Y and $Z_n, \ldots, Z_1$ are any random variables with finite second moments and that the means $\mu = \mathrm{E}Y$, $\mu_i = \mathrm{E}Z_i$ and covariance $\mathrm{Cov}(Y,Y)$, $\mathrm{Cov}(Y,Z_i)$, and $\mathrm{Cov}(Z_i,Z_j)$ are all known. Note that , this does not have to be related to a stationary process. Denote

$$\begin{aligned}
\boldsymbol{Z} =& (Z_n, \ldots, Z_1)^{\mathrm{T}}, \\
\boldsymbol{\mu}_Z =& (\mu_n, \ldots, \mu_1)^{\mathrm{T}}, \\
\boldsymbol{\gamma} =& (\mathrm{Cov}(Y, Z_n), \ldots, \mathrm{Cov}(Y, Z_1))^{\mathrm{T}}, \\
\boldsymbol{\Gamma} =& \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) = [\mathrm{Cov}(Z_{n+1-i}, Z_{n+1-j})]_{i,j=1}^{n}.
\end{aligned}$$

Then with the same argument,

$$\overline{\mathbb{P}}(Y \mid \boldsymbol{Z}, 1) = \mu_Y + \boldsymbol{a}^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{\mu}_Z)$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is any solution of

$$\boldsymbol{\Gamma a} = \boldsymbol{\gamma}.$$

And the mean squared error of this predictor is

$$\mathrm{E}[\{Y - \overline{\mathbb{P}}(Y \mid \boldsymbol{Z}, 1)\}^2] = \mathrm{Var}(Y) - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\gamma}.$$

Properties of the Prediction Operator of $\overline{\mathbb{P}}(\cdot \mid \boldsymbol{Z})$:
Suppose that $\mathrm{E}U^2 < \infty$, $\mathrm{E}V^2 < \infty$, $\boldsymbol{\Gamma} = \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z})$, and $\beta, \alpha_1, \ldots, \alpha_n$ are constants.

- $\overline{\mathbb{P}}(U \mid \boldsymbol{Z}) = \mathrm{E}U + \boldsymbol{a}^{\mathrm{T}}(\boldsymbol{Z} - \mathrm{E}\boldsymbol{Z})$, where $\boldsymbol{\Gamma a} = \mathrm{Cov}(U, \boldsymbol{Z})$.
- $\mathrm{E}[\{U - \overline{\mathbb{P}}(U \mid \boldsymbol{Z})\}\boldsymbol{Z}] = \boldsymbol{0}$ and $\mathrm{E}\{U - \overline{\mathbb{P}}(U \mid \boldsymbol{Z})\} = 0$
- $\mathrm{E}[\{U - \overline{\mathbb{P}}(U \mid \boldsymbol{Z})\}^2] = \mathrm{Var}(U) - \boldsymbol{a}^{\mathrm{T}}\mathrm{Cov}(U, \boldsymbol{Z})$
- $\overline{\mathbb{P}}(\alpha_1 U + \alpha_2 V + \beta \mid \boldsymbol{Z}) = \alpha_1 \overline{\mathbb{P}}(U \mid \boldsymbol{Z}) + \alpha_2 \overline{\mathbb{P}}(V \mid \boldsymbol{Z}) + \beta$
- $\overline{\mathbb{P}}\left(\sum_{i=1}^n \alpha_i Z_i + \beta \mid \boldsymbol{Z}\right) = \sum_{i=1}^n \alpha_i Z_i + \beta$
- $\overline{\mathbb{P}}(U \mid \boldsymbol{Z}) = \mathrm{E}U$ if $\mathrm{Cov}(U, \boldsymbol{Z}) = \boldsymbol{0}$.
- $\overline{\mathbb{P}}(U \mid \boldsymbol{Z}) = \overline{\mathbb{P}}\{\overline{\mathbb{P}}(U \mid \boldsymbol{Z}, \boldsymbol{V}) \mid \boldsymbol{Z}\}$ if $E(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})$ is finite.

These results comes directly from the standard projection mapping theory of a Hilbert space, in this case, this Hilber space is $\mathcal{H} = \{X : \mathrm{E}X^2 < \infty\}$ with $< X, Y >= \mathrm{E}(XY)$.

If $\mu = \mathrm{E}Y = 0$, $\mu_i = \mathrm{E}Z_i = 0$ (for example, we consider the zero-mean stationary process) We have

$$\begin{aligned}\boldsymbol{Z} =&(Z_n, \ldots, Z_1)^{\mathrm{T}},\\ \boldsymbol{\mu}_Z =&\mathbf{0},\\ \boldsymbol{\gamma} =&(\mathrm{Cov}(Y, Z_n), \ldots, \mathrm{Cov}(Y, Z_1))^{\mathrm{T}},\\ \boldsymbol{\Gamma} =&\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) = [\mathrm{Cov}(Z_{n+1-i}, Z_{n+1-j})]_{i,j=1}^{n}.\end{aligned}$$

It can be easily seen that

$$\overline{\mathbb{P}}(Y \mid \boldsymbol{Z}, 1) = \overline{\mathbb{P}}(Y \mid \boldsymbol{Z}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{Z}$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is any solution of

$$\boldsymbol{\Gamma a} = \boldsymbol{\gamma}.$$

And the mean squared error of this predictor is $\mathrm{Var}(Y) - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\gamma}$.

Example 5.1. For an AR(1) series: $X_t = \phi X_{t-1} + W_t$ where $|\phi| < 1$ and $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$. Find

(1) $\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_1, 1)$

(2) $\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2, 1)$

(3) $\overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2, 1)$

Solution: For Part (1) and (2), it suffices to find $\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_i)$. We have

$$\begin{aligned}\boldsymbol{Z} =&(X_n, \ldots, X_i)^{\mathrm{T}},\\ \boldsymbol{\gamma} =&\frac{\sigma^2}{1-\phi^2}(\phi, \phi^2, \ldots, \phi^{n-i})^{\mathrm{T}},\\ \boldsymbol{\Gamma} =&\frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-i} \\ \phi & 1 & \phi & \cdots & \phi^{n-i-1} \\ \vdots & & & & \vdots \\ \phi^{n-i} & \phi^{n-i-1} & \phi^2 & \cdots & 1 \end{pmatrix}.\end{aligned}$$

Equation $\boldsymbol{\Gamma a} = \boldsymbol{\gamma}$ yields that $\boldsymbol{a} = (\phi, 0, \ldots, 0)^{\mathrm{T}}$. Thus

$$\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_i, 1) = \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_i) = \phi X_n.$$

For Part (3), we have

$$\begin{aligned}
\boldsymbol{Z} =& (X_n, \ldots, X_2)^{\mathrm{T}}, \\
\boldsymbol{\gamma} =& \frac{\sigma^2}{1-\phi^2}(\phi^{n-1}, \phi^{n-2}, \ldots, \phi)^{\mathrm{T}}, \\
\boldsymbol{\Gamma} =& \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-2} \\ \phi & 1 & \phi & \cdots & \phi^{n-3} \\ \vdots & & & & \vdots \\ \phi^{n-2} & \phi^{n-3} & \phi^2 & \cdots & 1 \end{pmatrix}.
\end{aligned}$$

Equation $\boldsymbol{\Gamma a} = \boldsymbol{\gamma}$ yields that $\boldsymbol{a} = (0, \ldots, 0, \phi)^{\mathrm{T}}$. Thus

$$\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2, 1) = \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2) = \phi X_2.$$

Example 5.2. For the causal AR(p) process defined by

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t, \quad \{W_t\} \sim \mathrm{WN}(0, \sigma^2),$$

where W_t is uncorrelated with X_s for $s < t$. Then we have

$$\begin{aligned}
\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_1, 1) &= \overline{\mathbb{P}}(\phi_1 X_n + \cdots + \phi_p X_{n+1-p} + W_{n+k} \mid X_n, \ldots, X_1, 1) \\
&= \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + \overline{\mathbb{P}}(W_{n+1} \mid X_n, \ldots, X_1, 1) \\
&= \phi_1 X_n + \cdots + \phi_p X_{n+1-p}.
\end{aligned}$$

Example 5.3. For any zero-mean stationary process $\{X_t\}$, suppose we have

$$\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2) = \sum_{j=1}^{n-1} a_j X_{n+1-j},$$

then

$$\overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) = \sum_{j=1}^{n-1} a_{n-j} X_{n+1-j},$$

and

$$\mathrm{E}\{X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2)\}^2 = \mathrm{E}\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}^2 = \mathrm{E}\{X_n - \overline{\mathbb{P}}(X_n \mid X_{n-1}, \ldots, X_1)\}^2.$$

5.4 The Partial Autocorrelation Fucntion (PACF)

Like the autocorrelation function, the PACF is another tool that conveys vital information regarding the dependence structure of a stationary process and depends only on the second order properties of the process.

The partial autocorrelation function (PACF) $\alpha_X(\cdot)$ of a stationary time series is defined by

$$\alpha_X(1) = \text{Corr}(X_2, X_1) = \rho_X(1),$$

and

$$\alpha_X(k) = \text{Corr}\{X_{k+1} - \overline{\mathbb{P}}(X_{k+1} \mid X_k, \ldots, X_2, 1),\ X_1 - \overline{\mathbb{P}}(X_1 \mid X_k, \ldots, X_2, 1)\}.$$

The value of $\alpha_X(k)$ is known as the partial autocorrelation of $\{X_t\}$ at lag k.

The PCVF $\alpha_X(k)$ may be regarded as the correlation between X_1 and X_{k+1}, adjusted for the intervening observations $X_2, \ldots, X_k$.

Remark 5.3. The definition in the above box define $\alpha_X(k)$ based on $\{X_1, X_2, \ldots, X_k, X_{k+1}\}$. But, it is also equivalent as the one based on $\{X_{t+1}, X_{t+2}, \ldots, X_{t+k}, X_{t+k+1}\}$ for any $t > 0$; i.e.,

$$\alpha_X(k) = \text{Corr}\{X_{t+k+1} - \overline{\mathbb{P}}(X_{t+k+1} \mid X_{t+k}, \ldots, X_{t+2}, 1),\ X_{t+1} - \overline{\mathbb{P}}(X_{t+1} \mid X_{t+k}, \ldots, X_{t+2}, 1)\}.$$

Example 5.4. Let $\{X_t\}$ be the zero mean AR(1) process

$$X_t = \phi X_{t-1} + W_t.$$

Then

$$\alpha_X(1) = \text{Corr}(X_2, X_1) = \text{Corr}(\phi X_1 + W_2, X_1) = \phi.$$

Based on Example 5.1, we have $\overline{\mathbb{P}}(X_{k+1} \mid X_k, \ldots, X_2, 1) = \phi X_k$ and $\overline{\mathbb{P}}(X_1 \mid X_k, \ldots, X_2, 1) = \phi X_2$. Then for $k \geq 2$

$$\begin{aligned}\alpha_X(k) =& \text{Corr}(X_{k+1} - \phi X_k, X_1 - \phi X_2)\\ =& \text{Corr}(W_k, X_1 - \phi X_2)\\ =& 0.\end{aligned}$$

This says that the correlation between X_{k+1} and X_1

A **HW**: For the MA(1) process: $X_t = W_t + \theta W_{t-1}$, $|\theta| < 1$, $\{W_t\} \sim \text{WN}(0, \sigma^2)$, find its PACF.

Corollary 5.1. Let $\{X_t\}$ be a zero-mean stationary process with $\gamma_X(h)$ such that $\gamma_X(h) \to 0$ as $h \to 0$. Then

$$\overline{\mathbb{P}}(X_{k+1} \mid X_k, \ldots, X_1, 1) = \sum_{j=1}^{k} \phi_{kj} X_{k+1-j}.$$

Then from the equations

$$\mathrm{E}[\{X_{k+1} - \overline{\mathbb{P}}(X_{k+1} \mid X_k, \ldots, X_1, 1)\}X_j] = 0, \quad j = k, \ldots, 1.$$

We have

$$\begin{pmatrix} \rho_X(0) & \rho_X(1) & \rho_X(2) & \cdots & \rho_X(k-1) \\ \rho_X(1) & \rho_X(0) & \rho_X(1) & \cdots & \rho_X(k-2) \\ \vdots & & & & \vdots \\ \rho_X(k-1) & \rho_X(k-2) & \rho_X(k-3) & \cdots & \rho_X(0) \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho_X(1) \\ \rho_X(2) \\ \vdots \\ \rho_X(k) \end{pmatrix}. \tag{5.4}$$

The partial autocorrelation $\alpha_X(k)$ of $\{X_t\}$ at lag k is

$$\alpha_X(k) = \phi_{kk}, \quad k \geq 1,$$

Proof. We will prove this corollary later. □

The **sample partial partial autocorrelation** $\widehat{\alpha}_X(k)$ at lag k of $\{x_1, \ldots, x_n\}$ is defined, provided $x_i \neq x_j$ for some i and j, by

$$\widehat{\alpha}_X(k) = \widehat{\phi}_{kk}, \quad 1 \leq k \leq n,$$

where $\widehat{\phi}_{kk}$ is uniquely determined by (5.4) with each $\rho_X(j)$ replaced by the corresponding sample autocorrelation $\widehat{\rho}_X(j)$.

5.5 Recursive Methods for Computing Best Linear Predictors

In this section, we focus on zero-mean stationary processes. We establish two recursive algorithms for determining the one-step predictors,

$$\widehat{X}_{n+1} = \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_1)$$

and show how they can be used to compute the h-step predictors

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)$$

5.5.1 Recursive Prediction Using the Durbin-Levinson Algorithm

We can express $\widehat{X}_{n+1}$ in the form

$$\widehat{X}_{n+1} = \phi_{n1}X_n + \cdots + \phi_{nn}X_1, \quad n \geq 1.$$

And its mean squared error of prediction will be denoted by ν_n as

$$\nu_n = E(X_{n+1} - \widehat{X}_{n+1})^2, \quad n \geq 1.$$

Clearly, $\nu_0 = \gamma_X(0)$. The following proposition specified an algorithm, known as the Durbin-Levinson algorithm, which is a recursive shceme for computing $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})^{\mathrm{T}}$ and ν_n for $n = 1, 2, \ldots$.

Proposition 5.3. (**The Durbin-Levinson Algorithm**). If $\{X_t\}$ is a zero-mean stationary process with ACVF $\gamma_X(\cdot)$ such that $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$, then the coefficients ϕ_{nj} and mean squared errors ν_n as defined above satisfy $\phi_{11} = \gamma_X(1)/\gamma_X(0)$, $\nu_0 = \gamma_X(0)$,

$$\phi_{nn} = \left\{ \gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma_X(n-j) \right\} \nu_{n-1}^{-1},$$

$$\begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{nn} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}$$

and

$$\nu_n = \nu_{n-1}(1 - \phi_{nn}^2).$$

Proof. We consider the Hilbert space $\mathcal{H} = \{X : \mathrm{E}X^2 < \infty\}$ with inner produce $< X, Y >= \mathrm{E}(XY)$ with norm $\|X\|^2 =< X, X >$. By the definition of $\widehat{X}_{n+1}$, we can view $\widehat{X}_{n+1}$ is in the linear space of $\mathcal{H}$ spanned by $\{X_n, \ldots, X_1\}$, denoted by $\overline{\mathrm{sp}}\{X_n, \ldots, X_1\} \doteq \{Y : Y = a_1X_n + \cdots + a_nX_1 \text{ where } a_1, \ldots, a_n \in \mathbb{R}\}$. Since $X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)$ is orthogonal to $X_n, \ldots, X_2$; i.e.,

$$< X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2), X_k >= 0, \quad k = 2, \ldots, n.$$

We have

$$\begin{aligned} \overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1\} &= \overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\} \\ &= \overline{\mathrm{sp}}\{X_n, \ldots, X_2\} + \overline{\mathrm{sp}}\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}. \end{aligned}$$

Thus

$$\widehat{X}_{n+1} = \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2) + a\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}, \tag{5.5}$$

where

$$a = \frac{< X_{n+1}, X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) >}{\|X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\|^2}. \tag{5.6}$$

By stationary, we have

$$\overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{j+1} \tag{5.7}$$

$$\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2) = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{n+1-j} \tag{5.8}$$

and

$$\begin{aligned} \|X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\|^2 &= \|X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2)\|^2 \\ &= \|X_n - \overline{\mathbb{P}}(X_n \mid X_{n-1}, \ldots, X_1)\|^2 = \nu_{n-1}. \end{aligned} \tag{5.9}$$

Then Equations (5.5), (5.7) and (5.8) provide

$$\widehat{X}_{n+1} = aX_1 + \sum_{j=1}^{n-1} (\phi_{n-1,j} - a\phi_{n-1,n-j}) X_{n+1-j}, \tag{5.10}$$

where from Equation (5.6) and (5.7),

$$\begin{aligned} a &= \left(< X_{n+1}, X_1 > - \sum_{j=1}^{n-1} \phi_{n-1,j} < X_{n+1}, X_{j+1} > \right) \nu_{n-1}^{-1} \\ &= \left\{ \gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j} \gamma_X(n-j) \right\} \nu_{n-1}^{-1}. \end{aligned}$$

Remark 5.1 told us that when $\gamma_X(h) \to 0$ as $h \to \infty$ guarantees that the representation

$$\widehat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j} \tag{5.11}$$

is unique. And comparing coefficients in (5.10) and (5.11), we therefore deduce that

$$\phi_{nn} = a$$

and

$$\phi_{nj} = \phi_{n-1,j} - a\phi_{n-1,n-j}, \quad j = 1, \ldots, n-1.$$

Lastly,

$$\begin{aligned}
\nu_n &= \|X_{n+1} - \widehat{X}_{n+1}\|^2 \\
&= \|X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2) - a\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}\|^2 \\
&= \|X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2)\|^2 + a^2\|\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}\|^2 \\
&\quad - 2a < X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2), X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) > \\
&= \nu_{n-1} + a^2\nu_{n-1} - 2a < X_{n+1}, X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) >
\end{aligned}$$

Based on (5.6) and (5.9), we have

$$\nu_n = \nu_{n-1} + a^2\nu_{n-1} - 2a^2\nu_{n-1} = \nu_{n-1}(1 - a^2).$$

□

Now we prove Corollary 5.1: The partial autocorrelation $\alpha_X(k)$ of $\{X_t\}$ at lag k is

$$\alpha_X(k) = \phi_{kk}, \quad k \geq 1.$$

Proof. We have

$$\begin{aligned}
\phi_{nn} = a &= \frac{< X_{n+1}, X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) >}{\|X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\|^2} \\
&= \frac{< X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2), X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) >}{\|X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\|^2} \\
&= \frac{< X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2), X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2) >}{\|X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2)\| \cdot \|X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\|} \\
&= \mathrm{Corr}\{X_{n+1} - \overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_2), X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\} \\
&= \alpha_X(n).
\end{aligned}$$

□

5.5.2 Recursive Prediction Using the Innovations Algorithm

The Durbin-Levinson Algorithm is based on the decomposition of $\overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1\}$ into two orthogonal subspaces:

$$\begin{aligned}
\overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1\} &= \overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\} \\
&= \overline{\mathrm{sp}}\{X_n, \ldots, X_2\} + \overline{\mathrm{sp}}\{X_1 - \overline{\mathbb{P}}(X_1 \mid X_n, \ldots, X_2)\}.
\end{aligned}$$

The Innovation Algorithm is based on the decomposition of $\overline{\mathrm{sp}}\{X_n, \ldots, X_2, X_1\}$ to n orthogonal subspaces; i.e.,

$$\begin{aligned}
\overline{\mathrm{sp}}\{X_n, \ldots, X_1\} &= \overline{\mathrm{sp}}\{X_1 - \widehat{X}_1, X_2 - \widehat{X}_2, \ldots, X_n - \widehat{X}_n\} \\
&= \overline{\mathrm{sp}}\{X_1 - \widehat{X}_1\} + \overline{\mathrm{sp}}\{X_2 - \widehat{X}_2\} + \cdots + \overline{\mathrm{sp}}\{X_n - \widehat{X}_n\}
\end{aligned}$$

where noting that

$$\widehat{X}_i = \overline{\mathbb{P}}(X_i \mid X_{i-1}, \ldots, X_1) \quad \text{and} \quad \widehat{X}_1 = 0.$$

Thus, we have

$$\widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}).$$

We now establish the recursive scheme for computing $\{\theta_{nj}, j = 1, \ldots, n; \nu_n\}$ for $n = 1, 2, \ldots$ through the following proposition. Note that, this proposition is more generally applicable than the previous one since we allow $\{X_t\}$ to be a possible non-stationary with mean zero and autocovariance function

$$\kappa_X(i, j) =< X_i, X_j >= \mathrm{E}(X_i X_j).$$

Proposition 5.4. (**The Innovations Algorithm**). If $\{X_t\}$ is a process with mean zero and $\mathrm{E}(X_iX_j) = \kappa_X(i,j)$, where the matrix $[\kappa_X(i,j)]_{i,j=1}^n$ is non-singular for each $n = 1, 2, \ldots$, then the one-step predictors $\widehat{X}_{n+1}$, $n \geq 0$, and their mean squared errors ν_n, $n \geq 1$, are given by

$$\widehat{X}_{n+1} = \begin{cases} 0, & n = 0, \\ \sum_{j=1}^n \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & n \geq 1, \end{cases}$$

and

$$\begin{cases} \nu_0 = \kappa_X(1,1), \\ \theta_{11} = \nu_0^{-1}\kappa_X(2,1) \\ \nu_1 = \kappa_X(2,2) - \theta_{11}^2\nu_0 \\ \theta_{nn} = \nu_0^{-1}\kappa_X(n+1,1), n \geq 2 \\ \theta_{n,n-k} = \nu_k^{-1}\left\{\kappa_X(n+1,k+1) - \sum_{j=0}^{k-1}\theta_{k,k-j}\theta_{n,n-j}\nu_j\right\}, \quad k = 1, \ldots, n-1, n \geq 2 \\ \nu_n = \kappa_X(n+1,n+1) - \sum_{j=0}^{n-1}\theta_{n,n-j}^2\nu_j, n \geq 2. \end{cases}$$

Proof. By the orthogonality, we have

$$\begin{aligned} < \widehat{X}_{n+1}, X_{k+1} - \widehat{X}_{k+1} > &= < \sum_{j=1}^n \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), X_{k+1} - \widehat{X}_{k+1} > \\ &= \theta_{n,n-k} < X_{k+1} - \widehat{X}_{k+1}, X_{k+1} - \widehat{X}_{k+1} > = \theta_{n,n-k}\nu_k. \end{aligned}$$

Thus

$$\begin{aligned} \theta_{n,n-k} &= \nu_k^{-1} < \widehat{X}_{n+1}, X_{k+1} - \widehat{X}_{k+1} > \\ &= \nu_k^{-1} < X_{n+1}, X_{k+1} - \widehat{X}_{k+1} > \\ &= \nu_k^{-1}\left\{\kappa_X(n+1,k+1) - \sum_{j=1}^k \theta_{k,j} < X_{n+1}, X_{k+1-j} - \widehat{X}_{k+1-j} >\right\} \\ &= \nu_k^{-1}\left\{\kappa_X(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j} < X_{n+1}, X_{j+1} - \widehat{X}_{j+1} >\right\} \\ &= \nu_k^{-1}\left\{\kappa_X(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}\nu_j\right\}. \end{aligned}$$

Then

$$\begin{aligned}
\nu_n &= \|X_{n+1} - \widehat{X}_{n+1}\|^2 = \|X_{n+1}\|^2 - \|\widehat{X}_{n+1}\|^2 \\
&= \kappa_X(n+1, n+1) - \left\| \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}) \right\|^2 \\
&= \kappa_X(n+1, n+1) - \sum_{j=1}^{n} \theta_{nj}^2 \left\| X_{n+1-j} - \widehat{X}_{n+1-j} \right\|^2 \\
&= \kappa_X(n+1, n+1) - \sum_{j=1}^{n} \theta_{nj}^2 \nu_{n-j} \\
&= \kappa_X(n+1, n+1) - \sum_{k=0}^{n-1} \theta_{n,n-k}^2 \nu_k.
\end{aligned}$$

□

Example 5.5. (Prediction of an MA(1) Process Using the Innovations Algorithm). If $\{X_t\}$ is the process,

$$X_t = W_t + \theta W_{t-1}, \quad \{W_t\} \sim \text{WN}(0, \sigma^2),$$

Then

$$\begin{aligned}
\nu_0 &= (1+\theta^2)\sigma^2, \quad \theta_{11} = \nu_0^{-1}\kappa_X(2,1) = \nu_0^{-1}\theta\sigma^2, \\
\nu_1 &= (1+\theta^2)\sigma^2 - \theta_{11}^2\nu_0 = (1 + \theta^2 - \nu_0^{-1}\theta^2\sigma^2)\sigma^2, \\
\theta_{22} &= 0, \quad \theta_{21} = \nu_1^{-1}\theta\sigma^2, \quad \cdots \\
\nu_n &= (1 + \theta^2 - \nu_n^{-1}\theta^2\sigma^2)\sigma^2.
\end{aligned}$$

If we define $r_n = \nu_n/\sigma^2$, then we can write

$$\widehat{X}_{n+1} = \theta(X_n - \widehat{X}_n)/r_{n-1},$$

where $r_0 = 1 + \theta^2$ and $r_{n+1} = 1 + \theta^2 - \theta^2/r_n$.

5.5.3 Recursive Calculation of the h-Step Predictors, $h \geq 1$

Since $\overline{\text{sp}}\{X_n, \ldots, X_1\}$ is a linear subspace of $\overline{\text{sp}}\{X_{n+h-1}, \ldots, X_1\}$, and when $j < h$, $X_{n+h-j} - \widehat{X}_{n+h-j}$ is orthogonal to $\overline{\text{sp}}\{X_n, \ldots, X_1\}$, we have

$$\begin{aligned}
\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1) &= \overline{\mathbb{P}}\{\overline{\mathbb{P}}(X_{n+h} \mid X_{n+h-1}, \ldots, X_1) \mid X_n, \ldots, X_1\} \\
&= \overline{\mathbb{P}}(\widehat{X}_{n+h} \mid X_n, \ldots, X_1) \\
&= \overline{\mathbb{P}}\left\{ \sum_{j=1}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}) \mid X_n, \ldots, X_1 \right\} \\
&= \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}), \qquad (5.12)
\end{aligned}$$

$$(5.13)$$

Further the mean squared error can be expressed as

$$\begin{aligned}
E\{X_{n+h} - \overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)\}^2 &= \|X_{n+h}\|^2 - \|\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)\|^2 \\
&= \kappa_X(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2 \nu_{n+h-j-1}.
\end{aligned}$$

Remark 5.4. Note that, while the Durbin-Levison algorithm gives the coefficients $X_1, \ldots, X_n$ in the representation of $\widehat{X}_{n+1} = \sum_{j=1}^n \phi_{nj} X_{n+1-j}$, the Innovations algorithm gives the coefficients of the "innovations", $(X_j - \widehat{X}_j), j = 1, \ldots, n$, in the orthogonal expansion $\widehat{X}_{n+1} = \sum_{j=1}^n \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j})$. The latter expansion is extremely simple to use, especially, in the case of ARMA(p, q) processes.

5.6 Recursive Prediction of an ARMA(p, q) Process

For an causal ARMA(p, q) process $\{X_t\}$ defined by

$$\phi(B)X_t = \theta(B)W_t, \quad \{W_t\} \sim \text{WN}(0, \sigma^2).$$

Instead of applying the Innovations algorithm to $\{X_t\}$, we apply it to the transformed process

$$\begin{cases} Z_t = \sigma^{-1} X_t, & t = 1, \ldots, m \\ Z_t = \sigma^{-1}\phi(B)X_t = \sigma^{-1}(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}) = \sigma^{-1}\theta(B)W_t, & t > m, \end{cases}$$

where $m = \max(p, q)$. For notational convenience, we define $\theta_0 = 1$ and assume that $p \geq 1$ and $q \geq 1$.

It can be seen that

$$\overline{\text{sp}}\{X_1, \ldots, X_n\} = \overline{\text{sp}}\{Z_1, \ldots, Z_n\}, \quad n \geq 1.$$

For $n \geq 1$, we also use the notation $\widehat{X}_{n+1}$ and $\widehat{Z}_{n+1}$ to denote the predictor

$$\widehat{X}_{n+1} = \overline{\mathbb{P}}(X_{n+1} \mid X_1, \ldots, X_n) \quad \text{and} \quad \widehat{Z}_{n+1} = \overline{\mathbb{P}}(Z_{n+1} \mid Z_1, \ldots, Z_n),$$

respectively. Of course, we have $\widehat{X}_1 = \widehat{Z}_1 = 0$.

Now we apply the Innovations algorithm to $\{Z_t\}$. For $\{Z_t\}$, we have

$$\kappa_Z(i,j) = \begin{cases} \sigma^{-2}\gamma_X(i-j), & 1 \leq i,j \leq m, \\ \sigma^{-2}\left\{\gamma_X(i-j) - \sum_{r=1}^{p} \phi_r \gamma_X(r-|i-j|)\right\}, & \min(i,j) \leq m < \max(i,j) \leq 2m, \\ \sum_{r=0}^{q} \theta_r \theta_{r+|i-j|} & \min(i,j) > m, \\ 0, & \text{otherwise}, \end{cases} \tag{5.14}$$

where we set $\theta_j = 0$ for $j > q$.

Then based on the Innovations algorithm, we can obtained θ_{nj}s such that

$$\begin{cases} \widehat{Z}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(Z_{n+1-j} - \widehat{Z}_{n+1-j}), & 1 \leq n < m, \\ \widehat{Z}_{n+1} = \sum_{j=1}^{q} \theta_{nj}(Z_{n+1-j} - \widehat{Z}_{n+1-j}), & n \geq m, \end{cases}$$

and

$$r_n = \mathrm{E}(Z_{n+1} - \widehat{Z}_{n+1})^2.$$

It is worthwhile to point out that $\theta_{nj} = 0$ when both $n \geq m$ and $j > q$. Why?

Now, we show the relationship between $\widehat{X}_t$ and $\widehat{Z}_t$. When $t = 1, \ldots, m$

$$\widehat{Z}_t = \overline{\mathbb{P}}(Z_t \mid Z_{t-1}, \ldots, Z_1) = \overline{\mathbb{P}}(\sigma^{-1}X_t \mid X_{t-1}, \ldots, X_1) = \sigma^{-1}\widehat{X}_t.$$

For $t > m$, we have

$$\begin{aligned} \widehat{Z}_t &= \overline{\mathbb{P}}(Z_t \mid Z_{t-1}, \ldots, Z_1) = \overline{\mathbb{P}}(\sigma^{-1}\phi(B)X_t \mid X_{t-1}, \ldots, X_1) \\ &= \sigma^{-1}\overline{\mathbb{P}}(X_t - \phi X_{t-1} - \cdots - \phi^p X_{t-p} \mid X_{t-1}, \ldots, X_1) = \sigma^{-1}(\widehat{X}_t - \phi X_{t-1} - \cdots - \phi^p X_{t-p}) \\ &= \sigma^{-1}\{\widehat{X}_t + \phi(B)X_t - X_t\}. \end{aligned}$$

Thus

$$X_t - \widehat{X}_t = \sigma(Z_t - \widehat{Z}_t), \quad \forall t \geq 1.$$

Thus, when $1 \leq n < m$

$$\widehat{X}_{n+1} = X_{n+1} - \sigma(Z_{n+1} - \widehat{Z}_{n+1}) = (X_{n+1} - \sigma Z_{n+1}) + \sigma\widehat{Z}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}).$$

when $n \geq m$,

$$\begin{aligned}\widehat{X}_{n+1} &= X_{n+1} - \sigma(Z_{n+1} - \widehat{Z}_{n+1}) \\ &= X_{n+1} - (X_{n+1} - \phi_1 X_n - \cdots - \phi_p X_{n+1-p}) + \sigma \widehat{Z}_{n+1} \\ &= \phi X_n + \cdots + \phi^p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}).\end{aligned}$$

Thus, we have

$$\begin{cases} \widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & 1 \leq n < m, \\ \widehat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & n \geq m, \end{cases} \tag{5.15}$$

and further

$$\mathrm{E}(X_{n+1} - \widehat{X}_{n+1})^2 = \sigma^2 \mathrm{E}(Z_{n+1} - \widehat{Z}_{n+1})^2 = \sigma^2 r_n. \tag{5.16}$$

Equations (5.14), (5.15) and (5.16) provides a recursive calculation of one-step predictor $\overline{\mathbb{P}}(X_{n+1} \mid X_n, \ldots, X_1)$ for a general ARMA(p, q) process.

Remark 5.5. Note that, the covariance $\kappa_Z(i, j)$ depend only on $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ and not on σ^2. The same is therefore true of θ_{nj} and r_n.

Remark 5.6. The representation (5.15) is particularly convenient from a practical point of view, not only because of the simple recusion relations for the coefficients, but also because for $n \geq m$ it requires the storage of at most p past observations $X_n, \ldots, X_{n+1-p}$ and at most q past innovations $(X_{n+1-j} - \widehat{X}_{n+1-j})$, $j = 1, \ldots, q$, in order to predict X_{n+1}. Direct application of the Innovations algorithm to $\{X_t\}$ on the other hand leads to a representation of $\widehat{X}_{n+1}$ in terms of all the n preceding innovations $(X_j - \widehat{X}_j)$, $j = 1, \ldots, n$.

Example 5.6. (Prediction of an MA(q) Process). For an MA(q) process; i.e., $\{X_t\}$ is defined as

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}.$$

which can be viewed as ARMA$(1, q)$ process. Thus, we can apply (5.15) and obtain

$$\begin{cases} \widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & 1 \le n < q, \\ \widehat{X}_{n+1} = \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & n \ge q, \end{cases}$$

where the coefficients θ_{nj} are found by the Innovations algorithm using the covariance $\kappa_Z(i, j)$, where

$$\begin{cases} Z_t = \sigma^{-1} X_t, & t = 1, \ldots, q \\ Z_t = \sigma^{-1}\phi(B) X_t = \sigma^{-1} X_t, & t > q, \end{cases} .$$

Thus the process $\{Z_t\}$ and $\{\sigma^{-1} X_t\}$ are identical, and the covariances are simply and

$$\kappa_Z(i, j) = \sigma^{-2}\gamma_X(i - j) = \sum_{r=0}^{q-|i-j|} \theta_r \theta_{r+|i-j|}.$$

5.6.1 h-step prediction of an ARMA(p, q) process

When $h \ge 1$, similarly as in Section 5.5.3; based on (5.12), we have,

$$\begin{aligned} \overline{\mathbb{P}}(\widehat{Z}_{n+h} \mid Z_n, \ldots, Z_1) &= \overline{\mathbb{P}}(\widehat{Z}_{n+h} \mid X_n, \ldots, X_1) \\ &= \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(Z_{n+h-j} - \widehat{Z}_{n+h-j}) \\ &= \sigma^{-1} \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}). \end{aligned}$$

Then when $1 \le h \le m - n$,

$$\begin{aligned} \overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1) &= \overline{\mathbb{P}}\{X_{n+h} - \sigma(Z_{n+h} - \widehat{Z}_{n+h}) \mid X_n, \ldots, X_1\} \\ &= \sigma\overline{\mathbb{P}}\{\widehat{Z}_{n+h} \mid X_n, \ldots, X_1\} \\ &= \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}). \end{aligned}$$

When $h > m - n$

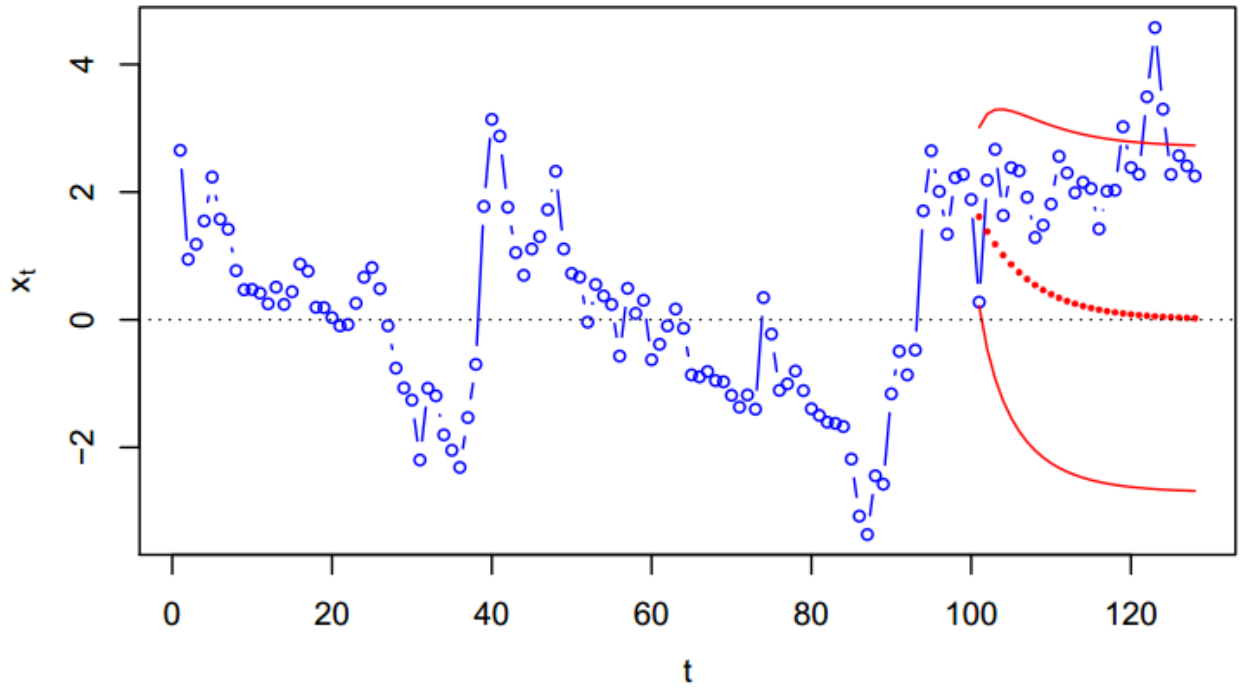
$$\begin{aligned}\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1) &= \overline{\mathbb{P}}\{X_{n+h} - \sigma(Z_{n+h} - \widehat{Z}_{n+h}) \mid X_n, \ldots, X_1\} \\ &= \overline{\mathbb{P}}\left(\sum_{i=1}^{p} \phi_i X_{n+h-i} \mid X_n, \ldots, X_1\right) + \sigma\overline{\mathbb{P}}\{\widehat{Z}_{n+h} \mid X_n, \ldots, X_1\} \\ &= \sum_{i=1}^{p} \phi_i \overline{\mathbb{P}}(X_{n+h-i} \mid X_n, \ldots, X_1) + \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}).\end{aligned}$$

Since $\theta_{nj} = 0$ when both $n \geq m$ and $j > q$, we have obtained the h-step predictor

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1) = \begin{cases} \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}), & 1 \leq h \leq m - n, \\ \sum_{i=1}^{p} \phi_i \overline{\mathbb{P}}(X_{n+h-i} \mid X_n, \ldots, X_1) & \\ + \sum_{h \leq j \leq q} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}), & h > m - n. \end{cases}$$

One last thing I would to say about prediction is that, under a Gassian assumption, we can use the prediction and the prediction mean squared error to construct confidence interval. For an unkonwn X_{n+h}, an 95% CI is

$$\overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1) \pm 1.96[E\{X_{n+h} - \overline{\mathbb{P}}(X_{n+h} \mid X_n, \ldots, X_1)\}^2]^{1/2}.$$



5.7 Miscellanea

Example 5.7. (Prediction of an AR(p) Process). For an AR(p) process; i.e., $\{X_t\}$ satisfies that

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t$$

and $\text{Cov}(W_t, X_s) = 0$ for $s < t$. With no difficulty, we can see that, when $n \geq p$,

$$\begin{aligned}\widehat{X}_{n+1} =& \overline{\mathbb{P}}(\widehat{X}_{n+1} \mid X_n, \ldots, X_1) \\ =& \overline{\mathbb{P}}(\phi_1 X_n + \cdots + \phi_p X_{n+1-p} + W_t \mid X_n, \ldots, X_1) \\ =& \phi_1 X_n + \cdots + \phi_p X_{n+1-p}.\end{aligned}$$

The one-step prediction if fully determined by its previous p observations. Further

$$E\{X_{n+1} - \widehat{X}_{n+1}\}^2 = EW_t^2 = \sigma^2.$$

Noting that, we can view that, conditional on $\{X_n, \ldots, X_1\}$, $X_{n+1} = \phi_1 X_n + c \cdots + \phi_p X_{n+1-p} + W_t$ follows a distribution with mean $\phi_1 X_n + \cdots + \phi_p X_{n+1-p}$ and variance σ^2. Recall that the optimal predictor of X given Y is $E(X \mid Y)$. Thus, in our AR(p) case, we have the optimal predictor being the same as the optimal linear predictor.

The other interesting thing is that, in this example, neither the Durbin-Levinson algorithm nor the Innovations algorithm is need for compute $\widehat{X}_{n+1}$. Notice that, in both algorithm, we need the autocovariance or more generally the covariance function. For example, in the Durbin-Levison algorithm, we have

$$\widehat{X}_{n+1} = \phi_{n1} X_n + \cdots + \phi_{nn} X_1, \quad n \geq 1.$$

and its mean squared error of prediction will be denoted by ν_n as

$$\nu_n = E(X_{n+1} - \widehat{X}_{n+1})^2, \quad n \geq 1.$$

where $\phi_{11} = \gamma_X(1)/\gamma_X(0)$, $\nu_0 = \gamma_X(0)$,

$$\phi_{nn} = \left\{ \gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j} \gamma_X(n-j) \right\} \nu_{n-1}^{-1},$$

$$\begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{nn} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}$$

and

$$\nu_n = \nu_{n-1}(1 - \phi_{nn}^2).$$

Thus, with the acknowledgement of $\gamma_X(h)$, the coefficients $\phi_{n,j}$s are fully determined. Now we know $\phi_{n,j}$s from the model definition directly. It does not make use of $\gamma_X(h)$ of $\{X_t\}$. Naturally, a question rises as: can we use the ϕs to find all the $\gamma_X(h)$s?

To do so, we look as the Step-Down Durbin Levinson algorithm: given $\phi_{n1}, \ldots, \phi_{nn}$ and ν_n,

$$\begin{aligned}\phi_{n-1,j} =& \frac{\phi_{nj} + \phi_{nn}\phi_{n,n-j}}{1-\phi_{nn}^2} \\ \nu_{n-1} =& \nu_n/(1-\phi_{nn}^2)\end{aligned}$$

Thus, start with $\phi_{p,1} = \phi_1, \ldots, \phi_{pp} = \phi_p$ and $\nu_p = \sigma^2$, we can step-down recursively to get

$$\phi_{p-1,j}s \ \& \ \nu_{p-1}, \phi_{p-2,j}s \ \& \ \nu_{p-2}, \ldots, \phi_{11} \ \& \ \nu_1.$$

Then we can find all the $\gamma_X(h)$ via

$$\begin{aligned}\gamma_X(0) =& \nu_0 = \nu_1/(1-\phi_{11}^2) \\ \gamma_X(1) =& \gamma_X(0)\phi_{11}\end{aligned}$$

and

$$\gamma_X(n) = \nu_{n-1}\phi_{nn} + \sum_{j=1}^{n-1}\phi_{n-1,j}\gamma_X(n-j)$$

Thus

$$\begin{aligned}\gamma_X(2) =& \phi_{22}\nu_1 + \phi_{11}\gamma_X(1) \\ \gamma_X(3) =& \phi_{33}\nu_2 + \phi_{32}\gamma_X(2) + \phi_{2,2}\gamma_X(1) \\ &\vdots \\ \gamma_X(p) =& \phi_{pp}\nu_{p-1} + \phi_{p-1,1}\gamma_X(p-1) + \cdots + \phi_{p-1,p-1}\gamma_X(1)\end{aligned}$$

Noting that, Once we have $\gamma_X(0), \ldots, \gamma_X(p)$, we have for $k \geq p+1$,

$$\gamma_X(k) = \phi_1\gamma_X(k+1) + \cdots + \phi_p\gamma_X(k-p).$$

Example 5.8. (Generating a realization of an ARMR(p,q) process.) How to generate exact realizations of ARMA processes? Let us consider generating stationary and causal Gassian AR(p) processes:

$$Y_t - \phi_1 Y_{t-1} - \cdots \phi_p Y_{t-p} = W_t, W_t \sim N(0, \sigma^2).$$

Recall that, for any $t \geq p+1$, we have

$$\widehat{Y}_t = \overline{\mathbb{P}}(Y_t \mid Y_{t+1}, \ldots, Y_1) = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}.$$

The innovations are

$$U_t = Y_t - \widehat{Y}_t = Z_t$$

with MSE being

$$\nu_{t-1} = \mathrm{Var} U_t = \sigma^2.$$

Now, we can use step-down L-D recursions to get coefficients for

$$\widehat{Y}_t = \phi_{t-1,1} Y_{t-1} + \cdots + \phi_{t-1,t-1} Y_1, t = 2, 3, \ldots, p$$

also the associated MSEs ν_{t-1}.

Now we have all the innovations $U_t = Y_t - \widehat{Y}_t$, for $t = 1, \ldots, p$, where

1. $\mathrm{E} U_t = 0$ and $\mathrm{Var} U_t = \nu_{t-1}$.
2. $U_1, U_2, \ldots, U_p$ are uncorrelated random variables, (by the Gaussian assumption, it means independent normal)

Thus, we can easily simulate U_ts, $t = 1, \ldots, p$. Then we unroll U_t to get simulations fof Y_ts for $t = 1, \ldots, p$:

$$\begin{aligned} U_1 &= Y_1 \\ U_2 &= Y_2 - \phi_{11} Y_1 \\ U_3 &= Y_3 - \phi_{21} Y 2 - \phi_{22} Y_1 \\ &\vdots \\ U_p &= Y_p - \phi_{p-1,1} Y_{p-1} - \cdots - \phi_{p-1,p-1} Y_1. \end{aligned}$$

Now, we have $\{Y_t, t = 1, \ldots, p\}$, we can start generating Y_ts for $t > p$ based on the definition of the AR(p) model.

Once we know how to simulate AR process $\phi(B) Y_t = Z_t$, we can easily simulate an ARMA process $\phi(B) X_t = \theta(B) Z_t$ based on

$$X_t = \theta(B) Y_t.$$

6 Estimation for ARMA Models

The goal of this chapter is to estimate the ARMA(p, q) model for observed time series $x_1, \ldots, x_n$. More specifically, we need determine p and q; i.e., order selection; need estimate all the unknown parameters, including the process mean, coefficients ϕ_j and/or θ_j, and white noise variance σ^2.

6.1 The Yule-Walker Equations and Parameter Estimation for Autoregressive Processes

Let $\{X_t\}$ be the zero-mean causal autoregressive process,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t, \quad \{W_t\} \sim \text{WN}(0, \sigma^2). \tag{6.1}$$

Our aim is to find estimators of the coefficient vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\text{T}}$ and the white noise variance σ^2 based on the observations $X_1, \ldots, X_n$, where of course, $n > p$.

The causality assumption allows us to write X_t in the form

$$X_t = \sum_{h=0}^{\infty} \psi_h W_{t-h},$$

where $\psi(z) = \sum_{h=0}^{\infty} \psi_h z^h = 1/\phi(z)$, $|z| \le 1$. Now we multiply each side of (6.1) by X_{t-j}, $j = 0, \ldots, p$ and then take exception. It leads to

$$\begin{aligned}
& E(X_{t-j}X_t - \phi_1 X_{t-j}X_{t-1} - \cdots - \phi_p X_{t-j}X_{t-p}) \\
&= E(X_{t-j}W_t) \\
&= E\left\{\sum_{h=0}^{\infty} \psi_h W_{t-j-h} W_t, \right\} \\
&= \sum_{h=0}^{\infty} \psi_h E\{W_{t-j-h}W_t\} \\
&= \sigma^2 I(j = 0), \text{ since } \psi_0 = 1
\end{aligned}$$

Thus

$$\begin{aligned}
\gamma_X(0) - \phi_1\gamma_X(1) - \cdots - \phi_p\gamma_X(p) &= \sigma^2, \\
\gamma_X(1) - \phi_1\gamma_X(0) - \cdots - \phi_p\gamma_X(p-1) &= 0, \\
&\vdots \\
\gamma_X(p) - \phi_1\gamma_X(p-1) - \cdots - \phi_p\gamma_X(0) &= 0.
\end{aligned}$$

Now, we obtain the Yule-Walker equations,

$$\boldsymbol{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p, \tag{6.2}$$

and

$$\sigma^2 = \gamma_X(0) - \boldsymbol{\phi}^{\mathrm{T}} \boldsymbol{\gamma}_p. \tag{6.3}$$

where $\boldsymbol{\Gamma}_p = [\gamma_X(i-j)]_{i,j=1}^p$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_p = (\gamma_X(1), \ldots, \gamma_X(p))^{\mathrm{T}}$. Recall that, in Section 2.6.2, we have proposed the estimator of γ_X and proved that the estimator $\widehat{\gamma}_X$ is nonnegative definite. Using that, we have our so-called Yule-Walker estimator $\widehat{\phi}$ and $\widehat{\sigma}^2$:

$$\widehat{\boldsymbol{\Gamma}}_p \widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\gamma}}_p, \tag{6.4}$$

$$\widehat{\sigma}^2 = \widehat{\gamma}_X(0) - \widehat{\boldsymbol{\phi}}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_p, \tag{6.5}$$

Noting that we have

$$\widehat{\gamma}_X(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \overline{X}_n)(X_t - \overline{X}_n) & \text{if } |h| < n \\ 0 & \text{if } |h| \geq n. \end{cases}$$

Based on Lemma 2.3, we know that $\widehat{\gamma}_X$ is nonnegative definite. Thus, $\widehat{\gamma}_X(\cdot)$ is the auto covariance function of some stationary process base on Theorem 2.3; i.e.,

"A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and non-negative definite."

According to Proposition 2.1; i.e.,

"If $\{X_t\}$ is a stationary $q-$correlated time series (i.e., $\mathrm{Cov}(X_s, X_t) = 0$ whenever $|s-t| > q$) with mean 0, then it can be represented as an MA(q) process."

this stationary process must be an MA($n-1$) process. Finally, based on Proposition 5.1; i.e.,

"For a stationary process, if $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$, then the covariance matrix $\boldsymbol{\Gamma}_n = [\gamma_X(i-j)]_{i,j=1}^n$ is positive definite for every n. "

if $\widehat{\gamma}_X(0) > 0$, then it can be shown that $\widehat{\boldsymbol{\Gamma}}_p$ is non-singular. Diving by $\widehat{\gamma}_X(0)$, we therefore obtain

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p \tag{6.6}$$

$$\widehat{\sigma}^2 = \widehat{\gamma}_X(0)(1 - \widehat{\boldsymbol{\rho}}_p^{\mathrm{T}} \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p), \tag{6.7}$$

where $\widehat{\boldsymbol{\phi}}_p = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p))^{\mathrm{T}} = \widehat{\boldsymbol{\gamma}}_p / \widehat{\gamma}_X(0)$.

Remark 6.1. One feature of the Yule-Walker estimator is that, the fitted model

$$X_t - \widehat{\phi}_1 X_{t-1} - \cdots - \widehat{\phi}_p X_{t-p} = W_t, \quad \{W_t\} \sim \mathrm{WN}(0, \widehat{\sigma}^2),$$

is also causal. And the fitted model's ACVF is $\widehat{\gamma}_X(h)$ for $h = 0, 1, \ldots, p$ (but in general different for higher lags).

Theorem 6.1. If $\{X_t\}$ is the causal AR(p) process with $\{W_t\} \sim \text{IID}(0, \sigma^2)$, then the Yule-Walker estimator $\widehat{\boldsymbol{\phi}}$ enjoys that

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{\Gamma}_p^{-1}),$$

and

$$\widehat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Remark 6.2. Noting that, the Yule-Walker estimator is based on moment matching method. Generally speaking, moment based estimator can be far less efficient than the MLE. However, one good thing of the Yule-Walker estimator is that, it is asymptotically the same as the MLE estimator.

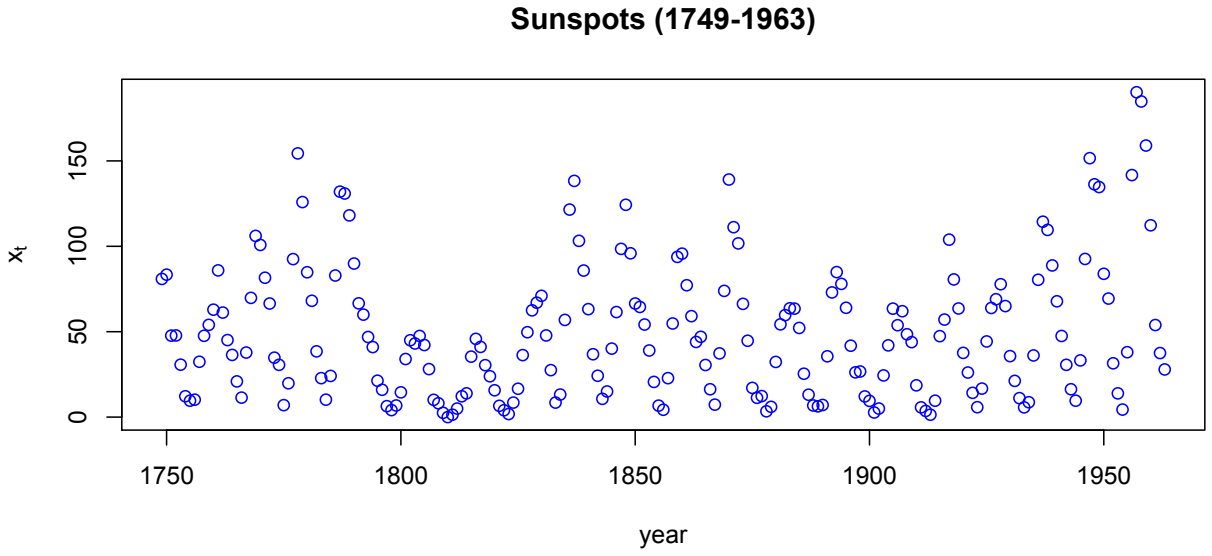
Remark 6.3. Approximated 95% confidence interval for ϕ_j is given by

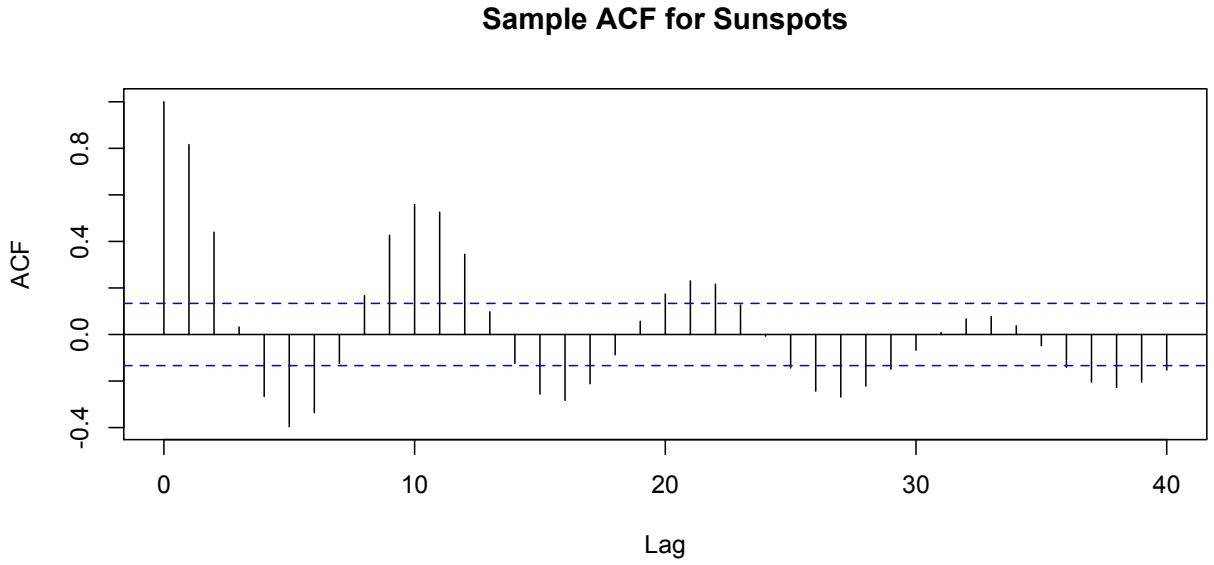
$$\left[\hat{\phi}_j \pm 1.96\sqrt{\widehat{v}_{jj}}/\sqrt{n}\right]$$

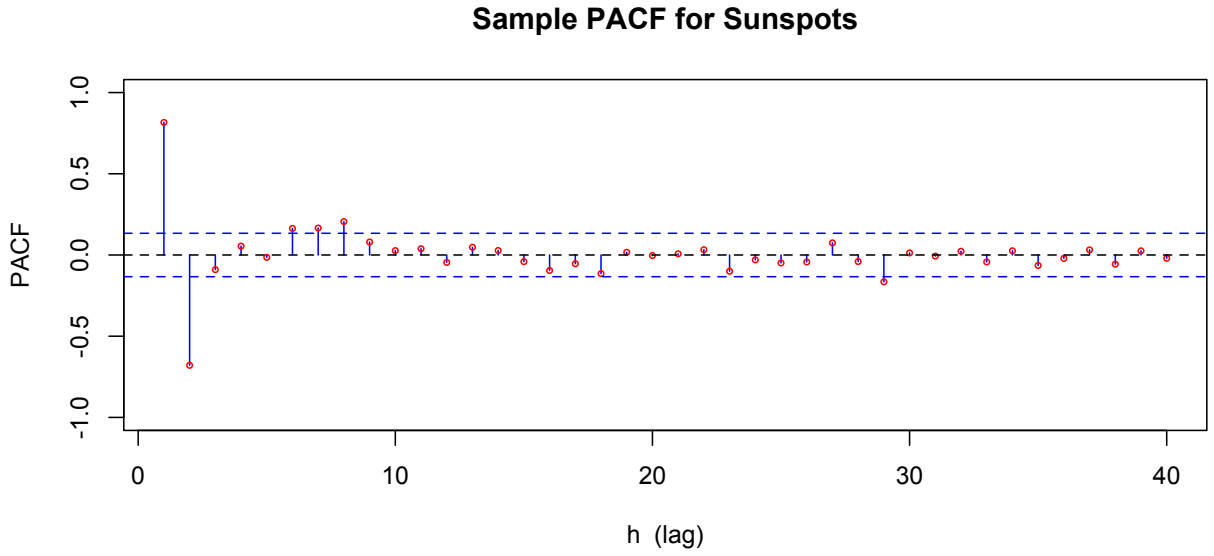
where $\widehat{v}_{jj}$ is the jth diagonal element of $\widehat{\sigma}^2 \widehat{\boldsymbol{\Gamma}}_p^{-1}$. And a 95% confidence region for $\boldsymbol{\phi}$ is then

$$(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^{\mathrm{T}} \widehat{\boldsymbol{\Gamma}} (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \leq \chi^2_{0.95}(p) \frac{\widehat{\sigma}^2}{n}.$$

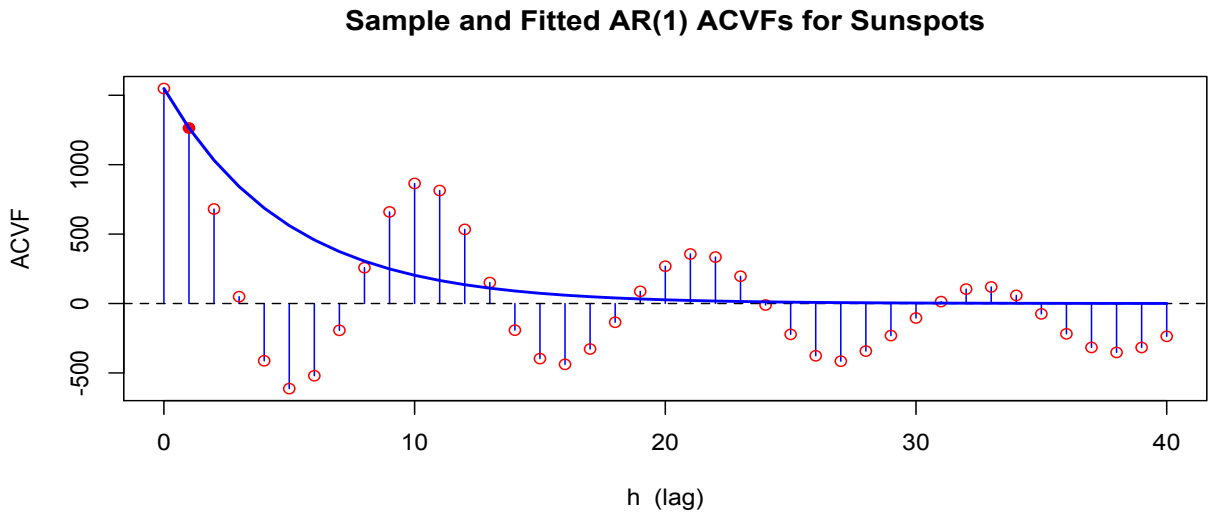
Example 6.1. Here is a dataset; number of sunspots observed each year between 1749 and 1963. The plot of the data is the following

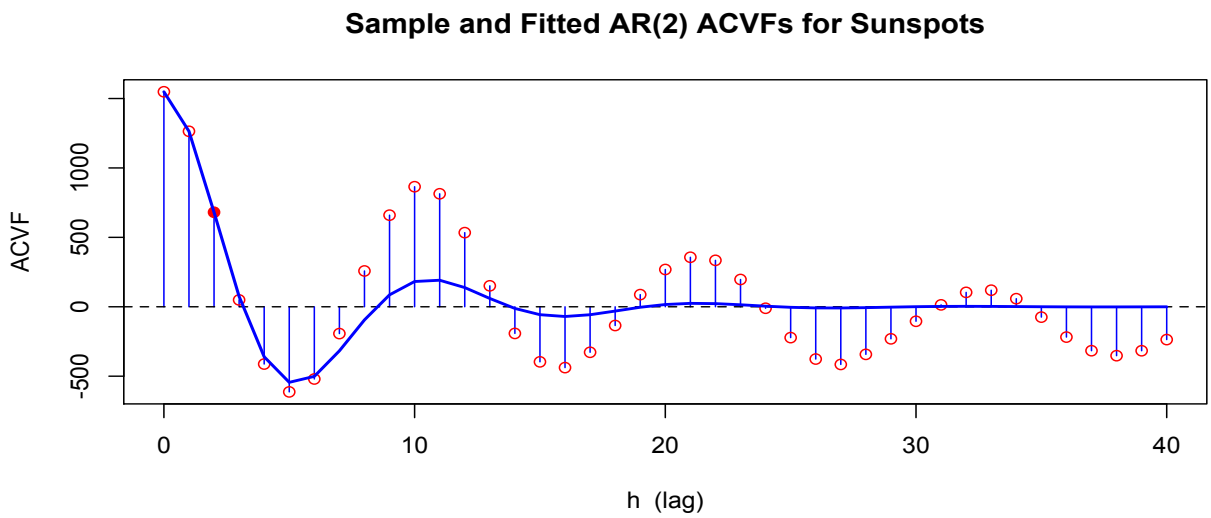


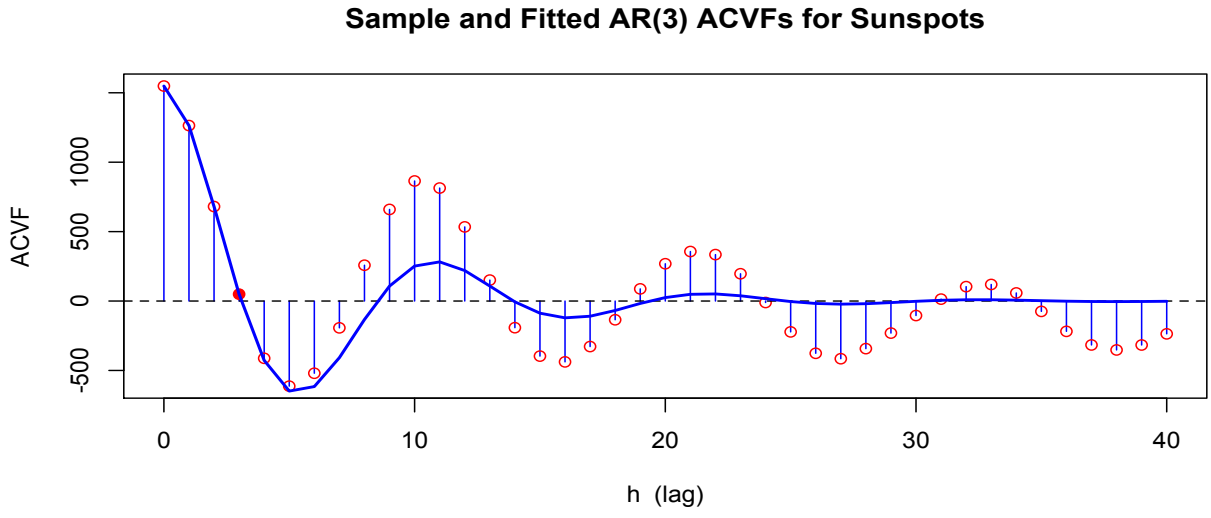


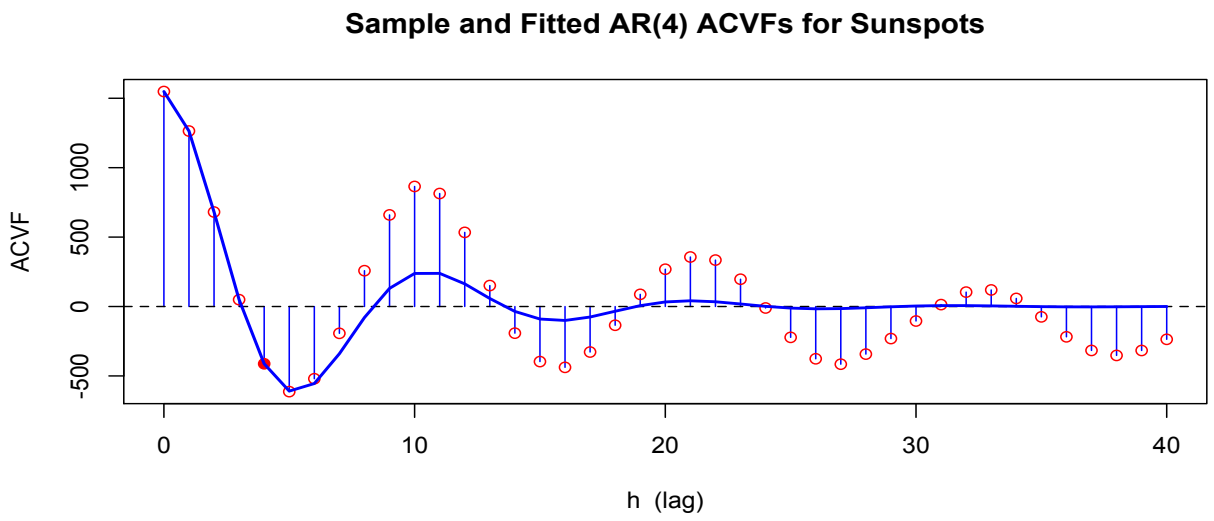


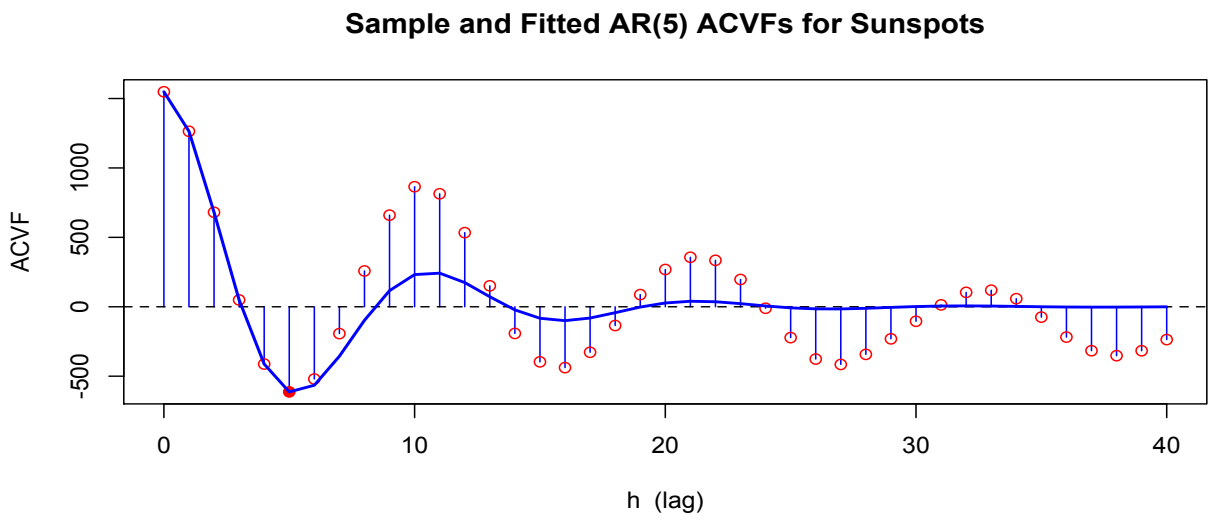
Fitting the series for AR(p), where $p = 1, 2, 3, 4, 5, 6, 7, 8, 29$, we have

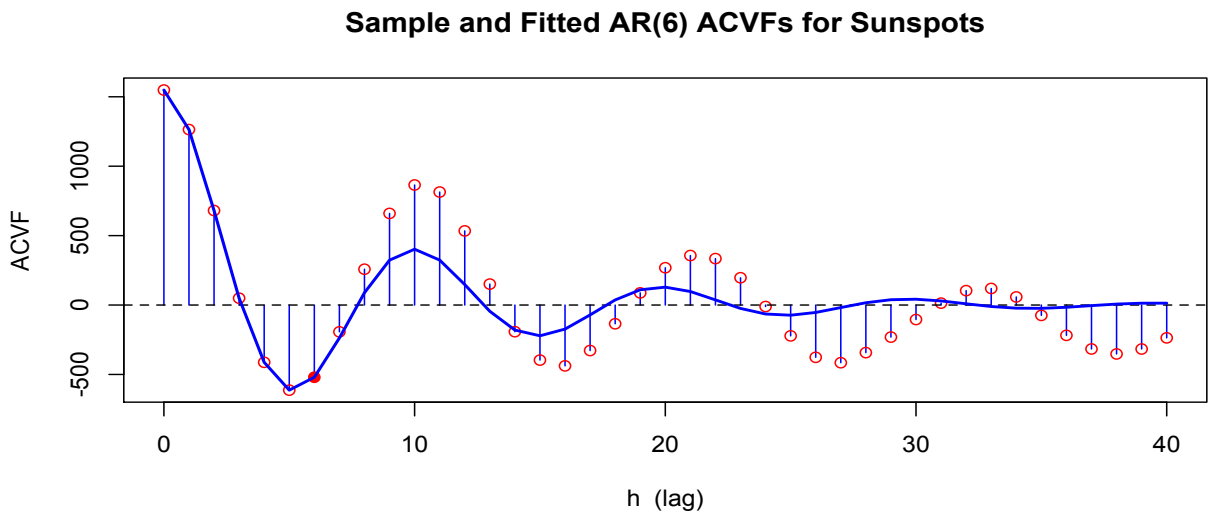

Sample and Fitted AR(1) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)

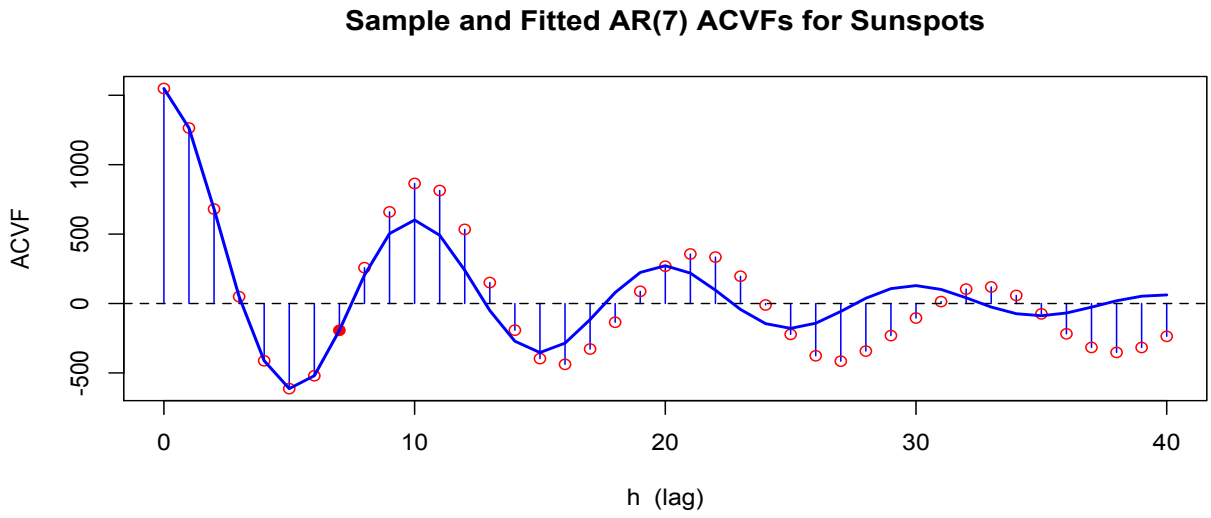

Sample and Fitted AR(2) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)

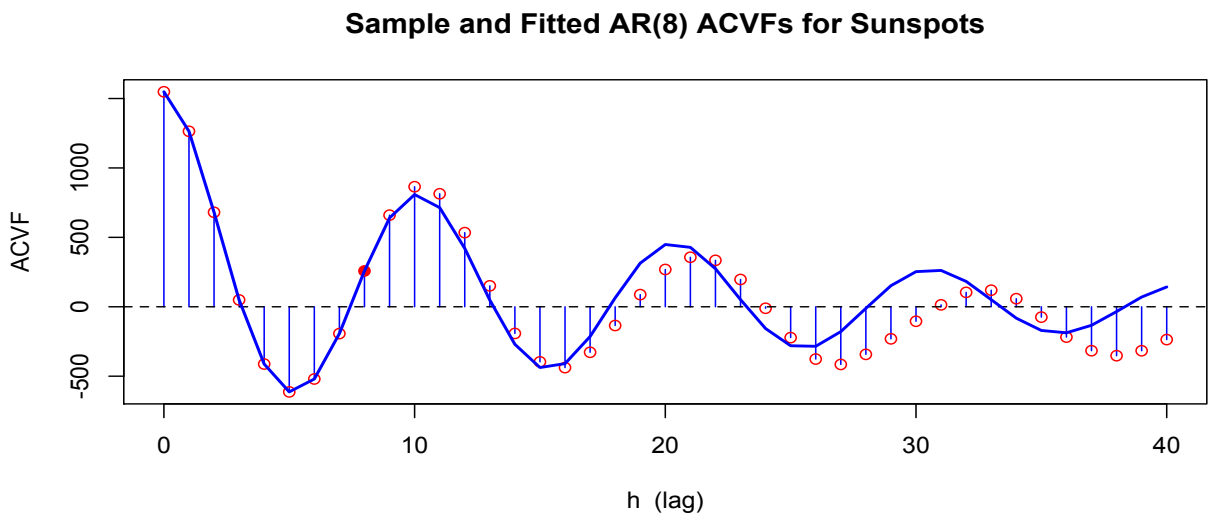

Sample and Fitted AR(3) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)

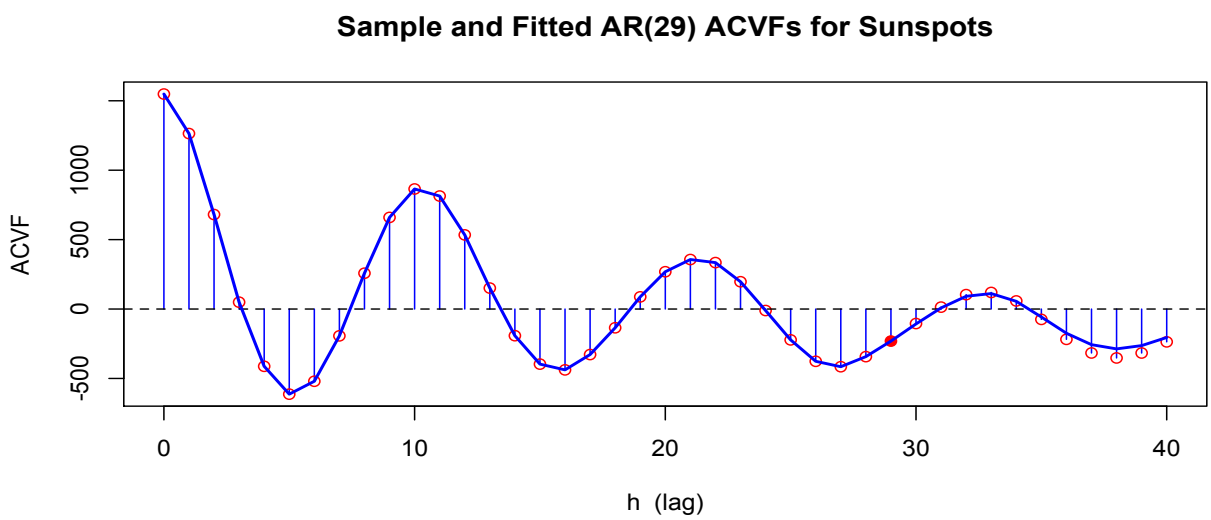

Sample and Fitted AR(4) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)


Sample and Fitted AR(5) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)


Sample and Fitted AR(6) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)


Sample and Fitted AR(7) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)


Sample and Fitted AR(8) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)


Sample and Fitted AR(29) ACVFs for Sunspots
ACVF
-500
0
500
1000
0
10
20
30
40
h (lag)

Theorem 6.2. If $\{X_t\}$ is the causal AR(p) process with $\{W_t\} \sim \text{IID}(0, \sigma^2)$ and if

$$\widehat{\boldsymbol{\phi}}_m = (\phi_{m1}, \ldots, \phi_{mm})^{\mathrm{T}} = \widehat{\boldsymbol{R}}_m^{-1} \widehat{\boldsymbol{\rho}}_m, m > p,$$

then

$$\sqrt{n}(\widehat{\boldsymbol{\phi}}_m - \boldsymbol{\phi}_m) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{\Gamma}_m^{-1})$$

where $\boldsymbol{\phi}_m$ is the coefficient vector of the best linear predictor $\boldsymbol{\phi}_m^{\mathrm{T}} \boldsymbol{X}_m$ of X_{m+1} based on $\boldsymbol{X}_m = (X_m, \ldots, X_1)^{\mathrm{T}}$; i.e., $\boldsymbol{\phi}_m = \boldsymbol{R}_m^{-1} \boldsymbol{\rho}_m$. In particular for $m > p$,

$$\sqrt{n} \widehat{\phi}_{mm} \xrightarrow{d} N(0, 1).$$

Remark 6.4. When fitting AR models, the order p will of course be unknown. If the true one is p, but we want to fit it by order m, then we should expect the estimated coefficient vector $\widehat{\boldsymbol{\phi}}_m = (\phi_{m1}, \ldots, \phi_{mm})^{\mathrm{T}}$ to have a small value of $\hat{\phi}_{mm}$ for each $m > p$.

Remark 6.5. Noting that ϕ_{mm} is the PACF of $\{X_t\}$ at lag m, recall, PACF is a good tool to identify AR series (while ACF is for MA processes). If $m > p$, we have $\phi_{mm} = 0$.

Remark 6.6. For order selection of p, our textbook suggest setting p to be the smallest p_0, such that

$$|\hat{\phi}_{mm}| < 1.96/\sqrt{n}, \quad \text{for } p_0 < m \leq H,$$

where H is the maximum lag for a reasonable estimator of γ_X; i.e., $H \leq n/4$ and $n \geq 50$.

6.2 Preliminary Estimation for Autoregressive Processes Using the Durbin-Levinson Algorithm

Suppose we have observations $x_1, \ldots, x_n$ of a zero-mean stationary time series. Provided $\widehat{\gamma}_x(0) > 0$ we can fit an autoregressive process of order $m < n$ to the data by means of the Yule-Walker equations. The fitted AR(m) process is

$$X_t - \widehat{\phi}_{m1} X_{t-1} - \cdots \widehat{\phi}_{mm} X_{t-m} = W_t, \quad W_t \sim \text{WN}(0, \widehat{\nu}_m) \tag{6.8}$$

where

$$\widehat{\boldsymbol{\phi}}_m = (\phi_{m1}, \ldots, \phi_{mm})^{\mathrm{T}} = \widehat{\boldsymbol{R}}_m^{-1} \widehat{\boldsymbol{\rho}}_m, \tag{6.9}$$

and

$$\widehat{\nu}_m = \widehat{\gamma}_X(0)\{1 - \widehat{\boldsymbol{\phi}}_m^{\mathrm{T}} \widehat{\boldsymbol{R}}^{-1} \widehat{\boldsymbol{\phi}}_m\}. \tag{6.10}$$

Based on Theorem 6.2, we know that $\widehat{\boldsymbol{\phi}}_m$ is the estimator of $\boldsymbol{\phi}_m$ which is the coefficient vector of the best linear predictor $\boldsymbol{\phi}_m^{\mathrm{T}} \boldsymbol{X}_m$ of X_{m+1} based on $\boldsymbol{X}_m = (X_m, \ldots, X_1)^{\mathrm{T}}$, further, we can see that $\widehat{\nu}_m$ is more like an estimator of the corresponding mean squared prediction error.

This leads us to the following proposition.

Proposition 6.1. (The Durbin-Levison Algorithm for Fitting Autogressive Models). If $\widehat{\gamma}_X(0) > 0$ then the fitted autoregressive model (6.8) for $m = 1, 2, \ldots, n-1$, can be determined recursively from the relations, $\widehat{\phi}_{11} = \widehat{\gamma}_X(1)/\widehat{\gamma}_X(0)$, $\widehat{\nu}_0 = \widehat{\gamma}_X(0)$,

$$\widehat{\phi}_{mm} = \left\{ \widehat{\gamma}_X(m) - \sum_{j=1}^{m-1} \widehat{\phi}_{m-1,j} \widehat{\gamma}_X(m-j) \right\} \widehat{\nu}_{m-1}^{-1},$$

$$\begin{pmatrix} \widehat{\phi}_{m1} \\ \vdots \\ \widehat{\phi}_{m,m-1} \end{pmatrix} = \begin{pmatrix} \widehat{\phi}_{m-1,1} \\ \vdots \\ \widehat{\phi}_{m-1,m-1} \end{pmatrix} - \widehat{\phi}_{mm} \begin{pmatrix} \widehat{\phi}_{m-1,m-1} \\ \vdots \\ \widehat{\phi}_{m-1,1} \end{pmatrix}$$

and

$$\widehat{\nu}_m = \widehat{\nu}_{m-1}(1 - \widehat{\phi}_{mm}^2).$$

Again, noting that, $\widehat{\phi}_{11}, \widehat{\phi}_{22}, \ldots, \widehat{\phi}_{mm}$ are the sample partial autocorrelation function at lags $1, 2, \ldots, m$. Using R function *acf* and selecting option *type*="partial" directly calculates the sample PACFs.

Example 6.2. We generated a sequence of AR(2) series:

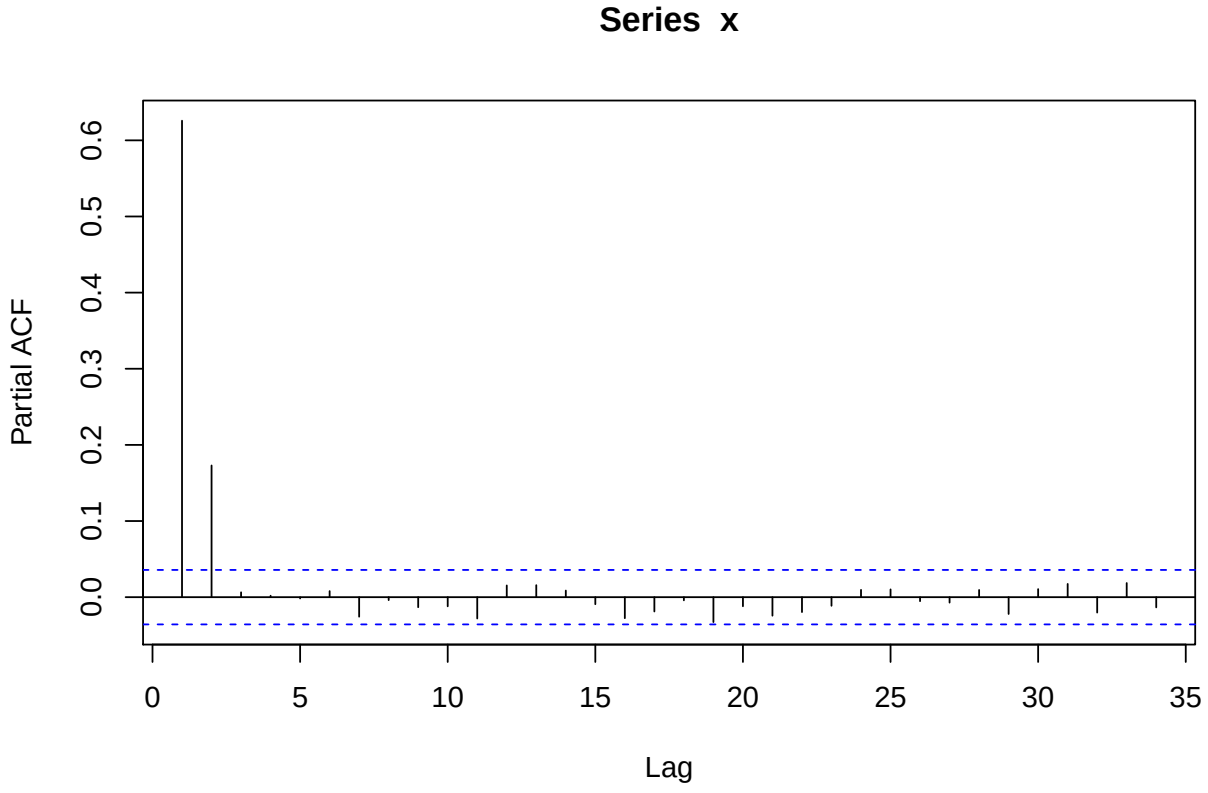
```
set.seed(720)
w=rnorm(5000,0,1.5)
x=filter(w,filter=c(0.5,0.2),method="recursive")
x=x[-(1:2000)]
pacf=acf(x,type="partial")
YW=ar(x,method="yule-walker")
YW

Call:
ar(x = x, method = "yule-walker")
Coefficients:
     1       2
0.5176  0.1730
Order selected 2  sigma^2 estimated as  2.327

YW[c(2,3,14)]
$ar
[1] 0.5175761 0.1729591
$var.pred
```

```
[1] 2.327124
$asy.var.coef
              [,1]          [,2]
[1,]  0.0003236854 -0.0002025678
[2,] -0.0002025678  0.0003236854

ar(x,method="yule-walker",aic=FALSE,3)[c(2,3,14)]
$ar
[1] 0.516461352 0.169623096 0.006445428
$var.pred
[1] 2.327804
$asy.var.coef
              [,1]          [,2]          [,3]
[1,]  3.337645e-04 -0.0001727485 -5.772761e-05
[2,] -1.727485e-04  0.0004131905 -1.727485e-04
[3,] -5.772761e-05 -0.0001727485  3.337645e-04
```



6.3 Preliminary Estimation for Moving Average Processes Using the Innovations Algorithm

We know that we can fit autoregressive models of orders $1, 2, \ldots,$ to the data $x_1, \ldots, x_n$ by applying the Durbin-Lesion algorithm based on the sample autocovariances. Just like this, we can also fit moving average models,

$$X_t = W_t + \widehat{\theta}_{m1} W_{t-1} + \cdots + \widehat{\theta}_{mm} W_{t-m}, \quad W_t \sim \mathrm{WN}(0, \widehat{\nu}_m),$$

of orders $m = 1, 2, \ldots,$ by means of the innovations algorithm, where $\widehat{\boldsymbol{\theta}_m} = (\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mm})^{\mathrm{T}}$ and white noise variance $\widehat{\nu}_m$, $m = 1, 2, \ldots,$ are specified as follows.

(Innovation Estimates of Moving Average Parameters). If $\widehat{\gamma}_X(0) > 0$, we define the innovation estimates $\widehat{\boldsymbol{\theta}}_m$, $\widehat{\nu}_m$ appearing above for $m = 1, \ldots, n-1$, by the recursion relations $\widehat{\nu}_0 = \widehat{\gamma}_X(0)$,

$$\widehat{\theta}_{m,m-k} = \widehat{\nu}_k^{-1} \left[\widehat{\gamma}_X(m-k) - \sum_{j=0}^{k-1} \widehat{\theta}_{m,m-j} \widehat{\theta}_{k,k-j} \widehat{\nu}_j \right], \quad k = 0, \ldots, m-1,$$

and

$$\widehat{\nu}_m = \widehat{\gamma}_X(0) - \sum_{j=0}^{m-1} \widehat{\theta}_{m,m-j}^2 \widehat{\nu}_j.$$

The asymptotic behavior can be proved more generally, not only for MA models (since for each observed sequence, we can calculate $\widehat{\gamma}_X(h)$, and then use the above recursive means to find $\widehat{\theta}$s).

Theorem 6.3. (The Asymptotic Behavior of $\widehat{\boldsymbol{\theta}}_m$). Let $\{X_t\}$ be the causal invertible ARMA process $\phi(B)X_t = \theta(B)W_t$, $\{W_t\} \sim \mathrm{IID}(0, \sigma^2)$, $\mathrm{E}W_t^4 < \infty$, and let $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, $|z| \le 1$, (withe $\psi_0 = 1$, and $\psi_j = 0$ for $j < 0$). Then for any sequence of positive integers $\{m(n) : n = 1, 2, \ldots\}$ such that $m < n$, $m \to \infty$ and $m = o(n^{1/3})$ as $n \to \infty$, we have for each k,

$$\sqrt{n}\big(\widehat{\theta}_{m1} - \psi_1, \widehat{\theta}_{m2} - \psi_2, \ldots, \widehat{\theta}_{mk} - \psi_k\big)^{\mathrm{T}} \xrightarrow{d} N(0, A),$$

where $A = [a_{ij}]_{i,j=1}^{k}$ and

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \psi_{i-r} \psi_{j-r}.$$

Moreover,

$$\widehat{\nu}_m \xrightarrow{p} \sigma^2.$$

Remark 6.7. Difference between this approach and the one based on the Durbin-Levsion algorithm. For an AR(p) process, the Yule-Walker estimator $\widehat{\boldsymbol{\phi}}_p = (\widehat{\phi}_{p1}, \ldots, \widehat{\phi}_{pp})^{\mathrm{T}}$ is consistent for $\boldsymbol{\phi}_p$ as the sample size $n \to \infty$. However for an MA(q) process the estimator $\widehat{\boldsymbol{\theta}}_q = (\widehat{\theta}_{q1}, \ldots, \widehat{\theta}_{qq})^{\mathrm{T}}$ is not consistent for the true parameter vector $\boldsymbol{\theta}_q$ as $n \to \infty$. For consistency it is necessary to use the estimator $(\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mq})^{\mathrm{T}}$ with $\{m(n)\}$ satisfying the conditions of the above theorem.

Example 6.3. Consider MA process $X_t = W_t - 1.5W_{t-1} + .5W_{t-2}$, $W_t \sim N(0, 2)$. Run the following codes, we can see that

```
set.seed(720)
w = rnorm(2000,0,sqrt(2)) # 500 N(0,1) variates
v = filter(w, sides=1, c(1,-1.5,.5)) # moving average
v=tail(v,500)
require(itsmr)
jj=c(1:10,10,10,10)
mm=c(1:10,20,50,100)
for(i in 1:13)
{
print(ia(v,jj[i],m=mm[i]))
}
```

m \ j	1	2	3	4	5	6	7	8	9	10	$\widehat{\nu}_m$
1	-0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.23
2	-0.88	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.75
3	-1.02	0.20	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.39
4	-1.10	0.27	-0.05	0.12	0.00	0.00	0.00	0.00	0.00	0.00	2.43
5	-1.14	0.28	-0.08	0.09	-0.10	0.00	0.00	0.00	0.00	0.00	2.26
6	-1.19	0.29	-0.09	0.09	-0.12	0.04	0.00	0.00	0.00	0.00	2.47
7	-1.22	0.32	-0.07	0.13	-0.10	0.10	0.03	0.00	0.00	0.00	2.22
8	-1.24	0.31	-0.10	0.11	-0.14	0.08	-0.02	-0.07	0.00	0.00	2.04
9	-1.28	0.32	-0.09	0.12	-0.14	0.10	-0.01	-0.05	0.07	0.00	2.06
10	-1.29	0.34	-0.08	0.14	-0.13	0.11	0.00	-0.04	0.10	-0.01	2.20
20	-1.32	0.36	-0.09	0.16	-0.14	0.11	0.01	-0.05	0.09	-0.06	2.05
50	-1.32	0.38	-0.10	0.16	-0.14	0.10	0.01	-0.04	0.08	-0.06	2.04
100	-1.34	0.41	-0.12	0.17	-0.15	0.11	0.02	-0.06	0.08	-0.04	2.05

6.4 Preliminary Estimation for ARMA(p, q) Processes

Theorem 6.3 basically says that, for a causal invertible ARMA process $\phi(B)X_t = \theta(B)W_t$, we can use the innovations algorithm to obtain $\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mk}$, for each k, which are consistent estimator of $\psi_1, \ldots, \psi_k$, where $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$. What more can we do using this results? More specifically? $\psi(z)$ comes form $\theta(z)$ and $\phi(z)$, can we use the estimator of $\psi(z)$ and go back to estimate $\theta(z)$ and $\phi(z)$?

Let $\{X_t\}$ be the zero-mean causal ARMA(p, q) process,

$$X_t - \phi_1 X_{t-1} - \cdots \phi_p X_{t-p} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}, \quad \{W_t\} \sim \text{WN}(0, \sigma^2).$$

The causality assumption ensures that

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

Based on

$$\sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)},$$

we can match the coefficients as

$$\phi_0 = 1$$

and

$$\psi_j = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \psi_{j-i}, \quad j = 1, 2, \ldots.$$

and by convention, $\theta_j = 0$ for $j > q$ and $\phi_j = 0$ for $j > p$. So when $j > p$, we actually have

$$\psi_j = \sum_{i=1}^{p} \phi_i \psi_{j-i}, \quad j = q+1, q+2, \ldots, q+p$$

We know how to estimate ψ_js. Replacing them with their estimators, we have

$$\widehat{\theta}_{mj} = \sum_{i=1}^{p} \phi_i \widehat{\theta}_{m,j-i}, \quad j = q+1, q+2, \ldots, q+p$$

Then, solving for ϕ_i, we obtain the estimator $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ (natural question: does the solution exist? If so, unique? Further, is the fitted ARMA process causal?). Then, the estimate of $\theta_1, \ldots, \theta_q$ is found easily from

$$\widehat{\theta}_j = \widehat{\theta}_{mj} - \sum_{i=1}^{\min(j,p)} \widehat{\phi}_i \widehat{\theta}_{m,j-i}, \quad j = 1, 2, \ldots, q.$$

Finally, the white noise variance σ^2 is estimated by

$$\widehat{\sigma}^2 = \widehat{\nu}_m.$$

By the consistency of $\widehat{\theta}_{mj} \xrightarrow{p} \psi_j$, where $m = m(n)$ satisfying the condition in Theorem 6.3, we have

$$\widehat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}, \quad \widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}, \quad \text{and } \widehat{\sigma}^2 \xrightarrow{p} \sigma^2 \quad \text{as } n \to \infty.$$

However, the efficiency (asymptotic variance) of this moment-matching type estimator is somewhat poor. In the next section, we discuss a more efficient estimation procedure (strictly more efficient if $q \geq 1$) of $(\boldsymbol{\phi}, \boldsymbol{\theta})$ based on maximization of the Gaussian likelihood. Noting that, when $q = 0$, we have AR process. And in the first section of this Chapter, we discussed that the Yule-Walker (based on moment-matching) estimator is the same efficient as the MLE.

6.5 Recursive Calculation of the Likelihood of an Arbitrary Zero-Mean Gaussian Process

All the previous discussed method are based on matching moments. This is very natural thoughts, since for a general stationary time series, it is basically determined by its first moment (mean) and its second moments (ACVF). Can we gain more if we assume more assumptions about the sequence? In this section, we assume $\{X_t\}$ to be a Gaussian process, more specifically, an arbitrary zero-mean Gaussian Process. In the next section, we focus on Gaussian ARMA processes.

Let $\{X_t\}$ be a Gaussian process with mean zero and covariance function $\kappa(i, j) = \mathrm{E}(X_i X_j)$. Denoting $\boldsymbol{X}_n = (X_1, \ldots, X_n)^{\mathrm{T}}$ and Γ_n as the covariance matrix of $\boldsymbol{X}_n$; i.e., $\Gamma_n = E(\boldsymbol{X}_n, \boldsymbol{X}_n^{\mathrm{T}})$ which is assumed to be non-singular. Then, we have

$$\boldsymbol{X}_n \sim N(\boldsymbol{0}, \Gamma_n).$$

Then the likelihood of $\boldsymbol{X}_n$ is

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp(-\boldsymbol{X}_n^{\mathrm{T}} \Gamma_n^{-1} \boldsymbol{X}_n / 2)$$

Evaluating $\det \Gamma_n$ and Γ_n^{-1} can be avoided by using the one-step predictors and their mean squared errors.

Denoting the one-step predictors of $\boldsymbol{X}$ as $\widehat{\boldsymbol{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)^{\mathrm{T}}$ where $\widehat{X}_1 = 0$ and $\widehat{X}_j = E(X_j \mid X_1, \ldots, X_{j-1})$, $j \geq 2$, and $\nu_{j-1} = E(X_j - \widehat{X}_j)^2$ for $j = 1, \ldots, n$.

We know that, from the innovations algorithm,

$$\widehat{X}_{n+1} = \begin{cases} 0, & n = 0, \\ \sum_{j=1}^{n} \theta_{nj} (X_{n+1-j} - \widehat{X}_{n+1-j}), & n \geq 1, \end{cases}$$

Now define

$$C = \begin{pmatrix} 1 & & & & \\ \theta_{11} & 1 & & & \\ \theta_{22} & \theta_{21} & 1 & & \\ \vdots & \vdots & \cdots & \ddots & \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \cdots & \theta_{n-1,1} & 1 \end{pmatrix}$$

Then, we can write

$$\widehat{\boldsymbol{X}}_n = (C - I)(\boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n),$$

where I is the n dimensional identity matrix. Hence

$$\boldsymbol{X}_n = \boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n + \widehat{\boldsymbol{X}}_n = C(\boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n).$$

Noting that

$$\boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n \sim N(\boldsymbol{0}, D),$$

where $D = \mathrm{diag}\{\nu_0, \ldots, \nu_{n-1}\}$. Thus,

$$\Gamma_n = CDC^{\mathrm{T}}.$$

Further, we have

$$\boldsymbol{X}_n^{\mathrm{T}}\Gamma_n^{-1}\boldsymbol{X}_n = (\boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n)^{\mathrm{T}}D^{-1}(\boldsymbol{X}_n - \widehat{\boldsymbol{X}}_n) = \sum_{j=1}^{n}(X_j - \widehat{X}_j)^2/\nu_j.$$

and

$$\det\Gamma_n = \det C \times \det D \times \det C = \nu_0\nu_1\cdots\nu_{n-1}.$$

Finally, we can rewrite the likelihood as

$$L(\Gamma_n) = (2\pi)^{-n/2}(\nu_0\cdots\nu_{n-1})^{-1/2}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\frac{(X_j - \widehat{X}_j)^2}{\nu_{j-1}}\right\}.$$

6.6 Maximum Likelihood Estimation for ARMA Processes

Using the MLE developed from last section, suppose, now we have a causal Gaussian ARMA(p, q) process,

$$\phi(B)X_t = \theta(B)W_t$$

The one step predictor is obtained through the Innovations algorithm; i.e.,

$$\widehat{X}_{i+1} = \sum_{j=1}^{i} \theta_{ij}(X_{i+1-j} - \widehat{X}_{i+1-j}), \quad 1 \le i < m = \max(p, q),$$

and

$$\widehat{X}_{i+1} = \phi_1 X_1 + \cdots + \phi_p X_{i+1-p} + \sum_{j=1}^{q} \theta_{ij}(X_{i+1-j} - \widehat{X}_{i+1-j}), i \ge m,$$

and

$$E(X_{i+1} - \widehat{X}_{i+1})^2 = \sigma^2 r_i.$$

Thus, the MLE can be derived as

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-n/2}(r_0 \cdots r_{n-1})^{-1/2} \exp\left[-\frac{1}{2}\sigma^{-2}\sum_{j=1}^{n}\frac{(X_j - \widehat{X}_j)^2}{r_{i-1}}\right]$$

Setting first derivative of $\log L$ w.r.t. σ^2 to be zero, we have

$$\widehat{\sigma}^2 = n^{-1}S(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}})$$

where

$$S(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}) = \sum_{j=1}^{n}\frac{(X_j - \widehat{X}_j)^2}{r_{j-1}}$$

and $\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}$ are the values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ which minimizes

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}) = \log\{n^{-1}S(\boldsymbol{\phi}, \boldsymbol{\theta})\} + n^{-1}\sum_{j=1}^{n}\log r_{j-1}.$$

This l function is referred to as the reduced likelihood, which is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ only. Note that, we start with the causal condition, so it is better to search for $\boldsymbol{\phi}$ that makes the sequence casual. However, invertible is not required, but you can also do that.

An intuitively appealing alternative estimation procedure is to minimize the weighted sum of squares

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{n}\frac{(X_j - \widehat{X}_j)^2}{r_{j-1}}$$

We refer the resulting estimator to as the least squares estimator $\widetilde{\boldsymbol{\phi}}$ and $\widetilde{\boldsymbol{\theta}}$. Further the least square estimate of σ^2 is

$$\widetilde{\sigma}^2 = \frac{S(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}})}{n-p-q}.$$

6.7 Asymptotic properties of the MLE

Denoting $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\phi}}^{\mathrm{T}}, \widehat{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}}$, if $\{W_t\} \sim \mathrm{IID}N(0, \sigma^2)$ and $\{X_t\}$ is causal and invertible, then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{V}(\boldsymbol{\beta}))$$

where for $p \geq 1$ and $q \geq 1$,

$$\boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \begin{bmatrix} \mathrm{E}\boldsymbol{U}_t\boldsymbol{U}_t^{\mathrm{T}} & \mathrm{E}\boldsymbol{U}_t\boldsymbol{V}_t^{\mathrm{T}} \\ \mathrm{E}\boldsymbol{V}_t\boldsymbol{U}_t^{\mathrm{T}} & \mathrm{E}\boldsymbol{V}_t\boldsymbol{V}_t^{\mathrm{T}} \end{bmatrix}^{-1}$$

where $\boldsymbol{U}_t = (U_t, \ldots, U_{t+1-p})^{\mathrm{T}}$ and $\boldsymbol{V}_t = (V_t, \ldots, V_{t+1-q})^{\mathrm{T}}$ and

$$\phi(B)U_t = W_t$$

and

$$\theta(B)V_t = W_t.$$

AICC Criterion: Choose p, q, $\boldsymbol{\phi}_p$ and $\boldsymbol{\theta}_q$ to minimize

$$\mathrm{AICC} = -2\log L\left\{\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \frac{S(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q)}{n}\right\} + 2\frac{(p+q+1)n}{(n-p-q-2)}.$$

6.8 Diagnostic checking

Based on data $\{X_1, \ldots, X_n\}$, we can fit an ARMA(p, q) model and obtain the MLE $\widehat{\boldsymbol{\phi}}$, $\widehat{\boldsymbol{\theta}}$ and σ^2. We denote the fitted process as

$$\widehat{\phi}(B)X_t = \widehat{\theta}(B)W_t.$$

Based on the fitted process, we can calculate the one-step predictor of X_t based on $X_1, \ldots, X_{t-1}$, given the values of $\widehat{\boldsymbol{\phi}}$, $\widehat{\boldsymbol{\theta}}$ and σ^2. We denote this predictor as

$$\widehat{X}_t(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}), t = 1, \ldots, n.$$

The residuals are defined by

$$\widehat{W}_t(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}) = \frac{X_t - \widehat{X}_t(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}})}{\sqrt{r_{t-1}(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}})}}, t = 1, \ldots, n.$$

If the fitted model is exactly true, then we could say that $\{\widehat{W}_t\} \sim \mathrm{WN}(0, \widehat{\sigma}^2)$. However, we are not that lucky, even we were, we do not believe we were. Nonetheless, $\widehat{W}_t$ should have properties that

are similar to those of the white noise sequence

$$W_t(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{X_t - \widehat{X}_t(\boldsymbol{\phi}, \boldsymbol{\theta})}{\sqrt{r_{t-1}(\boldsymbol{\phi}, \boldsymbol{\theta})}}, t = 1, \ldots, n.$$

which approximates the white noise term in the sense that

$$E\{W_t(\boldsymbol{\phi}, \boldsymbol{\theta}) - W_t\}^2 \to 0, \quad t \to \infty.$$

Consequently, the properties of the residuals $\{\widehat{W}_t\}$ should reflect those of the white noise sequence $\{W_t\}$ generating the underlying true process. In particular, the sequence $\{\widehat{W}_t\}$ should be approximated (i) uncorrelated if $W_t \sim \text{WN}(0, \sigma^2)$, (ii) independent if $\{W_t\} \sim \text{IID}(0, \sigma^2)$, and (iii) normally distributed if $W_t \sim N(0, \sigma^2)$.

7 Nonstationary process

7.1 Introduction of ARIMA process

If the data exhibits no apparent deviations from stationarity and has a rapidly decreasing autocorrelation function, we shall seek a suitable ARMA process to represent the mean-correlated data. If not, then we shall first look for a transformation of the data which generates a new series with the above properties. This can frequently be achieved by differencing, leading us to consider the class of ARIMA (autoregressive-integrated moving average) processes.

Definition of intrinsically stationary process: Stochastic process $\{X_t\}$ is said to be intrinsically stationary of integer order $d > 0$ if $\{X_t\}$, $\{\nabla X_t\}$, ..., $\{\nabla^{d-1} X_t\}$ are non-stationary, but $\{\nabla^d X_t\}$ is a stationary process.

Note that

$$\begin{aligned}
\nabla X_t &= (1-B)X_t = X_t - X_{t-1} \\
\nabla^2 X_t &= \nabla(\nabla X_t) = (1-B)^2 X_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2} \\
&\vdots \\
\nabla^d X_t &= (1-B)^d X_t = \sum_{k=0}^{d} \binom{d}{k} (-1)^k B^k X_t = \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k}
\end{aligned}$$

And any stationary process are intrinsically stationary of order $d = 0$.

Definition of the ARIMA(p, d, q) process: If d is a non-negative integer, then $\{X_t\}$ is said to be an ARIMA(p, d, q) process if

1. $\{X_t\}$ is intrinsically stationary of order d and
2. $\{\nabla^d X_t\}$ is a causal ARMA(p, q) process.

With $\{W_t\} \sim \text{WN}(0, \sigma^2)$, we can express the model as

$$\phi(B)(1-B)^d X_t = \theta(B) W_t$$

or equivalently, as

$$\phi^*(B) X_t = \theta(B) W_t$$

where $\phi^*(B) = \phi(B)(1-B)^d$, $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q, respectively and $\phi(z) \neq 0$ for $|z| \leq 1$. Noting that if $d > 0$, then $\phi^*(z)$ has a zero of order d at $z = 1$ (on the unit circle).

Example 7.1. A simplest example of ARIMA process is ARIMA$(0, 1, 0)$:

$$(1-B)X_t = X_t - X_{t-1} = W_t,$$

for which, assuming existence of X_0 and $t \geq 1$,

$$\begin{aligned}
X_1 &= X_0 + W_1 \\
X_2 &= X_1 + W_2 = X_0 + W_1 + W_2 \\
X_3 &= X_2 + W_3 = X_0 + W_1 + W_2 + W_3 \\
&\vdots \\
X_t &= X_0 + \sum_{k=1}^{t} W_k.
\end{aligned}$$

Above is a random walk starting from X_0. Assuming X_0 is uncorrelated with W_ts, we have

$$\begin{aligned}
\mathrm{Var}X_t &= \mathrm{Var}\left(X_0 + \sum_{k=1}^{t} W_k\right) \\
&= \mathrm{Var}X_0 + \sum_{k=1}^{t} \mathrm{Var}W_k \\
&= \mathrm{Var}X_0 + t\sigma^2
\end{aligned}$$

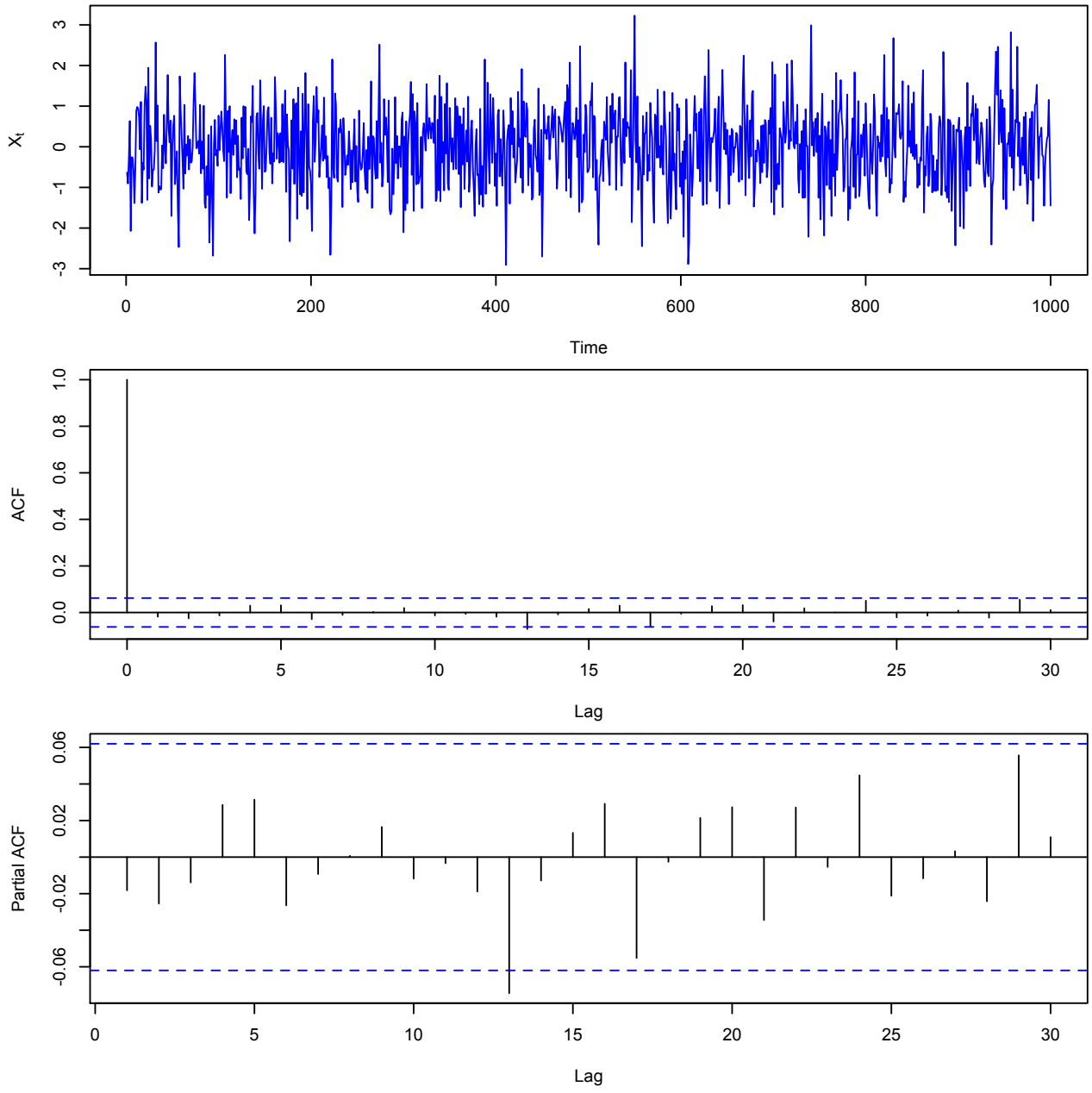
which is either time-dependent or infinite if $\mathrm{Var}X_0 = \infty$. Further

$$\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}\left(X_0 + \sum_{k=1}^{t+h} W_k, X_0 + \sum_{k=1}^{t} W_k\right) = \mathrm{Var}X_0 + \min(t, t+h)\sigma^2.$$

Thus, ARIMA$(0, 1, 0)$ is a non-stationary process. This same true for all ARIMA(p, d, q) process when d is a positive integer.

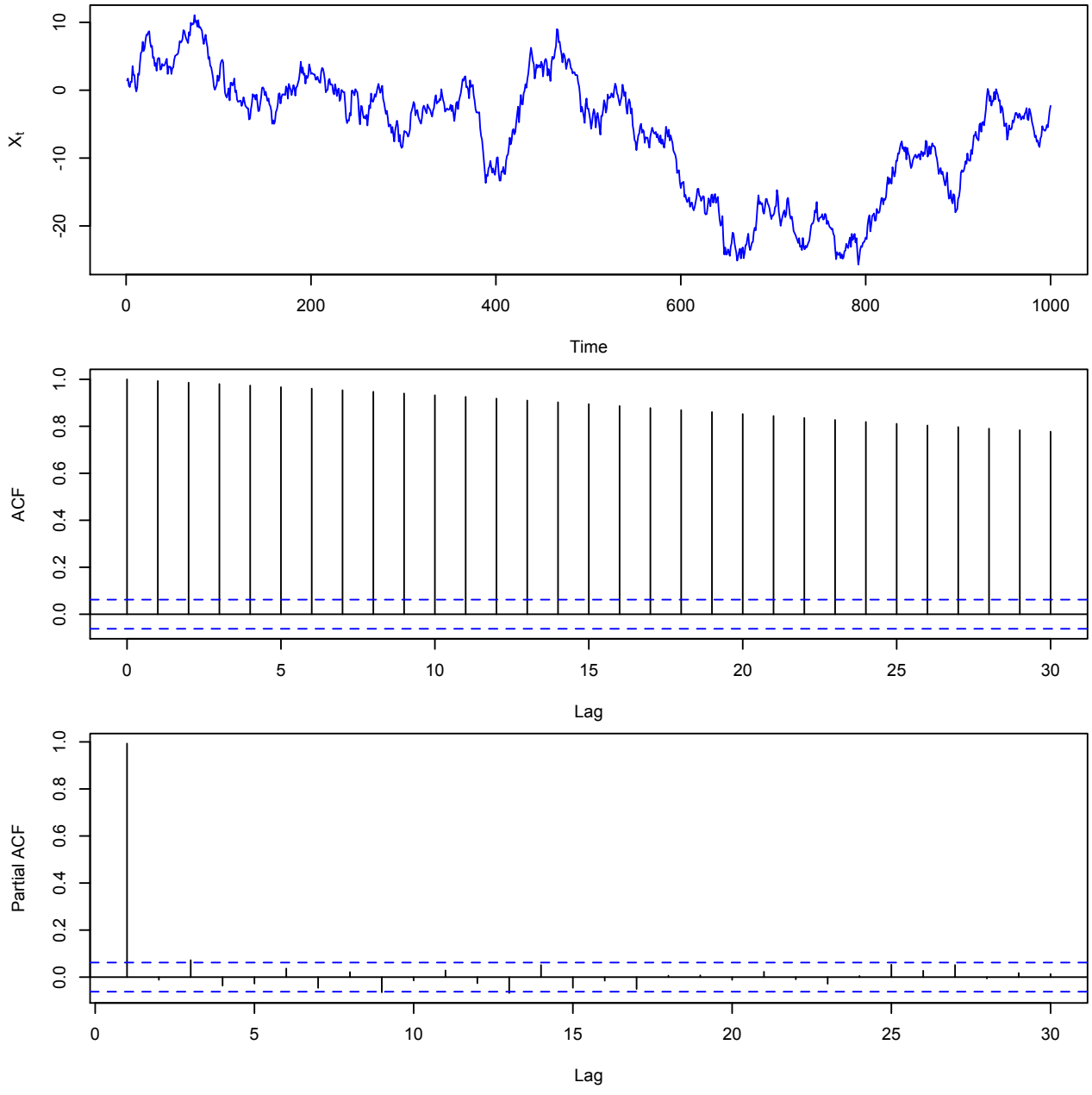
A realization of white noise.

```
N=1000; Wt=rnorm(N,0,1); par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1))
plot.ts(Wt,col="blue",ylab=expression(X[t]));acf(Wt,type="correlation");
acf(Wt, type="partial")
```



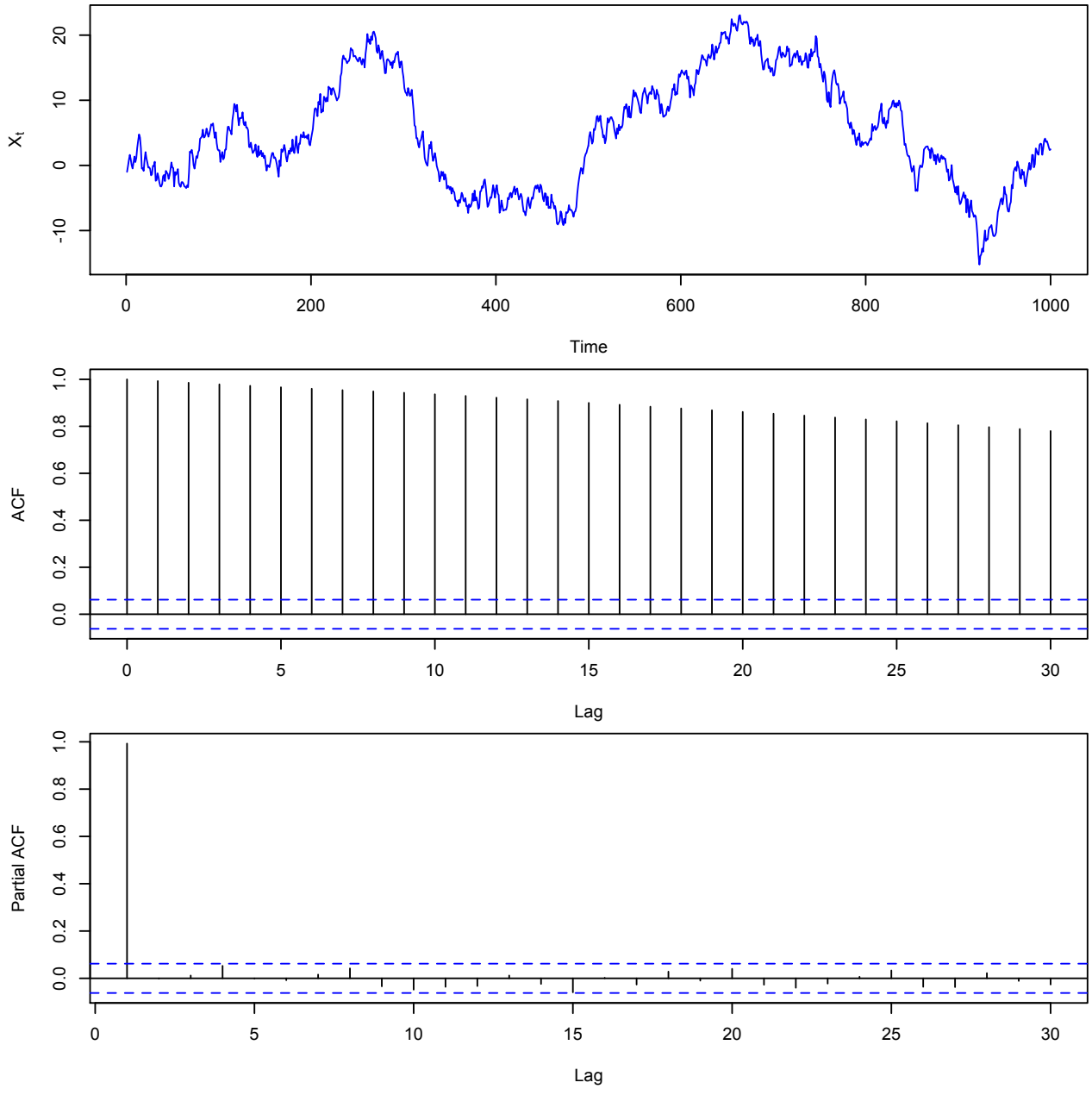
A realizaiton of random walk.

```
N=1000; Wt=rnorm(N,0,1); Xt=cumsum(Wt);
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt,col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



Another realization of random walk.

```
N=1000; Wt=rnorm(N,0,1); Xt=cumsum(Wt);
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt,col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



Example 7.2. $\{X_t\}$ is an ARIMA$(1,1,0)$ process if for some $\phi \in (-1,1)$,

$$(1-\phi B)(1-B)X_t = W_t, \quad \{W_t\} \sim \text{WN}(0,\sigma^2).$$

We can write $Y_t = (1-B)X_t$ which is an AR(1) and thus

$$Y_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}.$$

Back to $\{X_t\}$, we have

$$X_t - X_{t-1} = Y_t$$

and recursively,

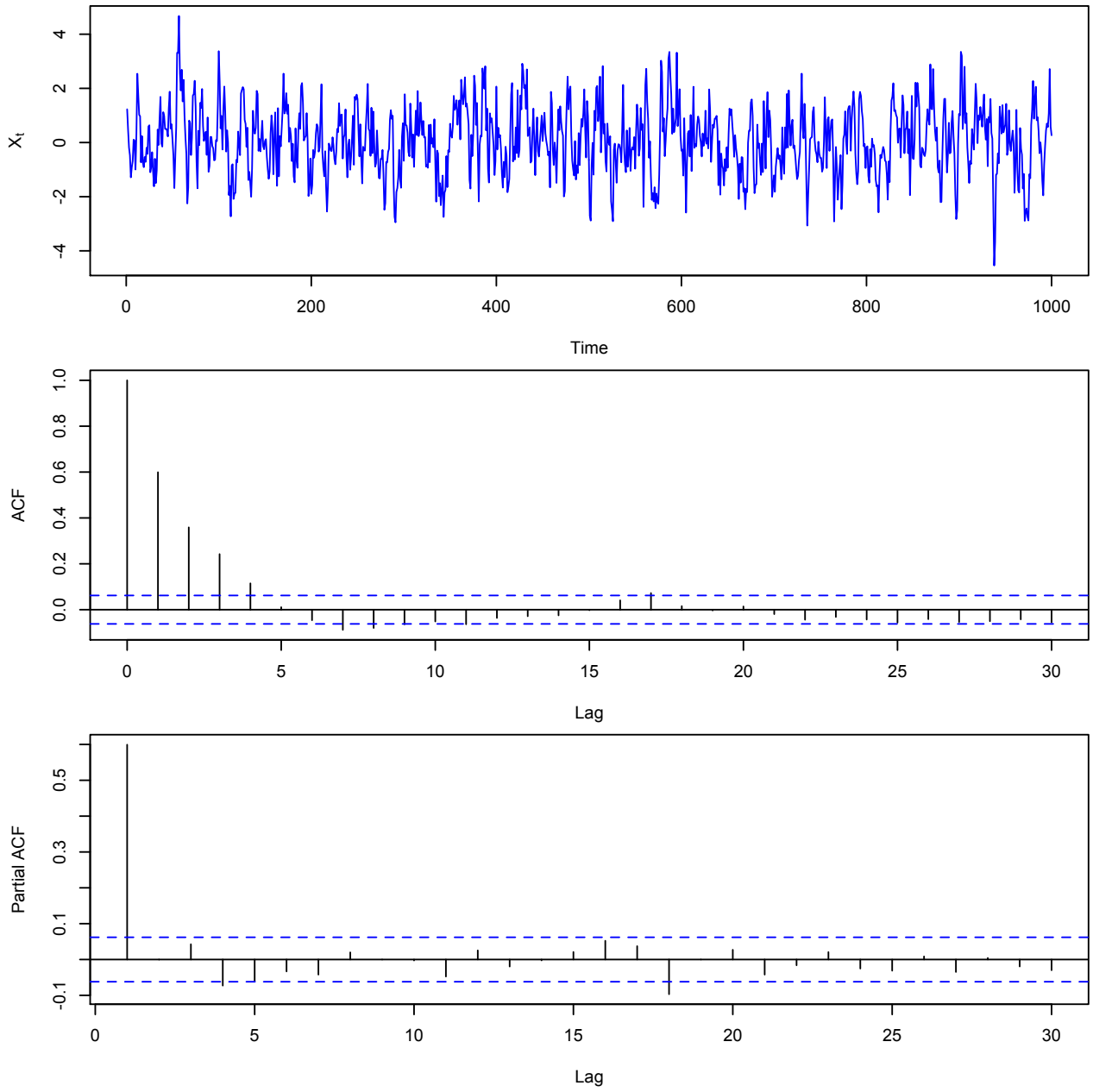
$$X_t = X_0 + \sum_{j=1}^{t} Y_j, \quad t \geq 1.$$

Further

$$\begin{aligned}
\text{Var}X_t &= \text{Var}X_0 + \text{Var}\left(\sum_{j=1}^{t} Y_j\right) \\
&= \text{Var}X_0 + \sum_{i=1}^{t}\sum_{j=1}^{t} \text{Cov}(Y_i, Y_j) \\
&= \text{Var}X_0 + \sum_{i=1}^{t}\sum_{j=1}^{t} \gamma_Y(i-j) \\
&= \text{Var}X_0 + t\gamma_Y(0) + 2\sum_{i=2}^{t}(t-i+1)\gamma_Y(i-1) \\
&= \text{Var}X_0 + \gamma_Y(0)\left\{t + 2\sum_{i=2}^{t}(t-i+1)\rho_Y(i-1)\right\} \\
&= \text{Var}X_0 + \frac{\sigma^2}{1-\phi^2}\left\{t + 2\sum_{i=2}^{t}(t-i+1)\phi^{i-1}\right\}
\end{aligned}$$

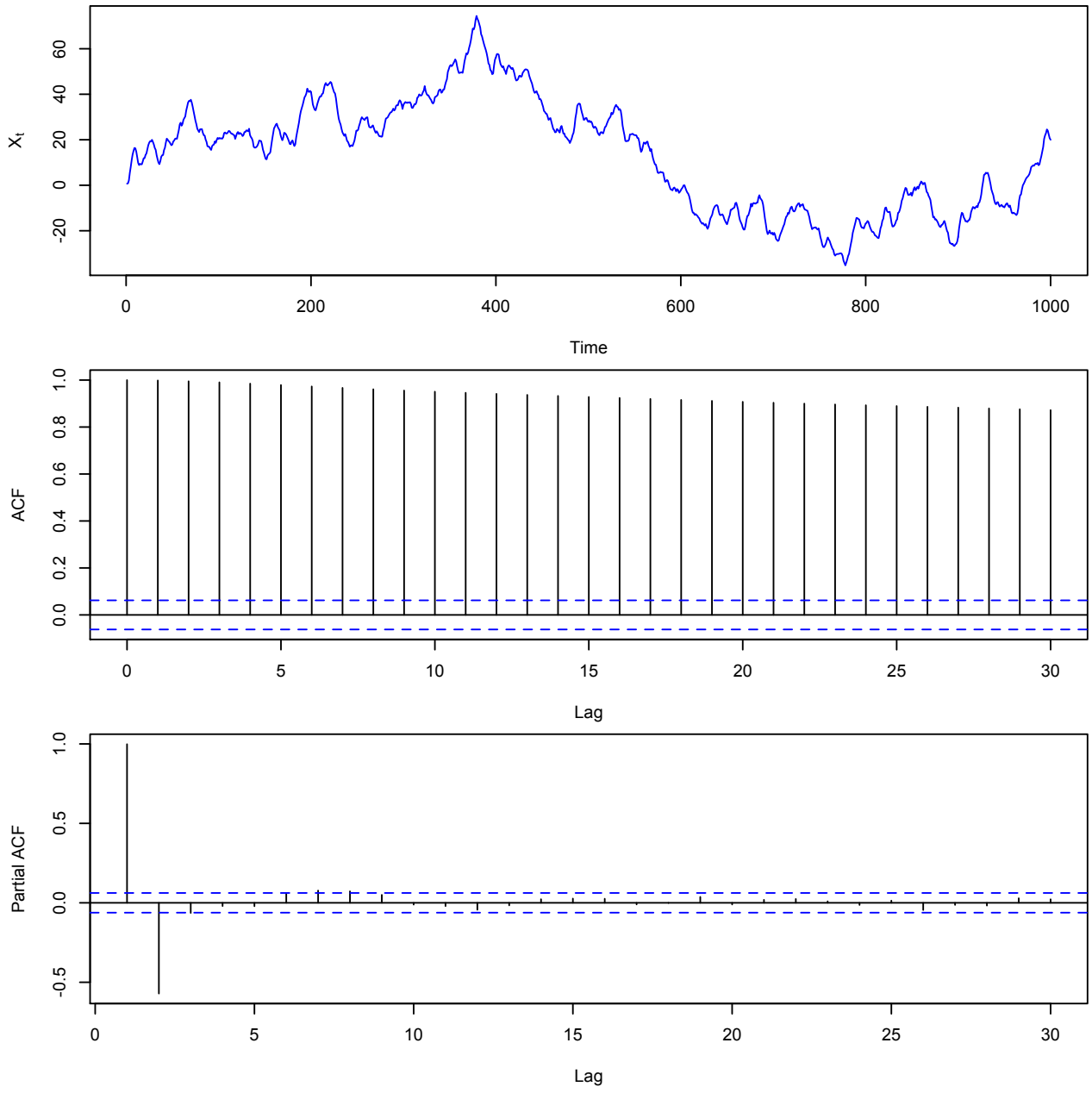
An realization of AR(1).

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, filter=c(.6), method="recursive")[-(1:50)]
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Yt, col="blue",ylab=expression(X[t]));
acf(Yt,type="correlation");acf(Yt, type="partial")
```



An realization of ARIMA$(1,1,0)$.

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, filter=c(.6), method="recursive")[-(1:50)];Xt=cumsum(Yt)
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt, col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



Example 7.3. $\{X_t\}$ is an ARIMA$(0,1,1)$ process, then

$$(1-B)X_t = (1+\theta B)W_t, \quad \{W_t\} \sim \mathrm{WN}(0,\sigma^2).$$

We can write $Y_t = (1+\theta B)W_t$ which is an MA(1), thus

$$X_t - X_{t-1} = Y_t$$

and recursively,

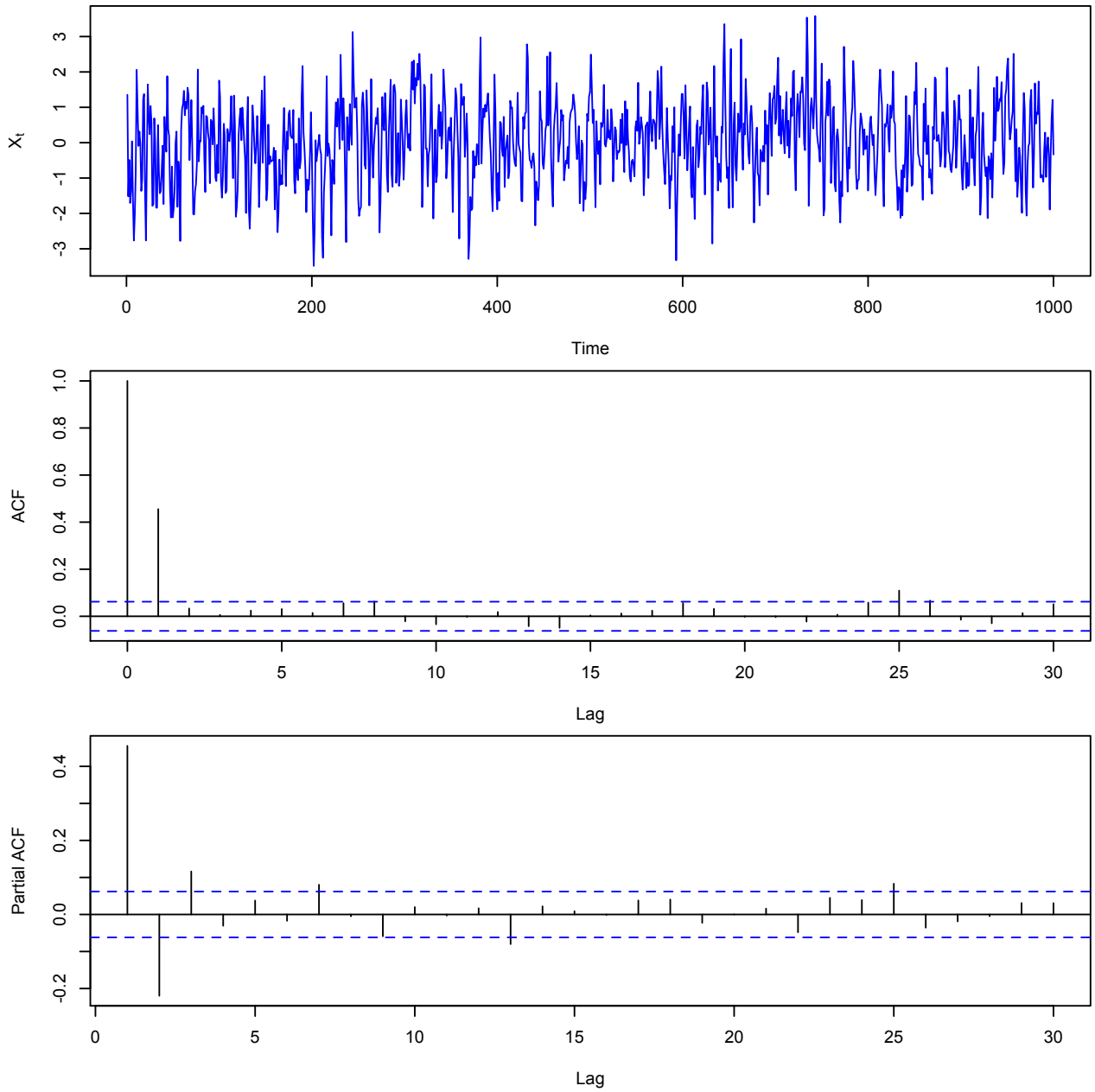
$$X_t = X_0 + \sum_{j=1}^{t} Y_j = X_0 + \sum_{j=1}^{t} (W_t + \theta W_{t-1}).$$

Then, if X_0 is uncorrelated with Y_ts,

$$\begin{aligned}
\mathrm{Var}X_t &= \mathrm{Var}X_0 + \mathrm{Var}\left(\sum_{j=1}^{t} Y_j\right) \\
&= \mathrm{Var}X_0 + \sum_{i=1}^{t}\sum_{j=1}^{t} \mathrm{Cov}(Y_i, Y_j) \\
&= \mathrm{Var}X_0 + \sum_{i=1}^{t}\sum_{j=1}^{t} \gamma_Y(i-j) \\
&= \mathrm{Var}X_0 + t\gamma_Y(0) + 2\sum_{i=2}^{t}(t-i+1)\gamma_Y(i-1) \\
&= \mathrm{Var}X_0 + t\gamma_Y(0) + 2(t-1)\gamma_Y(1) \\
&= \mathrm{Var}X_0 + t(1+\theta^2)\sigma^2 + 2(t-1)\theta\sigma^2.
\end{aligned}$$

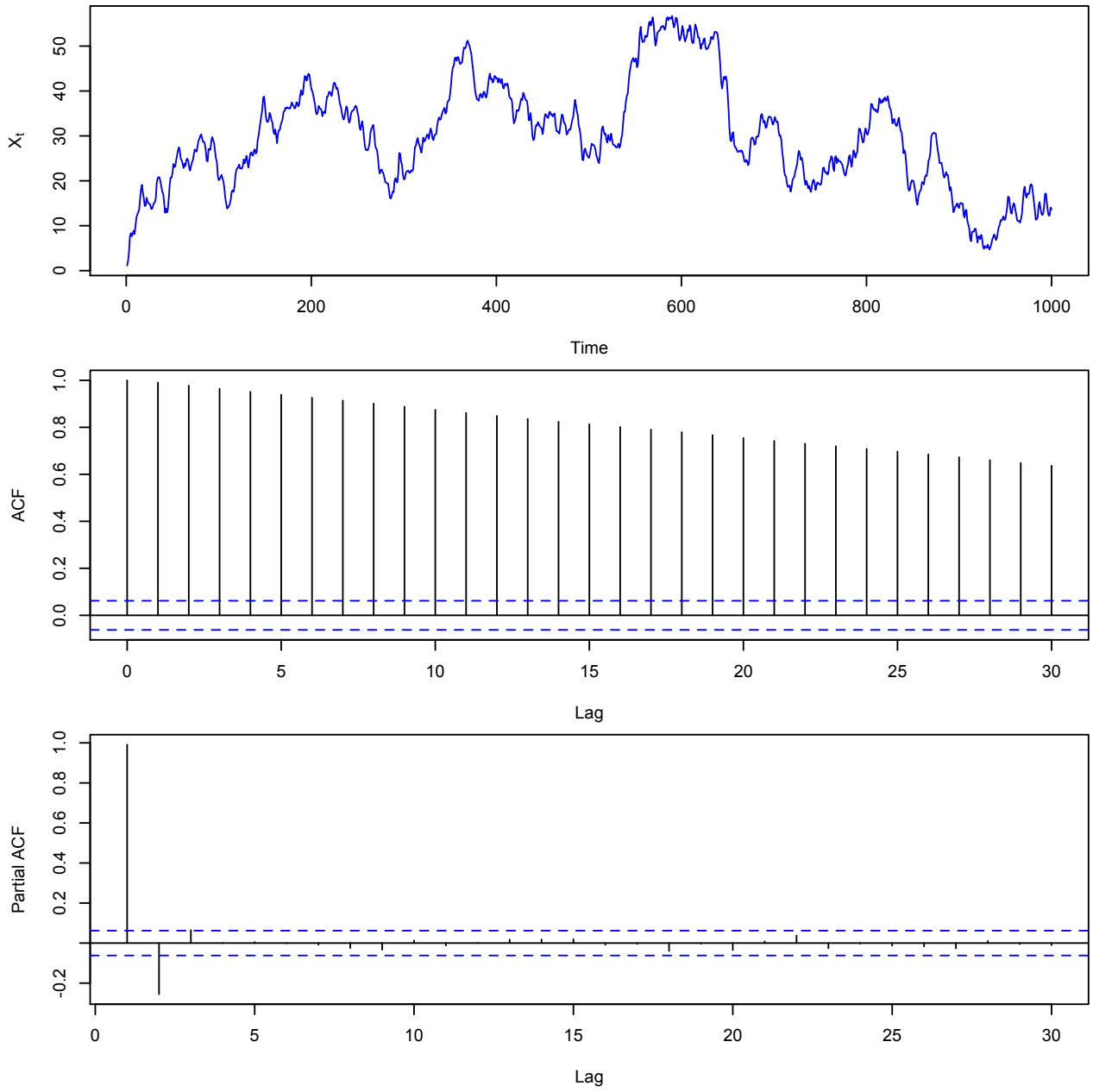
An realization of MA(1).

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, sides=1, c(1,.6))[-(1:50)];
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Yt, col="blue",ylab=expression(X[t]));
acf(Yt,type="correlation");acf(Yt, type="partial")
```



An realization of ARIMA$(0,1,1)$.

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, sides=1, c(1,.6))[-(1:50)]; Xt=cumsum(Yt)
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt, col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



Example 7.4. $\{X_t\}$ is an ARIMA$(0, 1, q)$ process, then

$$(1 - B)X_t = \theta(B)W_t, \quad \{W_t\} \sim \text{WN}(0, \sigma^2).$$

We can write $Y_t = \theta(B)W_t$ which is an MA(1), thus

$$X_t = X_0 + \sum_{j=1}^{t} \theta(B)W_j = X_0 + \sum_{j=1}^{t} (W_t + \theta W_{t-1}).$$

Example 7.5. Moreover, if $\{X_t\}$ is an ARIMA$(p, 1, q)$, based on the causality assumption, we can write $Y_t = (1 - B)X_t$ as

$$Y_t = \sum_{j=0}^{\infty} \psi^j W_{t-j},$$

where $\psi(z) = \theta(z)/\phi(z)$. Then recursively,

$$X_t = X_0 + \sum_{j=1}^{t} Y_j.$$

Example 7.6. What if $d = 2$? For ARIMA$(p, 2, q)$, we have

$$\phi(B)(1 - B)^2 X_t = \theta(B)W_t,$$

then

$$(1 - B)^2 X_t = Z_t = \psi(B)W_t, \quad \text{where } \psi(z) = \theta(z)/\phi(z),$$

is ARMA(p, q). Letting $(1 - B)Y_t = Z_t$, we have

$$(1 - B)Y_t = Z_t$$

Thus $\{Y_t\}$ is an ARIMA$(p, 1, q)$ and

$$Y_t = Y_0 + \sum_{j=1}^{t} Z_t.$$

Since $(1 - B)X_t = Y_t$, we further have

$$X_t = X_0 + \sum_{j=1}^{t} Y_t.$$

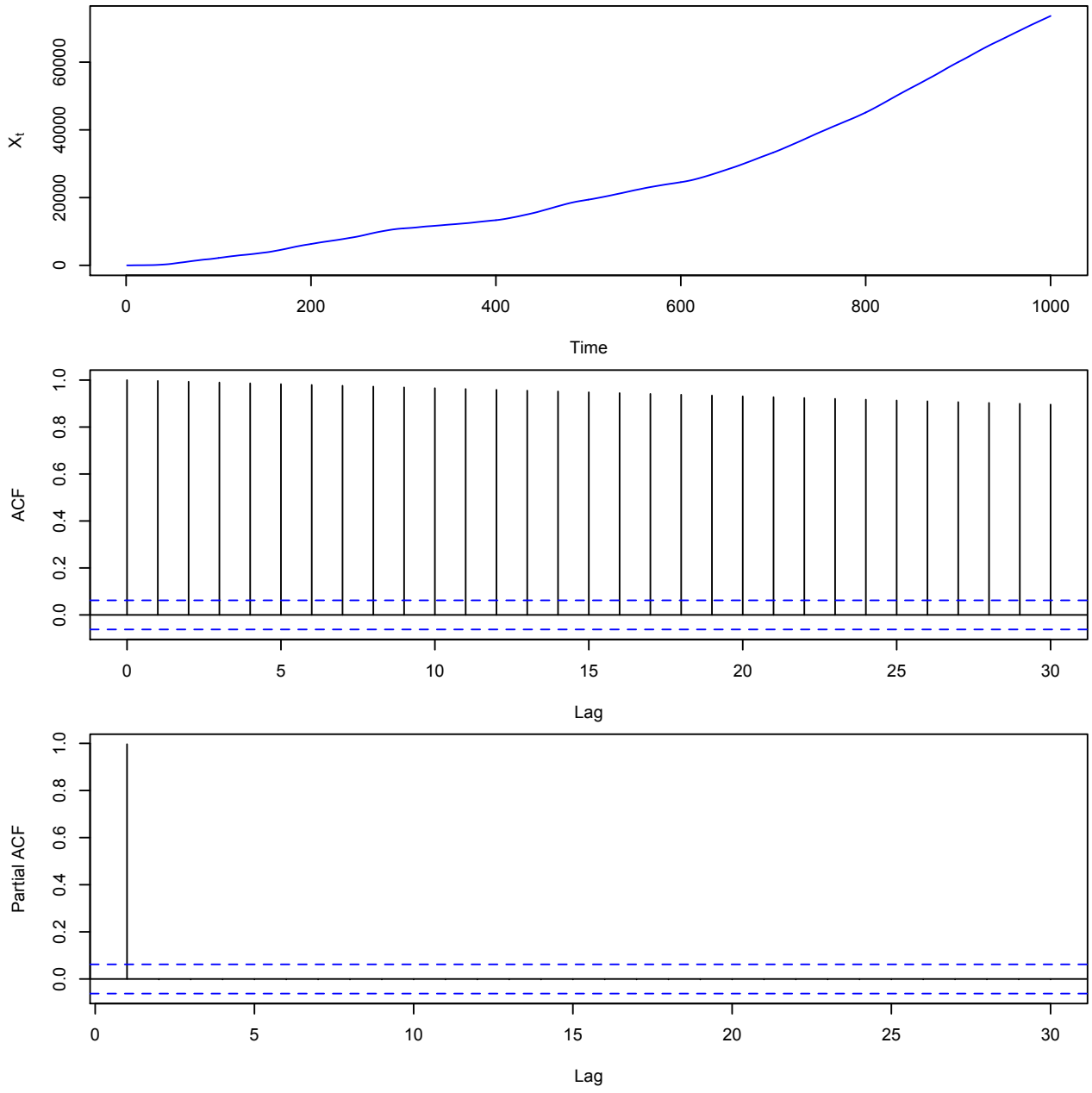
Notice that for any constant α_0 and α_1,

$$X_t^* = X_t + \alpha_0 + \alpha_1 t$$

is also a solution to the equation $\phi(B)(1 - B)^2 X_t = \theta(B)W_t$,.

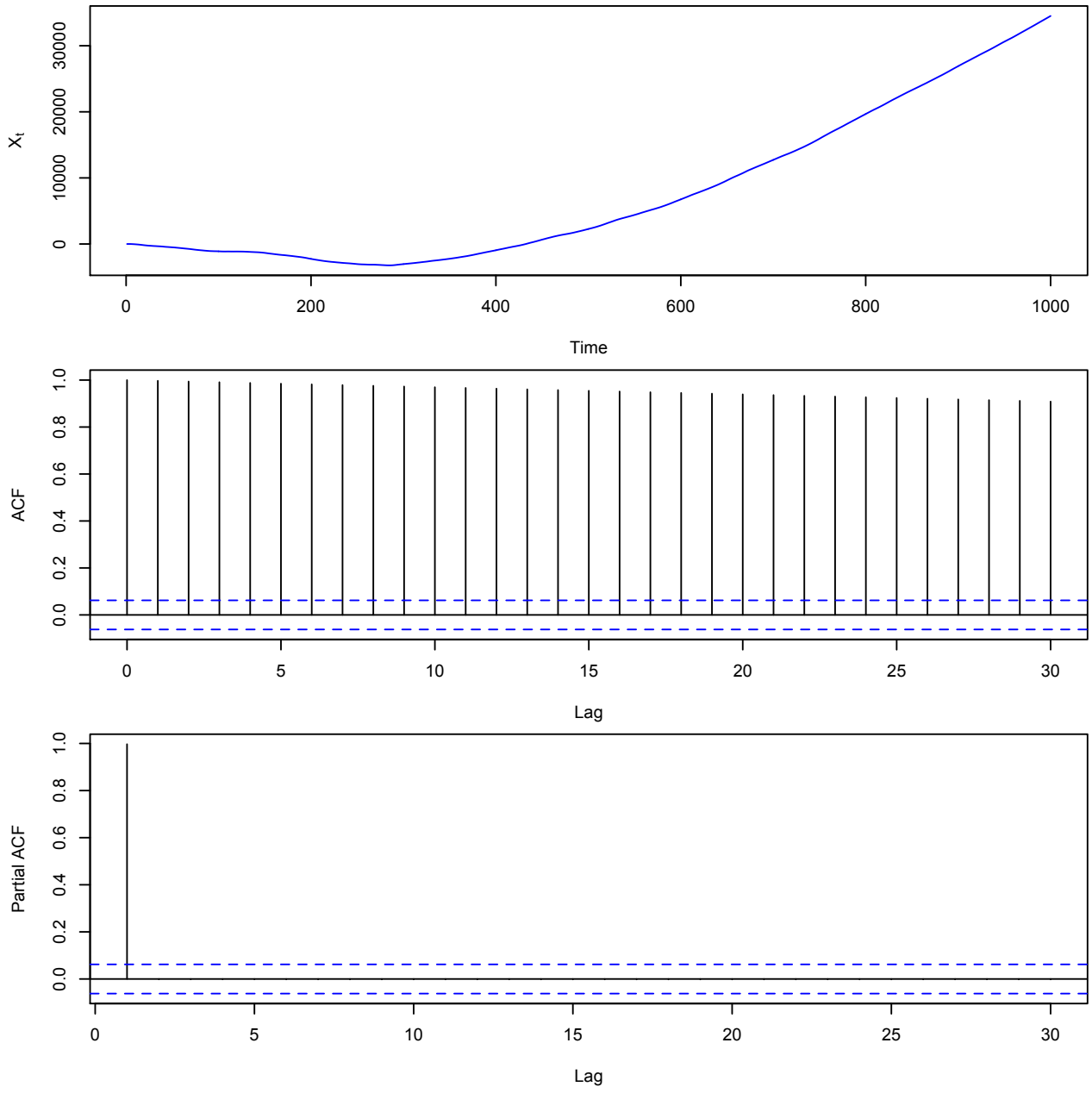
An realization of ARIMA$(1, 2, 0)$.

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, filter=c(.6), method="recursive")[-(1:50)];Xt=cumsum(cumsum(Yt))
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt, col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



An realization of ARIMA$(0, 2, 1)$.

```
N=1050;Wt= rnorm(N,0,1);
Yt = filter(Wt, sides=1, c(1,.6))[-(1:50)]; Xt=cumsum(cumsum(Yt))
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1));
plot.ts(Xt, col="blue",ylab=expression(X[t]));
acf(Xt,type="correlation");acf(Xt, type="partial")
```



7.2 Over-differencing?

While differencing a time series often seems to yield a series visually more amenable to modeling as a stationary process, over-differencing is a danger! If X_t is ARMA(p, q) already, and satisfies

$$\phi(B)X_t = \theta(B)W_t,$$

then one more differencing provides that

$$(1-B)\phi(B)X_t = (1-B)\theta(B)W_t,$$
$$\phi(B)(1-B)X_t = \theta(B)(1-B)W_t;$$

i.e., one more difference of X_t, denoted by $Y_t = \nabla X_t$ satisfies

$$\phi(B)Y_t = \theta^*(B)W_t$$

where $\theta(z) = \theta(z)(1-z)$. Noting that $\theta^*(z)$ has a root on the unit circle. Thus Y_t is a non-invertible ARMA$(p, q+1)$ process. Summary of evils of over-differencing:

1. ARMA$(p, q+1)$ model usually has more complex covariance structure than ARMA(p, q).
2. ARMA$(p, q+1)$ model has one more parameter to estimate than ARMA(p, q) model.
3. Sample size is reduced by 1 (from n to $n-1$, not a big issue you may think...)

```
N=5000; Wt=rnorm(N,0,1); par(mar=c(4.5,4.5,.1,.1));
design.mat=matrix(c(1:12),3,4);  layout(design.mat)

plot.ts(Wt, col="blue",ylab=expression(X[t])); acf(Wt,type="correlation",ylim=c(-1,1));
acf(Wt, type="partial",ylim=c(-1,1));

Xt=Wt[-1]-Wt[-length(Wt)]; plot.ts(Xt, col="blue",ylab=expression(W[t]))
acf(Xt,type="correlation",ylim=c(-1,1)); acf(Xt, type="partial",ylim=c(-1,1))

Xt=Xt[-1]-Xt[-length(Xt)]; plot.ts(Xt, col="blue",ylab=expression(W[t]))
acf(Xt,type="correlation",ylim=c(-1,1)); acf(Xt, type="partial",ylim=c(-1,1))

Xt=Xt[-1]-Xt[-length(Xt)];  plot.ts(Xt, col="blue",ylab=expression(W[t]))
acf(Xt,type="correlation",ylim=c(-1,1)); acf(Xt, type="partial",ylim=c(-1,1))
```

Unit root tests help determine if differencing is needed. For example, suppose X_t obeys an ARIMA$(0, 1, 0)$ model:

$$(1-B)X_t = W_t$$

This can be viewed as a special case of AR(1) process as

$$(1-\phi B)X_t = W_t, \quad \text{where } \phi = 1.$$

Thus, we can design a test for null hypothesis $\phi = 1$. However, the MLE inference only hold for $|\phi| < 1$; i.e., when the process is causal.

Dicey and Fuller (1979) designed an alternative test statistic. They use ordinary least square to estimate $\phi^* = \phi - 1$, and then test hypothesis $\phi^* = 0$. The method is based on the equation:

$$\begin{aligned}\nabla X_t &= X_t - X_{t-1} \\ &= (\phi X_{t-1} + W_t) - X_{t-1} \\ &= (\phi - 1)X_{t-1} + W_t \\ &= \phi^* X_{t-1} + W_t.\end{aligned}$$

Thus the Dickey-Fuller unit root test is to use ordinary least squares (OLS) to regress ∇X_t on X_{t-1}.

Note that, if the mean of X_t is μ but not 0, then we have

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + W_t,$$

where $\phi_0^* = \mu(1-\phi)$ and $\phi_1^* = \phi - 1$. The goal is now to test

$$H_0 : \phi_1^* = 0 \text{ versus } H_a : \phi_1^* < 0$$

Since we do not need consider $\phi_1^* > 0$ (which corresponding to non-causal AR(1) model).

In the standard regression model

$$Y_t = a + bZ_t + e_t, \quad t = 1, \ldots, m.$$

Minimizing the least square

$$\sum_t (Y_t - a - bZ_t)^2$$

yields the OLS estimator of b as

$$\begin{aligned}\widehat{b} &= \frac{\sum_t (Y_t - \bar{Y})(Z_t - \bar{Z})}{\sum_t (Z_t - \bar{Z})^2} = \frac{\sum_t Y_t (Z_t - \bar{Z})}{\sum_t (Z_t - \bar{Z})^2}, \\ \widehat{a} &= \bar{Y} - \widehat{b}\bar{Z},\end{aligned}$$

and the standard error of $\widehat{b}$ is taken to be

$$\widehat{\mathrm{SE}}(\widehat{b}) = \left\{ \frac{\sum_t (Y_t - \widehat{a} - \widehat{b}Z_t)^2}{(m-2)\sum_t (Z_t - \bar{Z})^2} \right\}^{1/2}.$$

where $\bar{Y}$ and $\bar{Z}$ are sample means of Y_ts and Z_ts, respectively.

Now, denote our model is

$$\nabla X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + W_t, \quad \text{for } t = 2, \dots n.$$

Define $Y_t = X_t - X_{t-1}$ and $Z_t = X_{t-1}$. We have a regression model

$$Y_t = \phi_0^* + \phi_1^* Z_t + W_t, \quad \text{for } t = 2, \dots n.$$

Then, we have

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

Thus, the OLS estimator of ϕ_1^* is

$$\begin{aligned}
\widehat{\phi}_1^* &= \frac{\sum_{t=2}^n Y_t (Z_t - \bar{Z})}{\sum_{t=2}^n (Z_t - \bar{Z})^2} = \frac{\sum_{t=2}^n (X_t - X_{t-1})(X_{t-1} - \bar{Z})}{\sum_{t=2}^n (X_{t-1} - \bar{Z})^2} \\
&= \frac{\sum_{t=2}^n \{(X_t - \bar{Z}) - (X_{t-1} - \bar{Z})\}(X_{t-1} - \bar{Z})}{\sum_{t=2}^n (X_{t-1} - \bar{Z})^2} \\
&= \frac{\sum_{t=2}^n (X_t - \bar{Z})(X_{t-1} - \bar{Z}) - \sum_{t=2}^n (X_{t-1} - \bar{Z})^2}{\sum_{t=2}^n (X_{t-1} - \bar{Z})^2} \\
&= \frac{\sum_{t=2}^n (X_t - \bar{Z})(X_{t-1} - \bar{Z})}{\sum_{t=2}^n (X_{t-1} - \bar{Z})^2} - 1.
\end{aligned}$$

And

$$\widehat{\mathrm{SE}}(\widehat{\phi}_1^*) = \left\{ \frac{\sum_{t=2}^n (\nabla X_t - \widehat{\phi}_0^* - \widehat{\phi}_1^* X_{t-1})^2}{(n-3) \sum_t (X_{t-1} - \bar{Z})^2} \right\}^{1/2}.$$

Further

$$\widehat{\phi}_0^* = \bar{Y} - \widehat{\phi}_1^* \bar{Z} = \frac{X_n - X_1}{n-1} - \widehat{\phi}_1^* \frac{\sum_{t=1}^{n-1} X_t}{n-1} = \frac{1}{n-1} \left(X_n - X_1 - \widehat{\phi}_1^* \sum_{t=1}^{n-1} X_t \right).$$

Then test statistic for

$$H_0 : \phi_1^* = 0 \text{ versus } H_a : \phi_1^* < 0$$

is t-like ratio

$$t = \frac{\widehat{\phi}_1^*}{\widehat{\mathrm{SE}}(\widehat{\phi}_1^*)}$$

However, this t does not obey a t-distribution. We reject null (have a unit root) in favor of alternative (AR(1) is appropriate) at level α is t falls below $100(1-\alpha)$ percentage point established for Dickey-Fuller test statistic under assumption that n is large. The 1%, 5% and 10% critical points are -3.43, -2.86, and -2.57, respectively.

7.3 Seasonal ARIMA Models

The classical decomposition of the time series

$$X_t = m_t + s_t + Y_t.$$

where m_t is the trend component, s_t is the seasonal component, and Y_t is the random noise component. We have learned to use differencing to eliminate the trend m_t. (Why? One easy way to interpret that is that, we can view m_t as a smooth function of t. Then any smooth function can be approximated by a polynomial; i.e., there exits d, $\alpha_0, \ldots, \alpha_d$ such that

$$m_t \approx \sum_{k=0}^{d} \alpha_k t^d.$$

Then

$$X_t = \sum_{k=0}^{d} \alpha_k t^d + S_t + Y_t.$$

Taking d-order differencing, we have

$$\nabla^d X_t = \alpha_d + \nabla^d S_t + \nabla^d Y_t.$$

Further

$$\nabla^{d+1} X_t = \nabla^{d+1} S_t + \nabla^{d+1} Y_t.$$

Thus, trend has been eliminated.) In this section, we learn how to handle seasonal components.

7.3.1 Seasonal ARMA models

Seasonal models allow for randomness in the seasonal pattern from one cycle to the next. For example, we have r years of monthly data which we tabulate as follows:

Year	Jan.	Feb.	$\cdots$	Dec.
1	X_1	X_2	$\cdots$	X_{12}
2	X_{13}	X_{14}	$\cdots$	X_{24}
3	X_{25}	X_{26}	$\cdots$	X_{36}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	$X_{1+12(r-1)}$	$X_{2+12(r-1)}$	$\cdots$	$X_{12+12(r-1)}$

Each column in this table may itself be viewed as a realization of a time series. Suppose that each one of these twelve time series is generated by the same ARMA(P, Q) model, or more specifically that the series corresponding to the jth month,

$$\{X_{j+12t}, t = 0, 1, \ldots, r-1\}$$

satisfies

$$X_{j+12t} = \Phi_1 X_{j+12(t-1)} + \cdots + \Phi_p X_{j+12(t-P)} + U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \cdots + \Theta_Q U_{j+12(t-Q)}$$

where

$$\{U_{j+12t}, t = \cdots, -1, 0, 1, \cdots\} \sim \text{WN}(0, \sigma_U^2)$$

holds for each $j = 1, \ldots, 12$. Since the same ARMA(P, Q) model is assumed to apply to each month, we can rewrite is as

$$X_t = \Phi_1 X_{t-12} + \cdots + \Phi_P X_{t-12P} + U_t + \Theta_1 U_{t-12} + \cdots + \Theta_Q U_{t-12Q}$$

Or

$$\Phi(B^{12}) X_t = \Theta(B^{12}) U_t, \tag{7.1}$$

where $\Phi(z) = 1 - \Phi_1 z - \cdots \Phi_p z^p$, $\Theta(z) = 1 + \Theta_1 z + \cdots \Theta_Q z^Q$, and $\{U_{j+12t}, t = \cdots, -1, 0, 1, \cdots\} \sim$ WN$(0, \sigma_U^2)$ for each j.

Now, for simplicity, we assume $U_t \sim \text{WN}(0, \sigma^2)$, then we have X_t as an ARMA process:

$$\Phi(B^{12}) X_t = \Theta(B^{12}) W_t, \quad W_t \sim \text{WN}(0, \sigma^2)$$

Or more generally,

$$\Phi(B^s) X_t = \Theta(B^s) W_t, \quad W_t \sim \text{WN}(0, \sigma^2)$$

We call this model as Seasonal ARMA (SARMA) process with period s (an integer), denoted by $ARMA(P, Q)_s$.

If $\Phi(z)$ and $\Theta(z)$ make a stationary ARMA(P, Q), then X_t is also a stationary ARMA process.

What would an ARMA$(1, 1)_4$ look like

$$(1 - \Phi_1 B^4) X_t = (1 + \Theta_1 B^4) W_t$$

which is

$$X_t = \Phi_1 X_{t-4} + W_t + \Theta_1 W_{t-4};$$

i.e., a regular ARMA$(4, 4)$ model (with certain coefficients being zero). Now, let us think about a monthly model and a seasonal MA; i.e., ARMA$(0, 1)_{12}$. The model can be written as

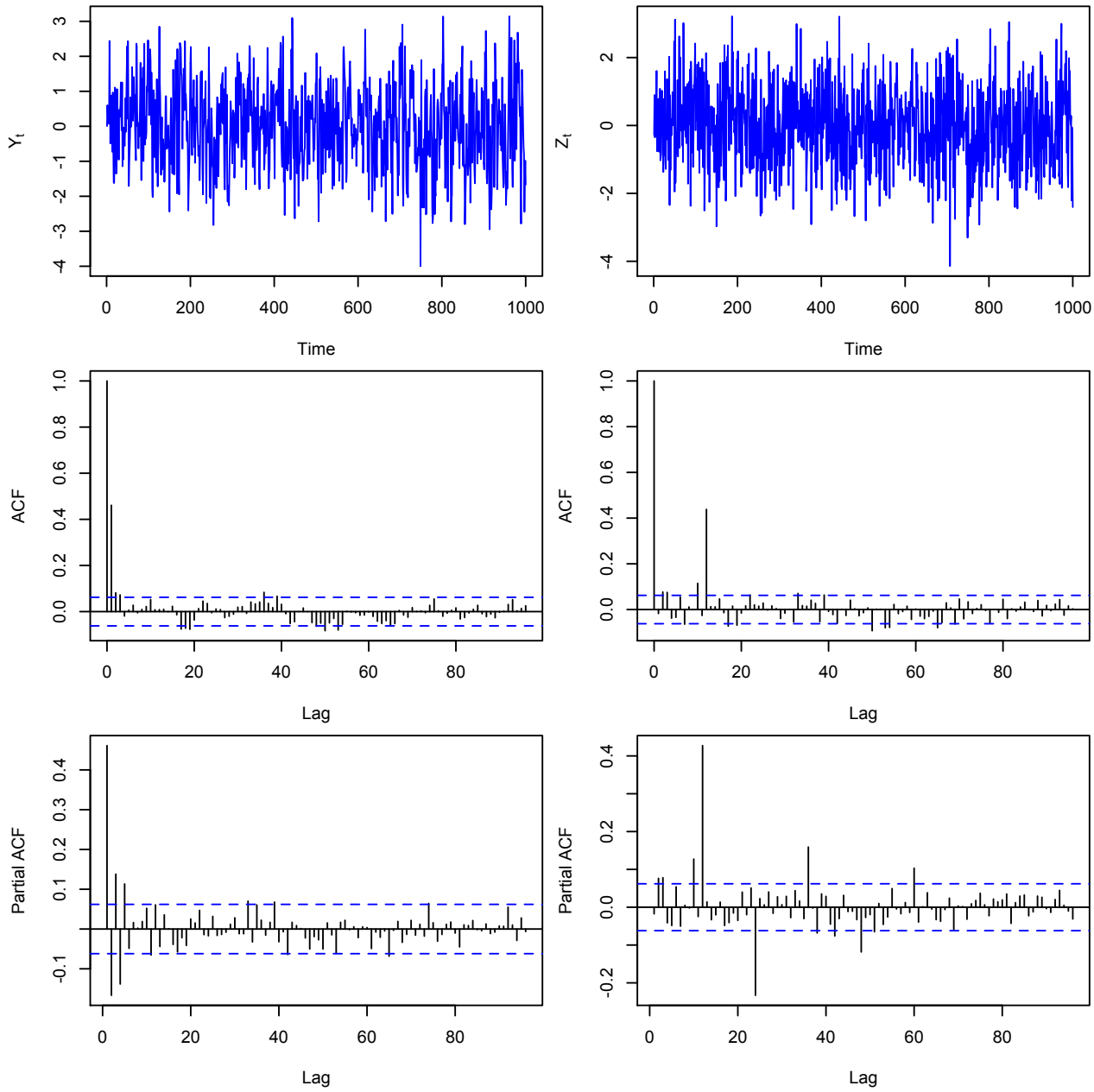
$$X_t = W_t + \Theta_1 W_{t-12}.$$

Obviously, it is causal and stationary. Then

$$\gamma_X(h) = E(X_t X_{t+h}) = E(W_t + \Theta_1 W_{t-12})(W_{t+h} + \Theta_1 W_{t+h-12}) = (1 + \Theta_1^2)\sigma^2 I(h = 0) + \Theta_1 \sigma^2 I(h = 12).$$

Thus, we should see one spike at lag 0 and also at 12.

```
N=1050;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1));
design.mat=matrix(c(1:6),3,2);  layout(design.mat);
Yt = filter(Wt, sides=1, c(1,.6))[-(1:50)];
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=96);acf(Yt, type="partial",lag=96)
Zt= filter(Wt, sides=1, c(1,rep(0,11),.6))[-(1:50)];
plot.ts(Zt, col="blue",ylab=expression(Z[t]));
acf(Zt,type="correlation",lag=96);acf(Zt, type="partial",lag=96)
```



Now, let us check a seasonal AR; i.e., ARMA$(1,0)_{12}$; i.e.,

$$X_t = \Phi_1 X_{t-12} + W_t.$$

Similarly as in the analysis of $AR(1)$ process, we have, recursively,

$$\begin{aligned} X_t &= \Phi_1 X_{t-12} + W_t = \Phi_1^2 X_{t-24} + \phi_1 W_{t-12} + W_t \\ &= \cdots \\ &= \sum_{j=0}^{\infty} \Phi_1^j W_{t-12j}. \end{aligned}$$

Thus,

$$\gamma_X(0) = \sum_{j=0}^{\infty} \Phi_1^{2j} \sigma^2 = \sigma^2 \frac{1}{1-\Phi_1^2}$$

What about $\gamma_X(1), \gamma_X(2), \ldots, \gamma_X(11)$? They are all zero. And

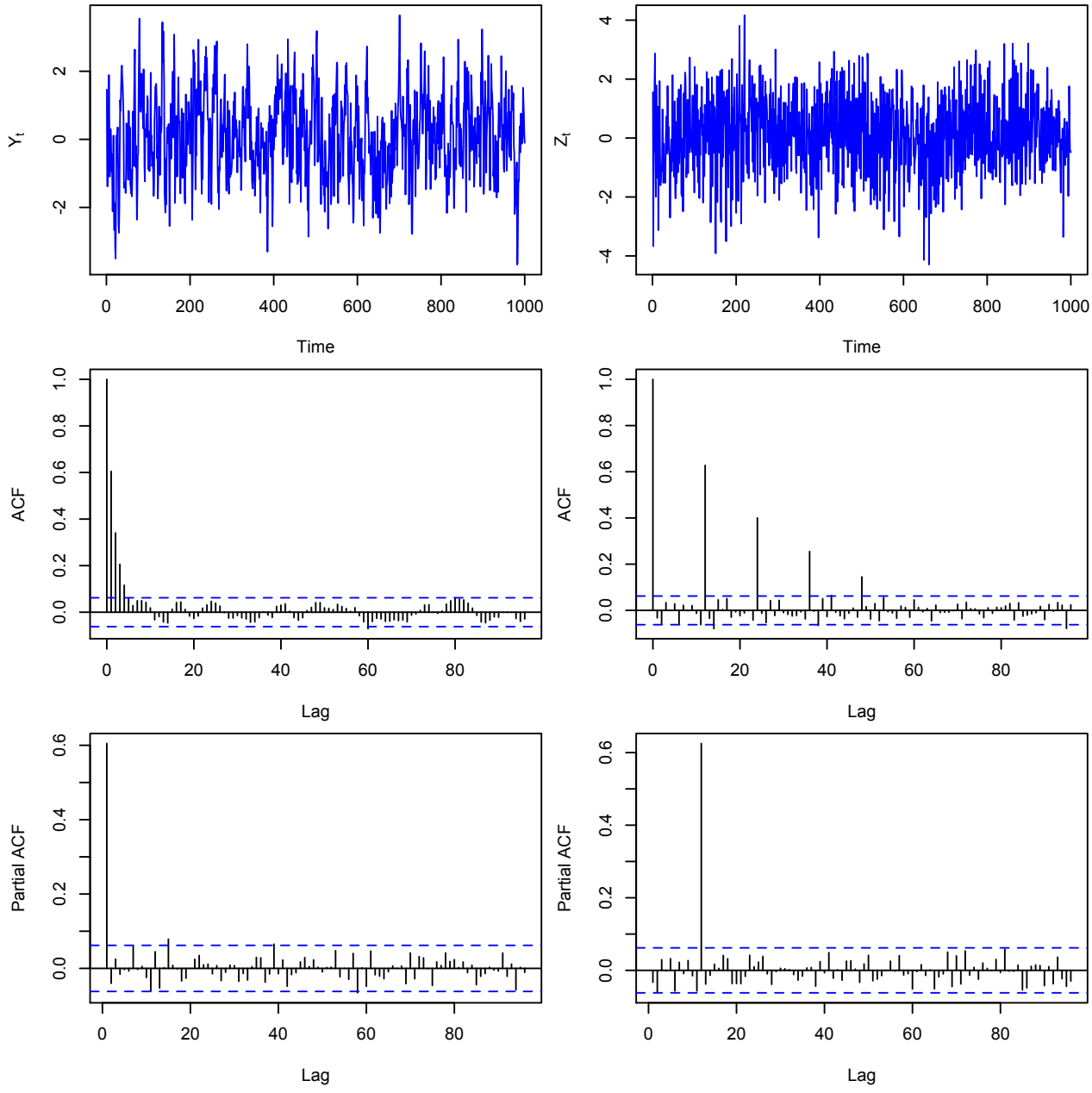
$$\begin{aligned} \gamma_X(12) &= E\left(\sum_{j=0}^{\infty} \Phi_1^j W_{t-12j}\right)\left(\sum_{j=0}^{\infty} \Phi_1^j W_{t+12-12j}\right) \\ &= E\left(\sum_{j=0}^{\infty} \Phi_1^j W_{t-12j}\right)\left(W_{t+12} + \Phi_1 \sum_{j=1}^{\infty} \Phi_1^{j-1} W_{t-12(j-1)}\right) \\ &= E\left(\sum_{j=0}^{\infty} \Phi_1^j W_{t-12j}\right)\left(W_{t+12} + \Phi_1 \sum_{j=0}^{\infty} \Phi_1^j W_{t-12j}\right) \\ &= \Phi_1 \sum_{j=0}^{\infty} \Phi_1^{2j} \sigma^2 = \sigma^2 \frac{\Phi_1}{1-\Phi_1^2}, \end{aligned}$$

Further

$$\gamma_X(h) = \sigma^2 \sum_{\text{integer } k} \frac{\Phi_1^k}{1-\Phi_1^2} I(h = 12k).$$

This is very similar to an AR(1), however, we are ignoring all the pages between multiples of 12. We are not worried about what is going on at 1,2,3,4,..., we are only interested in the 12,24,36,.... You can certainly generalized it to a general period, like ARMA$(1,0)_s$, for any integer s.

```
N=1050;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1));
design.mat=matrix(c(1:6),3,2);  layout(design.mat);
Yt = filter(Wt, filter=c(.6), method="recursive")[-(1:50)];
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=96);acf(Yt, type="partial",lag=96)
Zt = filter(Wt, filter=c(rep(0,11),.6), method="recursive")[-(1:50)];
plot.ts(Zt, col="blue",ylab=expression(Z[t]));
acf(Zt,type="correlation",lag=96);acf(Zt, type="partial",lag=96)
```



It is unlikely that the 12 series (in column) corresponding to the different months are uncorrelated. To incorporate dependence between these series we assume now that the $\{U_t\}$ sequence follows an ARMA(p, q) model,

$$\phi(B)U_t = \theta(B)W_t, \quad W_t \sim \text{WN}(0, \sigma^2).$$

Then we have

$$\begin{aligned} \Phi(B^{12})X_t &= \Theta(B^{12})U_t \\ &= \Theta(B^{12})\phi^{-1}(B)\theta(B)W_t \\ \Rightarrow \qquad \phi(B)\Phi(B^{12})X_t &= \theta(B)\Theta(B^{12})W_t, \quad W_t \sim \text{WN}(0, \sigma^2). \end{aligned}$$

More generally, we have the SARMA model as

$$\text{ARMA}(p, q) \times (P, Q)_s$$

written as

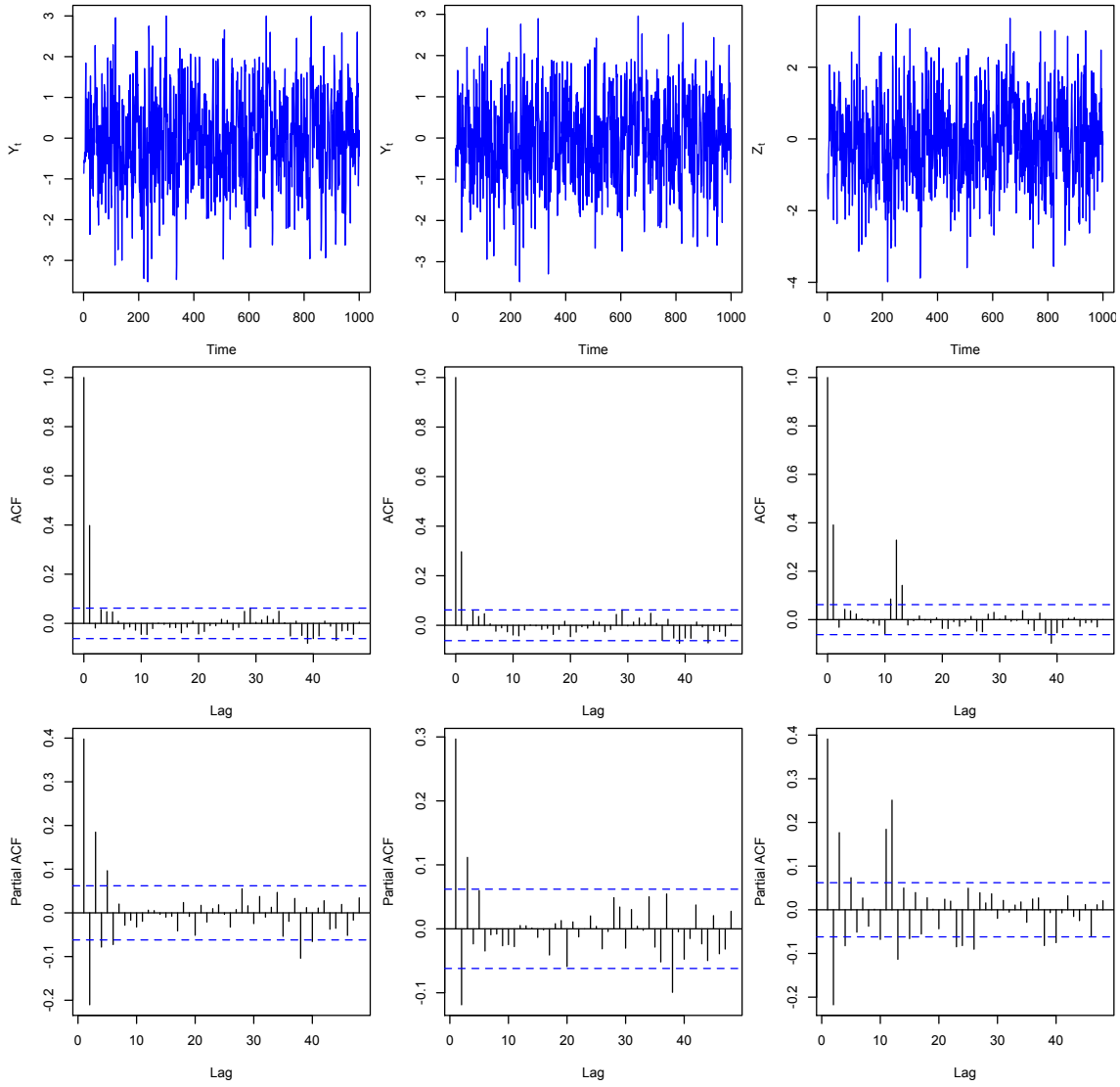
$$\phi(B)\Phi(B^s)X_t = \theta(B)\Theta(B^s)W_t, \quad W_t \sim \text{WN}(0, \sigma^2).$$

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_p z^p$, $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ and $\Theta(z) = 1 + \Theta_1 z + \cdots \Theta_Q z^Q$.

For example, we consider ARMA$(0, 1) \times (0, 1)_{12}$; i.e.,

$$X_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})W_t = W_t + \theta_1 W_{t-1} + \Theta_1 W_{t-12} + \theta_1 \Theta_1 W_{t-13}.$$

```
N=1050;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1));
design.mat=matrix(c(1:9),3,3);  layout(design.mat);
#MA(1), theta=0.6
Yt = filter(Wt, sides=1, c(1,.6))[-(1:50)];
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)
#MA(1), theta=0.4
Yt = filter(Wt, sides=1, c(1,.4))[-(1:50)];
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)
#SARMA(0,1)*(0,1)_{12}, theta=0.6, Theta=0.4
Zt = filter(Wt, sides=1, c(1,.6,rep(0,10),.4,.24))[-(1:50)];
plot.ts(Zt, col="blue",ylab=expression(Z[t]));
acf(Zt,type="correlation",lag=48);acf(Zt, type="partial",lag=48)
```



What is an ARMA$(1,1) \times (1,1)_{12}$?

$$(1-\phi_1 B)(1-\Phi_1 B^{12})X_t = (1+\theta_1 B)(1+\Theta_1 B^{12})W_t$$

as

$$X_t - \phi_1 X_{t-1} - \Phi_1 X_{t-12} + \phi_1\Phi_1 X_{t-13} = W_t + \theta_1 W_{t-1} + \Theta_1 W_{t-12} + \theta_1\Theta_1 W_{t-13}.$$

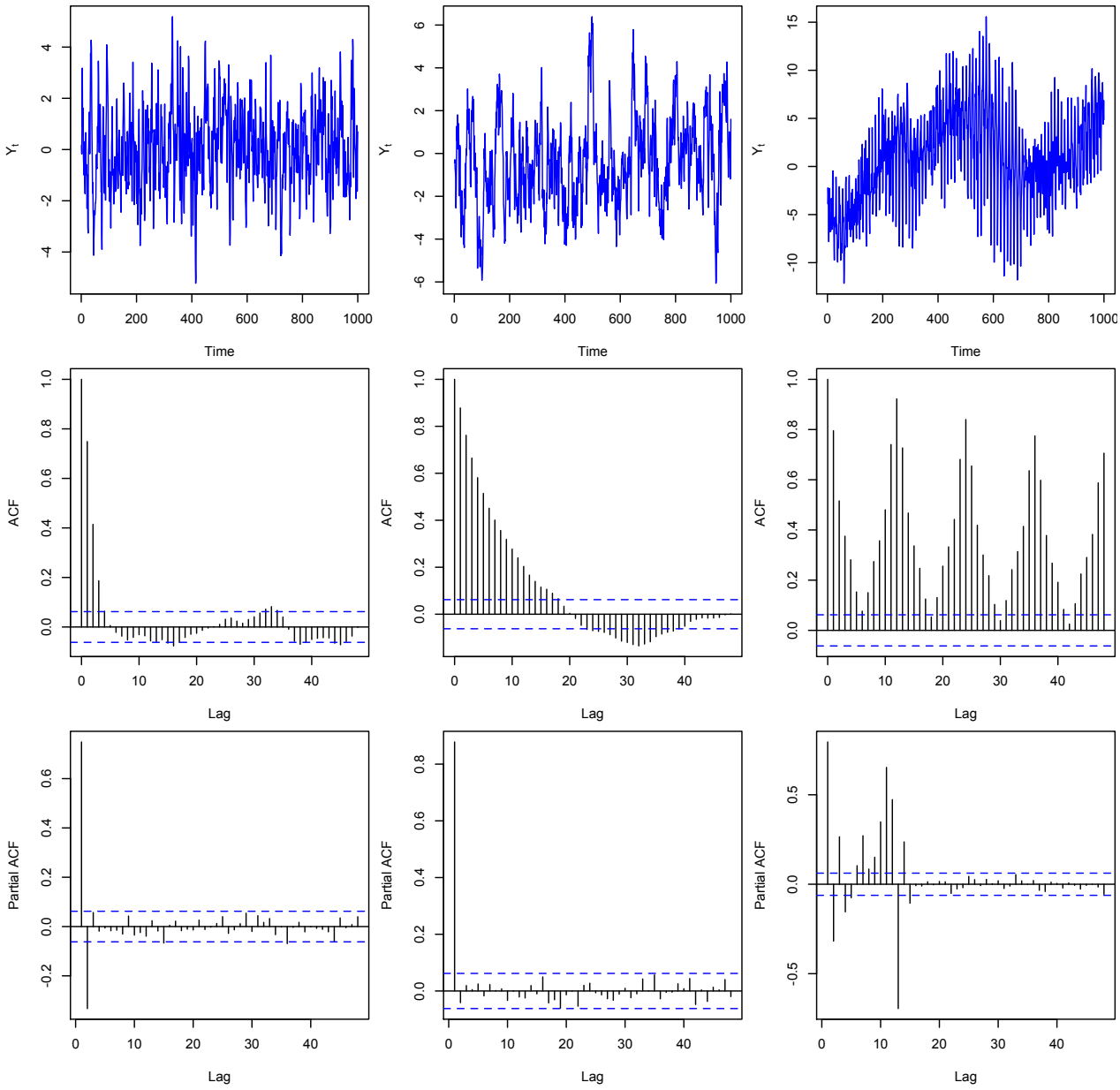
Things gets mess, specially when you start getting into higher values.

```
N=1000;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1));
design.mat=matrix(c(1:9),3,3);layout(design.mat);

# ARMA(1,1), phi=0.6, theta=0.4
Yt=arima.sim(model=list(ar=c(.6),ma=c(.4)),n=N)
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)

# ARMA(1,1), phi=0.9, theta=0.1
Yt=arima.sim(model=list(ar=c(.9),ma=c(.1)),n=N)
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)

# ARMA(1,1)*(1,1)_12, phi=0.6, Phi=0.9,theta=0.4, Theta=0.1
Yt=arima.sim(model=list(ar=c(.6,rep(0,10),.9,-.54),ma=c(.4,rep(0,10),.1,.04)),n=N)
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)
```


Y_t
Time
ACF
Lag
Partial ACF

7.3.2 Seasonal ARIMA Models

In equation (7.1), we have

$$\Phi(B^{12})X_t = \Theta(B^{12})U_t,$$

We start with U_t to be white noise, then relaxing this a little bit by putting U_t to be an ARMA(p, q) process, we have the Seasonal ARMA(p, q) model. Now we keep relaxing the assumption. What if we put U_t to be an ARIMA(p, d, q) process? This makes sense. Think about a monthly temperature dataset across many years. The period is 12. But, there is a trend among the months. From winter to summer, it is warmer and warmer, then decreases when getting back to winter. This makes the stationarity of U_t quite unreasonable. Putting an ARIMA model on U_t can somehow fix this. That is

$$\phi(B)(1-B)^d U_t = \theta(B)W_t, \quad W_t \sim \text{WN}(0, \sigma^2). \tag{7.2}$$

Then, X_t becomes

$$\Phi(B^{12})X_t = \Theta(B^{12})U_t \quad \Rightarrow \quad \phi(B)\Phi(B^{12})(1-B)^d X_t = \Theta(B^{12})\theta(B)W_t.$$

Or more generally, we have X_t as

$$\phi(B)\Phi(B^s)\{(1-B)^d X_t\} = \theta(B)\Theta(B^s)W_t, \quad W_t \sim \text{WN}(0, \sigma^2).$$

Can we relax it more? On the part of U_t, based on what we have learned, the assumption of U_t which assumes U_t is ARIMA is the best we have relax. However about on X_t? We can do more! In equation (7.1), we have X_t satisfies

$$\Phi(B^{12})X_t = \Theta(B^{12})U_t,$$

which is a more like stationary structure. What if the structure of X_t is already non-stationary regardless of the choice of U_t? We put a similar ARIMA structure on X_t, since X_t is seasonal, we assume

$$\Phi(B^{12})\{(1-B^{12})^D X_t\} = \Theta(B^{12})U_t.$$

Then combining with (7.2), we have

$$\begin{aligned} &\Phi(B^{12})\{(1-B^{12})^D X_t\} = \Theta(B^{12})U_t \\ \Rightarrow\ &\phi(B)\Phi(B^{12})(1-B)^d(1-B^{12})^D X_t = \theta(B)\Theta(B^{12})W_t. \end{aligned}$$

Generalizing the period form $s = 12$ to a general s, we finally have the definition of a SARIMA model as

The SARIMA$(p,d,q)\times(P,D,Q)_s$ Process: If d and D are non-negative integers, then $\{X_t\}$ is said to be a seasonal ARIMA$(p,d,q)\times(P,D,Q)_s$ process with period s if the differenced process $Y_t=(1-B)^d(1-B^s)^D X_t$ is a causal ARMA process,

$$\phi(B)\Phi(B^s)Y_t=\theta(B)\Theta(B^s)W_t,\quad W_t\sim \mathrm{WN}(0,\sigma^2),$$

where $\phi(z)=1-\phi_1 z-\cdots-\phi_p z^p$, $\Phi(z)=1-\Phi_1 z-\cdots-\Phi_p z^p$, $\theta(z)=1+\theta_1 z+\cdots+\theta_q z^q$ and $\Theta(z)=1+\Theta_1 z+\cdots\Theta_Q z^Q$.

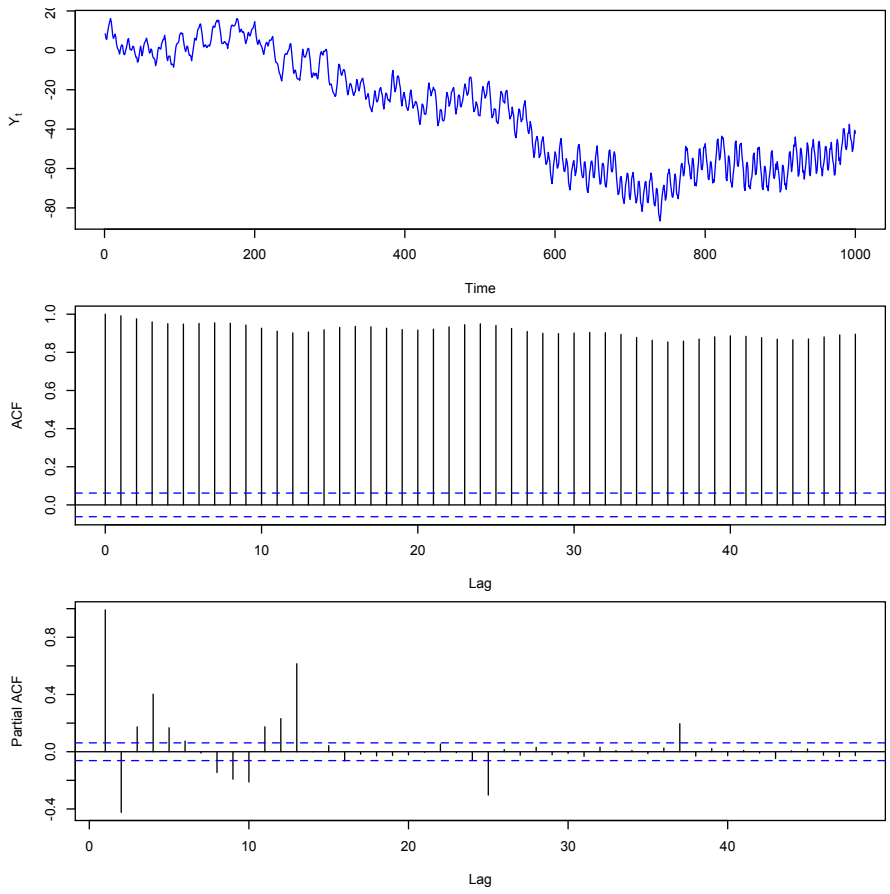
Now, we check basic examples. Consider SARIMA$(0,1,1)\times(0,1,1)_{12}$; i.e.,

$$(1-B)(1-B^{12})X_t=(1+\theta_1 B)(1+\Theta_1 B^{12})W_t.$$

Then,

$$X_t-X_{t-1}-X_{t-12}+X_{t-13}=W_t+\theta_1 W_{t-1}+\Theta_1 W_{t-12}+\theta_1\Theta_1 W_{t-13}.$$

```
N=1100;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1)); par(mfrow=c(3,1))
Yt=filter(Wt, sides=1, c(1,.4,rep(0,10),.6,.24))[-(1:50)]
Yt=filter(Yt, filter=c(1,rep(0,10),-1,1),method="recursive")[-(1:50)]
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)
```



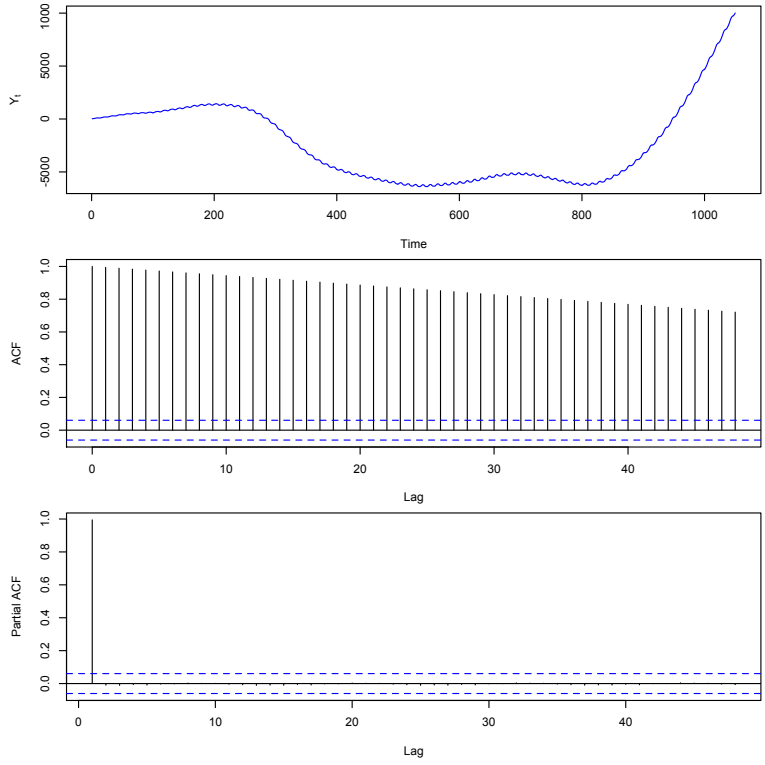
Then SARIMA$(1,1,0)\times(1,1,0)_{12}$; i.e.,

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})X_t = W_t.$$

Then,

$$\begin{aligned} W_t = X_t &- (1+\phi_1)X_{t-1} + \phi_1 X_{t-2} \\ &- (1+\Phi_1)X_{t-12} + (1+\phi_1+\Phi_1+\phi_1\Phi_1)X_{t-13} - (\phi_1+\phi_1\Phi_1)X_{t-14} \\ &+ \Phi_1 X_{t-24} - (\Phi_1+\phi_1\Phi_1)X_{t-25} + \phi_1\Phi_1 X_{t-26} \end{aligned}$$

```
N=1100;Wt= rnorm(N,0,1);par(mar=c(4.5,4.5,.1,.1)); par(mfrow=c(3,1)); phi=.9; Phi=.8;
Yt=filter(Wt, filter=c(1+phi,-phi,rep(0,9),1+Phi,-1-phi-Phi-phi*Phi,phi+phi*Phi,
rep(0,9),-Phi,Phi+phi*Phi,-phi*Phi),method="recursive")[-(1:50)];
plot.ts(Yt, col="blue",ylab=expression(Y[t]));
acf(Yt,type="correlation",lag=48);acf(Yt, type="partial",lag=48)
```



Typically, period s is common set by users from the background of the dataset. For example, years weather dataset usually has $s = 12$ (12 month a year) ; financial data has $s = 4$ (four quarters a year); etc. For the specification of D, in application, it is rarely more than one and P , Q are commonly less than 3. For given values of p, d, q, P, D, Q, we need estimate the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}$ and σ^2. The estimation is very trivial. Since, once d and D are given, we have $Y_t = (1-B)^d(1-B^s)^D X_t$ constitute an ARMA$(p+sP, q+sQ)$ process in which some of the coefficients are zero and the rest are function.

7.4 Regression with stationary errors

Recall classical decomposition model for time series Y_t, namely,

$$Y_t = m_t + s_t + Z_t \tag{7.3}$$

where m_t is trend; s_t is periodic with known period s (i.e. $s_{t-s} = s_t$ for all integer t) satisfying $\sum_{j=1}^{s} s_j = 0$; and Z_t is a stationary process with zero mean. Often, m_t and s_t are taken to be deterministic (no randomness). As we have seen, SARIMA model can help us capture stochastic s_t, and can also handle deterministic m_t and s_t through differencing. However, differencing to eliminate deterministic m_t and/or s_t can lead to undesirable overdifferencing of Z_t. As an alternative approach, one can treat model (7.3) as a regression model with stationary errors.

Standard linear regression models take the form of

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}$$

where

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is vector of responses;
- $\boldsymbol{X}$ is a $n \times k$ matrix and $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is the so-called regression coefficients;
- $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ is vector of $\mathrm{WN}(0, \sigma^2)$ random variables.

For example, we can take

$$Y_t = \beta_1 + \beta_2 t + Z_t$$

where m_t is model by a line $\beta_1 + \beta_2 t$. Or, we have

$$Y_t = \beta_1 + \beta_2 \cos(2\pi\delta t) + \beta_3 \sin(2\pi\delta t) + Z_t.$$

which counts for the seasonality. Both of them can be written in the form of

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}.$$

The standard approach is through the use of the so-called ordinary least squares (OLS) estimator, denoted by $\widehat{\boldsymbol{\beta}}_{ols}$, which is the minimizer of the sum of squared errors:

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} (Y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Setting the derivative of $S(\boldsymbol{\beta})$ to be zero, we obtained the normal equations:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$

and hence

$$\widehat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

if $\boldsymbol{X}'\boldsymbol{X}$ has full rank. If Z_ts are white noises, we have $\widehat{\boldsymbol{\beta}}_{old}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ (best in that, if $\widehat{\boldsymbol{\beta}}$ is any other unbiased estimator, then

$$\mathrm{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\beta}}_{ols}) \leq \mathrm{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\beta}})$$

for any vector $\boldsymbol{c}$ of constants) and further the covariance matrix for $\widehat{\boldsymbol{\beta}}_{ols}$ is given by $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

However, for time series, uncorrelated errors Z are usually unrealistic. More realistically, we have the error term Z_t to be random variables from a stationary process with zero mean. For example, $\{Z_t\}$ could be a causal ARMA(p, q) process:

$$\phi(B)Z_t = \theta(B)W_t, \quad W_t \sim \mathrm{WN}(0, \sigma^2).$$

What should we do? The solution is the generalized least square (GLS) estimator, denoted by $\widehat{\boldsymbol{\beta}}_{gls}$ which minimizes

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})'\boldsymbol{\Gamma}_n^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Thus, the solution is

$$\widehat{\boldsymbol{\beta}}_{gls} = (\boldsymbol{X}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{Y}$$

Why? Recall in the developing of MLE of ARMA(p, q) processes, we have derived that

$$\boldsymbol{\Gamma}_n = CDC'.$$

Then multiplying $D^{-1/2}C^{-1}$ to the model lead to

$$D^{-1/2}C^{-1}\boldsymbol{Y} = D^{-1/2}C^{-1}\boldsymbol{X}\boldsymbol{\beta} + D^{-1/2}C^{-1}\boldsymbol{Z}$$

Then

$$\begin{aligned}
\mathrm{Cov}(D^{-1/2}C^{-1}\boldsymbol{Z}) &= D^{-1/2}C^{-1}\mathrm{Cov}(\boldsymbol{Z})(D^{-1/2}C^{-1}\boldsymbol{Z})' \\
&= D^{-1/2}C^{-1}\boldsymbol{\Gamma}_n(D^{-1/2}C^{-1}\boldsymbol{Z})' = D^{-1/2}C^{-1}CDC'C'^{-1}D^{-1/2} \\
&= \boldsymbol{I}_n
\end{aligned}$$

Thus, we can view that

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{Z}}, \quad \{Z_t\} \sim \mathrm{WN}(0, 1).$$

where $\widetilde{\boldsymbol{Y}} = D^{-1/2}C^{-1}\boldsymbol{Y}$, and $\widetilde{\boldsymbol{X}} = D^{-1/2}C^{-1}\boldsymbol{X}$ and $\widetilde{\boldsymbol{Z}} = D^{-1/2}C^{-1}\boldsymbol{Z}$. Then based on the OLS estimator we have the estimator as

$$(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}_{gls}.$$

In principle, we can use ML under a Gaussian assumption to estimate all parameters in model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}$. Note that, all parameters include $\boldsymbol{\beta}$ and the ARMA(p, q) model for $\boldsymbol{Z}$. However, in practice, the following iterative scheme for parameter estimation often works well

1. Fit the model by OLS and obtain $\widehat{\boldsymbol{\beta}}_{ols}$;
2. Compute the residuals $Y_t - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_{ols}$;
3. fit AMAR(p, q) or other stationary model to residuals
4. using fitted model, compute $\widehat{\boldsymbol{\beta}}_{gls}$ and form residuals $Y_t - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_{gls}$
5. Fit same model to residuals again
6. repeat steps 4 and 5 until parameter estimates have stabilized.

8 Multivariate Time Series

In this section, we consider m times series $\{X_{ti}, t = 0, \pm 1, \pm 2, \dots\}, i = 1, 2, \dots, m$ jointly.

8.1 Second order properties of multivariate time series

Denote

$$\boldsymbol{X}_t = (X_{t1}, \dots, X_{tm})^{\mathrm{T}}, \quad t = 0, \pm 1, \pm 2, \dots$$

The second-order properties of the multivariate time series $\{\boldsymbol{X}_t\}$ are then specified by the mean vectors

$$\boldsymbol{\mu}_t = \mathrm{E}\boldsymbol{X}_t = (\mu_{t1}, \dots, \mu_{tm})^{\mathrm{T}},$$

and covariance matrices

$$\Gamma(t+h, t) = \mathrm{Cov}(\boldsymbol{X}_{t+h}, \boldsymbol{X}_t) = E\{(\boldsymbol{X}_{t+h} - \boldsymbol{\mu}_{t+h})(\boldsymbol{X}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\} = [\gamma_{ij}(t+h, t)]_{i,j=1}^{m}.$$

(Stationary Multivariate Time Series). The series $\{\boldsymbol{X}_t\}$ is said to be stationary if $\boldsymbol{\mu}_t$ and $\Gamma(t + h, t), h = 0, \pm 1, \pm 2, \dots$, are independent of t.

For a stationary series, we shall use the notation

$$\boldsymbol{\mu} = \mathrm{E}\boldsymbol{X}_t$$

and

$$\Gamma(h) = E\{(\boldsymbol{X}_{t+h} - \boldsymbol{\mu}_{t+h})(\boldsymbol{X}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\} = [\gamma_{ij}(h)]_{i,j=1}^{m}$$

to represent the mean of the series and the covariance matrix at lag h, respectively.

Note that, for each i, $\{X_{ti}\}$ is stationary with covariance function $\gamma_{ii}(\cdot)$ and mean function μ_i. The function $\gamma_{ij}(\cdot)$ where $i \neq j$ is called the cross-covariance function of the two series $\{X_{ti}\}$ and $\{X_{tj}\}$. It should be noted that $\gamma_{ij}(\cdot)$ is not in general the same as $\gamma_{ji}(\cdot)$.

Further, the correlation matrix function $R(\cdot)$ is defined by

$$R(h) = \left[\gamma_{ij}(h)/\sqrt{\gamma_{ii}(0)\gamma_{jj}(0))}\right]_{i,j=1}^{m} = [\rho_{ij}(h)]_{i,j=1}^{m}.$$

The function $R(\cdot)$ is the covariance matrix function of the normalized series obtained by subtracting $\boldsymbol{\mu}$ from $\boldsymbol{X}_t$ and then dividing each component by its standard deviation.

Lemma 8.1. The covariance matrix function $\Gamma(\cdot) = [\gamma_{ij}(\cdot)]_{i,j=1}^{m}$ of a stationary time series $\{\boldsymbol{X}_t\}$ has the properties

1. $\Gamma(h) = \Gamma'(-h)$;
2. $|\gamma_{ij}(h)| \leq \sqrt{\gamma_{ii}(0), \gamma_{jj}(0)}, i, j = 1, \dots, m$;
3. $\gamma_{ii}(\cdot)$ is an autocovariance function, $i = 1, \dots, m$;

4. $\sum_{j,k=1}^{n} \boldsymbol{a}_j' \Gamma(j-k) \boldsymbol{a}_k \geq 0$ for all $n \in \{1, 2, \dots\}$ and $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{R}^m$.

And the correlation matrix R satisfies the above four and further

$$\rho_{ii}(0) = 1.$$

Example 8.1. Consider the bivariate stationary process $\{\boldsymbol{X}_t\$\}$ defined by

$$\begin{aligned} X_{t1} &= W_t \\ X_{t2} &= W_t + 0.75 W_{t-10} \end{aligned}$$

where $\{W_t\} \sim \mathrm{WN}(0, 1)$. Then

$$\boldsymbol{\mu} = \mathbf{0}$$

and

$$\begin{aligned} \Gamma(-10) &= \mathrm{Cov}\left\{ \begin{pmatrix} X_{t-10,1} \\ X_{t-10,2} \end{pmatrix}, \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} \right\} \\ &= \mathrm{Cov}\left\{ \begin{pmatrix} W_{t-10} \\ W_{t-10} + 0.75 W_{t-20} \end{pmatrix}, \begin{pmatrix} W_t \\ W_t + 0.75 W_{t-10} \end{pmatrix} \right\} \\ &= \begin{pmatrix} 0 & 0.75 \\ 0 & 0.75 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \Gamma(0) &= \mathrm{Cov}\left\{ \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix}, \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} \right\} \\ &= \mathrm{Cov}\left\{ \begin{pmatrix} W_t \\ W_t + 0.75 W_{t-10} \end{pmatrix}, \begin{pmatrix} W_t \\ W_t + 0.75 W_{t-10} \end{pmatrix} \right\} \\ &= \begin{pmatrix} 1 & 1 \\ 1 & 1.5625 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \Gamma(10) &= \mathrm{Cov}\left\{ \begin{pmatrix} X_{t+10,1} \\ X_{t+10,2} \end{pmatrix}, \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} \right\} \\ &= \mathrm{Cov}\left\{ \begin{pmatrix} W_{t+10} \\ W_{t+10} + 0.75 W_t \end{pmatrix}, \begin{pmatrix} W_t \\ W_t + 0.75 W_{t-10} \end{pmatrix} \right\} \\ &= \begin{pmatrix} 0 & 0 \\ 0.75 & 0.75 \end{pmatrix} \end{aligned}$$

otherwise $\Gamma(j) = 0$. The correlation matrix function is given by

$$R(-10) = \begin{pmatrix} 0 & 0.6 \\ 0 & 0.48 \end{pmatrix}, R(0) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, R(-10) = \begin{pmatrix} 0 & 0 \\ 0.6 & 0.48 \end{pmatrix},$$

and $R(j) = 0$ otherwise.

(Multivariate White Noise). The m-variate series $\{\boldsymbol{W}_t, t = 0, \pm 1, \pm 2, \dots\}$ is said to be white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ written

$$\{\boldsymbol{W}_t\} \sim \text{WN}(0, \boldsymbol{\Sigma}),$$

if and only if $\{\boldsymbol{W}_t\}$ is stationary with mean vector $\mathbf{0}$ and covariance matrix function,

$$\Gamma(h) = \begin{cases} \boldsymbol{\Sigma} & \text{if } h = 0 \\ 0 & \text{otherwise.} \end{cases}$$

If further, we have independence, then we write

$$\{\boldsymbol{W}_t\} \sim \text{IID}(0, \boldsymbol{\Sigma}),$$

Multivariate white noise is used as a building block from which can be constructed an enormous variety of multivariate time series. The linear process are those of the form

$$\boldsymbol{X}_t = \sum_{j=-\infty}^{\infty} C_j \boldsymbol{W}_{t-j}, \quad \{\boldsymbol{W}_t\} \sim \text{WN}(0, \boldsymbol{\Sigma}),$$

where $\{C_j\}$ is a sequence of matrices whose components are absolutely summable. It is easy to see that, this linear process have mean $\mathbf{0}$ and covariance matrix function

$$\Gamma(h) = \sum_{j=-\infty}^{\infty} C_{j+h} \boldsymbol{\Sigma} C_j', \quad h = 0, \pm 1, \pm 2, \dots$$

Estimation of $\boldsymbol{\mu}$. Based on the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, and unbiased estimate of $\boldsymbol{\mu}$ is given by the vector of sample means

$$\overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t.$$

This estimator is consistent and asymptotic normal with rate root$-n$.

Estimation of $\Gamma(h)$. Based on the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, as in the univariate case, a natural estimate of the covariance matrix $\Gamma(h)$ is

$$\widehat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\boldsymbol{X}_{t+h} - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_n)' & \text{for } 0 \le h \le n-1, \\ n^{-1} \sum_{t=-h+1}^{n} (\boldsymbol{X}_{t+h} - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_t - \overline{\boldsymbol{X}}_n)' & \text{for } -n+1 \le h < 0, \end{cases}$$

Denoting $\widehat{\Gamma}(h) = [\widehat{\gamma}_{ij}(h)]_{i,j=1}^m$, we estimate the cross correlation function by

$$\widehat{\rho}_{ij}(h) = \widehat{\gamma}_{ij}(h) / \sqrt{\widehat{\gamma}_{ii}(0)\widehat{\gamma}_{jj}(0)}$$

If $i = j$ this reduces to the sample autocorrelation function of the ith series.

Theorem 8.1. Let $\{\boldsymbol{X}_t\}$ be the bivariate time series

$$\boldsymbol{X}_t = \sum_{k=-\infty}^{\infty} C_k \boldsymbol{W}_{t-k}, \quad \{\boldsymbol{W}_t = (W_{t1}, W_{t2})'\} \sim \text{IID}(0, \boldsymbol{\Sigma})$$

where $\{C_k = [C_k(i,j)]_{i,j=1}^2\}$ is a sequence of matrices with $\sum_{k=-\infty}^{\infty} |C_k(i,j)| < \infty$, $i, j = 1, 2,$. Then as $n \to \infty$,

$$\widehat{\gamma}_{ij}(h) \xrightarrow{p} \gamma_{ij}(h)$$

and

$$\widehat{\rho}_{ij}(h) \xrightarrow{p} \rho_{ij}(h)$$

for each fixed $h \ge 0$ and for $i, j = 1, 2$.

Theorem 8.2. Suppose that

$$X_{t1} = \sum_{j=-\infty}^{\infty} \alpha_j W_{t-j,1}, \quad \{W_{t1}\} \sim \text{IID}(0, \sigma_1^2)$$

and

$$X_{t2} = \sum_{j=-\infty}^{\infty} \beta_j W_{t-j,2}, \quad \{W_{t2}\} \sim \text{IID}(0, \sigma_2^2)$$

where the two sequences $\{W_{t1}\}$ and $\{W_{t2}\}$ are independent, $\sum_j |\alpha_j| < \infty$ and $\sum_j |\beta_j| < \infty$. Then if $h \ge 0$,

$$\widehat{\rho}_{12}(h) \sim \text{AN}\left(0, n^{-1} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)\right).$$

If $h, k \ge 0$ and $h \ne k$, then

$$\begin{pmatrix} \widehat{\rho}_{12}(h) \\ \widehat{\rho}_{12}(k) \end{pmatrix} \sim \text{AN}\left\{\mathbf{0}, \begin{pmatrix} n^{-1}\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j) & n^{-1}\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h) \\ n^{-1}\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h) & n^{-1}\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j) \end{pmatrix}\right\}.$$

This theorem plays an important role in testing for correlation between two processes. If one of the two processes is white noise, then

$$\widehat{\rho}_{12}(h) \sim \text{AN}(0, n^{-1})$$

in which case it is straightforward to test the hypothesis that $\rho_{12}(h) = 0$. The rejection region is

$$|\widehat{\rho}_{12}(h)| > z_{\alpha/2}/\sqrt{n}$$

However, if neither process is white noise, then a value of $|\widehat{\rho}_{12}(h)|$ which is large relative to $n^{-1/2}$ does not necessarily indicate that $\rho_{12}(h)$ is different from zero. For example, suppose that $\{X_{t1}\}$ and $\{X_{t2}\}$ are two independent and identical AR(1) process with $\rho_{11}(h) = \rho_{22}(h) = 0.8^{|h|}$. Then the asymptotic variance of $\widehat{\rho}_{12}(h)$ is

$$n^{-1}\left\{1 + 2\sum_{k=1}^{\infty}(0.8)^{2k}\right\} = 4.556n^{-1}.$$

Thus, the rejection region is

$$|\widehat{\rho}_{12}(h)| > z_{\alpha/2}\sqrt{4.556}/\sqrt{n}$$

Thus, it would not be surprising to observe a value of $\widehat{\rho}_{12}(h)$ as large as $3n^{-1/2}$ even though $\{X_{t1}\}$ and $\{X_{t2}\}$ are independent. If no the other hand $\rho_{11}(h) = 0.8^{|h|}$ and $\rho_{22}(h) = (-0.8)^{|h|}$, then the asymptotic variance of $\widehat{\rho}_{12}(h)$ is $0.2195n^{-1}$ and an observed value of $3n^{-1/2}$ for $\widehat{\rho}_{12}(h)$ would be very unlikely.

8.2 Multivariate ARMA processes

(Multivariate ARMA(p, q) process). $\{\boldsymbol{X}_t, t = 0, \pm 1, \dots\}$ is an m-variate ARMA(p, q) process if $\{\boldsymbol{X}_t\}$ is a stationary solution of the difference equations,

$$\boldsymbol{X}_t - \Phi_1\boldsymbol{X}_{t-1} - \cdots - \Phi_p\boldsymbol{X}_{t-p} = \boldsymbol{W}_t + \Theta_1\boldsymbol{W}_{t-1} + \cdots + \Theta_q\boldsymbol{W}_{t-q},$$

where $\Phi_1, \dots, \Phi_p$, $\Theta_1, \dots, \Theta_q$ are real $m \times m$ matrix and $\{\boldsymbol{W}_t\} \sim \text{WN}(0, \boldsymbol{\Sigma})$.

Of course, we can write it in the more compact form

$$\Phi(B)\boldsymbol{X}_t = \Theta(B)\boldsymbol{W}_t, \quad \{W_t\} \sim \text{WN}(0, \boldsymbol{\Sigma}).$$

where $\Phi(z) = I - \Phi_1 z - \cdots - \Phi_p z^p$ and $\Theta(z) = I + \Theta_1 z + \cdots + \Theta_q z^q$ are matrix-valued polynomials, and I is the $m \times m$ identity matrix.

Example 8.2. (Multivariate AR(1) process). This process satisfies

$$\boldsymbol{X}_t = \Phi\boldsymbol{X}_{t-1} + \boldsymbol{W}_t, \quad \{W_t\} \sim \text{WN}(0, \boldsymbol{\Sigma}).$$

Same argument, we have

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \Phi^j \boldsymbol{W}_{t-j},$$

provided all the eigenvalues of Φ are less than 1 in absolute value, i.e., provided

$$\det(I - z\Phi) \neq 0 \text{ for all } z \in \mathbb{C} \text{ such that } |z| \leq 1.$$

Theorem 8.3. (Causality Criterion). If

$$\det \Phi(z) \neq 0 \text{ for all } z \in \mathbb{C} \text{ such that } |z| \leq 1.$$

then we have exactly one stationary solution,

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \Psi_j \boldsymbol{W}_{t-j}$$

where the matrices Ψ_j are determined uniquely by

$$\Psi(z) = \sum_{j=0}^{\infty} \Psi_j z^j = \Phi^{-1}(z)\Theta(z), \quad |z| \leq 1.$$

Theorem 8.4. (Invertibility Criterion). If

$$\det \Theta(z) \neq 0 \text{ for all } z \in \mathbb{C} \text{ such that } |z| \leq 1,$$

and $\{X_t\}$ is a stationary solution of the ARMA equation, then

$$\boldsymbol{W}_t = \sum_{j=0}^{\infty} \Pi_j \boldsymbol{X}_{t-j}$$

where the matrices Ψ_j are determined uniquely by

$$\Pi(z) = \sum_{j=0}^{\infty} \Pi_j z^j = \Theta^{-1}(z)\Phi(z), \quad |z| \leq 1.$$

9 State-Space Models

State-space model define a rich class of processes that have served well as time series models. It contains ARIMA and SARIMA models as special cases.

9.1 State-Space Models

In this section, we shall illustrate some of the many times series models which can be represented in linear state-space form. We start with a more general definition of noise.

- $\boldsymbol{W}_t \sim \text{UN}(\boldsymbol{\mu}_t, \boldsymbol{R}_t)$ denotes uncorrelated noise with mean vectors $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{R}_t$ (noting that they are allowed to change with t)
- $\boldsymbol{W}_t \sim \text{WN}(\boldsymbol{\mu}, \boldsymbol{R})$ denotes white noise as defined before.

The state-space model for w-dimensional time series $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$, consists of two equations

1. *Observation equation* takes form

$$\boldsymbol{Y}_t = \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \quad t = 1, 2, \ldots, \tag{9.1}$$

   where

   - $\boldsymbol{X}_t$ is v-dimensional *state vector* (stochastic in general)
   - $\boldsymbol{Y}_t$ is w-dimensional *output vector*
   - $\boldsymbol{G}_t$ is $w \times v$ *observation matrix* (deterministic)
   - $\boldsymbol{W}_t \sim \text{UN}(0, \boldsymbol{R}_t)$ is *observation noise*, which $\boldsymbol{W}_t$ & $\boldsymbol{R}_t$ having dimensions w & $w \times w$ (can be degenerate, i.e., $\det \boldsymbol{R}_t = 0$).

   Observation equation essentially says that we can observe linear combinations of variables in state vector, but only in presence of noise.

2. *State-transition equation* takes form

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 1, 2, \ldots, \tag{9.2}$$

   where

   - $\boldsymbol{F}_t$ is $v \times v$ state transition matrix (deterministic)
   - $\boldsymbol{V}_t \sim \text{UN}(0, \boldsymbol{Q}_t)$ is state transition noise, with $\boldsymbol{V}_t$ & $\boldsymbol{Q}_t$ having dimensions v & $v \times v$.

   Recursively, we have

$$\begin{aligned}
\boldsymbol{X}_t &= \boldsymbol{F}_{t-1}(\boldsymbol{F}_{t-2}\boldsymbol{X}_{t-2} + \boldsymbol{V}_{t-2}) + \boldsymbol{V}_{t-1} = \cdots \\
&= (\boldsymbol{F}_{t-1} \cdots \boldsymbol{F}_1)\boldsymbol{X}_1 + (\boldsymbol{F}_{t-1} \cdots \boldsymbol{F}_2)\boldsymbol{V}_1 + \cdots + \boldsymbol{F}_{t-1}\boldsymbol{V}_{t-2} + \boldsymbol{V}_{t-2} \\
&= f_t(\boldsymbol{X}_1, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{t-1})
\end{aligned}$$

And

$$\begin{aligned}\boldsymbol{Y}_t &= \boldsymbol{G}_t f_t(\boldsymbol{X}_1, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{t-1}) + \boldsymbol{W}_t \\ &= g_t(\boldsymbol{X}_1, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{t-1}, \boldsymbol{W}_t)\end{aligned}$$

Two additional assumptions

(a) $E\{\boldsymbol{W}_s \boldsymbol{V}_t'\} = 0$ for all s and t (here 0 is a $w \times v$ matrix of zeros; in words, every observation noise random variable is uncorrelated with every state-transition noise random variable).

(b) Assuming $E\{\boldsymbol{X}_t\} = 0$ for convenience, $E\{\boldsymbol{X}_1 \boldsymbol{W}_t'\} = 0$ and $\mathrm{E}\{\boldsymbol{X}_1 \boldsymbol{V}_t'\} = 0$ for all t (in words, initial state vector random variables are uncorrelated with observation and state-transition noise).

Example 9.1. To interpret the state-space model, we start with a very simple example: recall the classical decomposition model for time series Y_t, namely,

$$Y_t = m_t + s_t + W_t$$

where m_t is trend, s_t is periodic, W_t is a stationary process. m_t and s_t can be treated as deterministic or stochastic. Now, we consider the simple version, which is known as a *local level* model in which m_t is stochastic and $s_t = 0$:

$$\begin{aligned}Y_t &= X_t + W_t,\ \{W_t\} \sim \mathrm{WN}(0, \sigma_W^2) \\ X_{t+1} &= X_t + V_t,\ \{V_t\} \sim \mathrm{WN}(0, \sigma_V^2)\end{aligned}$$

where $E\{W_s V_t\} = 0$ for all s & t. This can be easily seen as a simple case of state space model:

$$\begin{aligned}\boldsymbol{Y}_t &= \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \\ Y_t &= X_t + W_t,\end{aligned}$$

where $\boldsymbol{Y}_t = Y_t$, $G_t = 1$, $\boldsymbol{X}_t = X_t$, $\boldsymbol{W}_t = W_t$ and $R_t = \sigma_W^2$. and

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \tag{9.3}$$

$$X_{t+1} = X_t + V_t \tag{9.4}$$

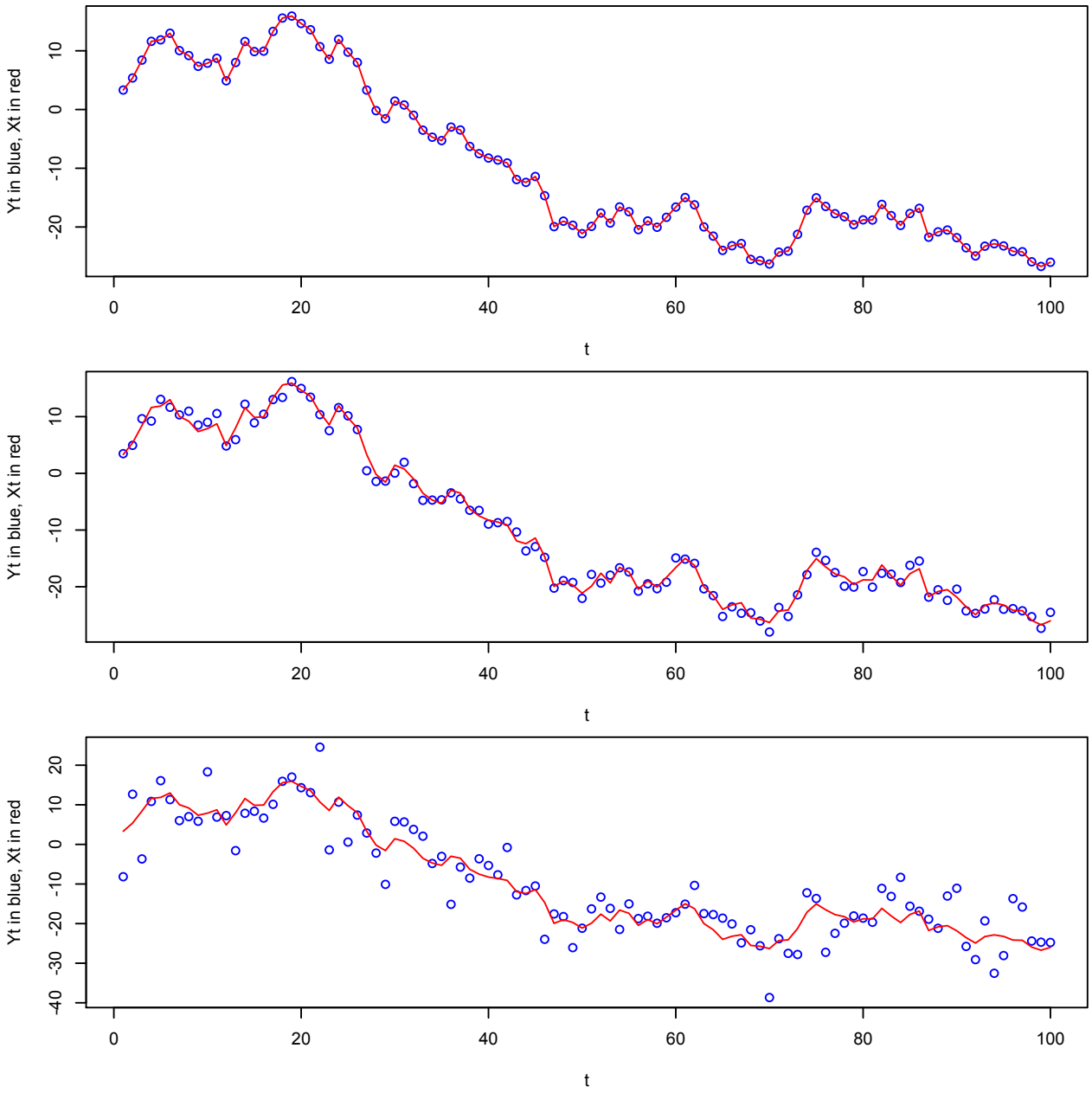
where $\boldsymbol{X}_{t+1} = X_{t+1}$, $\boldsymbol{F}_t = 1$, $\boldsymbol{V}_t = V_t$ and $Q_t = \sigma_V^2$.

To fully specify this state-space model, we need define initial state $\boldsymbol{X}_1 = X_t$ to be a random variable that is uncorrelated with W_t's and V_t's. In addition, we assume $\mathrm{E}(X_t) = m_1$ and $\mathrm{Var}(X_t) = P_1$. Thus, this model has 4 parameters $\sigma_W^2 = R_t$, $\sigma_V^2 = V_t$, $m - 1$ and P_1.

Since $X_{t+1} = X_t + V_t$, state variable X_t is a random walk starting form m_1 and then Y_t is the sequence of X_t which is corrupted by noise W_t.

```
par(mfrow=c(3,1));par(mar=c(4.5,4.5,.1,.1))
```

```
m1=1; P1=1; SigmaV=2;
X1=rnorm(1,m1,sqrt(P1));N=100
Vt=rnorm(N-1, 0, SigmaV);Xt=cumsum(c(X1,Vt));
for(SigmaW in c(0,1,5))
{
Yt=rnorm(N,Xt,SigmaW);
plot(1:N, Yt,col="blue",xlab="t", ylab="Yt in blue, Xt in red")
lines(1:N, Xt, col="red")
}
```



Example 9.2. AR(1) process as a State-Space Model: The AR(1) model has form

$$X_{t+1} = \phi X_t + W_{t+1}, \quad \{W_t\} \sim \text{WN}(0, \sigma^2)$$

This provided $X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}$ and consequently we have

$$X_1 = \sum_{j=0}^{\infty} \phi^j W_{1-j}.$$

Now, we can write it in the state-space model form

$$\begin{aligned} \boldsymbol{X}_{t+1} &= X_{t+1} \\ \boldsymbol{F}_t &= \phi \\ \boldsymbol{V}_t &= W_{t+1} \\ \boldsymbol{Y}_t &= X_t \\ \boldsymbol{G}_t &= 1 \\ \boldsymbol{W}_t &= 0. \end{aligned}$$

Example 9.3. How about ARMA(1,1)? let $\{X_t\}$ be the causal and invertible ARMA(1,1) process satisfying

$$X_t = \phi X_{t-1} + W_t + \theta W_{t-1}, \quad \{W_t\} \sim \text{WN}(0, \sigma^2).$$

To write it in state-space model form, we first observe that

$$\begin{aligned} X_t &= (1 - \phi B)^{-1}(1 + \theta B) W_t \\ &= (1 + \theta B)\{(1 - \phi B)^{-1} W_t\} \\ &= (1 + \theta B) Z_t \\ &= \begin{bmatrix} \theta & 1 \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Z_t \end{bmatrix} \end{aligned}$$

where $Z_t = (1 - \phi B)^{-1} W_t$; i.e.,

$$(1 - \phi B) Z_t = W_t \quad \text{or} \quad Z_t = \phi Z_{t-1} + W_t,$$

or

$$\begin{bmatrix} Z_t \\ Z_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Z_t \end{bmatrix} + \begin{bmatrix} 0 \\ W_{t+1} \end{bmatrix}$$

Further $Z_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}$. Thus, we have

$$\begin{aligned}
\boldsymbol{X}_{t+1} &= \begin{bmatrix} Z_t \\ Z_{t+1} \end{bmatrix} \\
\boldsymbol{F}_t &= \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \\
\boldsymbol{V}_t &= \begin{bmatrix} 0 \\ W_{t+1} \end{bmatrix} \\
\boldsymbol{Y}_t &= X_t \\
\boldsymbol{G}_t &= \begin{bmatrix} \theta & 1 \end{bmatrix} \\
\boldsymbol{W}_t &= 0
\end{aligned}$$

with

$$\boldsymbol{X}_1 = \begin{bmatrix} Z_0 \\ Z_1 \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{\infty} \phi^j W_{-j} \\ \sum_{j=0}^{\infty} \phi^j W_{1-j} \end{bmatrix}$$

Example 9.4. AR(p):

$$X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + W_t, \quad \{W_t\} \sim \text{WN}(0, \sigma^2).$$

Then we have

$$\boldsymbol{X}_t = (X_t)'$$

Four classical problems in State-Space Models: Given observations $Y_1, \ldots, Y_t$,

1. What is the best predictor of State X_t? (filtering).
2. What is the best predictor of State X_{t+1}? (forecasting).
3. What is the best predictor of State X_s for $s < t$? (smoothing).
4. What are the best estimates of model parameters? (estimation).