Time Series and Forecasting | 9

by **Patrick Boily**, inspired by **Rafal Kulik**

Many traditional statistical methods assume that observations are independently and identically distributed, which is unlikely to happen in real life. At best, this assumption may be sufficiently accurate to allow for some predictive power; at worst, it can lead to wildly inaccurate insights and predictions.

A time series is a sequence of values, measured at regular intervals over time. The motivation of time series analysis lies in the assumption that what happened in the past has an influence on what will happen in the future. Typically, time series are used for **trend analysis** and for **forecasting** future values when there are good reasons to suspect the existence of cycles in the data.* Generally speaking, the forecast horizon is the length of the prediction period: predictions at shorter horizons tend to be more reliable and accurate than predictions at longer horizons.

Ideally, the reporting periods used in time series analysis should be identical (e.g. daily, monthly, quarterly or yearly), the measurements should be taken over discrete (exclusive), consecutive periods, and the concepts and the measurement approach should be consistent over time. Detection of periodicity should be done by graphical representation of the data (and the frequency of data collection) using logic (e.g., is there an expectation of hourly, weekly, monthly, quarterly, and/or x-year cycles). More information is available in [2, 1, 5, 3, 4].

9.1 Introduction

Various time series analysis methods and tests are found in applications and in the literature, including:

- **auto-regressive** models (AR),
- **smoothing and filtering** models (such as moving averages (MA) and exponential smoothing (ES)),
- **detrending** models (such as ARMA, finite differences, etc.),
- **seasonal decomposition** models (such as X11, X12, X13, and ARIMA models), and
- **linear** and **non-linear forecasting** models (suc has Holt's Method, Winter's Method, GARCH models, etc.).

We start by providing examples and some of the basic concepts of the discipline.

* For instance, a time series analysis could be used to predict the number of passengers going through Canadian airports at various points in the future. Or an economist might be interested in forecasting the stock market, using time series analysis.

9.1.1 Simple Examples

White Noise Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance 1. Sometimes such a sequence is called a **white noise.** A sample white noise path consisting of 100 steps, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below.[1]

```
z = rnorm(100);
plot.ts(z)
```

Random Walk Let $\{Z_t\}$ be a sequence of i.i.d.[2] random variables with mean 0 and variance σ_Z^2. Define $X_t = \sum_{i=1}^{T} Z_i$, $t = 1, 2, \ldots$. A sample random walk of 100 steps, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below (see Figure 9.1 for the output).

```
z = rnorm(100);
x = cumsum(z);
plot.ts(x)
```

Model with Trend A linear or polynomial trend can sometimes be found in time series models. Consider, for instance, the time series

$$X_t = 1 + 2t + Z_t, \quad t = 1, 2, \ldots,$$

where $\{Z_t\}$ is a sequence of i.i.d. random variables. The linear trend is $m_t = 1 + 2t$. A 100-step realization of this model, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below (see Figure 9.1).

Linear trend

```
trend = 1+2*seq(1:100);
z = rnorm(100,0,10);
x = z+trend;
plot.ts(x)
```

For economics data, we may want to take into account an exponential inflation trend. If the interest rate r is assumed to be fixed, the nominal price X_t is actually the real (deflated) price P_t with respect to inflation:

$$X_t = P_t e^{rt}, \quad t = 1, 2, \ldots.$$

This phenomenon is illustrated in the quarterly earnings of Johnson & Johnson share (1960–80), as shown in Figure 9.1.

Exponential trend

```
require(stats);
x = JohnsonJohnson;
plot.ts(x)
```

[1] The output is shown in Figure 9.1. Note that the specific realization of the time series depends on the seed used to generate the pseudo-random numbers in R. In the absence of a `set.seed(...)` command, the realization will change after every call; with the command, the realization will be the same after every call. This comment should be kept in mind at all times when producing examples.

[2] Independent, identically distributed

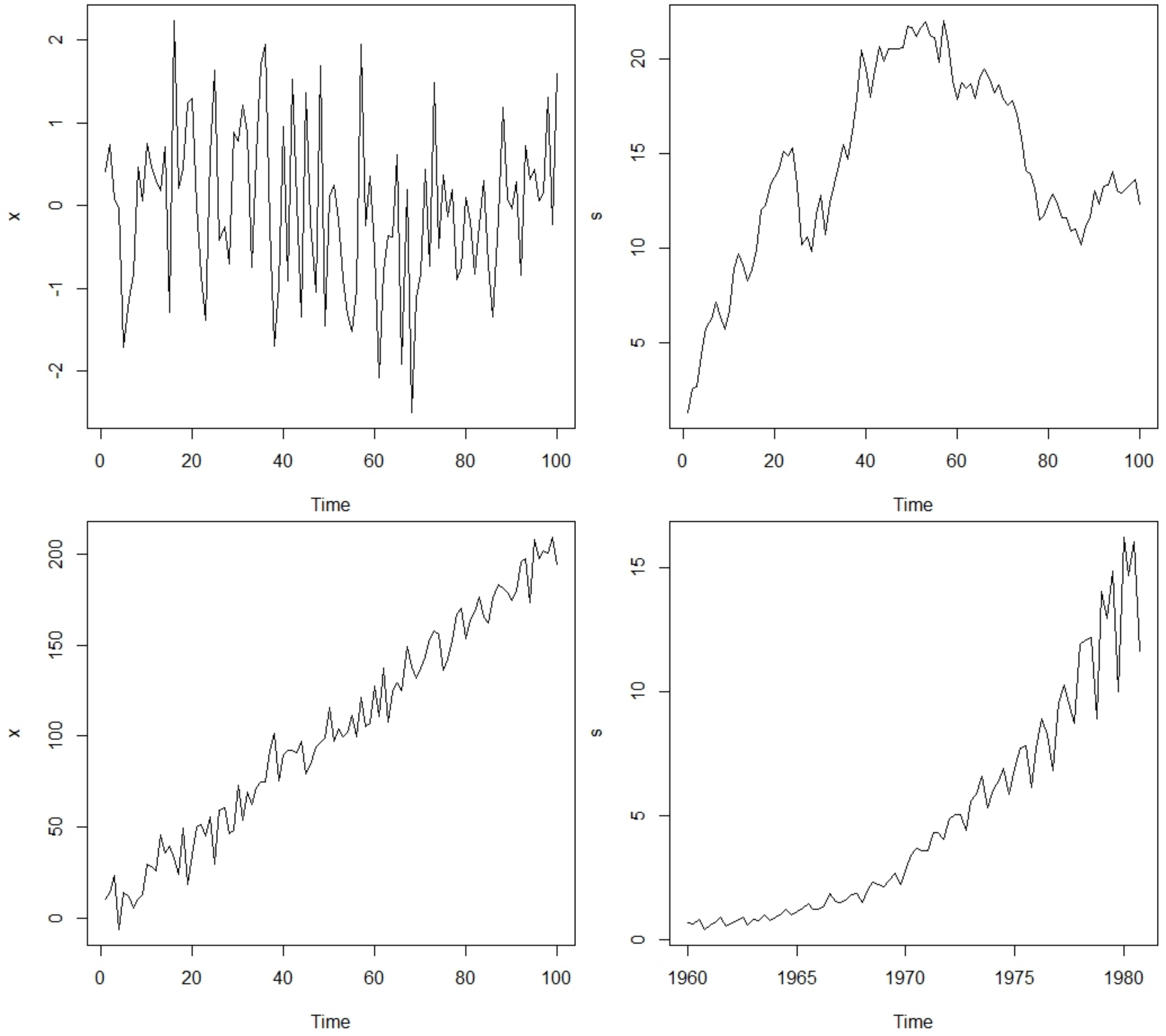


Figure 9.1: Simple time series: white noise (top left), random walk (top right), linear trend (bottom left), exponential trend (bottom right).

9.1.2 Pre-Processing

Component decomposition is central to time series analysis. Displaying the components of a time series is also helpful in understanding the data. Each of the components represents a category of patterns.

Generally speaking, there are three common components of time series: **trend**, **seasonality**, and **irregular**. We briefly discuss other potential components, but for the sake of simplicity, only the first two of these will be discussed in this chapter:

- the **trend** component describes the overall "changing direction" of the data, either increase or decrease or flat, which is a long-term effect and not necessarily linear;[3]
- the **seasonal** component reveals the seasonal effect on a series of data, such as that passengers in the airport will increase during summer vacation season;[4]
- The **irregular** (anomalous) component is a short-term effect, which can vary considerably from period to period, and includes measurement errors, unseasonal change, etc. – once the trend, seasonal,

[3] For example, in the linear trend time series model of Figure 9.1, the bottom left graph shows the trend going up, and so we expect X_t to increase with t.

[4] If the monthly deaths of lung disease in London, UK, shows peaks occurring at the beginning of each year, say, then we conclude that winter is a harsher time for such deaths than summer is, in general.

and cyclical effects are removed, we use the residual of the time series to identify the irregular contributions;

- **cyclical** components usually lasts at least two years – note that, in general, the exact length of an ongoing cycle cannot be predicted;[5]
- **other** components may include calendar effect (trading day, leap year, etc.), government policies, strike actions, exceptional events, inclement weather, etc.

Decomposition Models Traditionally, decomposition follows one of three models: **multiplicative**, **additive**, and **pseudo-additive**.

The **additive** approach assumes that:

1. the seasonal component S_t and the irregular component I_t are independent of the trend behaviour m_t;
2. the seasonal component S_t remains stable from year to year; and
3. the seasonal fluctuations are such that $\sum_{j=1}^{n} S_{t+j} = 0$.[6]

Mathematically, the model is expressed as:

$$X_t = m_t + S_t + I_t.$$

All components share the same dimensions and units. After seasonality adjustment, the seasonality adjusted series is:

$$SA_t = X_t - S_t = m_t + I_t.$$

The **multiplicative** approach assumes that:

1. the magnitude of the seasonal spikes/troughs increases when the trend increases (and vice versa);
2. the trend m_t has the same dimensions as the original series X_t, and the seasonal component S_t and the irregular component I_t are dimensionless and centered around 1;
3. the seasonal fluctuations are such that $\sum_{j=1}^{n} S_{t+j} = 0$, and
4. the original series X_t does not contain zero values.

Mathematically, the model is expressed as:

$$X_t = m_t \times S_t \times I_t.$$

All components share the same units. After seasonality adjustments, the seasonality adjusted series is

$$SA_t = \frac{X_t}{S_t} = m_t \times I_t$$

To transform a multiplicative model into an additive model, we could take a logarithmic transformation, such as:

$$\log X_t = \log m_t + \log S_t + \log I_t,$$

assuming that none of the component values are non-positive.

The **pseudo-additive** approach assumes that some of the values of the original series X_t are 0 (or very close to 0) and that:

[5] For example, the global financial crisis in 2008 lasted about 5 years. The difference between seasonal and cyclical is that the former displays the change over a fixed time period.

[6] For daily series, $n = 365$; for monthly series, $n = 12$; for quarterly series, $n = 4$, and so on.

1. the seasonal component S_t and the irregular component I_t are both dependent on the trend level m_t, but independent of each other, and
2. the trend m_t has the same dimensions as the original series X_t, and the seasonal component S_t and the irregular component I_t are dimensionless and centered around 1.

Mathematically, the model is expressed as:

$$X_t = m_t + m_t \times (S_t - 1) + m_t \times (I_t - 1) = m_t \times (S_t + I_t - 1).$$

All components share the same units. After seasonality adjustment, the seasonality adjusted series is:

$$SA_t = X_t - m_t \times (S_t - 1) - m_t \times (D_t - 1) = m_t \times I_t$$

The **choice** of a model is driven by data behaviour and assumptions. The analyst needs to plot the time series graph and test a range of models, selecting the one which stabilized the seasonal component.

The simplest way to determine whether to use multiplicative or additive decomposition, is by graphing the time series. If the size of the seasonal variation increases/decreases over time, multiplicative decomposition should be used (such as in the last chart of Figure 9.1).

On the other hand, if the seasonal variation seems to be constant over time, an additive model should be used (bottom left, Figure 9.1).[7]

Illustration We illustrate the process of decomposition with an arbitrary time series recording the monthly number of hours for a variable called CV, whose values are shown in the Figure 9.2.

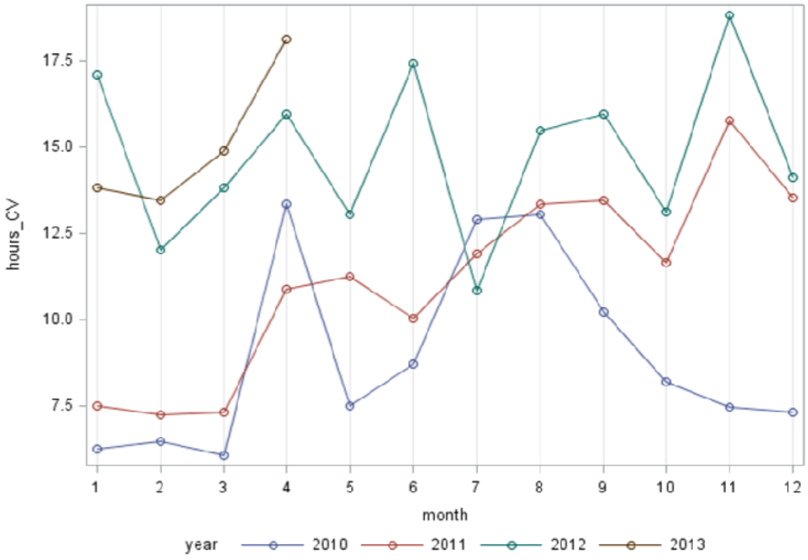


Figure 9.2: Time series; CV by year.

The continuous plot, Figure 9.3 ,shows that the size of the peaks and troughs does not seem to follow changing trends: the additive model is thus selected. The SAS procedure X12 agrees with that assessment, and further suggests no data transformation.

[7] A pseudo-additive model should be used when the data exhibits the characteristics of the multiplicative series, but with some X_t values near zero.

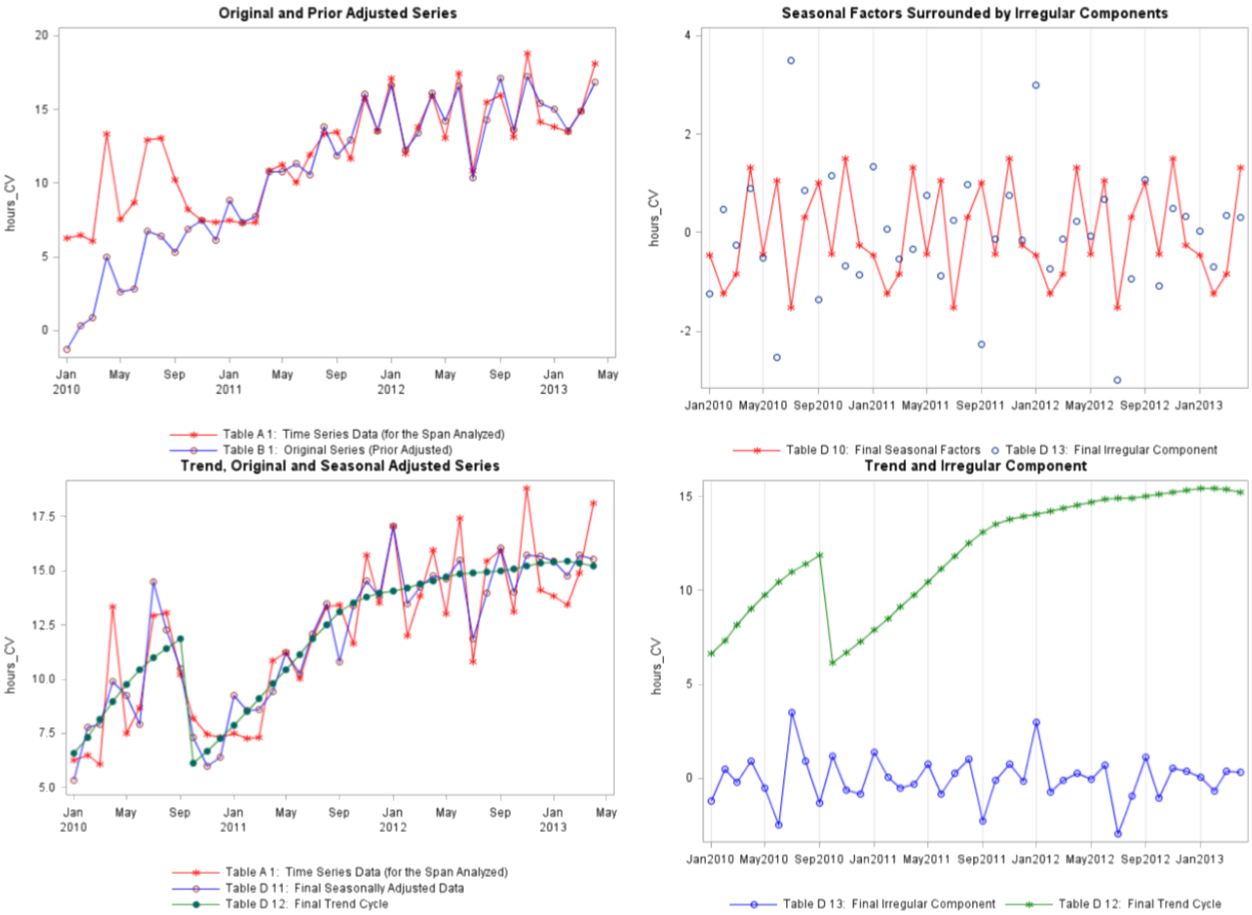
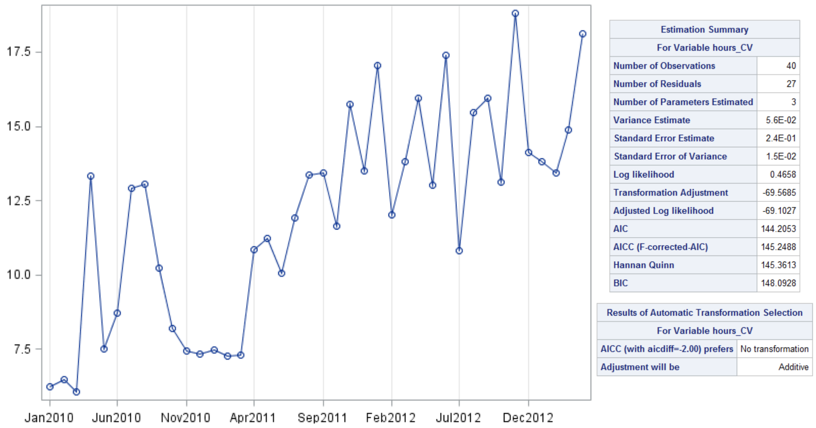


Figure 9.4: Diagnostic plots (top row) and adjusted plots (bottom row). Note that the analysis of a time series starts with estimation of the effects of festivals and trading days. These pre-calculated estimates are then used for prior adjustment of the series. The prior adjusted original series is subsequently analyzed using the seasonal adjustment.



Estimation Summary	
For Variable hours_CV	
Number of Observations	40
Number of Residuals	27
Number of Parameters Estimated	3
Variance Estimate	5.6E-02
Standard Error Estimate	2.4E-01
Standard Error of Variance	1.5E-02
Log likelihood	0.4658
Transformation Adjustment	-69.5685
Adjusted Log likelihood	-69.1027
AIC	144.2053
AICC (F-corrected-AIC)	145.2488
Hannan Quinn	145.3613
BIC	148.0928

Results of Automatic Transformation Selection	
For Variable hours_CV	
AICC (with aicdiff=-2.00) prefers	No transformation
Adjustment will be	Additive

Figure 9.3: Continuous CV; estimation summary.

The diagnostic plots are shown in Figure 9.4: the 2010 CV series is prior-adjusted from the beginning until OCT2010 after the detection of a level shift. The SI (Seasonal-Irregular) chart shows that there are more than one irregular component which exhibits volatility. The adjusted series is shown at the bottom of Figure 9.4 (the trend and irregular components are shown separately for readability).

Roll-Back In this chapter, however, we will focus on time series whose **structure** can be broken down into three additive components,

$$X_t = m_t + Y_t + S_t,$$

where:

- m_t is the trend;
- S_t is the seasonal component;
- Y_t is the **stationary** component (to be defined shortly).

In order to analyse time series, we first need to eliminate both the trend and the seasonal component.[8] We present a few ways to accomplish this, assuming that there is no seasonal component, i.e. $S_t \equiv 0$.

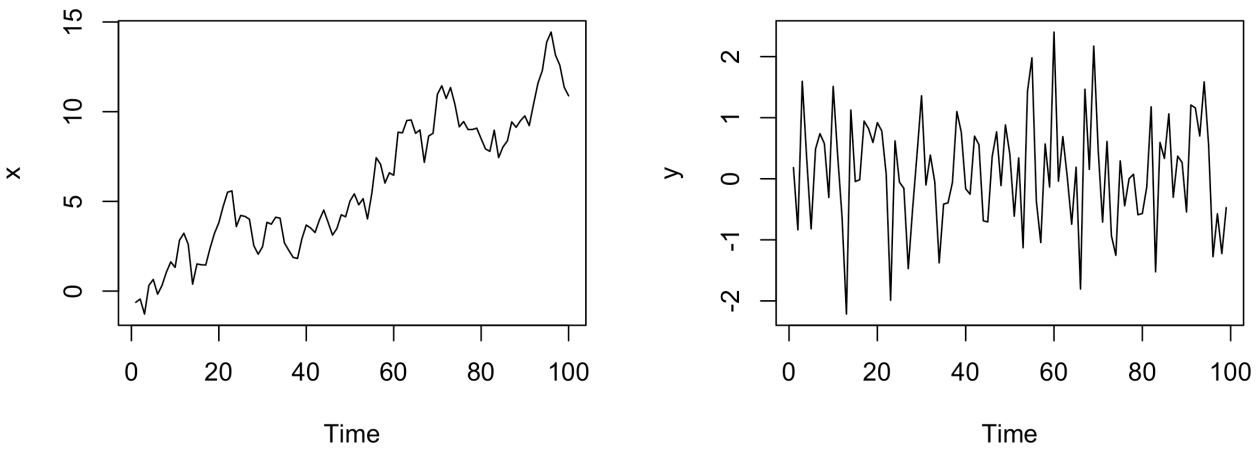
Differencing For the time series $\{X_t, t = 1, \dots, n\}$, we may calculate

$$\nabla X_t = X_t - X_{t-1}, \quad t = 2, \dots, n.$$

Depending on the nature of the trend in the original time series, the differenced time series may exhibit no trend.

Differencing a random walk

```
set.seed(1)
z = rnorm(100)
x = cumsum(z)
y = diff(x)
par(mfrow=c(1,2))
plot.ts(x)
plot.ts(y)
```

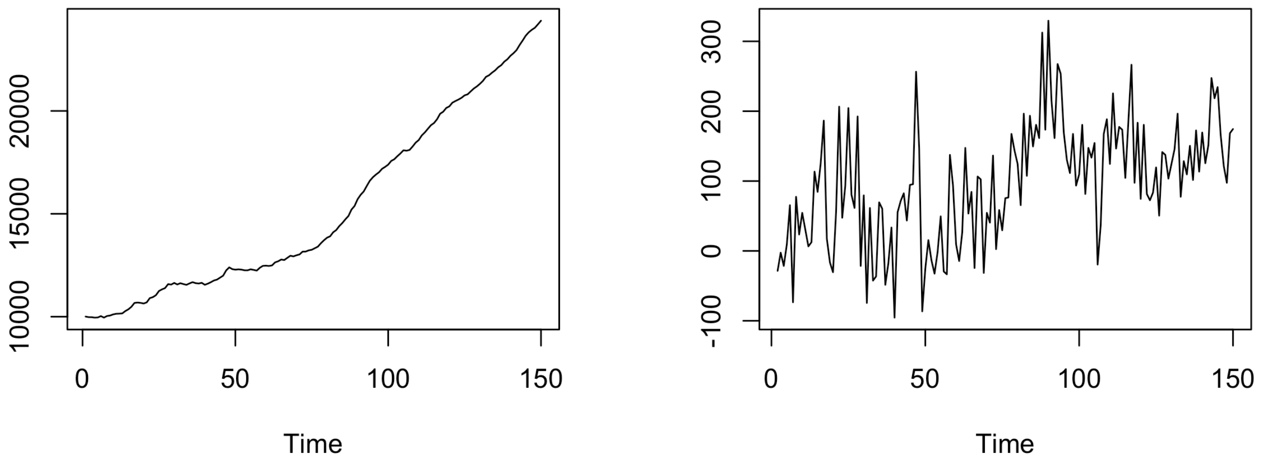


In a sense, differencing a time series is akin to **differentiating** a function $f : \mathbb{R} \to \mathbb{R}$; if the underlying trend is roughly linear, we expect the differenced time series to have **white noise** characteristics.[9]

But if the underlying trend is not linear, differencing only once might not detrend the original series, as can be seen below, where the trend has a clear (positive) slope.

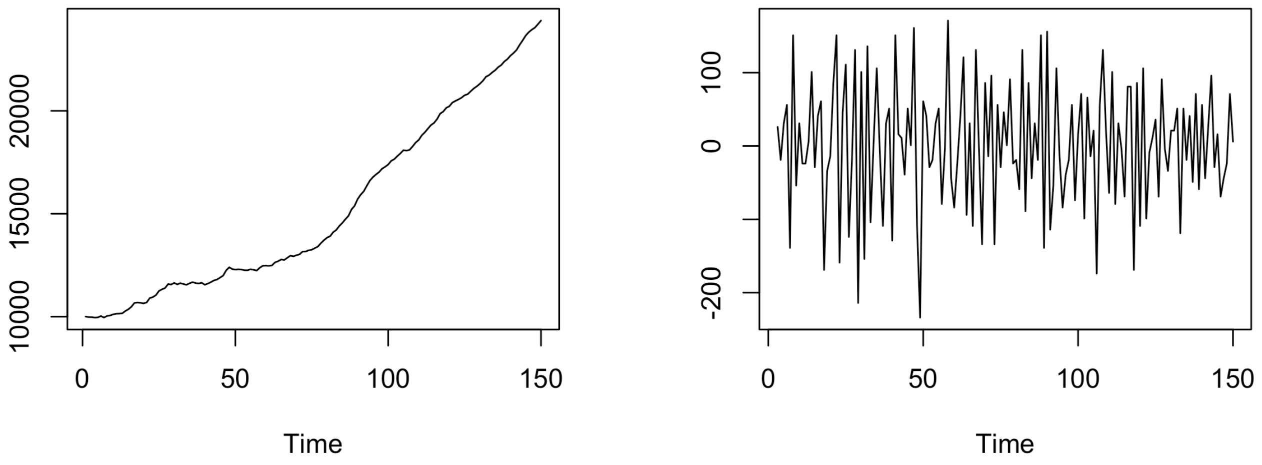
[8] Collectively, these are known as the **non-stationarities** of the time series.

[9] Which is to say, that the trend is horizontal.



Given that the original time series trend is concave up, differencing a second time could be a good strategy:[10]

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2},\ t = 3, \dots, n.$$



Polynomial Fitting When a linear trend is clearly visible ($m_t = a + bt$), then we can estimate the parameters a, b by minimizing

$$\sum_{t=1}^{n}(X_t - a - bt)^2.$$

This is a simple regression problem (see Chapter 8), where the independent variable is time t and the dependent variable is the time series itself. Consequently, the trend is estimated by

$$\widehat{m}_t = \widehat{a} + \widehat{b}t,$$

where $\widehat{a}$ and $\widehat{b}$ are the **least squares estimators** of a and b, respectively. In this case, the **detrended time series** is

$$\widehat{Y}_t = X_t - \widehat{m}_t,\ \ t = 1, \dots, n.$$

If the trend m_t would be better described by another polynomial, the process is similar; note however that it is not in general easy to justify using a non-linear polynomial trend.

[10] Since, by analogy, the second derivative of a quadratic function is the zero function.

As an example, consider the following time series, whose trend is linear by construction.

```
set.seed(11)
n=89; a=4; b=10;
Time=c(1:n);
X = a + b*Time + 20*rnorm(n)
```

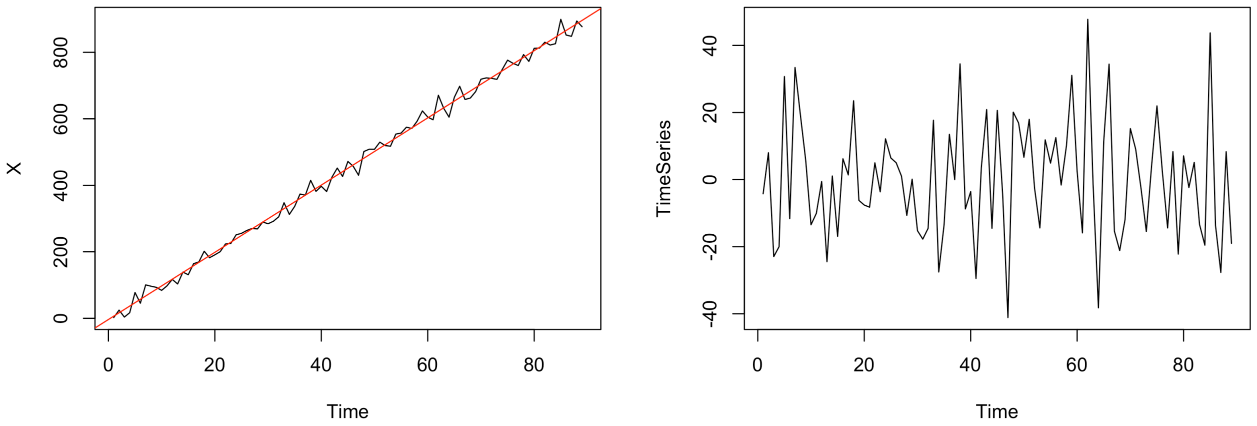
We can find the least squares estimates as follows:

```
estimation = lm(X~Time);
a.est = estimation$coefficients[1]; # Estimated intercept
b.est = estimation$coefficients[2]; # Estimated slope
c(a.est,b.est)
```

```
[1] -3.695528  10.10823
```

We plot the time series with its linear trend and compute the stationary part by removing the linear trend.

```
Fitted.Lin.Trend=a.est+b.est*Time;
TimeSeries=X-Fitted.Lin.Trend;

par(mfrow=c(1,2))
plot.ts(X)
abline(a=a.est,b=b.est, col="red", lwd=1);
plot.ts(TimeSeries);
```



Exponential Smoothing Let $\alpha \in (0, 1)$. We can estimate the trend *via*:

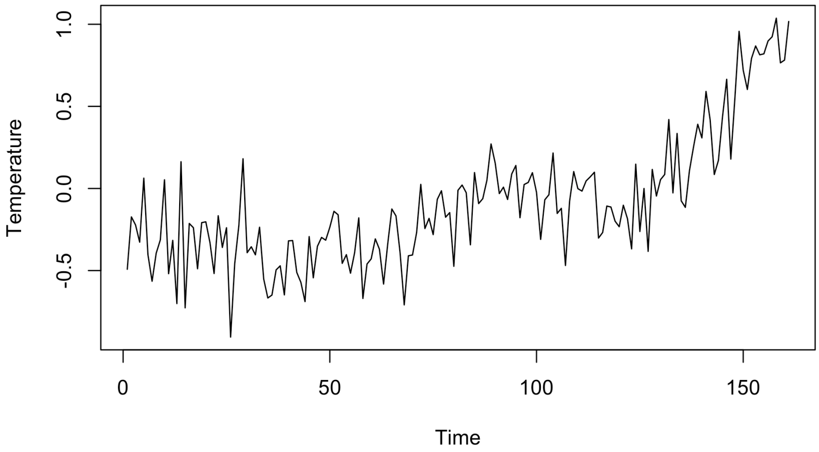
$$\widehat{m}_1 = X_1, \qquad \widehat{m}_t = \alpha X_t + (1-\alpha)\widehat{m}_{t-1}, \ \ t = 2, \dots, n.$$

In other words, at any time t, we assign weights α and $1-\alpha$ to the current observation and the preceding smoothed data. The detrended time series is

$$\widehat{Y}_t = X_t - \widehat{m}_t, \ \ t = 1, \dots, n.$$

Let us take a look at an example.

```
Temperature = c(-0.492, -0.173, -0.222, -0.327,  0.063,
         -0.403, -0.565, -0.394, -0.313,  0.053, -0.519,
         -0.316, -0.701,  0.163, -0.727, -0.213, -0.239,
         -0.489, -0.208, -0.203, -0.329, -0.518, -0.166,
         -0.359, -0.239, -0.905, -0.456, -0.223,  0.181,
         -0.391, -0.355, -0.404, -0.236, -0.551, -0.667,
         -0.649, -0.496, -0.471, -0.648, -0.319, -0.317,
         -0.511, -0.572, -0.689, -0.293, -0.544, -0.352,
         -0.298, -0.315, -0.236, -0.139, -0.160, -0.456,
         -0.403, -0.516, -0.391, -0.179, -0.670, -0.460,
         -0.429, -0.307, -0.370, -0.582, -0.339, -0.125,
         -0.167, -0.393, -0.709, -0.410, -0.405, -0.268,
          0.025, -0.244, -0.182, -0.281, -0.066, -0.014,
         -0.175, -0.147, -0.474, -0.011,  0.021, -0.026,
         -0.343,  0.097, -0.092, -0.062,  0.050,  0.271,
          0.155, -0.031,  0.008, -0.067,  0.088,  0.140,
         -0.178,  0.024,  0.037,  0.096, -0.024, -0.310,
         -0.069, -0.038,  0.216, -0.152, -0.121, -0.469,
         -0.078,  0.103, -0.001, -0.016,  0.046,  0.071,
          0.099, -0.302, -0.268, -0.107, -0.113, -0.199,
         -0.233, -0.102, -0.184, -0.368,  0.148, -0.262,
          0.000, -0.383,  0.116, -0.046,  0.054,  0.085,
          0.420, -0.027,  0.335, -0.075, -0.115,  0.110,
          0.256,  0.391,  0.308,  0.591,  0.418,  0.085,
          0.171,  0.438,  0.665,  0.179,  0.555,  0.957,
          0.720,  0.603,  0.792,  0.868,  0.814,  0.820,
          0.898,  0.924,  1.037,  0.765,  0.782,  1.017)
plot.ts(Temperature)
```

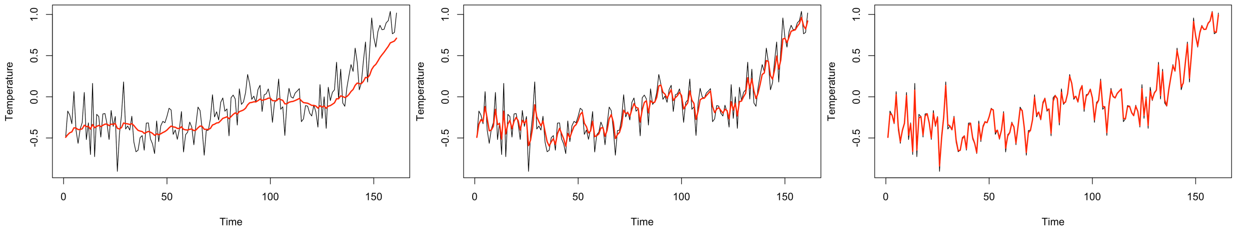


This times series is not stationary, so we need to remove its trend. Exponential smoothing is implemented in the following R function.

```
ExpSmooth <- function(x,alpha){
  # x: data
  # alpha: smoothing parameter
  n = length(x)
  Data = c(rep(0,n))
  Data[1] = x[1]
  for(i in 2:n){
    Data[i] = alpha*x[i] + (1-alpha)*Data[i-1]
  }
  out <- Data
}
```

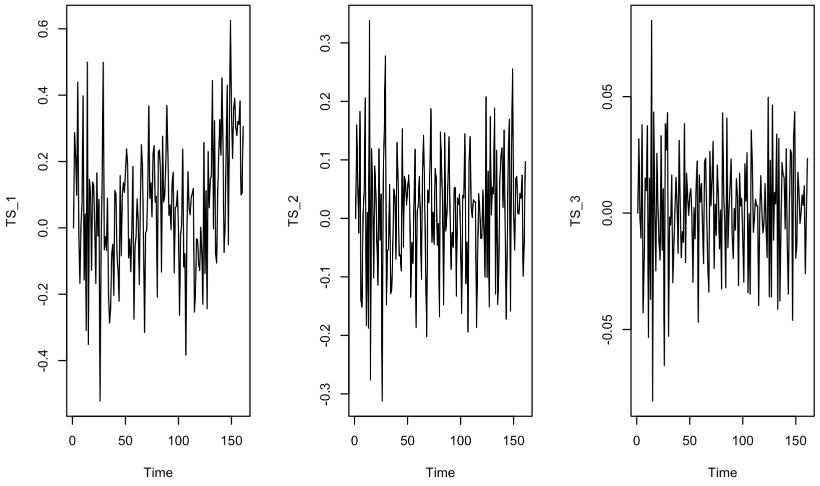
What effect does the parameter α have on the outcome? In general, the smaller α is, the smoother the trend is; here, we try $\alpha = 0.1, 0.5, 0.9$.

```
plot.ts(Temperature)
MySmoothedTS1 = ExpSmooth(Temperature,0.1)
points(MySmoothedTS1,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS2 = ExpSmooth(Temperature,0.5)
points(MySmoothedTS2,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS3 = ExpSmooth(Temperature,0.9)
points(MySmoothedTS3,col="red",type="l", lwd=2)
```



Using $\alpha = 0.1$ (left) indeed achieves the smoothest trend; $\alpha = 0.9$ (right) shows barely any smoothing. Detrending the series, we obtain:

```
TS_1 = Temperature-MySmoothedTS1
TS_2 = Temperature-MySmoothedTS2
TS_3 = Temperature-MySmoothedTS3

par(mfrow=c(1,3))
plot.ts(TS_1); plot.ts(TS_2); plot.ts(TS_3)
```



The outcome of the procedure is a time series (in this example, either TS_1, TS_2, or TS_3), which we hope can be treated as stationary.[11] Of course, different smoothing parameters α lead to different stationary time series – experience will inform the choice of α. The main thrust is that the exponential smoothing should not follow the data too closely while preserving the trend and the trend-removed dependence structure.

Moving Average Smoothing Another detrending approach requires us to pick a **window size** q (a positive integer). Then the trend is estimated *via*

$$\widehat{m}_t = (2q+1)^{-1}\sum_{j=-q}^{q} X_{t+j}, \quad q+1 \leq t \leq n-q.$$

The detrended time series is

$$\widehat{Y}_t = X_t - \widehat{m}_t, \quad t = q+1, \ldots, n-q.$$

Why does this method work? By assumption, we have $X_t = m_t + Y_t$. We assume further that $\mathrm{E}[Y_t] = 0$.[12] Then

$$(2q+1)^{-1}\sum_{j=-q}^{q} X_{t+j} = (2q+1)^{-1}\sum_{j=-q}^{q} m_{t+j} + (2q+1)^{-1}\sum_{j=-q}^{q} Y_{t+j}.$$

If the trend is linear ($m_t = a + bt$) Then

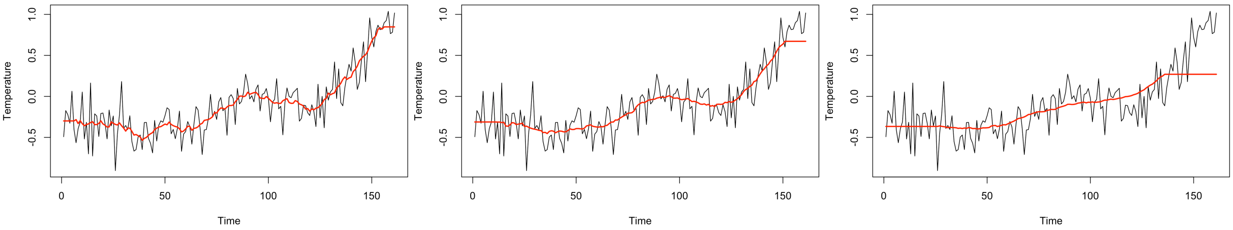
$$(2q+1)^{-1}\sum_{j=-q}^{q} m_{t+j} = (2q+1)^{-1}\sum_{j=-q}^{q} \{a + b(t+j)\} = a + bt.$$

We apply this approach to the Temperature data from the previous method, using $q = 5, 10, 25$.

```
MASmooth<-function(x,Q){
  # x: data set
  # Q: MA window size
  n = length(x)
  Smooth = c(rep(0,n))
  for(i in Q+1:(n-Q)){Smooth[i] = mean(x[(i-Q):(i+Q)])}
  for(i in 1:Q){Smooth[i] = Smooth[Q+1]}
  for(i in (n-Q+1):n){Smooth[i] = Smooth[(n-Q)]}
  out <- Smooth }

plot.ts(Temperature)
MySmoothedTS1 = MASmooth(Temperature,5)
points(MySmoothedTS1,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS2 = MASmooth(Temperature,10)
points(MySmoothedTS2,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS3 = MASmooth(Temperature,25)
points(MySmoothedTS3,col="red",type="l", lwd=2)
```

[11] These are the time series that will be analysed using the methods we discuss in this chapter.

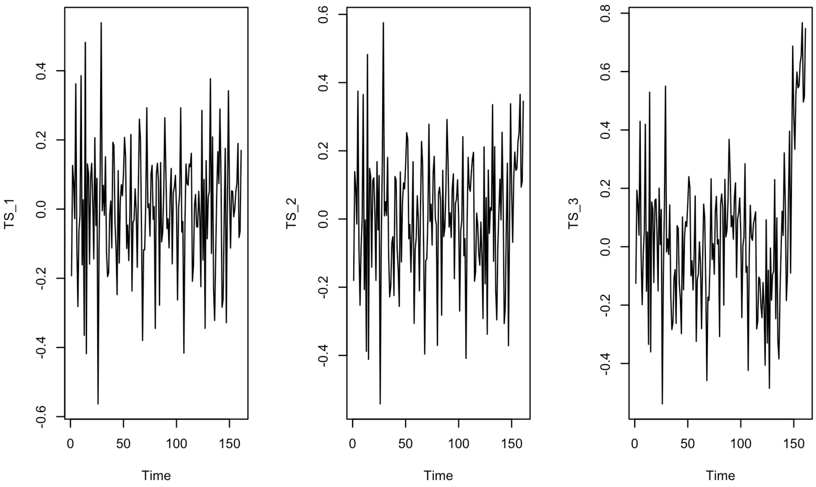
[12] If this is not the case, the non-zero mean can be always incorporated into the trend.



Note the flattening of the trend at the extremities.

The detrended time series are displayed below.

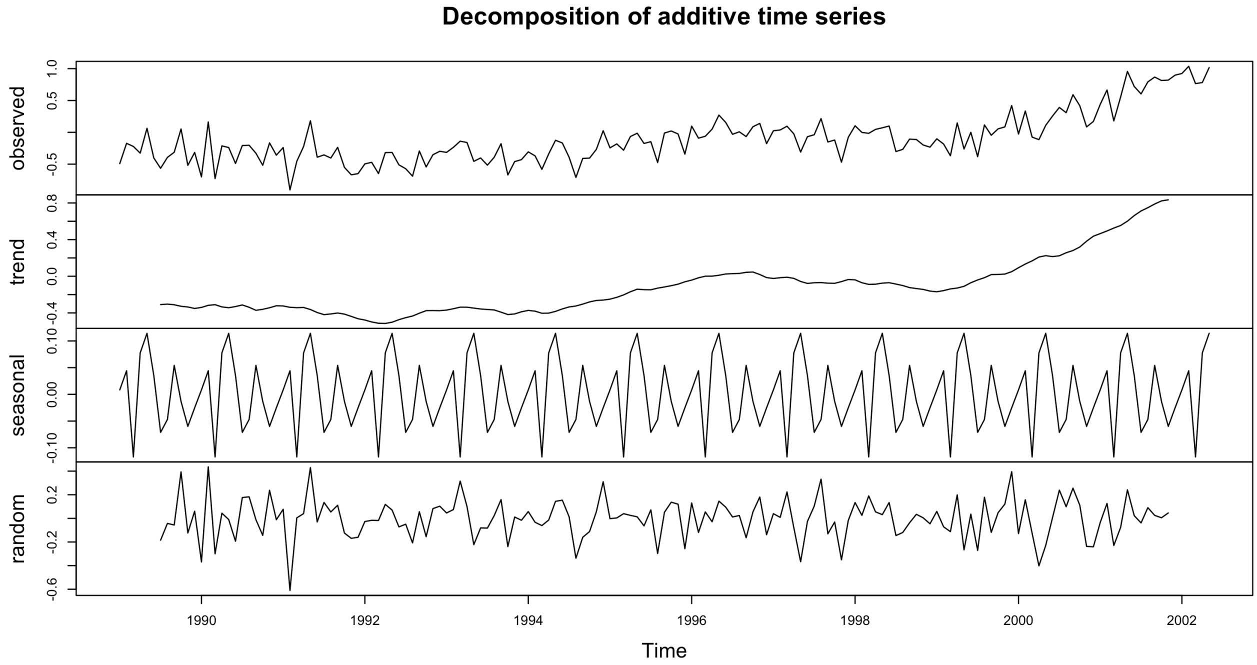
```
TS_1 = Temperature-MySmoothedTS1
TS_2 = Temperature-MySmoothedTS2
TS_3 = Temperature-MySmoothedTS2

par(mfrow=c(1,3))
plot.ts(TS_1)
plot.ts(TS_2)
plot.ts(TS_3)
```



Built-In Decomposer Most statistical analysis tools have built-in functions that can decompose time series according to some model.

For instance, if the temperature data is a monthly time series, starting in 1989 (and assuming that there is a seasonal component S_t), then `tseries`'s `decompose()` function can extract the stationary component (named `random` in this implementation) using an additive model and a moving average approach.

```
library(tseries)
Temperature.ts <- ts(Temperature, start=1989, freq=12)
plot(decompose(Temperature.ts))
```



The components can be isolated by calling:

- `decompose(Temperature.ts)$trend`,
- `decompose(Temperature.ts)$seasonal`, and
- `decompose(Temperature.ts)$random`.

9.1.3 Stationary Models, Autocovariance, and Autocorrelation

We now introduce the fundamental notions of time series analysis. [13]

Definitions and Properties

Let $\{X_t\}$ be a time series with $\mathrm{E}[X_t^2] < \infty$ for each t.

The expectation $\mu_X(t) = \mathrm{E}[X_t]$ is a function of t, the **mean function**. The **(auto)covariance function** of the time series is defined as

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

Note that $\gamma_X(t,t) = \mathrm{Var}(X_t)$.[14]

From our perspective, the most important properties of the covariance are that it is:

- **symmetric**
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X);$$
- **multilinear**
$$\mathrm{Cov}\left(\sum_{k=1}^{K} a_k X_k, \sum_{\ell=1}^{L} b_\ell Y_\ell\right) = \sum_{k=1}^{K}\sum_{\ell=1}^{L} a_k b_\ell \mathrm{Cov}(X_k, Y_\ell),$$
- and $\mathrm{Cov}(X,a) = 0$ for all $a \in \mathbb{R}$.

[13] Throughout this chapter, **time series** are sequences $\{X_t \mid t = t_0, \dots\}$ of random variables.

[14] When the context is clear, we will denote the mean function and the autocovariance function simply by μ and γ, respectively.

Cauchy's Inequality: if X, Y are r.v., then

$$\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

Proof: we may assume that $\mathrm{E}[X] = \mathrm{E}[Y] = 0$.[15] Define the function

$$g(t) = \mathrm{E}[(X+tY)^2] = t^2\mathrm{Var}(Y) + 2t\mathrm{Cov}(X,Y) + \mathrm{Var}(X), \quad t \in \mathbb{R}.$$

By construction, $g(t) \geq 0$ for all t. Since it is quadratic in t, it has at most one root, which is to say that its discriminant is non-positive. In other words

$$\Delta = 4(\mathrm{Cov}(X,Y))^2 - 4\mathrm{Var}(X)\mathrm{Var}(Y) \leq 0,$$

which implies the result. ■

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

- $\mu_X(t) \equiv \mu_X$, and
- $\gamma_X(t,s) = f_X(t-s)$ for some function f_X.

In particular, for such a time series, we must have $\sigma^2\ \{()\ X_t) \equiv \sigma^2_X$ and

$$\begin{aligned}\mathrm{Cov}(X_t, X_{t+1}) &= \gamma_X(t,t+1) = f_X(t+1-t) = f_X(1)\\ \mathrm{Cov}(X_{t+1}, X_{t+2}) &= \gamma_X(t+1,t+2) = f_X(t+2-(t+1)) = f_X(1)\\ &\vdots\\ \mathrm{Cov}(X_{t+k}, X_{t+k+1}) &= \gamma_X(t+k,t+k+1) = f_X(1), \quad k \geq 0.\end{aligned}$$

Lemma: assume that $\{X_t\}$ is a (weakly) stationary time series. Then the covariance function $\gamma_X(t,s)$ is a **non-negative definite function**.[16] **Proof:** we have

$$\begin{aligned}0 \leq \mathrm{Var}\left(\sum_{j=1}^n a_j X_j\right) &= \mathrm{Cov}\left(\sum_{i=1}^n a_j X_j, \sum_{j=1}^n a_j X_j\right)\\ &= \sum_{i=1}^n\sum_{j=1}^n a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n\sum_{j=1}^n a_i a_j \gamma_X(i,j).\end{aligned}$$

This completes the proof. ■

Under the same hypothesis as above, then $\gamma_X(t,s) = f_X(h)$, $h = t-s$; for simplicity's sake, we often write $\gamma_X(t-s)$ or $\gamma_X(h)$ for the covariance.[17]

The **(auto)correlation function** (ACF) of $\{X_t \mid t = 1,\ldots,n\}$ is given by:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{\mathrm{Cov}(X_1, X_{h+1})}{\mathrm{Var}(X_1)}.$$

Note that $\rho_X(0) = 1$.

Examples and Illustrations

White Noise Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance 1.

[15] Otherwise, set $X' = X - \mathrm{E}[X]$ and $Y' = Y - \mathrm{E}[Y]$ and work with X', Y' instead of X, Y. This can be done since the covariance and the variance are invariant under translation by a constant (see properties above).

[16] For all non-negative integers n and all real numbers $a_1, \ldots, a_n$ we have
$$\sum_{i=1}^n\sum_{j=1}^n a_i a_j \gamma_X(i,j) \geq 0.$$

[17] When the context is un-ambiguous.

Then $\mu_Z(t) = \mathrm{E}[Z_t] = 0$ and $\gamma_Z(t,t) = f_Z(0) = \mathrm{Var}(Z_t) = 1$ for all t, while $\gamma_Z(t,s) = f_Z(h) = 0$ for all $t \neq s \implies h \neq 0$. Since γ_Z only depends on $h = t - s$ and $\mu_Z \equiv 0$, $\{Z_t\}$ is (weakly) stationary.

Random Walk Let $\{Z_t\}$ be a sequence of i.i.d. random variables with mean 0 and variance σ_Z^2. Define $S_t = \sum_{i=1}^{T} Z_i$. Then $\mathrm{E}[S_t] = 0$, and

$$\begin{aligned}
\gamma_S(t,t+h) &= \mathrm{Cov}(S_t, S_{t+h}) = \mathrm{Cov}(S_t, S_t + Z_{t+1} + \cdots + Z_{t+h}) \\
&= \mathrm{Cov}(S_t, S_t) + \mathrm{Cov}(S_t, Z_{t+1} + \cdots + Z_{t+h}) \\
&= \mathrm{Cov}(S_t, S_t) + \mathrm{Cov}(Z_1 + \cdots + Z_t, Z_{t+1} + \cdots + Z_{t+h}) \\
&= \mathrm{Cov}(S_t, S_t) + \sum_{i=1}^{T}\sum_{j=1}^{h} \mathrm{Cov}(Z_i, Z_{t+j}) = \mathrm{Cov}(S_t, S_t) + 0 = \mathrm{Var}(S_t).
\end{aligned}$$

Since

$$\mathrm{Var}(S_t) = \mathrm{Var}(Z_1 + \cdots + Z_t) = \mathrm{Var}(Z_1) + \cdots \mathrm{Var}(Z_t) = \sigma_Z^2 + \cdots + \sigma_Z^2 = t\sigma_Z^2,$$

the autocovariance function depends on t (and not on $h = t - s$), and the sequence is not (weakly) stationary.

Model with Trend We revisti the model $X_t = 1 + 2t + Z_t$, $t = 1, 2, \ldots,$ where $\{Z_t\}$ is a sequence of i.i.d. random variables with mean $\mu_Z = \mathrm{E}[Z_t]$. Then

$$\mathrm{E}[X_t] = \mathrm{E}[1 + 2t + Z_t] = 1 + 2t + \mu_Z.$$

The mean function depends on t; the model is not (weakly) stationary.

"Multiplicative" Model Let $\{Z_t\}$ be i.i.d. with mean 0 and variance σ_Z^2. Define

$$X_t = Z_t Z_{t-1} Z_{t-2}, \quad t \geq 3.$$

Because $\mathrm{E}[Z_t] = 0$, we have

$$\sigma_Z^2 = \mathrm{Var}(Z_t) = \mathrm{E}[Z_t^2] - \mathrm{E}^2[Z_t] = \mathrm{E}[Z_t^2].$$

Since the Z_t are independent of one another, we have

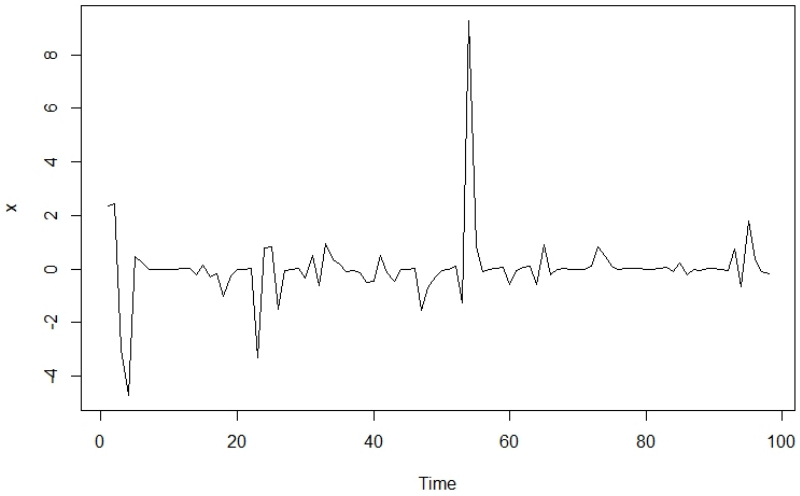
$$\begin{aligned}
\mathrm{E}[X_t] &= \mathrm{E}[Z_t Z_{t-1} Z_{t-2}] = \mathrm{E}[Z_t]\mathrm{E}[Z_{t-1}]\mathrm{E}[Z_{t-2}] = 0, \quad \text{and} \\
\mathrm{Var}(X_t) &= \mathrm{E}[X_t^2] = \mathrm{E}[Z_t^2 Z_{t-1}^2 Z_{t-2}^2] = \mathrm{E}[Z_t^2]\mathrm{E}[Z_{t-1}^2]\mathrm{E}[Z_{t-2}^2] = \sigma_Z^6
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(X_t, X_{t+1}) &= \mathrm{E}[X_t X_{t+1}] - \mathrm{E}[X_t]\mathrm{E}[X_{t+1}] \\
&= \mathrm{E}[\{Z_t Z_{t-1} Z_{t-2}\}\{Z_{t+1} Z_t Z_{t-1}\}] - 0 \\
&= \mathrm{E}[Z_{t+1}]\mathrm{E}[Z_t^2]\mathrm{E}[Z_{t-1}^2]\mathrm{E}[Z_{t-2}] = 0.
\end{aligned}$$

Similarly, we have $\mathrm{Cov}(X_t, X_s) = 0$ for $t \neq s$; the model is thus (weakly) stationary.

```
z=rnorm(100)
n=length(z)
zt=z[3:n]
zt1=z[2:(n-1)]
zt2=z[1:(n-2)]
x=zt*zt1*zt2
plot.ts(x)
```



MA(1) Let $\{Z_t\}$ be a sequence of independent random variables with $\mu_Z \equiv 0$ and variance $\sigma_Z^2 = \text{Var}(Z)$, and $\theta \in \mathbb{R}$. The MA(1) model is:

$$X_t = Z_t + \theta Z_{t-1}, \quad t \geq 2.$$

We see that $\text{E}[X_t] = \text{E}[Z_t + \theta Z_{t-1}] = \text{E}[Z_t] + \theta\text{E}[Z_{t-1}]$, and that

$$\begin{aligned}\text{Var}(X_t) &= \text{E}[X_t^2] = \text{E}[\{Z_t + \theta Z_{t-1}\}^2] \\ &= \text{E}[Z_t^2] + \theta^2\text{E}[Z_{t-1}^2] + 2\theta \underbrace{\text{E}[Z_t Z_{t-1}]}_{=0} = \sigma_Z^2 + \theta^2\sigma_Z^2 = \sigma_Z^2(1+\theta^2).\end{aligned}$$

Thus the autocovariance of MA(1) is

$$\gamma_X(t, t+h) = \gamma_X(h) = \begin{cases} \sigma_Z^2(1+\theta^2) & h = 0; \\ \sigma_Z^2\theta & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

Note that $\gamma_X(t, t+h) = \gamma_X(h)$ depends only on h and so a MA(1) time series is (weakly) stationary. Furthermore,

$$\rho_X(t, t+h) = \rho_X(h) = \begin{cases} 1 & h = 0; \\ \theta/(1+\theta^2) & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

The ACF then also only depends on h:

$$\rho_X(t, t+h) = \rho_X(h).$$

The set $\mathcal{T}_n$ of stationary time series of length n is a vector "subspace" over $\mathbb{R}$ of the set of all independent time series.[18]

[18] For a generous definition of subspace.

Indeed,

1. $\{0_t \mid t = 1, \ldots, n\} \in \mathcal{T}_n$;
2. if $\{X_t \mid t = 1, \ldots, n\} \in \mathcal{T}_n, \lambda \in \mathbb{R}$, then $\{\lambda X_t \mid t = 1, \ldots, n\} \in \mathcal{T}_n$;
3. if $\{X_t \mid t = 1, \ldots, n\}, \{Y_t \mid t = 1, \ldots, n\} \in \mathcal{T}_n$ are **independent** time series, then $\{W_t = X_t + Y_t \mid t = 1, \ldots, n\} \in \mathcal{T}_n$.

We only prove the third of these statements (the other two are left as exercises).

Let $\{X_t\}, \{Y_t\} \in \mathcal{T}_n$ be independent time series, with means μ_X, μ_Y and autocovariance functions γ_X and γ_Y, respectively. Set $W_t = X_t + Y_t$. Then

$$\mu_W(t) = \mathrm{E}[W_t] = \mathrm{E}[X_t + Y_t] = \mathrm{E}[X_t] + \mathrm{E}[Y_t] = \mu_X + \mu_Y (:= \mu_W)$$

and

$$\begin{aligned}
\gamma_W(t, t+h) &= \mathrm{E}[W_t W_{t+h}] - \mathrm{E}[W_t]\mathrm{E}[W_{t+h}] \\
&= \mathrm{E}[(X_t + Y_t)(X_{t+h} + Y_{t+h})] - \mu_W^2 \\
&= \mathrm{E}[X_t X_{t+h}] + \mathrm{E}[Y_t Y_{t+h}] + \underbrace{\mathrm{E}[X_t Y_{t+h}]}_{=\mathrm{E}[X_t]\mathrm{E}[Y_t]} + \underbrace{\mathrm{E}[Y_t X_{t+h}]}_{=\mathrm{E}[X_t]\mathrm{E}[Y_t]} - \mu_W^2 \\
&= \gamma_X(h) + \mu_X^2 + \gamma_Y(h) + \mu_Y^2 + \mu_X\mu_Y + \mu_X\mu_Y - (\mu_X + \mu_Y)^2 \\
&= \gamma_X(h) + \gamma_Y(h).
\end{aligned}$$

That is to say, $\{W_t\} \in \mathcal{T}_n$. ■

9.1.4 Partial Autocorrelation (PACF)

Let $\{X_t\} \in \mathcal{T}_n$ with $\mu_X = 0$. The **partial (auto)covariance** between X_t and X_{t+k} is the covariance between X_t and X_{t+k}, where we "condition out" the intermediate time series $X_{t+1}, \ldots, X_{t+k-1}$.

Assume that the random variables X_1 and X_3 from the stationary time series have the following relationship:

$$X_1 = \beta_{1,3} X_3 + Z,$$

where $\mu_Z = 0$, and Z is independent of both X_1, X_3. Then

$$\begin{aligned}
X_1 X_3 = \beta_{1,3} X_3^2 + Z X_3 &\implies \mathrm{E}[X_1 X_3] = \beta_{1,3}\mathrm{E}[X_3^2] + \mathrm{E}[Z X_3] \\
\implies \gamma_X(2) = \beta_{1,3}\gamma_X(0) + \mathrm{E}[Z]\mathrm{E}[X_3] &\implies \gamma_X(2) = \beta_{1,3}\gamma_X(0),
\end{aligned}$$

and so

$$\beta_{1,3} = \frac{\gamma_X(2)}{\gamma_X(0)} = \rho_X(2).$$

If $Z \sim \mathcal{N}(0, \sigma_Z^2)$, we recognize $\beta_{1,3}$ as the **OLS regression parameter** when regressing X_1 against X_3.[19] Similarly, if we further assume that

$$X_2 = \beta_{2,3} X_3 + V,$$

[19] Strictly speaking, if Z is not normal, the OLS qualifier does not apply but the rest of the argument still works.

where $V \sim \mathcal{N}(0, \sigma_V^2)$ is independent of both X_2, X_3, then the OLS regression parameter when regressing X_2 against X_3 is

$$\beta_{2,3} = \frac{\gamma_X(1)}{\gamma_X(0)} = \rho_X(1).$$

The **partial (auto)correlation** (PACF) between X_1 and X_2 is the correlation between X_1 and X_2, removing the effect of X_3:

$$\rho_{1,2;3} = \text{Corr}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3).$$

Hence,

$$\rho_{1,2;3} = \frac{\text{Cov}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3)}{\sqrt{\text{Var}(X_1 - \beta_{1,3}X_3)}\sqrt{\text{Var}(X_2 - \beta_{2,3}X_3)}}.$$

But we have

$$\begin{aligned}
&\text{Cov}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3)\\
&= \text{Cov}(X_1, X_2) + \text{Cov}(\beta_{1,3}X_3, \beta_{2,3}X_3) - \text{Cov}(X_1, \beta_{2,3}X_3) - \text{Cov}(\beta_{1,3}X_3, X_2)\\
&= \gamma_X(1) + \beta_{1,3}\beta_{2,3}\text{Cov}(X_3, X_3) - \beta_{2,3}\text{Cov}(X_1, X_3) - \beta_{1,3}\text{Cov}(X_3, X_2)\\
&= \gamma_X(1) + \beta_{1,3}\beta_{2,3}\gamma_X(0) - \beta_{2,3}\gamma_X(2) - \beta_{1,3}\gamma_X(1)\\
&= \gamma_X(1) + \rho_X(2)\rho_X(1)\gamma_X(0) - \rho_X(1)\gamma_X(2) - \rho_X(2)\gamma_X(1)\\
&= \gamma_X(1) + \rho_X(2)\gamma_X(1) - \rho_X(1)\gamma_X(2) - \rho_X(2)\gamma_X(1)\\
&= \gamma_X(1) + \rho_X(2)\gamma_X(1) - \tfrac{\gamma_X(1)}{\gamma_X(0)}\gamma_X(2) - \rho_X(2)\gamma_X(1)\\
&= \gamma_X(1) + [\rho_X(2)\gamma_X(1) - \gamma_X(1)\rho_X(2)] - \rho_X(2)\gamma_X(1) = \gamma_X(1)(1 - \rho_X(2)).
\end{aligned}$$

We also have:

$$\text{Var}(X_1 - \beta_{1,3}X_3) = \gamma_X(0)\left(1 - \rho_X^2(2)\right) \quad \text{and} \quad \text{Var}(X_2 - \beta_{2,3}X_3) = \gamma_X(0)\left(1 - \rho_X^2(1)\right).$$

Thus, the partial correlation is

$$\begin{aligned}
\rho_{1,2;3} &= \frac{\gamma_X(1)(1 - \rho_X(2))}{\gamma_X(0)\sqrt{\left(1 - \rho_X^2(2)\right)\left(1 - \rho_X^2(1)\right)}} = \frac{\rho_X(1) - \rho_X(1)\rho_X(2)}{\sqrt{\left(1 - \rho_X^2(2)\right)\left(1 - \rho_X^2(1)\right)}}\\
&= \frac{\text{Corr}(X_1, X_2) - \text{Corr}(X_2, X_3)\cdot\text{Corr}(X_1, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_3)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}}.
\end{aligned}$$

Note: $\gamma_X(1)$, $\text{Cov}(X_1, X_2)$, and $\text{Cov}(X_2, X_3)$ are interchangeable because the time series $\{X_t\}$ is stationary; thus we have $\text{Corr}(X_1, X_2) = \text{Corr}(X_2, X_3)$.

Similarly, the partial (auto)correlation between X_1 and X_3 is the correlation between X_1 and X_3, removing the effect of X_2:

$$\rho_{1,3;2} = \frac{\text{Corr}(X_1, X_3) - \text{Corr}(X_1, X_2)\cdot\text{Corr}(X_2, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_2)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}}.$$

The PACF Given a time series $\{X_t\}$, the partial autocorrelation at lag h, denoted $\alpha_X(h)$,[20] is the autocorrelation between X_t and X_{t+h}, removing the linear dependence of X_t on $X_{t+1}, \ldots, X_{t+h-1}$; the function α_X is called the **partial autocorrelation function** (PACF).

[20] Or $\alpha(h)$ if the context is clear.

Note that:

1. $\alpha(1) = \rho_X(1)$,
2. $\alpha(2) = \rho_{1,3;2}$,
3. $\alpha(3) = \rho_{1,4;2,3}$,
4. and so on.

A non-negligible aspect of the discipline involves computing the PACF for different models; we anticipate the task by providing some some calculations for a special case: the MA(1) model.

MA(1) Let $\{Z_t\}$ be a sequence of independent random variables with $\mu_Z \equiv 0$ and variance $\sigma_Z^2 = \text{Var}(Z)$, and $\theta \in \mathbb{R}$. The MA(1) model is $X_t = Z_t + \theta Z_{t-1}$, $t \geq 2$. We have seen that

$$\rho_X(h) = \begin{cases} 1 & h = 0; \\ \theta/(1+\theta^2) & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

Thus,

$$\begin{aligned} \alpha(2) &= \frac{\text{Corr}(X_1, X_3) - \text{Corr}(X_1, X_2) \cdot \text{Corr}(X_2, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_2)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}} \\ &= \frac{\rho_X(2) - \rho_X^2(1)}{\sqrt{1 - \rho_X^2(1)}\sqrt{1 - \rho_X^2(1)}} \\ &= \frac{0 - \dfrac{\theta^2}{(1+\theta^2)^2}}{1 - \dfrac{\theta^2}{(1+\theta^2)^2}} = \frac{-\theta^2}{1 + \theta^2 + \theta^4}. \end{aligned}$$

9.2 Estimating Model Parameters

In practice, we typically work with one of the time series' **realizations**, that is to say, the true $\mu(\cdot)$, $\gamma(\cdot)$ and $\alpha(\cdot)$ are not available to us.

9.2.1 Sample Statistics

As is usually the case, in statistical analysis, we can use the data at our disposal in order to estimate the model's parameters. As always, assume that $\{X_t\} \in \mathcal{T}_n$ is stationary.

Sample Mean The mean $\mu = \mu_X \equiv \text{E}[X_t]$ can be estimated by the **sample mean**:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Sample Variance The variance $\sigma_X^2 \equiv \mathrm{Var}(X_t) = \mathrm{E}[(X_t - \mu)^2]$ can be estimated by the **sample variance**:

$$\widehat{\sigma}_X^2 = \widehat{\gamma}_X(0) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Sample (Auto)Covariance The covariance $\gamma_X(h) = \mathrm{E}[(X_t - \mu)(X_{t+h} - \mu)]$ (ACVF) can be estimated by the **sample (auto)covariance**:

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

Sample (Auto)Correlation The (auto)correlation $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$ is estimated by the **sample autocorrelation** (**sample ACF**):

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

Sample PACF The PACF is estimated by the **sample PACF**; for instance, since

$$\alpha(2) = \frac{\rho_X(2) - \rho_X^2(1)}{\sqrt{1-\rho_X^2(1)}\sqrt{1-\rho_X^2(1)}} = \frac{\rho_X(2) - \rho_X^2(1)}{1-\rho_X^2(1)},$$

then

$$\widehat{\alpha}(2) = \frac{\widehat{\rho}_X(2) - \widehat{\rho}_X^2(1)}{1-\widehat{\rho}_X^2(1)}.$$

9.2.2 Examples

White Noise Recall that white noise $\{Z_t\}$ is a sequence of independent random variables with mean 0 and variance 1. Then $\gamma_X(0) = \rho_X(0) = 1$ and $\gamma_X(h) = \rho_X(h) = 0$ for $h \neq 0$.

We prepare a realization of the white noise time series.

```
set.seed(1)
z = rnorm(100)
n = length(z)
(muz = mean(z))
gamma0 = sum((z-muz)^2)/(n-1)
var(z)
```

```
[1] 0.1088874
[1] 0.8067621
```

We see that the sample mean and the sample variance are near 0 and 1, respectively. We can exhibit the sample ACF using the `acf()` function.

```
zt = z[2:n]; zt1 = z[1:(n-1)]
(corr = acf(z))
```

```
autocorrelations of series ‘z’, by lag

     0      1      2      3      4      5      6      7      8 
 1.000 -0.004 -0.027 -0.107 -0.113 -0.093 -0.125  0.065  0.043 
     9     10     11     12     13     14     15     16     17 
 0.026  0.025 -0.032 -0.042  0.053 -0.038 -0.022 -0.140  0.063 
    18     19     20 
-0.023 -0.084 -0.112 
```

For instance, we can extract $\widehat{\rho}(1)$ using the following call:

```
corr$acf[2]
```

```
[1] -0.003651251
```

But we can also compute it directly:

```
gamma1 = sum((zt1-muz)*(zt-muz))/(n-1)
(rho1 = gamma1/gamma0)
```

```
[1] -0.003651251
```

The sample PACF can be obtained *via* the `pacf()` function.

```
(partial.corr = pacf(z))
```

```
Partial autocorrelations of series ‘z’, by lag

     1      2      3      4      5      6      7      8      9 
-0.004 -0.027 -0.108 -0.116 -0.105 -0.153  0.023 -0.002 -0.025 
    10     11     12     13     14     15     16     17     18 
-0.005 -0.046 -0.052  0.069 -0.042 -0.039 -0.157  0.035 -0.053 
    19     20 
-0.121 -0.190 
```

For instance, we can extract $\widehat{\alpha}(2)$ using the following call:

```
partial.corr$acf[2]
```

```
[1] -0.02703468
```

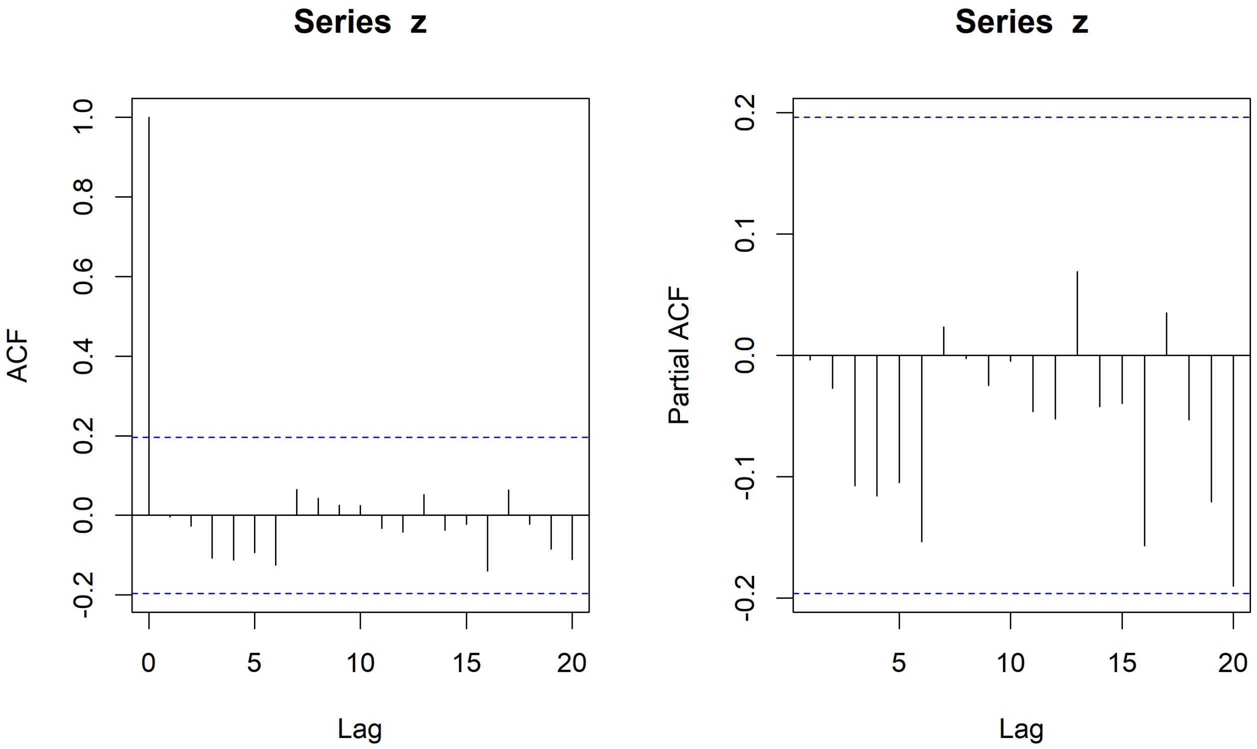
But we can also compute it directly:

```
(alpha2 = (corr$acf[3]-(corr$acf[2])^2)/(1-(corr$acf[2])^2))
```

```
[1] -0.02703468
```

Finally, we plot the sample ACF and sample PACF of the white noise time series against the lag h.[21]

```
par(mfrow=c(1,2))
acf(z); pacf(z)
```



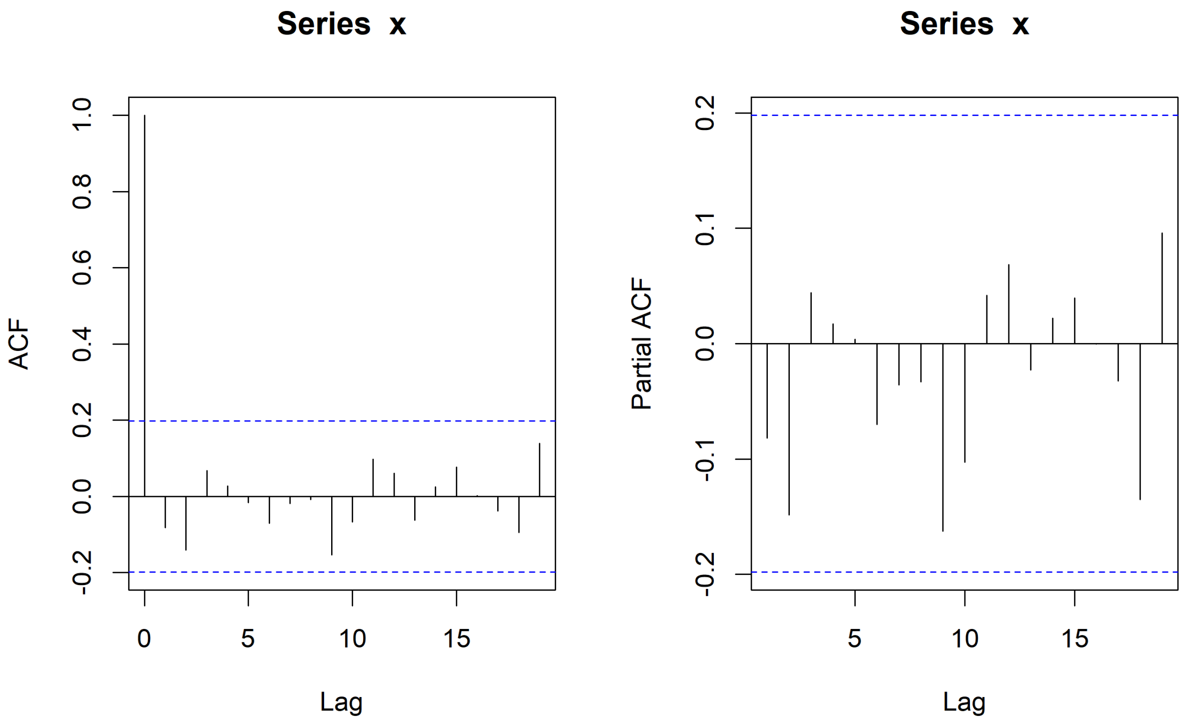
"Multiplicative" Model Let $\{Z_t\}$ be i.i.d. with mean 0 and variance σ_Z^2. Define

$$X_t = Z_t Z_{t-1} Z_{t-2}, \quad t \geq 3.$$

We prepare a realization of this time series, assuming that $Z_t \sim \mathcal{N}(0,1)$, and display its sample ACF and sample PACF.

```
set.seed(2)
z = rnorm(100)
n = length(z)
zt = z[3:n]; zt1 = z[2:(n-1)]; zt2 = z[1:(n-2)];
x = zt*zt1*zt2
par(mfrow=c(1,2))
acf(x)
pacf(x)
```

[21] The dotted blue lines in the ACF and PACF chart indicate the thresholds beyond which the recorded values can be seen as statistically different rom zero. These lines are located at a height of $\pm\frac{1.96}{\sqrt{n}}$ (see Section 9.6 for an explanation).



Are the results fundamentally different than those of the white noise time series?[22]

MA(1) Recall MA(1) model

$$X_t = Z_t + \theta Z_{t-1},$$

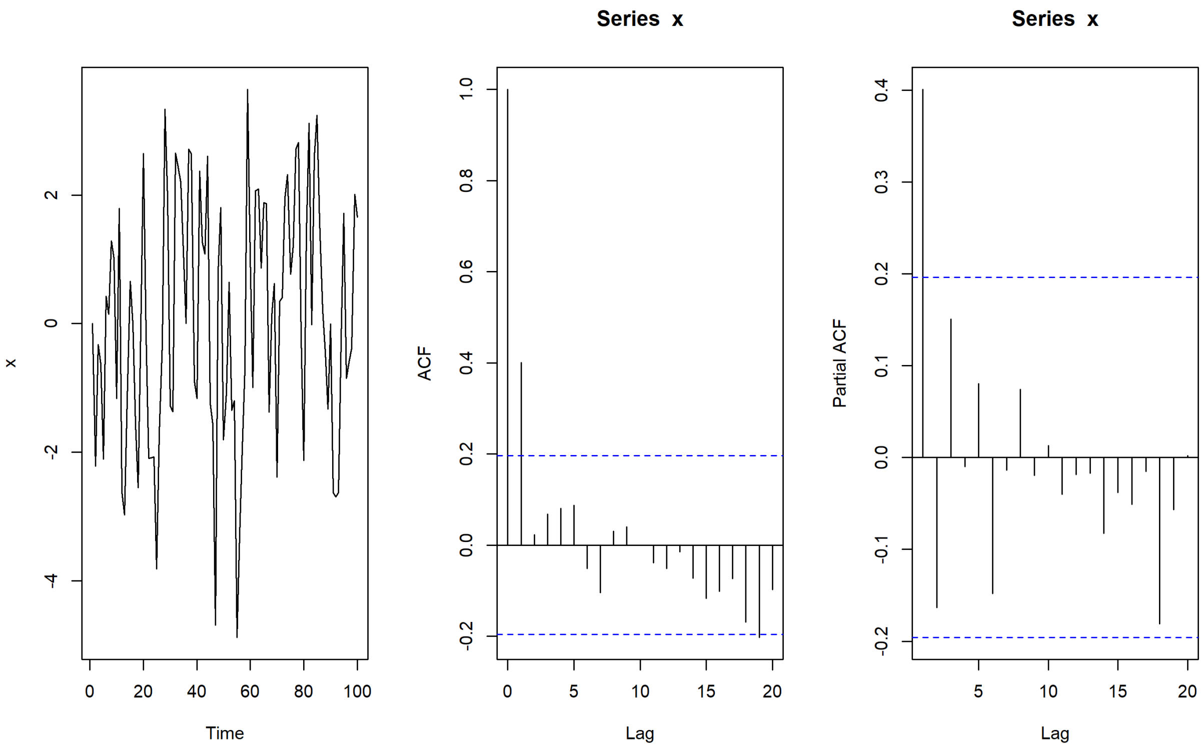
We have derived the ACF of this model previously: $\rho_X(0) = 1$, $rho_X(1) = \theta/(1+\theta^2)$, and $rho_X(h) = 0$ for $h > 1$. We prepare a realization of MA(1) as follows:

```
set.seed(3)
z = rnorm(100,0,1)
n = length(z)
x = rep(0,n)
theta = 2
for(i in 2:n){
    x[i] = z[i] + theta*z[i-1]
}
```

Theoretically, the only non-zero values of the ACF are at $h = 0$ and $h = 1$; is that also going to be the case in the sample ACF?

```
par(mfrow=c(1,3))
plot.ts(x)
corr = acf(x)
pacf(x)
```

[22] Keeping in mind that we are working with (potentially) different realizations of the respective time series.



It is not exactly so, obviously, but $\widehat{\rho}_X(0)$ and $\widehat{\rho}_X(1)$ are substantially larger than the remaining $\widehat{\rho}_X(h)$.

The theoretical value of $\rho_X(1)$ can be computed exactly:

```
(rho1 = theta/(1+theta^2))
```

```
[1] 0.4
```

How does that compare to the sample estimate $\widehat{\rho}_X(1)$?

```
corr[1]
```

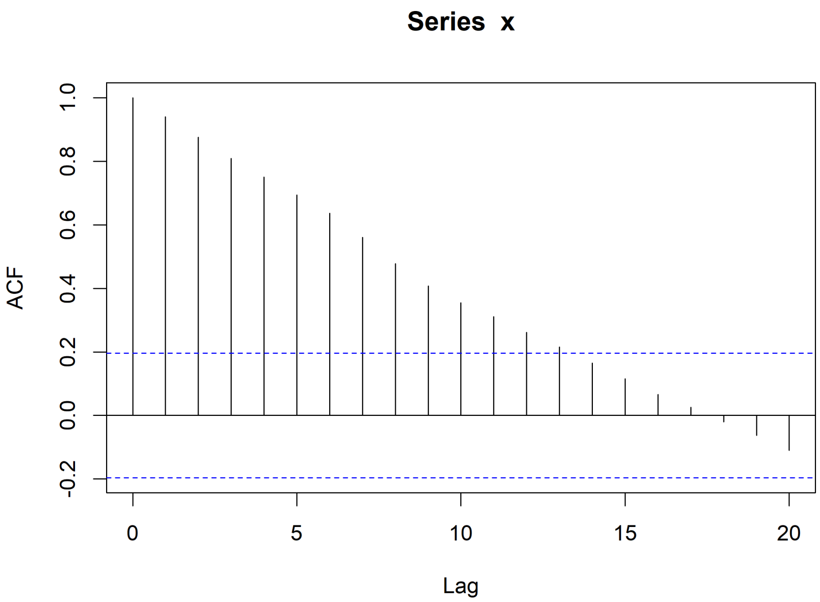
```
autocorrelations of series ‘x’, by lag

    1
0.401
```

Pretty darn close, we’d say.

Random Walk Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance σ_Z^2, and set $X_t = \sum_{i=1}^{T} Z_i$.

We prepare a realization of a random walk and display its sample ACF.

```
set.seed(4)
z=rnorm(100)
x=cumsum(z)
acf(x)
```



Well, that is certainly rather different than the other sample ACF we have studied so far... but perhaps it should not come as a surprise when we remember that random walks are **not** stationary.

Time series analysis, then, requires first that the time series be decomposed into its

- **stationary** (random) and
- **non-stationary** components (trend, level shifts, seasonality, etc.).

Next, we try to identify the nature of the random component *via* a model (using tools like the sample ACF and the sample PACF).

We will discuss commonly-encountered models in the following sections.

9.3 ARMA Models

In this section, we assume that the time series $\{X_t\} \in \mathcal{T}_n$ is stationary. We will discuss the simplest of the non-trivial time series analysis models, the **auto-regressive moving average** model (ARMA).

9.3.1 Linear Processes/Moving Averages

Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance $\text{Var}(Z_t) = \text{E}[Z_t^2] = \sigma_Z^2$.[23] Let ψ_j, $j \geq 0$, be a sequence of

[23] In the rest of this section, the assumptions on $\{Z_t\}$ will be taken for granted.

constants such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Then

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

is called a **linear process** or a **moving average**.[24]

The condition $\sum_{j=0}^{\infty} |\psi_j| < \infty$ ensures that the infinite series converges:

$$\mathrm{E}[|X_t|] \leq \sum_{j=0}^{\infty} |\psi_j| \mathrm{E}[|Z_{t-j}|] = \mathrm{E}[|Z_0|] \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Note that this condition is not necessary, however.[25]

Lemma: a linear process is a stationary time series with $\mathrm{E}[X_t] = 0$ and

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

Proof: if we assume that the convergence of the infinite sum of random variables is "uniform", then since $\mathrm{E}[Z_t] \equiv 0$, we have

$$\mathrm{E}[X_t] = \mathrm{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j}\right] = \sum_{j=0}^{\infty} \psi_j \mathrm{E}[Z_{t-j}] = 0;$$

that this is indeed the case is not trivial to show.[26]

We interchange $\sum$ and $\mathrm{E}[\cdot]$ once more,[27] to obtain:

$$\begin{aligned}\gamma_X(h) = \mathrm{E}[X_t X_{t+h}] - \mathrm{E}[X_t]\mathrm{E}[X_{t+h}] &= \mathrm{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j} \sum_{i=0}^{\infty} \psi_i Z_{t+h-i}\right] - 0 \\ &= \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \psi_j \psi_i \mathrm{E}[Z_{t-j} Z_{t+h-i}].\end{aligned}$$

Since the noise variables Z_t are independent, the only terms that contributes to the double sum are those for which $j = i - h$. Hence, the double sum collapses to a single sum:

$$\gamma_X(h) = \sum_{j=0}^{\infty}\sum_{j=0}^{\infty} \psi_j \psi_{j+h} \mathrm{E}[Z_{t-j}^2] = \sum_{j=0}^{\infty}\sum_{j=0}^{\infty} \psi_j \psi_{j+h} (\mu_Z^2 + \sigma_Z^2).$$

As $\mu_Z = 0$, we obtain the desired conclusion. ■

AR(1) The auto-regressive model of order 1, AR(1), with parameter ϕ takes the form

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

If $|\phi| < 1$, AR(1) is the linear process with $\psi_j = \phi^j$; according to the preceding lemma, we have

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} = \sigma_Z^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma_Z^2 \phi^h \sum_{j=0}^{\infty} (\phi^2)^j = \sigma_Z^2 \frac{\phi^h}{1-\phi^2},$$

[24] The terms **causal moving average** or **one-sided moving average** are also used, to indicate that the sum starts at a finite index j; a **non-causal linear process** would take the form $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, but we need a bi-directional sequence $\{Z_t \mid t \in \mathbb{Z}\}$ of independent random variables with mean 0 and variance σ_Z^2 for this to make sense.

[25] $\sum_{j=0}^{\infty} |\psi_j| < \infty \implies \sum_{j=0}^{\infty} \psi_j^2 < \infty$.

[26] The proof is outside the scope of these notes; we will take it as valid, sight unseen.

[27] Again, because of the L_2-convergence of the $\psi-$series.

using the formula for the sum of a geometric series.[28]

MA(q) The moving average model of order q, MR(q), with parameter vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)$ takes the form

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

This is a linear process with $\psi_0 = 1$, $\psi_1 = \theta_1, \ldots, \psi_q = \theta_q$, and $\psi_j = 0$, for all $j > q$;[29] according to the preceding lemma, we have

$$\gamma_X(h) = \begin{cases} \sigma_Z^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+h} = \sigma_Z^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & h = 0, \ldots, q \\ 0 & h > q. \end{cases}$$

9.3.2 ARMA in General

In order to define the general ARMA model, we introduce a crucial element of time series analysis.

Backward Shift Operator Recall that the difference operator ∇ acts on a time series $\{X_t\}$ according to

$$\nabla X_t = X_t - X_{t-1}, \qquad \text{as long as } X_{t-1} \text{ exists.}$$

The **backward shift operator** B is defined by

$$BX_t = (1 - \nabla)X_t = X_t - (X_t - X_{t-1}) = X_{t-1}.$$

It is easy to show (by induction, say) that $B^k X_t = X_{t-k}$, for all k for which X_{t-k} exists.

AR(1) If

$$X_t = \phi X_{t-1} + Z_t,$$

then, by formal manipulations of the expressions, we have

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t, \\ &= \phi^3 X_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t = \cdots \\ &= \cdots + \phi^4 Z_{t-4} + \phi^3 Z_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t, \end{aligned}$$

which we recognize as the AR(1) process.[30]

Equivalently, if we set $\phi(x) = 1 - \phi x$, then AR(1) rewrites as:

$$X_t - \phi B X_t = Z_t \iff (1 - \phi B) X_t = Z_t \iff \phi(B) X_t = Z_t.$$

MA(1) Recall that MA(1) is the linear process

$$X_t = Z_t + \theta Z_{t-1},$$

[28] Note that the sum does not converge for $|\phi| \geq 1$.

[29] We set $\theta_0 = 1$, by convention.

[30] Convergence still requires $|\phi| < 1$.

where the Z_t are as in AR(1) above. If we set $\theta(z) = 1 + \theta z$, then MA(1) rewrites as:

$$X_t = Z_t + \theta B Z_t \iff X_t = (1 + \theta B)Z_t \iff X_t = \theta(B)Z_t.$$

ARMA$(1,1)$ We can use $\phi(x)$ and $\theta(z)$ to define a new model:

$$\phi(B)X_t = \theta(B)Z_t,$$

which upon expansion becomes

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}.$$

This model combines the AR(1) and MA(1) models, which is why we call it an **auto-regressive moving average model of order** $(1,1)$.

ARMA(p,q) Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance $\text{Var}(Z_t) = \text{E}[Z_t^2] = \sigma_Z^2$. A time series $\{X_t\}$ is an **auto-regressive moving average model of order** (p,q), denoted ARMA(p,q), if it solves the equation

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots \theta_q Z_{t-q}.$$

Equivalently,

$$\phi(B)X_t = \theta(B)Z_t,$$

where

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

are the **auto-regressive** and **moving average** polynomials, respectively.

The statement "ARMA(p,q) solves the equation" means that we can write X_t as a stationary linear process

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

where the coefficients ψ_j depend on the model parameters $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$.

While ARMA models do not need to be causal, we will only be interested in causal models:

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

9.3.3 Stationarity and Causality

A **stationary solution** for ARMA(p,q) exists whenever the auto-regressive polynomial

$$\phi(z) = 1 - \phi_1 x - \cdots - \phi_p x^p$$

has no root on the complex unit circle, which is to say that of ϕ's roots satisfy $|x| \neq 1$.

A causal solution for ARMA(p,q) exists whenever the roots of the auto-regressive polynomial auto-regressive polynomial $\phi(x)$ **all lie outside the complex open unit disk**, which is to say that all of ϕ's roots satisfy $|x| > 1$.

Examples

1. The auto-regressive polynomial of the AR(1) model
$$X_t - 1.1X_{t-1} = Z_t$$
is $\phi(x) = 1 - 1.1x$; its only root is at $x_0 = 1/1.1$, for which $|x_0| < 1$. Thus we can write X_t as a stationary linear process, but there are no causal solution.
2. The model $X_t - 0.1X_{t-1} = Z_t$ is both stationary and causal.
3. The model $X_t - X_{t-1} = Z_t$ is causal but non-stationary; its auto-regressive polynomial $\phi(x)$ only has a root at $x = 1$.
4. Consider the AR(2) process $X_t - 0.1X_{t-1} - 0.4X_{t-2} = Z_t$. Equivalently, we can write $X_t - 0.1BX_t - 0.4B^2X_t = Z_t$; its auto-regressive polynomial is thus
$$\phi(x) = 1 - 0.1x - 0.4x^2,$$
whose roots are $x_1 \approx 1.46$ and $x_2 \approx -1.71$. Both of these roots have modulus larger than one, so the process is causal and there is a stationary solution.
5. Consider the AR(2) process $(1-B-B^2)X_t = Z_t$. The auto-regressive polynomial is
$$\phi(x) = 1 - x - x^2,$$
whose roots are $x_{1,2} = (-1 \pm i\sqrt{3})/2$. The modulus is 1 and so there are no stationary solution (but the process is causal).
6. Consider the AR(2) process $X_t - 0.1X_{t-1} + 0.4X_{t-2} = Z_t$. The auto-regressive polynomial is
$$\phi(x) = 1 - 0.2x + 0.4x^2,$$
whose only roots are imaginary:
$$x_{1,2} = \frac{0.1 \pm i\sqrt{1.56}}{0.8} = 0.25 \pm 0.1561249500i.$$
Both roots have the same modulus which is ≈ 1.58; this is larger than 1 so the linear process is stationary and causal.
7. Consider the AR(2) process $X_t - \phi X_{t-1} - \phi X_{t-2} = Z_t$; its auto-regressive polynomial is
$$\phi(x) = 1 - \phi x - \phi x^2,$$
whose roots are
$$x_{1,2}(\phi) = -\frac{\phi \pm \sqrt{\phi^2 + 4\phi}}{2\phi}.$$
Then $\Delta = \phi^2 + 4\phi = \phi(\phi + 4) > 0$ if $\phi < -4$ and $\phi > 0$, so the roots are real when $\phi \notin [-4, 0]$; over $(-4, 0)$, the roots are complex

conjugates, with

$$|x_{1,2}(\phi)| = \left|\frac{1}{2} \pm i\frac{\sqrt{-\phi^2-4\phi}}{2\phi}\right| = \sqrt{\frac{1}{4}+\frac{(-\phi^2-4\phi)}{4\phi^2}} = \sqrt{-\frac{1}{\phi}}.$$

We seek the instances where $|x_{1,2}(\phi)| = 1$.

a) When $\phi \notin [-4,0]$, $x_{1,2}(\phi) = \pm 1$ if and only if

$$-\frac{\phi \pm \sqrt{\phi^2+4\phi}}{2\phi} = \pm 1 \iff \phi \pm \sqrt{\phi^2+4\phi} = \pm 2\phi$$
$$\iff \phi \pm 2\phi = \pm\sqrt{\phi^2+4\phi},$$

that is, $-\phi = \pm\sqrt{\phi^2+4\phi}$ or $3\phi = \pm\sqrt{\phi^2+4\phi}$. Squaring on both sides yields $\phi^2 = \phi^2+4\phi$ or $9\phi^2 = \phi^2+4\phi$; this becomes $\phi = 0$, which we must reject as it is not in the domain of $x_{1,2}(\phi)$, or $\phi = 1/2$, which is.

b) When $\phi \in (-4,0)$, $|x_{1,2}(\phi)| = 1$ if and only if $\sqrt{-1/\phi} = 1$, so that $-1/\phi = 1$, or $\phi = -1$.

The situation is summarized in Figure 9.5.

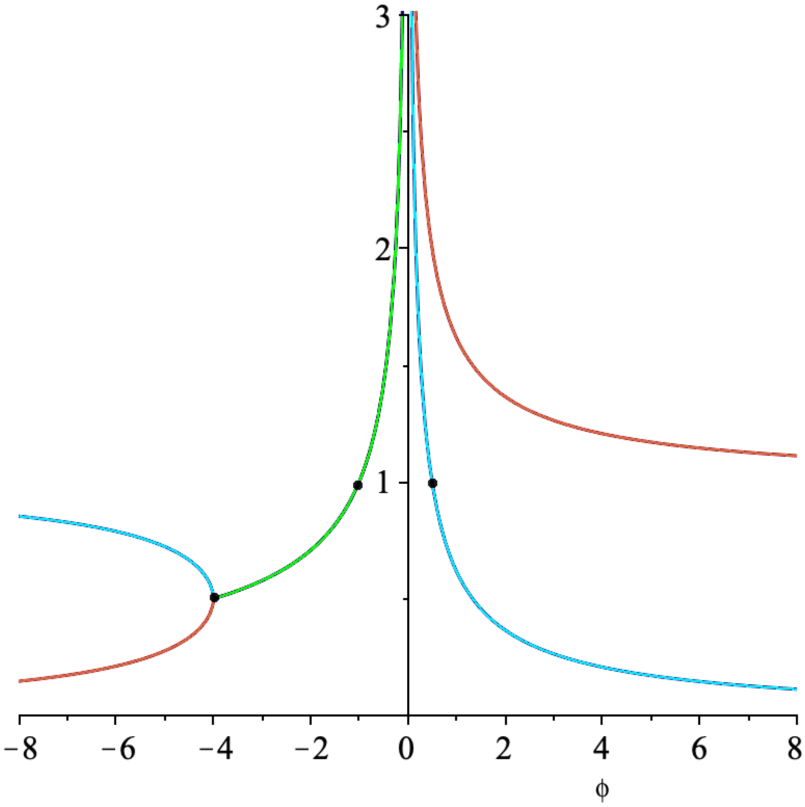


Figure 9.5: Modulus of the roots of the quadratic polynomial $\phi(x) = 1-\phi x-\phi x^2$ as a function of ϕ; the roots are real and distinct when $\phi < -4$ or $\phi > 0$ (red, blue); they are complex conjugates when $-4 < \phi < 0$ (green). The corresponding linear process is causal and stationary when the modulus is larger than or equal to 1 for both roots; by piecewise continuity of the moduli, we see that this is the case for $\phi \in [-2,0) \cup (0,1/2]$.

9.3.4 Linear Representation

Given an ARMA(p,q) model, how do we **represent** it as a linear process? There is no easy way to do this in the general case, but we will study some basic models.

MA(q) If $p = 0$, then an ARMA$(0, q)$ model is simply an MA(q) model, and its linear representation is trivial:

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

with $\psi_0 = 1, \psi_1 = \theta_1, \ldots, \psi_q = \theta_q$, and $\psi_k = 0$ for all $k > q$.

As q is finite,

$$\sum_{j=0}^{\infty} |\psi_j| = 1 + |\theta_1| + \cdots + |\theta_q| < \infty.$$

AR(1) The simplest auto-regressive model is obtained by setting $p = 1$ and $q = 0$ in ARMA(p, q):

$$\phi(B)X_t = Z_t,$$

where the auto-regressive polynomial is $\phi(x) = 1 - \phi x$. Define

$$\chi(x) = \frac{1}{\phi(x)} = \frac{1}{1 - \phi x}.$$

This function has a power series expansion:

$$\chi(x) = \frac{1}{1 - \phi x} = \sum_{j=0}^{\infty} \phi^j x^j,$$

which we know converges whenever $|\phi| < 1$. Multiplying the original model on both sides by $\chi(B)$ yields:

$$\chi(B)\phi(B)X_t = \chi(B)Z_t \implies X_t = \chi(B)Z_t,$$

since $\chi(x)\phi(x) = 1$ for all x, by construction. Thus, the linear representation of AR(1) is

$$X_t = \chi(B)Z_t = \sum_{j=0}^{\infty} \phi^j B^j Z_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j},$$

a formula we have seen before.

We note that the formal computation above only yields a **causal** linear representation when $|\phi| < 1$.[31]

ARMA$(1, 1)$ What can we say if $p = 1$ and $q = 1$, that is, if

$$\phi(B)X_t = \theta(B)Z_t,$$

where $\phi(x) = 1 - \phi x$ and $\theta(z) = 1 + \theta z$?

We once again define

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

Multiplying the original model on both sides by $\chi(B)$ yields:

$$\chi(B)\phi(B)X_t = \chi(B)\theta(B)Z_t, \implies X_t = \chi(B)\theta(B)Z_t,$$

[31] If $|\phi| > 1$, we one can still represent the process linearly, but it is not causal.

since $\chi(x)\phi(x) = 1$ for all x. In other words,

$$X_t = \chi(B)\theta(B)Z_t = \sum_{j=0}^{\infty}\phi^j B^j(1+\theta B)Z_t = \sum_{j=0}^{\infty}\phi^j B^j Z_t + \theta\sum_{j=0}^{\infty}\phi^j B^{j+1}Z_t$$
$$= \sum_{j=0}^{\infty}\phi^j Z_{t-j} + \theta\sum_{j=0}^{\infty}\phi^j Z_{t-(j+1)}.$$

But we would like X_t to take the form $\sum_{j=0}^{\infty}\psi_j Z_{t-j}$, that is, we want:

$$\sum_{j=0}^{\infty}\psi_j Z_{t-j} = \sum_{j=0}^{\infty}\phi^j Z_{t-j} + \theta\sum_{j=0}^{\infty}\phi^j Z_{t-j-1}.$$

We rewrite this equation as:

$$\psi_0 Z_t + \sum_{j=1}^{\infty}\psi_j Z_{t-j} = \phi^0 Z_t + \sum_{j=1}^{\infty}\phi^j Z_{t-j} + \theta\sum_{j=1}^{\infty}\phi^{j-1})Z_{t-j}$$
$$= \phi^0 Z_t + \sum_{j=1}^{\infty}(\phi^j + \theta\phi^{j-1})Z_{t-j}.$$

The linear representation of ARMA(1,1) is thus

$$\psi_0 = 1, \qquad \psi_j = \phi^{j-1}(\phi+\theta), \quad j \geq 1;$$

This formula was obtained under the assumptions that $|\phi| < 1$,[32] and that $\phi + \theta \neq 0$.[33]

ARMA($1, q$) The procedure for ARMA($1, q$) works in much the same way as it did for ARMA($1, 1$).

AR(p) The general procedure for AR(p), $p \geq 2$, is much more involved; we will not discuss it.

9.3.5 Autocovariance Function

The simplest ways to obtain the ACVF of an ARMA model either use the model's linear representation or a recursive method.

MA(q) and AR(1) The linear representation of the MA(q) model is trivial; for AR(1), we use the linear representation from Section 9.3.1. In both cases, we used the Lemma in that section to compute each model's ACVF (see p. 9.3.2).

ARMA($1, 1$) For this special case (and for ARMA(1,q) in general), we also use the linear representation from Section 9.3.4 and the Lemma from Section 9.3.1 to obtain the ACVF.

[32] To insure the convergence of the power series representation of $\chi(x)$.

[33] Otherwise, $X_t = Z_t$ for all t.

Specifically, since $\psi_0 = 1$, $\psi_j = \phi^{j-1}(\phi+\theta)$, $j \geq 1$, and $|\phi| < 1$, we have

$$\begin{aligned}\gamma_X(0) &= \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma_Z^2\psi_0^2 + \sigma_Z^2 \sum_{j=1}^{\infty} \psi_j^2 \\ &= \sigma_Z^2 + \sigma_Z^2 \sum_{j=1}^{\infty} (\phi^{j-1})^2(\phi+\theta)^2 \\ &= \sigma_Z^2 \left[1 + (\phi+\theta)^2 \sum_{j=1}^{\infty} \phi^{2(j-1)}\right] = \sigma_Z^2 \left[1 + \frac{(\phi+\theta)^2}{1-\phi^2}\right].\end{aligned}$$

Similarly,

$$\begin{aligned}\gamma_X(1) &= \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j\psi_{j+1} = \sigma_Z^2\psi_1 + \sigma_Z^2 \sum_{j=1}^{\infty} \psi_j\psi_{j+1} \\ &= \sigma_Z^2(\phi+\theta) + \sigma_Z^2 \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)\phi^j(\phi+\theta) \\ &= \sigma_Z^2 \left[(\phi+\theta) + \frac{1}{\phi}(\phi+\theta)^2 \sum_{j=1}^{\infty} \phi^{2j}\right] = \sigma_Z^2 \left[(\phi+\theta) + \phi\frac{(\phi+\theta)^2}{1-\phi^2}\right].\end{aligned}$$

For a general $h \geq 1$, note first that

$$\psi_0\psi_h = \psi_h = \phi^{h-1}(\phi+\theta) = \phi^{h-1}\phi^{1-1}(\phi+\theta) = \phi^{h-1}\psi_1 = \phi^{h-1}\psi_0\psi_1;$$

if $j \geq 1$, we also have

$$\begin{aligned}\psi_j\psi_{j+h} &= \phi^{j-1}(\phi+\theta)\phi^{j+h-1}(\phi+\theta) = \phi^{h-1}\left[\phi^{j-1}(\phi+\theta)\phi^j(\phi+\theta)\right] \\ &= \phi^{h-1}\psi_j\psi_{j+1}.\end{aligned}$$

Thus, $\gamma_X(h) = \phi^{h-1}\gamma_X(1)$ for $h \geq 1$, and so

$$\gamma_X(h) = \begin{cases} \sigma_Z^2\left[1 + \dfrac{(\phi+\theta)^2}{1-\phi^2}\right] & h = 0, \\ \sigma_Z^2\phi^{h-1}\left[(\phi+\theta) + \phi\dfrac{(\phi+\theta)^2}{1-\phi^2}\right] & h \geq 1. \end{cases}$$

AR(1) We can obtain $\gamma_X(h)$ for AR(1) by setting $\theta = 0$ in the the ACVF for the ARMA(1, 1) model, but we will illustrate a **recursive method** that generalizes to AR(p) or general ARMA(p, q) models with $p \geq 2$.

Let $h \in \mathbb{N}$. We start by multiplying the AR(1) equation $X_t = \phi X_{t-1} + Z_t$ by X_{t-h} on both sides and applying the expectation operator to obtain:

$$\mathrm{E}[X_t X_{t-h}] = \phi\mathrm{E}[X_{t-1}X_{t-h}] + \mathrm{E}[Z_t X_{t-h}].$$

By definition, $\gamma_X(h) = \mathrm{E}[X_t X_{t-h}] - \mathrm{E}[X_t]\mathrm{E}[X_{t-h}]$. But $\mathrm{E}[X_t] = 0$ for all t as $\{X_t\}$ is assumed to be stationary; thus $\mathrm{E}[X_t X_{t-h}] = \gamma_X(h)$ and $\mathrm{E}[X_{t-1}X_{t-h}] = \gamma_X(h-1)$.

For all $h \geq 1$ we know that Z_t is independent of X_{t-h}, which is most easily seen with the linear representation of AR(1): $X_{t-h} = \sum_{j=0}^{\infty} \phi^j Z_{t-h-j}$.[34]

Thus, $\mathrm{E}[Z_t X_{t-h}] = \mathrm{E}[Z_t]\mathrm{E}[X_{t-h}] = 0$, and the AR(1) equation is equiva-

[34] Note that this would not be the case if we had multiplied by X_{t+h} to start with.

lent to the recursive formula:

$$\gamma_X(h) = \mathrm{E}[X_t X_{t-h}] = \phi \mathrm{E}[X_{t-1} X_{t-h}] = \phi\gamma_X(h-1), \quad h \geq 1,$$

or, by induction:

$$\gamma_X(h) = \phi^{h-1}\gamma_X(0), \qquad h \geq 1.$$

We start the recursion by computing $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$. We have

$$\mathrm{Var}(X_t) = \phi^2 \mathrm{Var}(X_{t-1}) + \mathrm{Var}(Z_t),$$

again, since X_{t-1} and Z_t are independent.

As X_t is stationary, $\mathrm{Var}(X_t) = \mathrm{Var}(X_{t-1})$ for all t and we have

$$\sigma_X^2 = \phi^2\sigma_X^2 + \sigma_Z^2.$$

Solving for σ_X^2 yields:

$$\sigma_X^2 = \frac{\sigma_Z^2}{1-\phi^2}.$$

Finally

$$\gamma_X(h) = \phi^h\gamma_X(0) = \sigma_Z^2\frac{\phi^h}{1-\phi^2},$$

which agrees with the ACVF that was calculated in Section 9.3.1.

AR(2) This model's equation is $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. We use a similar approach: we multiply both sides by X_{t-h} and apply the expectation operator to obtain:

$$\mathrm{E}[X_t X_{t-h}] = \phi_1 \mathrm{E}[X_{t-1}X_{t-h}] + \phi_2 \mathrm{E}[X_{t-2}X_{t-h}] + \mathrm{E}[Z_t X_{t-h}].$$

An argument similar to the one presented for AR(1) yields the AR(2) recursion formula:

$$\gamma_X(h) = \phi_1\gamma_X(h-1) + \phi_2\gamma_X(h-2), \quad h \geq 2.$$

We start the recursion by computing $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$ and $\gamma_X(1)$.

To do so, we multiply the AR(2) equation by X_{t-1} and once again apply the expectation operator to get:

$$\mathrm{E}[X_t X_{t-1}] = \phi_1 \mathrm{E}[X_{t-1}^2] + \phi_2 \mathrm{E}[X_{t-2}X_{t-1}] + \underbrace{\mathrm{E}[Z_t X_{t-1}]}_{=0},$$

so that

$$\gamma_X(1) = \phi_1\gamma_X(0) + \phi_2\gamma_X(1) \implies \gamma_X(1)\frac{1-\phi_2}{\phi_1} = \gamma_X(0).$$

Next, we multiply the AR(2) equation by X_t and apply the expectation operator one last time to get:

$$\mathrm{E}[X_t^2] = \phi_1 \mathrm{E}[X_{t-1}X_t] + \phi_2 \mathrm{E}[X_{t-2}X_t] + \mathrm{E}[Z_t X_t].$$

But Z_t and X_t are **not** independent; in fact,

$$\mathrm{E}[Z_t X_t] = \mathrm{E}[Z_t(\phi_1 X_{t-1}+\phi_2 X_{t-2}+Z_t)] = \phi_1 \mathrm{E}[Z_t X_t]+\phi_2 \mathrm{E}[Z_t X_{t-2}]+\mathrm{E}[Z_t^2] = \sigma_Z^2,$$

and so

$$\gamma_X(0) = \phi_1\gamma_X(1) + \phi_2\gamma_X(2) + \sigma_Z^2.$$

However, we know that

$$\gamma_X(2) = \phi_1\gamma_X(1) + \phi_2\gamma_X(0)$$

from the AR(2) recursion formula, with $h = 2$; we can substitute this expression into the equation for $\gamma_X(0)$ to obtain:

$$\gamma_X(0) = \phi_1\gamma_X(1) + \phi_2\left\{\phi_1\gamma_X(1) + \phi_2\gamma_X(0)\right\} + \sigma_Z^2,$$

which yields:

$$\gamma_X(h) = \phi_1\gamma_X(h-1) + \phi_2\gamma_X(h-2), \qquad h \geq 2,$$

$$\gamma_X(1) = \sigma_Z^2\frac{\phi_1}{(1+\phi_2)\left\{(1-\phi_2)^2-\phi_1^2\right\}},$$

$$\gamma_X(0) = \sigma_Z^2\frac{1-\phi_2}{(1+\phi_2)\left\{(1-\phi_2)^2-\phi_1^2\right\}}.$$

We can perform a sanity check, by letting $\phi_2 = 0$, $\phi_1 = \phi$; the last two formulas reduce to $\gamma_X(0)$ and $\gamma_X(1)$ for AR(1).[35]

9.3.6 Partial Autocorrelation Function

The **partial autocorrelation** of a time series $\{X_t\}$ **at lag** h, denoted by $\alpha(h)$, is the autocorrelation between X_t and X_{t+h}, after removing the linear dependence of X_t on $X_{t+1}, \ldots, X_{t+h-1}$.

MA(1) We have already calculated $\alpha(2)$ for MA(1); for a general $h \in \mathbb{N}$, it can be shown that the PACF is:

$$\alpha(h) = \frac{-(-\theta)^h}{1+\theta^2+\cdots+\theta^{2h}}.$$

Since the denominator is always positive, we see that MA(1)'s PACF has an oscillating behaviour, but that it tapers to 0 when $h \longrightarrow \infty$.

AR(1) The PACF for the AR(1) model $X_t = \phi X_{t-1} + Z_t$ is such that

$$\alpha(1) = \rho_X(1) = \phi, \quad \alpha(2) = \mathrm{Corr}(X_t, X_{t+2} - \phi X_{t+1}) = \mathrm{Corr}(X_t, Z_{t+2}) = 0.$$

It turns out that this PACF behaviour is typical of AR(p) models.

[35] It is easy to see that the recursive formula for the ACVF of AR(p) takes the form:

$$\gamma_X(h) = \sum_{j=1}^{p}\phi_j\gamma_X(h-j).$$

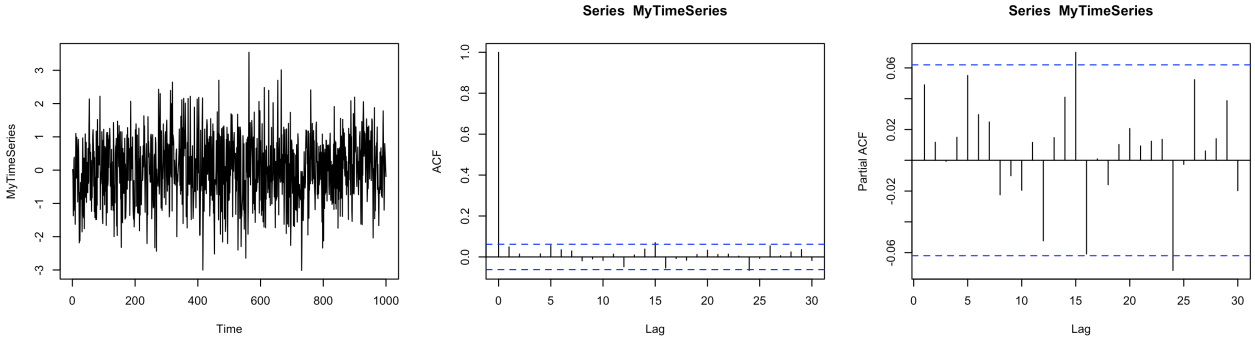
Theorem: consider a stationary AR(p) time series. Then

$$\alpha(h) = 0, \quad h = p+1, p+2, \ldots .$$

Examples In what follows, we generate a realization of various ARMA(p, q) models through package `tseries`' `arima()` function, and display the sample ACF and sample PACF plots.[36] Do the graphs have the expected characteristics?

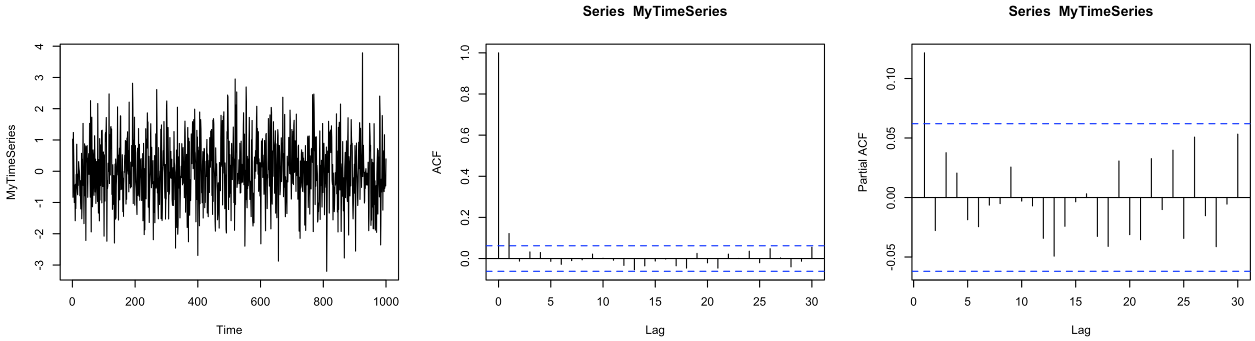
White Noise

```
library(tseries)
set.seed(10)
MyTimeSeries = arima.sim(model = list(ar = c()),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



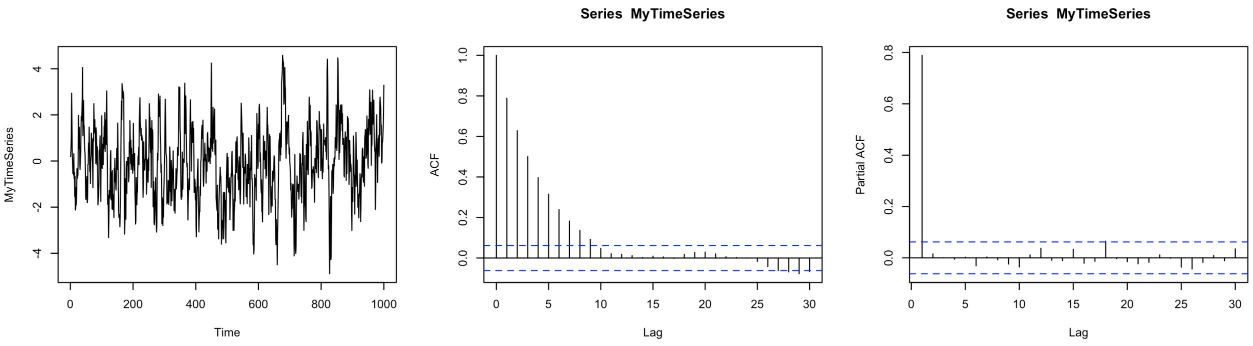
AR(1)

```
set.seed(11)
MyTimeSeries = arima.sim(model = list(ar = c(0.1)),
                         n = 1000,
                         rand.gen = rnorm);
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



[36] The examples will also showcase the syntax of the simulation function.

```
set.seed(12)
MyTimeSeries = arima.sim(model = list(ar = c(0.8)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

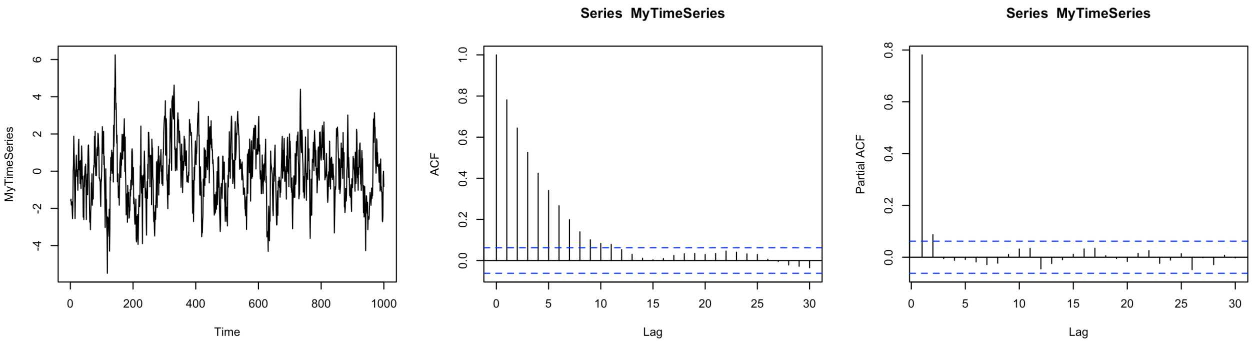


```
set.seed(14)
MyTimeSeries = arima.sim(model = list(ar = c(1.1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

```
Error in arima.sim(model = list(ar = c(1.1)), n = 1000, rand.gen = rnorm) :
  'ar' part of model is not stationary
```

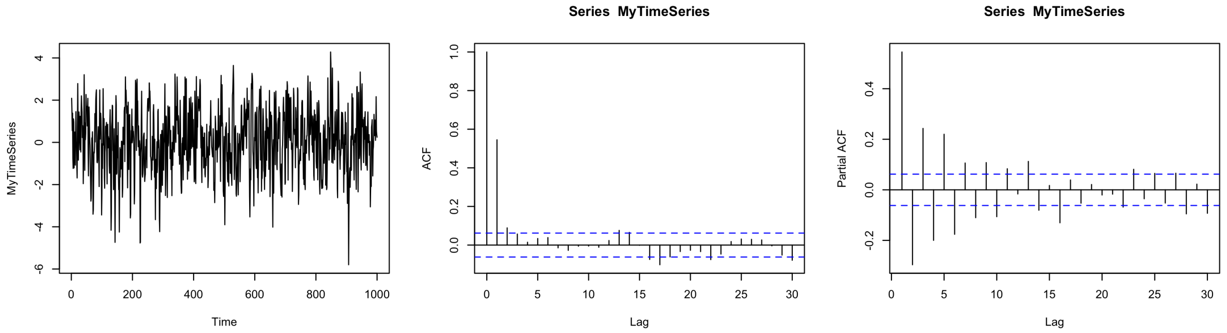
AR(2)

```
set.seed(13)
MyTimeSeries = arima.sim(model = list(ar = c(0.7,0.1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



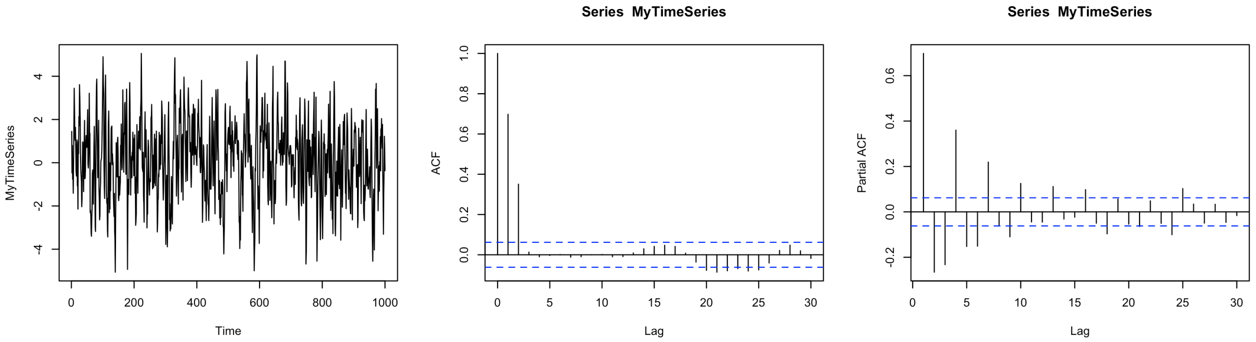
MA(1)

```
set.seed(15)
MyTimeSeries = arima.sim(model = list(ma = c(1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



MA(2)

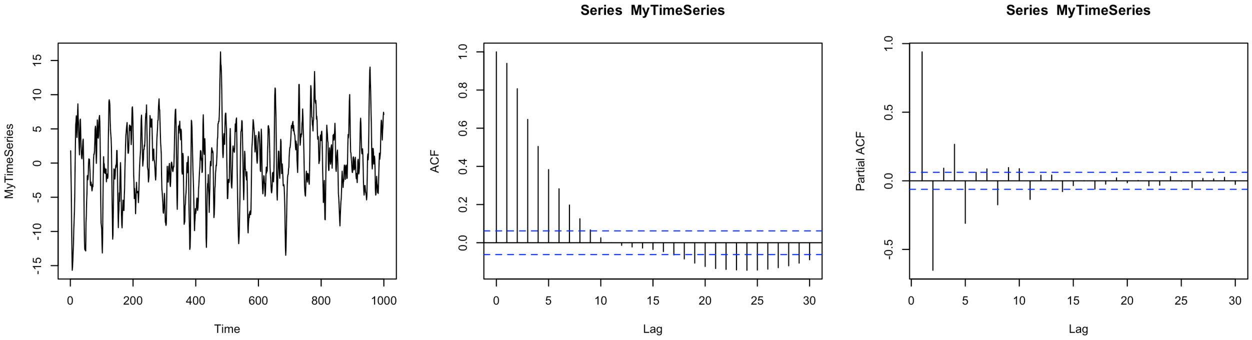
```
set.seed(16)
MyTimeSeries = arima.sim(model = list(ma = c(1,1)),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



ARMA(1, 2)

```
set.seed(17)
MyTimeSeries = arima.sim(model = list(ar = c(0.8),
                                      ma = c(1,1)),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
```

```
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



Summary:

- for AR(p) models $\gamma_X(h) \neq 0$ for any h.
- for MA(q) models $\gamma_X(h) = 0$ for any $|h| > q$;
- for AR(p) models $\alpha(h) = 0$ for any $|h| > p$;
- for MA(q) models $\alpha_X(h) \neq 0$ for any h.

9.4 Forecasting with Stationary Time Series

In practice, once of the main objectives of time series analysis is to **predict** (or **forecast**) X_{n+k} for some $k \geq 1$, having observed $\{X_1, \ldots, X_n\}$ from a time series with **known** mean μ and ACVF $\gamma_X(k)$, $k \geq 0$.

Consider a stationary sequence with mean $\mu = \mathrm{E}[X_t]$ and covariance $\gamma_X(h)$. Denote by $P_n X_{n+k}$ a prediction for X_{n+k}, given the n observations $X_1, \ldots, X_n$.

We will restrict ourselves to **linear predictors**, that is to say, predictors of the form:

$$P_n X_{n+k} = a_0 + a_1 X_n + \cdots + a_n X_1 = a_0 + \sum_{i=1}^{n} a_i X_{n+1-i},$$

where $a_0, a_1, \ldots, a_n \in \mathbb{R}$.

As is usually the case in statistical applications, this can be recast as an optimization problem. We seek values $\mathbf{a} = (a_0, \ldots, a_n)$ which minimize the expected **mean squared error** (MSE):

$$\mathrm{E}\left[(X_{n+k} - P_n X_{n+k})^2\right],$$

One challenge is that we cannot minimize $(X_{n+k} - P_n X_{n+k})^2$ directly since, there would be no reason to predict X_{n+k} if we already knew it.[37]

9.4.1 Yule-Walker Procedure

Let

$$S(\mathbf{a}) = \mathrm{E}\big[(X_{n+k} - P_n X_{n+k})^2\big] = \mathrm{E}\Big[(X_{n+k} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i})^2\Big].$$

[37] While the whole entreprise is reminiscent of OLS regression, there are some important differences, chief among them being that the predictors X_{n+1-i} are typically correlated with one another.

We minimize S by finding its critical points, i.e. by solving $\nabla S(\mathbf{a}) = \mathbf{0}$.

The partial derivative of S with respect to a_0 is

$$\mathrm{E}\Big[2(X_{n+k} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}) \cdot 1\Big] = 2\Big(\mu - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\Big);$$

setting it equal to 0 yields

$$a_0 = \mu\Big(1 - \sum_{i=1}^{n} a_i\Big).$$

If $\{X_t\}$ is assumed to be stationary, then $\mu = 0$, and so $a_0 = 0$.

The partial derivatives with respect to $a_1, \ldots, a_n$ are thus:

$$\mathrm{E}\Big[-2\Big(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i}\Big)X_{n+1-j}\Big], \qquad j = 1, \ldots, n.$$

Setting each of these to 0 yields:

$$\mathrm{E}[X_{n+k}X_{n+1-j}] - \sum_{i=1}^{n} a_i \mathrm{E}[X_{n+1-i}X_{n+1-j}] = 0, \qquad j = 1, \ldots, n.$$

Since $E[X_t] = \mu = 0$, the above expectations are the covariances of $\{X_t\}$ at lags $n + k - (n + 1 - j) = k - 1 + j$ and $n + 1 - i - (n + 1 - j) = i - j$, and we can thus write the system of equations as:

$$\gamma_X(k - 1 + j) = \sum_{i=1}^{n} a_i \gamma_X(i - j), \qquad j = 1, \ldots, n. \tag{9.1}$$

Define the matrix

$$\Gamma_n = [\gamma_X(|i - j|)]_{i,j=1}^{n}$$

and the column vectors

$$\boldsymbol{\gamma}(n; k) = (\gamma_X(k), \ldots, \gamma_X(k + n - 1))^\top, \qquad \underline{\mathbf{a}}_n = (a_1, \ldots, a_n)^\top.$$

We recognize Γ_n as the **variance-covariance matrix** of $(X_1, \ldots, X_n)$, whose diagonal entries are $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$.

If $n = 1$, for instance, then $\Gamma_1 = \gamma_X(0)$; if $n = 2$, then

$$\Gamma_2 = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(|-1|) & \gamma_X(0) \end{bmatrix} = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{bmatrix}.$$

We can write the system of n equations in n unknowns from (9.1) in a matrix-vector notation:

$$\Gamma_n \underline{\mathbf{a}}_n = \boldsymbol{\gamma}(n; k),$$

whose solution, assuming that Γ_n is invertible, is the **Yule-Walker forecasting formula**:

$$\underline{\mathbf{a}}_n = \Gamma_n^{-1} \boldsymbol{\gamma}(n; k).$$

Note that it is **model-independent**.[38]

[38] Well, the formula for $\underline{\mathbf{a}}_n$ is, at any rate. It only really assumes that the time series is stationary. But it does depend on the autocovariances of the time series; with a model, it is usually rather straightforward to compute these. Without a model, we have to use the sample autocovariances.

MSPE The above procedure guarantees that the **mean squared prediction error**

$$\mathrm{MSPE}_n(k) = \mathrm{E}\Big[(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i})^2\Big]$$

is minimized when $\underline{\mathbf{a}}$ is chosen according to the Yule-Walker procedure. Can we calculate the MSPE value?

Recall that the $\mathrm{E}[X_t] \equiv 0$ by stationarity. Thus,

$$\begin{aligned}
&\mathrm{E}\Big[(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i})^2\Big] \\
&= \mathrm{E}\left[X_{n+k}^2\right] - 2\sum_{i=1}^{n} a_i \mathrm{E}[X_{n+k}X_{n+1-i}] + \mathrm{E}\Big[\big(\sum_{i=1}^{n} a_i X_{n+1-i}\big)^2\Big] \\
&= \gamma_X(0) - 2\sum_{i=1}^{n} a_i \gamma_X(k+i-1) + \mathrm{E}\Big[\sum_{i,j=1}^{n} a_i X_{n+1-i}X_{n+1-j}a_j\Big] \\
&= \gamma_X(0) - 2\sum_{i=1}^{n} a_i \gamma_X(k+i-1) + \sum_{i,j=1}^{n} a_i \gamma_X(i-j) a_j \\
&= \gamma_X(0) - 2\underline{\mathbf{a}}_n^\top \boldsymbol{\gamma}(n;k) + \underline{\mathbf{a}}_n^\top \Gamma_n \underline{\mathbf{a}}_n = \gamma_X(0) - \underline{\mathbf{a}}_n^\top \boldsymbol{\gamma}(n;k).
\end{aligned}$$

An important remark is that the MSPE formula depends on k; in particular, it is possible that, given a set of observations $X_1, \dots, X_n$, predictions further in the future (i.e., having a larger k) may have a larger prediction error than those nearer $t = n$.[39]

Example: AR(1) Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and Z_t are i.i.d. with mean 0 and variance σ_Z^2. We have already seen that $\{X_t\}$ is stationary, and so that $\mu = \mathrm{E}[X_t] \equiv 0$.

Recall that the autocovariances for this model are:

$$\gamma_X(h) = \phi^h \frac{\sigma_Z^2}{1-\phi^2}, \qquad h \geq 0.$$

If we are interested in predicting X_{n+1}, then we need:

$$\boldsymbol{\gamma}(n;k) = \boldsymbol{\gamma}(n;1) = (\gamma_X(1), \dots, \gamma_X(n))^\top = \frac{\sigma_Z^2}{1-\phi^2}(\phi, \dots, \phi^n)^\top.$$

The Yule-Walker forecasting equation in this case becomes

$$\frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix} \phi \\ \vdots \\ \phi^n \end{pmatrix}.$$

We can show that the determinant of of Γ_n is

$$\det(\Gamma_n) = (-1)^{n-1}(\phi-1)^{n-1}(\phi+1)^{n-1}\left(\frac{\sigma_Z^2}{1-\phi^2}\right)^n \neq 0$$

since $|\phi| < 1$. There is thus a unique forecasting solution $\underline{\mathbf{a}}_n$.

[39] Of course, it could also be the other way around – but the point is that we should not expect $\mathrm{MSPE}_n(k)$ to be constant with k.

But

$$\Gamma_n\begin{pmatrix}\phi\\0\\\vdots\\0\end{pmatrix}=\frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix}1\cdot\phi+0\cdot(\ldots)\\\phi\cdot\phi+0\cdot(\ldots)\\\vdots\\\phi^{n-1}\cdot\phi+0\cdot(\ldots)\end{pmatrix}=\frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix}\phi\\\phi^2\\\vdots\\\phi^n\end{pmatrix},$$

and so $\underline{\mathbf{a}}_n=(\phi,0,\ldots,0)^\top$ is the unique Yule-Walker forecast vector for the AR(1) model.

The Yule-Walker prediction for X_{n+1} is thus

$$P_nX_{n+1}=a_1X_n+a_2X_{n-1}+\cdots+a_nX_1=\phi X_n,$$

while the MSPE is

$$\begin{aligned}\mathrm{MSPE}_n(1)&=\gamma_X(0)-\underline{\mathbf{a}}\,{:}^{\top}_{n}\,\boldsymbol{\gamma}(n;1)=\gamma_X(0)-\phi\gamma_X(1)-0\cdot\gamma_X(2)-\cdots-0\cdot\gamma_X(n)\\&=\frac{\sigma_Z^2}{1-\phi^2}-\phi^2\frac{\sigma_Z^2}{1-\phi^2}=\sigma_Z^2.\end{aligned}$$

Note, however, that these formulas cannot yet be used in a practical setting since they involve the unknown parameters ϕ and σ_Z^2.

9.4.2 Durbin-Levinson Algorithm

In the AR(1) prediction example, we were lucky that the solution $\underline{\mathbf{a}}_n$ was provided *in extremis*; there is a way to find the best linear predictor without having to compute the inverse of Γ_n. But it comes at a price: the approach only allows **one-step** prediction to P_nX_{n+1}.

We assume that $\mu=\mathrm{E}[X_t]\equiv 0$ and $a_0=0$, as in the Yule-Walker procedure.

We re-write the linear predictor as

$$P_nX_{n+1}=\phi_{n,1}X_n+\cdots+\phi_{n,n}X_1.$$

That is, $a_1=\phi_{n,1},\ldots,a_n=\phi_{n,n}$.

- If $n=1$, we seek to find $P_1X_2=\phi_{1,1}X_1$ which minimizes

  $$\mathrm{E}\left[(X_2-P_1X_2)^2\right]=\mathrm{E}\left[(X_2-\phi_{1,1}X_1)^2\right].$$

  We differentiate with respect to $\phi_{1,1}$ and set equal to 0 to find the critical point:

  $$\mathrm{E}\left[2(X_2-\phi_{1,1}X_1)(-X_1)\right]=0\implies\mathrm{E}[X_1X_2]=\phi_{1,1}\mathrm{E}[X_1^2],$$

  which is to say that

  $$\phi_{1,1}=\frac{\gamma_X(1)}{\gamma_X(0)}=\rho_X(1).$$

- If $n=2$, we seek to find $\phi_{2,1}$ and $\phi_{2,2}$ in

  $$P_2X_3=\phi_{2,1}X_2+\phi_{2,2}X_2.$$

As in the Yule-Walker procedure we minimize

$$\mathrm{E}[(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)^2].$$

Taking derivatives with respect to $\phi_{2,1}$ and $\phi_{2,2}$ leads to:

$$\begin{aligned}\mathrm{E}[-2X_2(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)] &= 0\\ \mathrm{E}[-2X_1(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)] &= 0;\end{aligned}$$

equivalently, since the mixed expectations are covariances and the squared ones are variances, this can be written as:

$$\begin{aligned}\gamma_X(1) - \phi_{2,1}\gamma_X(0) - \phi_{2,2}\gamma_X(1) &= 0\\ \gamma_X(2) - \phi_{2,1}\gamma_X(1) - \phi_{2,2}\gamma_X(0) &= 0.\end{aligned}$$

We divide both equations by $\gamma_X(0)$ and re-organize the terms to obtain:

$$\begin{aligned}\phi_{2,1} &= \rho_X(1) - \phi_{2,2}\rho_X(1) = \rho_X(1) - \phi_{2,2}\phi_{1,1}, \quad \text{by step } n = 1;\\ 0 &= \rho_X(2) - \phi_{2,1}\rho_X(1) - \phi_{2,2}\end{aligned}$$

Solving for $\phi_{2,1}$ and $\phi_{2,2}$, we arrive at

$$\begin{aligned}\phi_{2,2} &= \frac{\rho_X(2) - \phi_{1,1}\rho_X(1)}{1 - \phi_{1,1}\rho_X(1)},\\ \phi_{2,1} &= \rho_X(1) - \phi_{2,2}\phi_{1,1}.\end{aligned}$$

We use either $\phi_{1,1}$ or $\rho_X(1)$, solely based on convenience (since they are equal). In the last system of equations, the coefficients $\phi_{2,2}$ and $\phi_{2,1}$ are computed using sample autocorrelations, as well as $\phi_{1,1}$ (from the step $n = 1$).

This recursive procedure can be extended for a general n.

Durbin-Levinson Algorithm The coefficients $\phi_{n,1}, \ldots, \phi_{n,n}$ in the best linear prediction $P_n X_{n+1}$ can be computed recursively as:

$$\phi_{n,n} = \left[\gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma_X(n-j)\right] v_{n-1}^{-1};$$

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

and

$$v_n = v_{n-1}[1 - \phi_{n,n}^2], \qquad v_0 = \gamma_X(0), \qquad \phi_{1,1} = \rho_X(1).$$

Note that the Durbin-Levinson algorithm and the Yule-Walker procedure lead to the same results for $P_n X_{n+1}$; indeed, in both cases we compute the coefficents of the linear prediction $P_n X_{n+1}$ using the mean squared error criterion, the difference being that we approach the problem from two different angles.

AR(1) Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where Z_t are i.i.d. with mean 0 and variance σ_Z^2.

We know the ACVF and ACF of $\{X_t\}$ are

$$\gamma_X(h) = \phi^h \frac{\sigma_Z^2}{1-\phi^2}, \quad \text{and} \quad \rho_X(h) = \gamma_X(h)/\gamma_X(0) = \phi^h.$$

Using the Durbin-Levinson algorithm, we find the linear coefficients and predictors as follows:

$$\begin{aligned} \phi_{1,1} = \phi, &\quad P_1X_2 = \phi X_1; \\ \phi_{2,1} = \phi, \phi_{2,1} = 0, &\quad P_2X_3 = \phi X_2; \\ &\quad \vdots \quad \vdots \\ \phi_{n,1} = \phi, \phi_{n,2} = \cdots = \phi_{n,n} = 0, &\quad P_nX_{n+1} = \phi X_n. \end{aligned}$$

Partial Autocovariance function (PACF) As a by-product of the Durbin-Levinson algorithm, we obtain the PACF *via*:

$$\alpha(0) = 1; \qquad \alpha(h) = \phi_{h,h}, \quad h \geq 1.$$

9.4.3 Forecast Limits and Prediction Intervals

We obtained **model-independent** formulas for (linearly) predicted time series values in the preceding sections, depending solely on the sample autocovariances.[40] Discussions of **accuracy**, however, require model assumptions.

Let $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ be a causal linear process with $\mathrm{E}[Z_t] = 0$ and $\mathrm{Var}(Z_t) = \sigma_Z^2$, and $k \geq 1$ an integer.

It can be shown that the mean squared prediction error at P_nX_{n+k} is:

$$\mathrm{MSPE}_n(k) = \mathrm{E}[(X_{n+k} - P_nX_{n+k})^2] = \sigma_Z^2 \sum_{j=0}^{k-1} \psi_j^2.$$

The theoretical forecast limits of the $100(1-\alpha)\%$ **prediction interval** are thus:

$$P_nX_{n+k} \pm z_{\alpha/2}\sqrt{\mathrm{MSPE}_n(k)} = P_nX_{n+k} \pm z_{\alpha/2}\sigma_Z \sqrt{\sum_{j=0}^{k-1} \psi_j^2},$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.[41] Note that MSPE (and so the coefficients ψ_j) are **model-dependent**: no model, no prediction interval!

9.4.4 Example: Currency Conversion Data

We illustrate the notions presented in this section with an example, using the quarterly mean exchange rate between British pounds (UK) and New Zealand dollar (NZD), from Jan 1991 to Mar 2000 (prepared by Darrin Speegler).

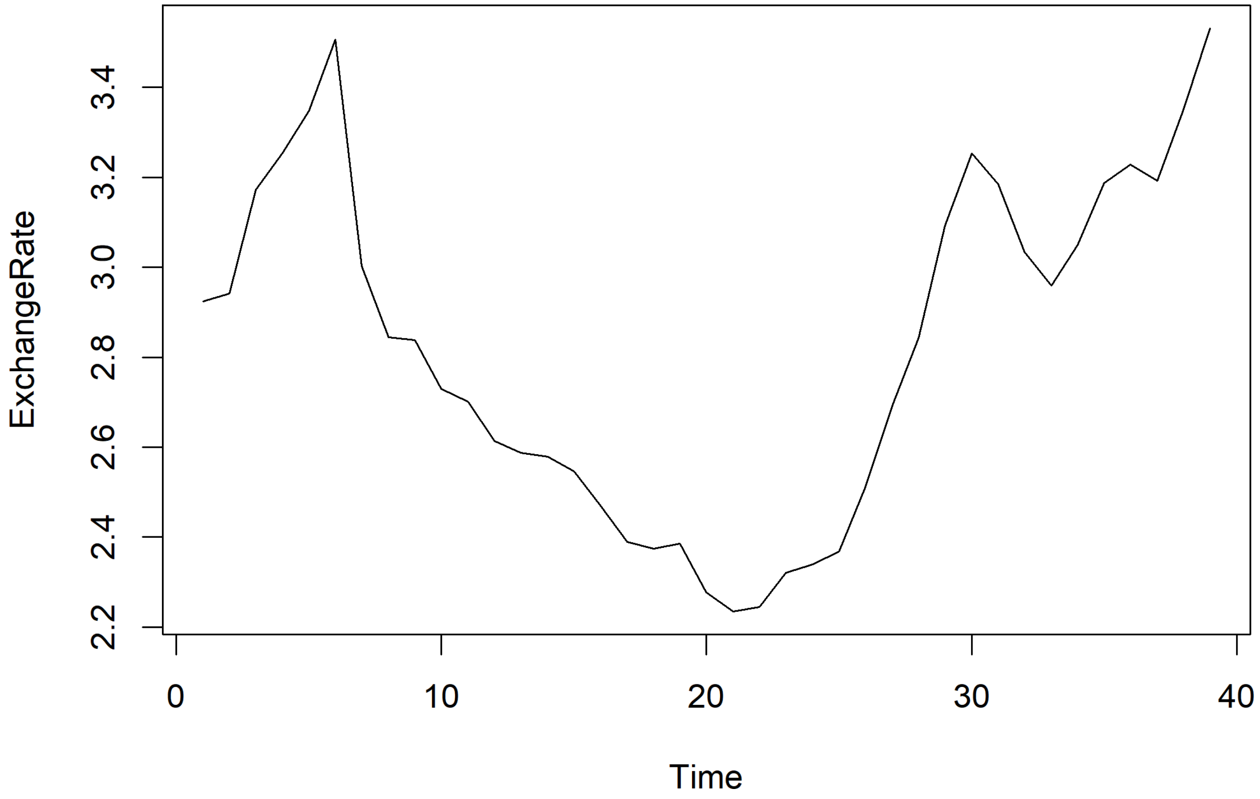
[40] Although we can use a model if one is available.

[41] You know the one: if $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$.

```
ExchangeRate = c(2.9243,2.9422,3.1719,3.2542,3.3479,
                 3.5066,3.0027,2.8440,2.8378,2.7301,
                 2.7008,2.6138,2.5874,2.5787,2.5470,
                 2.4701,2.3895,2.3705,2.3859,2.2766,
                 2.2351,2.2450,2.3208,2.3390,2.3687,
                 2.5120,2.6917,2.8435,3.0922,3.2528,
                 3.1852,3.0340,2.9593,3.0498,3.1869,
                 3.2286,3.1925,3.3522,3.5310)
```

The time series plot tells a better story.

```
plot.ts(ExchangeRate)
```



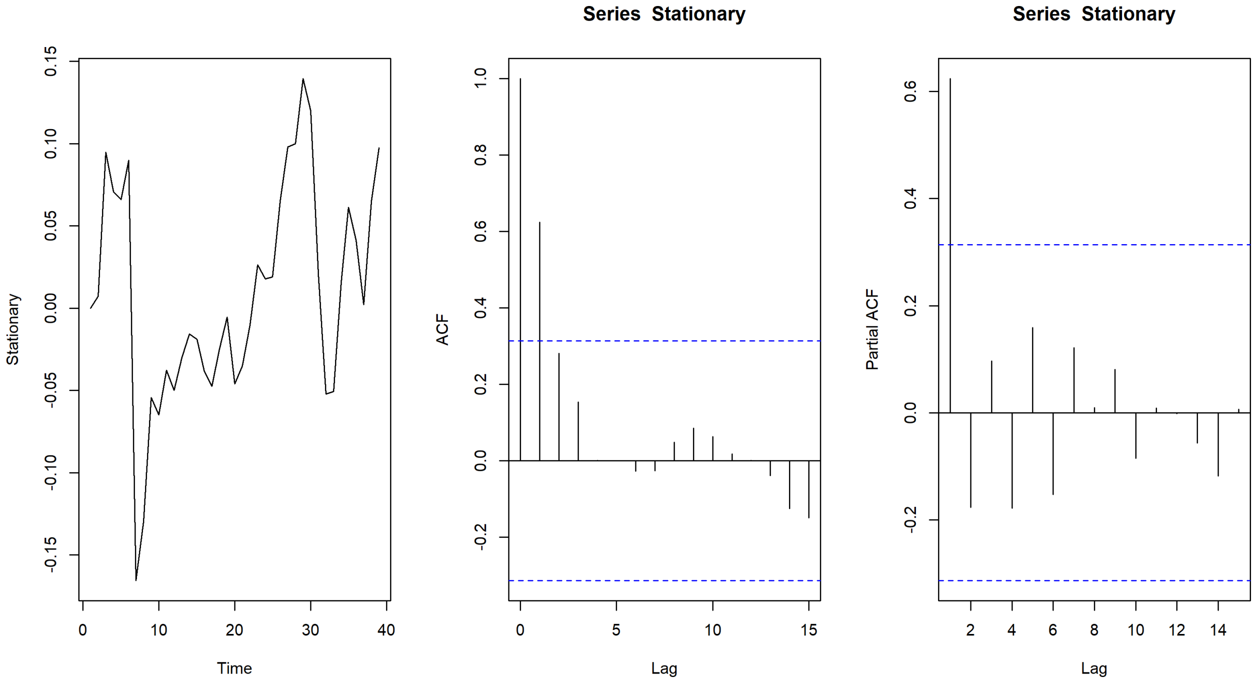
The model is clearly not stationary.

We detrend the data *via* the exponential smoother `ExpSmooth` of Section 9.1.2, with $\alpha = 0.6$.

```
alpha = 0.6
ExchangeRate.smoothed <- ExpSmooth(ExchangeRate,alpha)
Stationary = ExchangeRate - ExchangeRate.smoothed
```

The ACF and PACF of the stationary components are found below.

```
par(mfrow=c(1,3))
plot.ts(Stationary)
acf(Stationary)
pacf(Stationary)
```



The detrended time series looks like AR(1).[42] We centre the time series and use the Yule-Walker method to verify that this is indeed an appropriate model – we will be discussing this further in Section 9.5.3.

```
MyTimeSeries = Stationary
n = length(MyTimeSeries)
mean = mean(MyTimeSeries)
MyTimeSeries.centered = MyTimeSeries-mean(MyTimeSeries)
(fit.ar <- ar(MyTimeSeries.centered,method="yule-walker"))
```

```
Coefficients:
     1
0.6241

Order selected 1  sigma^2 estimated as  0.002842
```

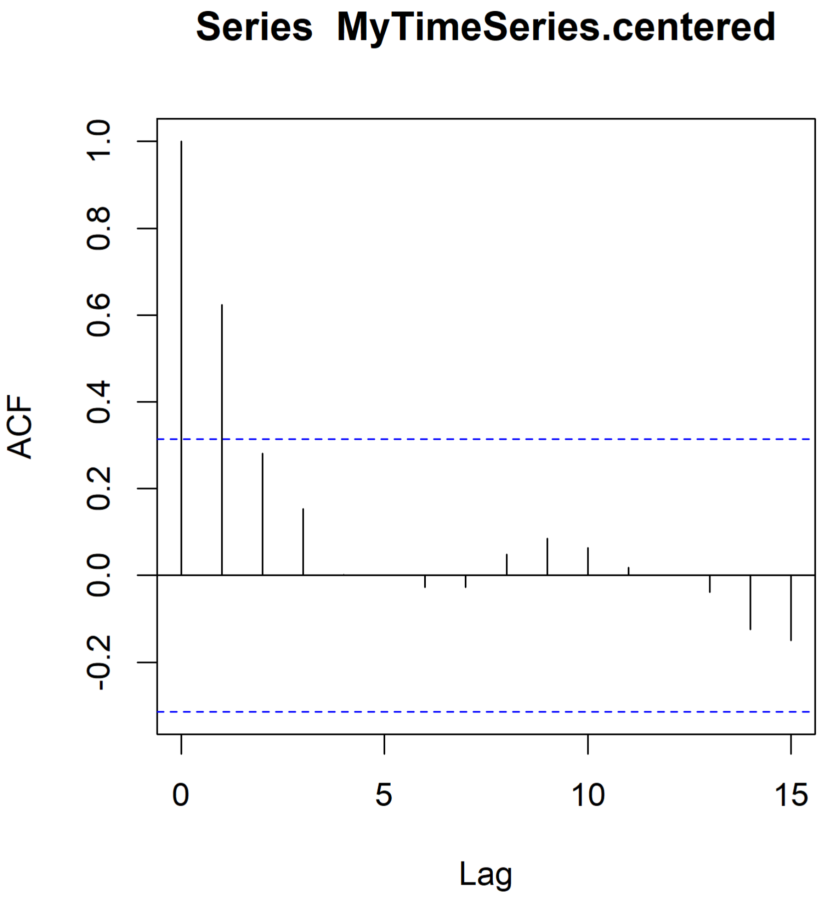
The Yule-Walker estimates of the selected AR(1) model are $\hat{\phi} = 0.6241$, $\sigma_X^2 = 0.002842$, respectively.

We can verify the Yule-Walker output by comparing with the ACF.

```
par(mfrow=c(1,1))
(ACF <- acf(MyTimeSeries.centered))
```

```
Autocorrelations of series ‘MyTimeSeries.centered’, by lag

     0      1      2      3      4      5      6      7      8
 1.000  0.624  0.281  0.154  0.001  0.000 -0.027 -0.027  0.048
     9     10     11     12     13     14     15
 0.085  0.063  0.018  0.001 -0.039 -0.125 -0.149
```

[42] Does it? How could you tell?



The second entry is indeed 0.624, the estimator of ϕ, which can also be accessed as follows.

```
phi = acf(MyTimeSeries.centered)$acf[2]
```

The sample variance of the centered data is:

```
(v = var(MyTimeSeries.centered))
```

```
[1] 0.004532399
```

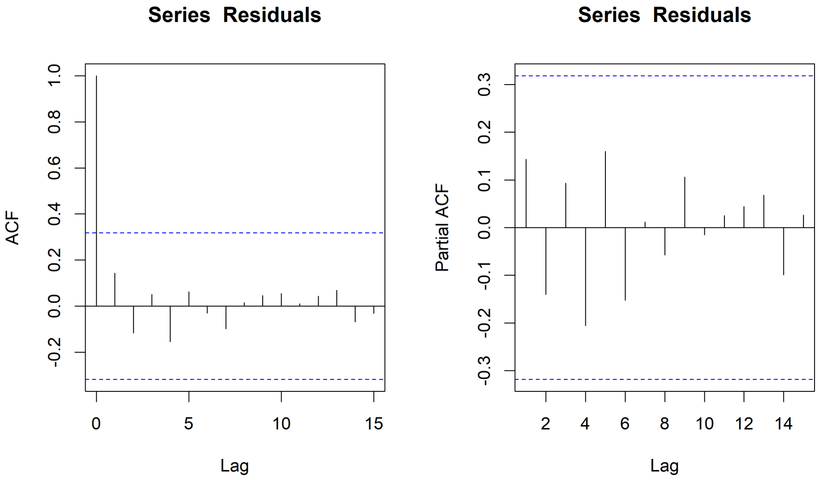
The estimator of σ_Z^2 is:

```
v-phi^2*v
```

```
[1] 0.002767046
```

How can we tell if the AR(1) fit is appropriate? We can compute the "residuals" of the $X_t - \hat{\phi}X_{t-1}$ and compare it to Z_t, which is to say an i.i.d. random variable with mean 0 and variance σ_Z^2. What do the residual time series ACF and PACF look like?

```
Residuals <- MyTimeSeries.centered[2:n] -
            phi*MyTimeSeries.centered[1:(n-1)]
par(mfrow=c(1,2))
acf(Residuals)
pacf(Residuals)
```



It certainly seems as though there is little dependence left in the residuals time series. We can apply the Ljung-Box test (which we will discuss in Section 9.6).

```
Box.test(Residuals,type="Ljung",lag=1,fitdf=1)
```

```
Box-Ljung test

data:  Residuals
X-squared = 30.799, df = 1, p-value = 2.862e-08
```

The outcome is compatible with the notion that the residuals are i.i.d. random variables.

We can also extract the residuals directly.

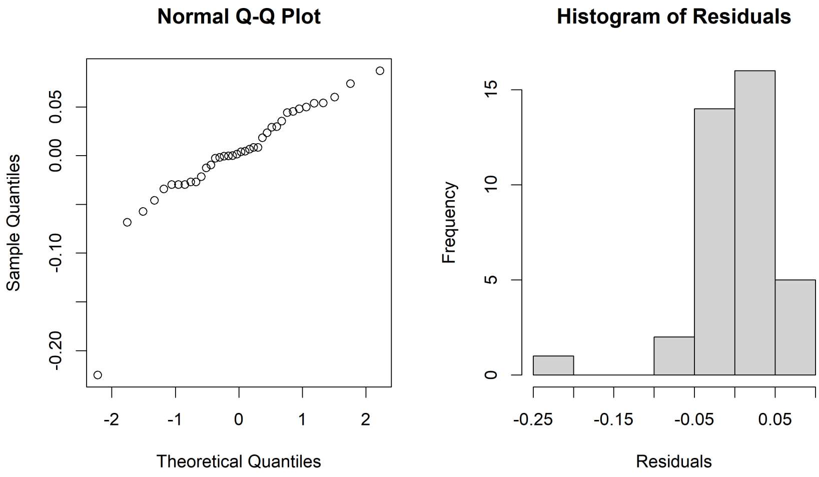
```
fit.ar$resid;
```

```
 [1]            NA  3.887364e-03  8.700283e-02  8.415571e-03
 [5]  1.833739e-02  4.546024e-02 -2.249571e-01 -2.963355e-02
 [9]  2.332064e-02 -3.416757e-02 -4.645337e-04 -2.963498e-02
[13] -2.659019e-03  8.326911e-05 -1.243842e-02 -2.978541e-02
[17] -2.692054e-02  1.589073e-03  6.651807e-03 -4.574392e-02
[21] -9.575653e-03  8.526158e-03  2.929546e-02 -1.888002e-03
[25]  4.617796e-03  4.978927e-02  5.405892e-02  3.551993e-02
[29]  7.382926e-02  2.972301e-02 -5.720633e-02 -6.845055e-02
[33] -2.147845e-02  4.429304e-02  4.800134e-02 -3.084927e-04
[37] -2.693691e-02  6.015361e-02  5.375058e-02
```

Note that this produces one "NA", as the first residual corresponds to $X_1 - \hat{\phi} X_0$, but X_0 does not exist in the original stationary time series.

The normality of the residuals (as well as their mean) can be visually assessed as follows.

```
par(mfrow=c(1,2))
qqnorm(Residuals);
hist(Residuals)
```



There are some off-the-beaten-track values, but for the most part, the data is compatible with the idea of the residuals being normally distributed, with mean 0 and variance $\widehat{\sigma}^2_Z$.

We can predict the next value of `MyTimeSeries`, and get the MSPE and its prediction interval as follows.

```
(prediction.next <- mean*(1-phi) + phi*MyTimeSeries[n])
(MSPE = (v-phi^2*v))
```

```
[1] 0.06405712
[1] 0.002767046
```

MSPE can also be obtained by typing `fit.ar$var.pred` at the prompt.

```
alpha=0.05
quantile = qnorm(1-alpha/2)
c(prediction.next - quantile*sqrt(MSPE),
  prediction.next + quantile*sqrt(MSPE))
```

```
[1] -0.03904232  0.16715655
```

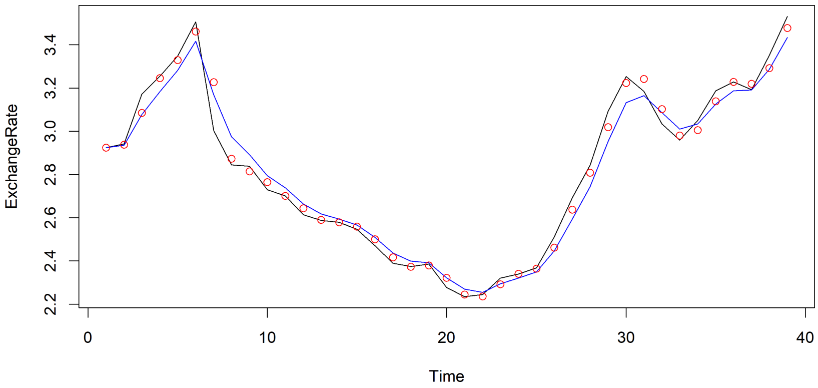
But to make a prediction in the original data, we need to take the last value in the smoothed time series and add the prediction for the stationary component; this serves as the prediction of the next observation for the original time series.

```
(Prediction.Exchange.Rate.next <-
    ExchangeRate.smoothed[n] + prediction.next)
```

```
[1] 3.497661
```

We can also determine the quality of the model fit by "predicting" past values of the original time series using the same process as above (black: original; blue: smoothed model: red: predictions).

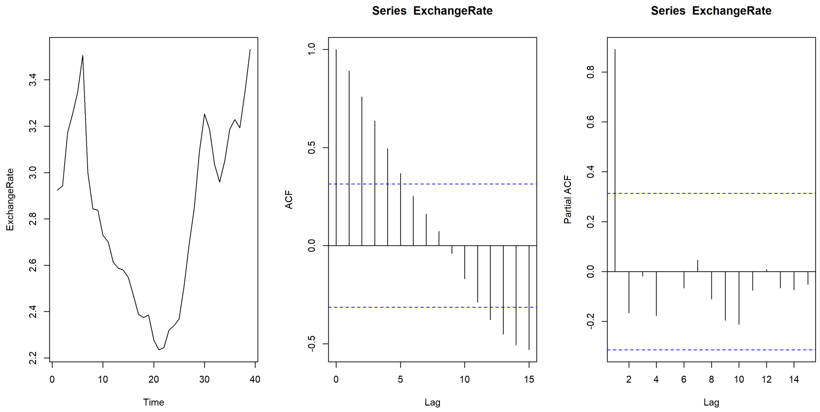
```
prediction <- mean*(1-phi) + phi*(MyTimeSeries)
prediction <- c(MyTimeSeries[1],prediction[1:n-1])
Prediction.Exchange.Rate <- ExchangeRate.smoothed +
        prediction[1:n]

par(mfrow=c(1,1))
plot.ts(ExchangeRate)
points(ExchangeRate.smoothed,type="l",col="blue")
points(Prediction.Exchange.Rate,type="p",col="red")
(Squared.Error =
    sum((Prediction.Exchange.Rate - ExchangeRate)^2))
```



```
[1] 0.1020082
```

What happens if we ignore the non-stationary behaviour and work on the original data itself instead of the stationary component? The Yule-Walker method says the data follows an AR(1) model, but with different $\widehat{\phi}$ and $\widehat{\sigma}_X^2$ values.

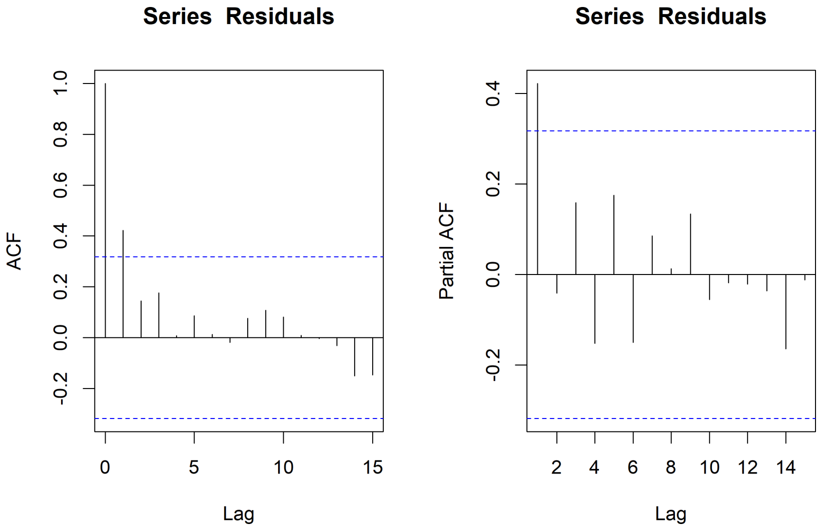
```
par(mfrow=c(1,3))
plot.ts(ExchangeRate)
acf(ExchangeRate)
pacf(ExchangeRate)

mean = mean(ExchangeRate)
ExchangeRate.centered = ExchangeRate - mean(ExchangeRate);
(fit.ar <- ar(ExchangeRate.centered,method="yule-walker"))
```



```
Coefficients:
     1
0.8903

Order selected 1  sigma^2 estimated as  0.03125
```

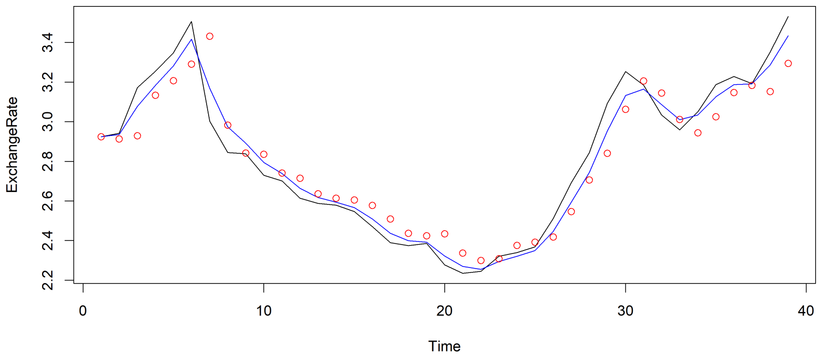
This fit's residuals do not appear to form an i.i.d. sequence.

```
phi = fit.ar$ar
Residuals <- ExchangeRate.centered[2:n] -
    phi*ExchangeRate.centered[1:(n-1)]

par(mfrow=c(1,2))
acf(Residuals)
pacf(Residuals)
```



Note, in particular, the large value of $\widehat{\rho}_X(1) \approx 0.5$. The fitted AR(1) model is the best of the AR models for the data, but it is unlikely to be correct. Nothing is stopping us from predicting new values on the (false) assumption that it was correct, unfortunately.

```
prediction <- mean*(1-phi) + phi*ExchangeRate
prediction <- c(ExchangeRate[1],prediction[1:n-1])
Prediction.Exchange.Rate.Wrong <- prediction[1:n]

par(mfrow=c(1,1))
plot.ts(ExchangeRate)
points(ExchangeRate.smoothed,type="l",col="blue")
points(Prediction.Exchange.Rate.Wrong,type="p",col="red")
(Squared.Error.Wrong =
    sum((Prediction.Exchange.Rate.Wrong-ExchangeRate)^2))
```



```
[1] 0.7490375
```

The predictions are clearly not as accurate as they were in our first attempt at analyzing the data – the squared error is seven times larger now than it was then.[43]

9.5 Estimation of ARMA Models

Let's assume that we have observations $\{X_1, \ldots, X_n\}$ from a time series and that we have also identified that a model ARMA(p, q) from which they could conceivably arise. How can we best estimate the parameters $\phi_1, \ldots, \phi_p$ and/or $\theta_1, \ldots, \theta_q$?

9.5.1 Mean: I.I.D. Case

Assume first that $X_1, \ldots, X_n$ are i.i.d. In practice, the mean of such a sequence is not typically 0. We estimate $\mu \equiv \mathrm{E}[X_t]$ by the **method of moments**, using the sample mean $\overline{X}$:

$$\mathrm{E}[\overline{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}n\mu = \mu.$$

Using the **independence** of the X_t, we have:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{\gamma_X(0)}{n}.$$

[43] This example highlights the importance of **understanding** the process; it is not sufficient to know how to produce new predictions from a time series data – we also need to know not to apply the procedure when the time series is not stationary, or when the model is a poor fit to the data.

This computation leads to the **Central Limit Theorem**.

Lemma: assume that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $\gamma_X(0)$. Then

$$\sqrt{n}\left\{\frac{\overline{X}-\mu}{\sqrt{\gamma_X(0)}}\right\} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

that is

$$\lim_{n\to\infty} P\left(\sqrt{n}\left\{\frac{\overline{X}-\mu}{\sqrt{\gamma_X(0)}}\right\} \leq x\right) = \Phi(x),$$

where Φ is the standard normal **cumulative distribution function**.

This allows us to construct a **95% confidence interval** for the mean μ:

$$\text{C.I.}(\mu; 0.95) \equiv \left(\overline{X} - 1.96\frac{\sqrt{\gamma_X(0)}}{\sqrt{n}}, \overline{X} + 1.96\frac{\sqrt{\gamma_X(0)}}{\sqrt{n}}\right).$$

This confidence interval involves the unknown $\gamma_X(0)$, which can be estimated with the sample variance.

9.5.2 Mean: Time Series

When the time series $\{X_1, \ldots, X_n\}$ does not consist of i.i.d. random variables but arises from a stationary time series, the estimate for μ remains valid, but the variance computation has to be modified.

Instead, we have

$$\begin{aligned}
\text{Var}(\overline{X}) &= \text{Cov}\left(\overline{X}, \overline{X}\right) = \text{Cov}\left(\frac{X_1+\cdots+X_n}{n}, \frac{X_1+\cdots+X_n}{n}\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}(X_i, X_j) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j)\\
&= \frac{1}{n^2}\sum_{h=-(n-1)}^{n-1}(n-|h|)\gamma_X(h) = \frac{1}{n^2}\sum_{h=-n}^{n}(n-|h|)\gamma_X(h)\\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|).
\end{aligned}$$

As an illustration, assume that $n = 3$. Then

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) &= 3\gamma_X(0) + 2\gamma_X(1) + 2\gamma_X(-1) + \gamma_X(2) + \gamma_X(-2)\\
&= \sum_{h=-2}^{2}(3-|h|)\gamma_X(h).
\end{aligned}$$

Assume now that $\gamma_X(|h|) \to 0$ as $|h| \to \infty$. Then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|) = \lim_{n\to\infty}\frac{1}{n}\sum_{h=-n}^{n}\gamma_X(|h|) = 0,$$

and

$$\begin{aligned}\lim_{n\to\infty} n\mathrm{Var}(\overline{X}) &= \lim_{n\to\infty} n\frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|)\\ &= \lim_{n\to\infty}\sum_{h=-n}^{n}\gamma_X(|h|) = \sum_{h=-\infty}^{\infty}\gamma_X(|h|) = \gamma_X(0)+2\sum_{h=1}^{\infty}\gamma_X(h)\end{aligned}$$

as long as $\{X_t\}$ is **short-range dependent** ($\Sigma_{-\infty}^{\infty}|\gamma_X(|h|)| < \infty$).

This computation is one of the main steps to establish the Central Limit Theorem in the general case.

Lemma: assume that $X_1,\ldots,X_n$ is a stationary short-range dependent time series with mean μ, variance $\gamma_X(0)$, and covariance function $\gamma_X(h)$. Then

$$\sqrt{n}\left\{\frac{\overline{X}-\mu}{\nu}\right\}\xrightarrow{\text{d}} N(0,1),$$

that is

$$\lim_{n\to\infty} P\left(\sqrt{n}\left\{\frac{\overline{X}-\mu}{\nu}\right\}\leq x\right) = \Phi(x),$$

where Φ is as above, and

$$\nu^2 = \gamma_X(0)+2\sum_{h=1}^{\infty}\gamma_X(h).$$

This allows us to construct a **95% confidence interval for the mean** μ:

$$\text{C.I.}(\mu;0.95) \equiv \left(\overline{X}-1.96\frac{\nu}{\sqrt{n}},\overline{X}+1.96\frac{\nu}{\sqrt{n}}\right).$$

This confidence interval involves the unknown ν.

Example Recall that the AR(1) model is $X_t = \phi X_{t-1}+Z_t$, with the usual assumptions on Z_t.[44] Then $\gamma_X(h) = \sigma_Z^2\frac{\phi^h}{1-\phi^2}$, and so

$$\nu^2 = \gamma_X(0)+2\sum_{h=1}^{\infty}\gamma_X(h) = \sigma_Z^2\frac{1}{1-\phi^2}+2\sigma_Z^2\frac{1}{1-\phi^2}\frac{\phi}{1-\phi} = \sigma_Z^2\frac{1}{(1-\phi)^2}.$$

9.5.3 Yule-Walker Estimators

The method we present now has similarities with Yule-Walker forecasting; it works quite well for AR(p) models.

Assume a stationary and causal AR(1) model: $X_t = \phi X_{t-1}+Z_t$, where $|\phi|<1$, $\mathrm{E}[Z_t]\equiv 0$, and $\mathrm{Var}(Z_t)\equiv\sigma_Z^2$. Multiply both sides of the equation, once by X_{t-1} and another time by X_t, to get

$$\begin{aligned}X_tX_{t-1} &= \phi X_{t-1}X_{t-1}+Z_tX_{t-1},\\ X_t^2 &= \phi X_tX_{t-1}+X_tZ_t.\end{aligned}$$

[44] In order to obtain the linear representation of the model, we need to have $\mu = 0$. If the data is not centered ($\mu\neq 0$), consider instead the shifted model

$$X_t-\mu = \phi(X_{t-1}-\mu)+Z_t.$$

The stationary solution will then be

$$X_t = \mu+\sum_{j=0}^{\infty}\phi^jZ_{t-j}.$$

We apply the expectation operator on both of these new equations (recall that $\mathrm{E}[X_t] = 0$ and that X_{t-1} is independent of Z_t because the time series is causal) to obtain:

$$\begin{aligned} \gamma_X(1) &= \phi\gamma_X(0) + 0, \\ \gamma_X(0) &= \phi\gamma_X(1) + \mathrm{E}[X_t Z_t]. \end{aligned}$$

That last term evaluates to

$$\mathrm{E}[X_t Z_t] = \mathrm{E}[(\phi X_{t-1} + Z_t)Z_t] = \phi\mathrm{E}[X_{t-1}Z_t] + \mathrm{E}[Z_t^2] = \sigma_Z^2.$$

Hence, the system reduces to:

$$\begin{aligned} \gamma_X(0)\phi &= \gamma_X(1) \\ \sigma_Z^2 &= \gamma_X(0) - \phi\gamma_X(1). \end{aligned}$$

Now, consider $p = 2$: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Multiply both sides of that equation, once each by by X_{t-2}, X_{t-1}, and X_t, to obtain:

$$\begin{aligned} X_t X_{t-2} - \phi_1 X_{t-1}X_{t-2} - \phi_2 X_{t-2}^2 &= Z_t X_{t-2}, \\ X_t X_{t-1} - \phi_1 X_{t-1}^2 - \phi_2 X_{t-2}X_{t-1} &= Z_t X_{t-1}, \\ X_t^2 - \phi_1 X_{t-1}X_t - \phi_2 X_{t-1}X_t &= X_t Z_t. \end{aligned}$$

We once again apply the expectation operator on each of these new equations to obtain:

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

As in section 9.4.1 we consider the variance-covariance matrix

$$\Gamma_p = [\gamma_X(i-j)]_{i,j=1}^p,$$

and the vectors

$$\boldsymbol{\phi}_p = \left(\phi_1, \dots, \phi_p\right)^\top \quad \text{and} \quad \boldsymbol{\gamma}(p;1) = \left(\gamma_X(1), \dots, \gamma_X(p)\right)^\top.$$

For $p = 1$, $\Gamma_1 = \gamma_X(0)$; for $p = 2$,

$$\Gamma_2 = \begin{pmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(-1) & \gamma_X(0) \end{pmatrix} = \begin{pmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{pmatrix}.$$

We can thus re-write the AR(1) and AR(2) systems above as:

$$\Gamma_p\boldsymbol{\phi}_p = \boldsymbol{\gamma}(p;1), \qquad \sigma_Z^2 = \gamma_X(0) - \boldsymbol{\phi}_p^\top\boldsymbol{\gamma}(p;1).$$

Equivalently, we obtain the **Yule-Walker equations**

$$\boldsymbol{\phi}_p = \Gamma_p^{-1}\boldsymbol{\gamma}(p;1), \qquad \sigma_Z^2 = \gamma_X(0) - \boldsymbol{\phi}_p^\top\boldsymbol{\gamma}(p;1),$$

which are very similar to the Yule-Walker forecast equations.[45] It is not hard to see that the equations hold for a general AR(p).

[45] Note that they do involve unknown autocovariances.

We can combine them with the method of moments,[46] to obtain the **Yule-Walker estimators**:

$$\widehat{\boldsymbol{\phi}}_p = \widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}(p;1), \qquad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\boldsymbol{\phi}}_p^\top \widehat{\boldsymbol{\gamma}}(p;1),$$

where $\widehat{\Gamma}_p$ and $\widehat{\boldsymbol{\gamma}}(p;1)$ are obtained by substituting γ_X by $\widehat{\gamma}_X$.

Theorem: for a large-enough sample size n, the Yule-Walker estimators are approximately normal, with

$$\widehat{\boldsymbol{\phi}}_p \sim \mathcal{N}\left(\boldsymbol{\phi}_p, \frac{1}{n}\sigma_Z^2\Gamma_p^{-1}\right).$$

In particular, for $p = 1$,

$$\widehat{\phi} \sim \mathcal{N}\left(\phi, \frac{1}{n}\sigma_Z^2\gamma_X^{-1}(0)\right).$$

That is, $\text{Var}(\widehat{\phi}) \sim \frac{1}{n}\sigma_Z^2\gamma_X^{-1}(0)$.

Confidence interval for AR(1) The **theoretical** confidence interval for the parameter ϕ of AR(1) is

$$\text{C.I.}_\alpha(\phi) \equiv \widehat{\phi} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sigma_Z\sqrt{\gamma_X^{-1}(0)},$$

where $z_{\alpha/2}$ is the standard normal quantile. Since σ_Z^2 and $\gamma_X(0)$ are unknown, we replace them with estimators to obtain the **empirical** (practical) confidence interval

$$\text{C.I.}_\alpha(\phi) \approx \widehat{\phi} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\widehat{\sigma}_Z\sqrt{\widehat{\gamma}_X^{-1}(0)},$$

where

$$\widehat{\phi} = \frac{\widehat{\gamma}_X(1)}{\widehat{\gamma}_X(0)} \quad \text{and} \quad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\phi}\widehat{\gamma}_X(1).$$

Confidence interval for AR(2) The limiting variance-covariance matrix for the Yule-Walker estimators $\widehat{\phi}_1$, $\widehat{\phi}_2$ is

$$\sigma_Z^2\Gamma_2^{-1} = \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix}.$$

Indeed, we have

$$\Gamma_2 = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{bmatrix} \implies \Gamma_2^{-1} = \frac{1}{\gamma_X^2(0)-\gamma_X^2(1)}\begin{bmatrix} \gamma_X(0) & -\gamma_X(1) \\ -\gamma_X(1) & \gamma_X(0) \end{bmatrix}.$$

Previously, we saw that

$$\gamma_X(1) = \sigma_Z^2\frac{\phi_1}{(1+\phi_2)\left\{(1-\phi_2)^2-\phi_1^2\right\}} \quad \text{and} \quad \gamma_X(0) = \sigma_Z^2\frac{1-\phi_2}{(1+\phi_2)\left\{(1-\phi_2)^2-\phi_1^2\right\}}.$$

Substituting these in the expression for Γ_2^{-1} yields the desired result.

[46] We simply replace the mean with the sample mean and the autocovariances with the sample autocovariances.

In particular, $\text{Var}(\widehat{\phi}_1) \sim \frac{1}{n}(1-\phi_2^2)$ and $\text{Var}(\widehat{\phi}_2) \sim \frac{1}{n}(1-\phi_2^2)$. Consequently,

$$\text{C.I.}_\alpha(\phi_1) \equiv \widehat{\phi}_1 \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{1-\widehat{\phi}_2^2} \quad \text{and} \quad \text{C.I.}_\alpha(\phi_2) \equiv \widehat{\phi}_2 \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{1-\widehat{\phi}_2^2},$$

where $\widehat{\phi}_1$ and $\widehat{\phi}_2$ are obtained from the Yule-Walker estimators.

9.5.4 Example

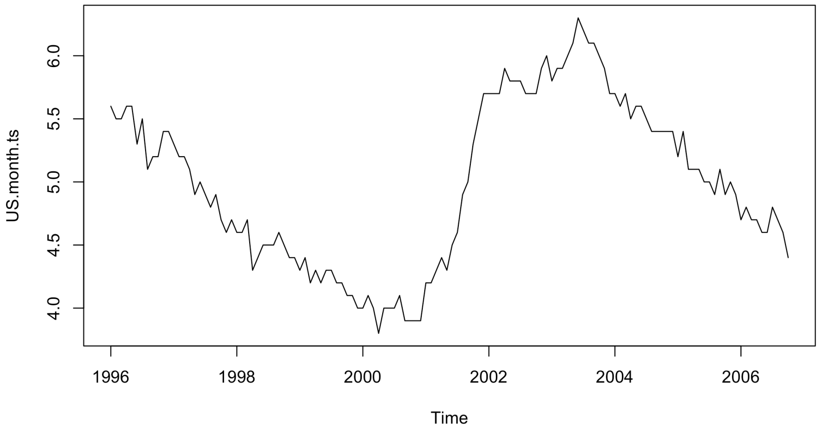
We illustrate this last concept with a simple example.

US Unemployment Data The United States' monthly unemployment rate starting with January 1996 is collected in `USunemp.txt` [3].

```
US.month <- c(5.6,5.5,5.5,5.6,5.6,5.3,5.5,5.1,5.2,5.2,
              5.4,5.4,5.3,5.2,5.2,5.1,4.9,5.0,4.9,4.8,
              4.9,4.7,4.6,4.7,4.6,4.6,4.7,4.3,4.4,4.5,
              4.5,4.5,4.6,4.5,4.4,4.4,4.3,4.4,4.2,4.3,
              4.2,4.3,4.3,4.2,4.2,4.1,4.1,4.0,4.0,4.1,
              4.0,3.8,4.0,4.0,4.0,4.1,3.9,3.9,3.9,3.9,
              4.2,4.2,4.3,4.4,4.3,4.5,4.6,4.9,5.0,5.3,
              5.5,5.7,5.7,5.7,5.7,5.9,5.8,5.8,5.8,5.7,
              5.7,5.7,5.9,6.0,5.8,5.9,5.9,6.0,6.1,6.3,
              6.2,6.1,6.1,6.0,5.9,5.7,5.7,5.6,5.7,5.5,
              5.6,5.6,5.5,5.4,5.4,5.4,5.4,5.4,5.2,5.4,
              5.1,5.1,5.1,5.0,5.0,4.9,5.1,4.9,5.0,4.9,
              4.7,4.8,4.7,4.7,4.6,4.6,4.8,4.7,4.6,4.4)
```

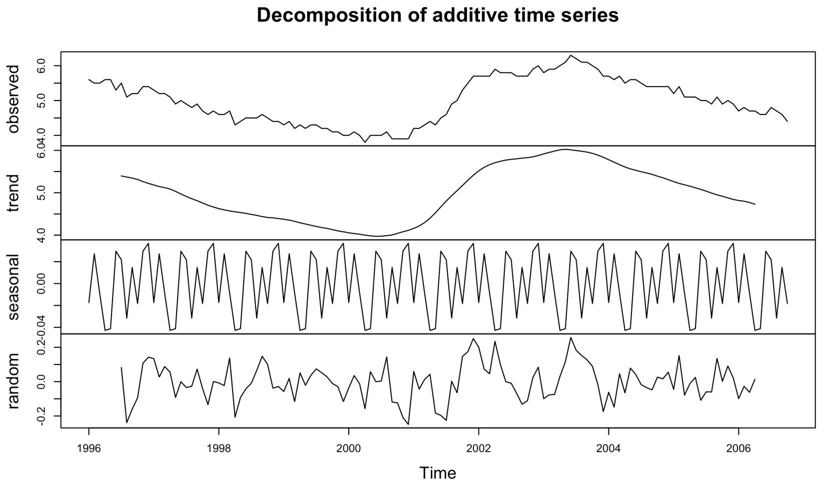
We put the data in a `ts` object and plot the data.

```
US.month.ts <- ts(US.month,start=c(1996,1), freq=12)
plot.ts(US.month.ts)
```



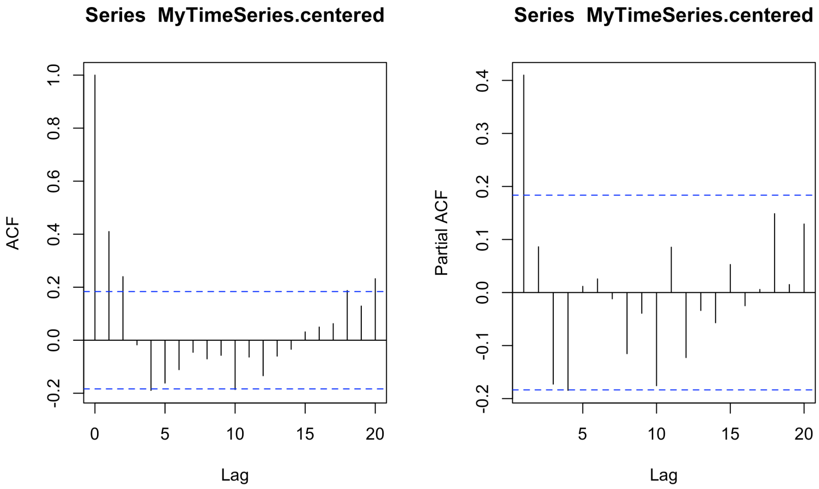
The time series is clearly not stationary, so we decompose it.

```
plot(decompose(US.month.ts))
```



We recover the stationary part from this decomposition and analyse it as below.[47]

```
Stationary <- decompose(US.month.ts)$random
MyTimeSeries = Stationary[7:120]
mean = mean(MyTimeSeries);
MyTimeSeries.centered = MyTimeSeries-mean
par(mfrow=c(1,2))
acf(MyTimeSeries.centered)
pacf(MyTimeSeries.centered)
```



The ACVF/ACF has non-zero values at various lags h (outside the band); the PACF has all zero values for $h > 1$ (inside the band); the eye test suggests an AR(1) model.

But a formal test (using the Yule-Walker) method suggests instead that the order of the model is more likely to be $p = 4$.

[47] The default smoother for the `decompose()` function is a moving average.

```
n = length(MyTimeSeries)
fit.ar <- ar(MyTimeSeries.centered,method="yule-walker")
fit.ar$order
```

```
[1] 4
```

The Yule-Walker estimates for the coefficients $\phi_1, \phi_2, \phi_3, \phi_4$ and for the random component variance σ_Z^2 are given by:

```
fit.ar$ar
fit.ar$var.pred
```

```
[1] 0.3576   0.1788  -0.1008  -0.1845
[1] 0.009106
```

We compute the limiting variance covariance matrix $\sigma_Z^2\Gamma_4^{-1}$ as follows.

```
rho = acf(MyTimeSeries.centered)$acf
gamma.0 = var(MyTimeSeries.centered)
sigma.2.Z = fit.ar$var.pred
gamma.h = rho * gamma.0
Gamma.4 = matrix(c(gamma.h[1],gamma.h[2],gamma.h[3],gamma.h[4],
         gamma.h[2],gamma.h[1],gamma.h[2],gamma.h[3],
         gamma.h[3],gamma.h[2],gamma.h[1],gamma.h[2],
         gamma.h[4],gamma.h[3],gamma.h[2],gamma.h[1]),4,4)
Gamma.4.inv = solve(Gamma.4)
(limit.V_CV = sigma.2.Z*Gamma.4.inv)
```

```
           [,1]       [,2]       [,3]       [,4]
[1,]  1.0014029 -0.3900340 -0.1511252  0.1729498
[2,] -0.3900340  1.1234465 -0.3050721 -0.1511252
[3,] -0.1511252 -0.3050721  1.1234465 -0.3900340
[4,]  0.1729498 -0.1511252 -0.3900340  1.0014029
```

Note that we can obtain the matrix directly from the `fit.ar` object.

```
(n-1)*fit.ar$asy.var.coef
```

Finally, we simply apply the formulas to obtain approximate 95% confidence intervals on the AR(4) coefficients.

```
rbind(fit.ar$ar - 1.96/sqrt(n)*sqrt(diag(limit.V_CV)),
      fit.ar$ar + 1.96/sqrt(n)*sqrt(diag(limit.V_CV)))
```

```
          [,1]        [,2]        [,3]          [,4]
[1,] 0.1739213 -0.01581274 -0.29541378 -0.3682125028
[2,] 0.5413204  0.37333082  0.09372978 -0.0008134319
```

9.6 Diagnostic Tests

Assume that an AR(1) model $X_t = \phi X_{t-1} + Z_t$ is fit to the data, i.e., we estimate ϕ by $\widehat{\phi}$ and σ^2_Z by $\widehat{\sigma}^2_Z$. We can now compute the time series of **residuals**

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

Note that $\widehat{Z}_t \neq Z_t$, in general, but we would expect them to be near one another if the fit is good. As such, the properties of Z_t should be similar to those of $\widehat{Z}_t$.

It is important to ensure that the model is an **adequate fit** to the data – in particular, the residuals should not exhibit significant autocorrelations at lags $|h| \geq 1$.

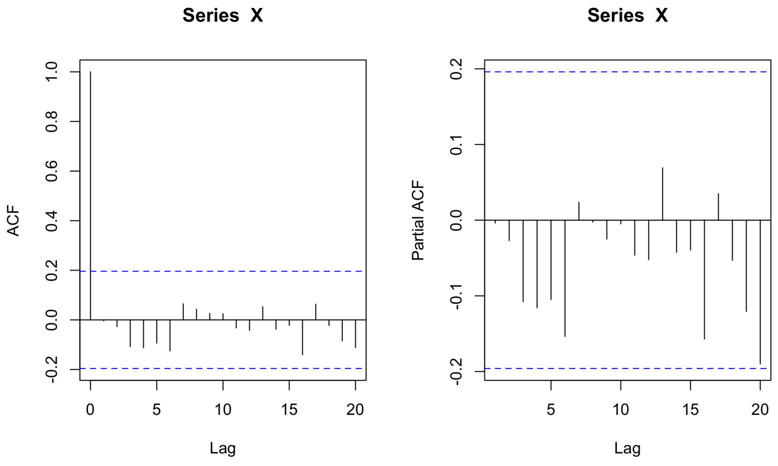
If the random variables $\widehat{Z}_t$ are i.i.d., then the correlations $\rho_X(|h|) = 0$ at any lag $h \neq 0$ zero. However, the sample correlations are typically not zero, since there usually are random fluctuations in the data. In general, for large n, the sample correlation at any lag is normally distributed with mean zero and variance $1/n$. This provides a 95% confidence interval for the sample autororrelations: $\pm 1.96/\sqrt{n}$.[48]

White Noise The 95% threshold for a white noise time series with $\mu = 0$ and $\sigma^2 = 1$, with $n = 100$ observations is computed below.

```
n = 100
set.seed(1)
X = rnorm(n)
(threshold = 1.96/sqrt(n))
```

```
[1] 0.196
```

```
par(mfrow=c(1,2))
acf(X)
pacf(X)
```



[48] This corresponds to the blue lines seen on the ACF plot. Whenever the sample ACF is within the confidence intervals, the rule-of-thumb is to treat the corresponding auto-correlation as zero.

9.6.1 Ljung-Box Test

That is not the only approach, however. Let h be a positive integer (the **lag**) and define

$$Q_h = n \sum_{j=1}^{h} \frac{\widehat{\gamma}_X(j)}{\widehat{\gamma}_X(0)}.$$

Under the null hypothesis that the residuals are i.i.d. , the statistic Q_h has a χ^2 distribution with h degrees of freedom. A large value of Q_h suggests that the sample autocorrelations are too large for the data to arise from the draw of an i.i.d. sequence. We would therefore reject the i.i.d. hypothesis at confidence level α if $Q_h > \chi^2_{1-\alpha}(h)$.

White Noise We can conduct the Ljung-Box test on the white noise time series from the previous section, with $h = 2$, say.

```
Box.test(X,type="Ljung",lag=2,fitdf=0)
```

```

        Box-Ljung test

data:  X
X-squared = 0.077367, df = 2, p-value = 0.9621
```

Thus, we conclude that the data is compatible with X being i.i.d., at confidence level $\alpha = 0.05$

AR(1) Model This time, we simulate an auto-regressive model (so the time series not i.i.d.) and repeat the procedure.

```
set.seed(1)
MyTimeSeries = arima.sim(model=list(ar=c(0.8)),
                         n=1000,rand.gen=rnorm)
Box.test(MyTimeSeries,type="Ljung",lag=2,fitdf=0)
```

```

        Box-Ljung test

data:  MyTimeSeries
X-squared = 904.66, df = 2, p-value < 2.2e-16
```

We see that the i.i.d. assumption is correctly rejected.

The Ljung-Box test is applied to the residuals. The parameter `fitdf` is the number of the parameters that need to be estimated. In an ARMA(p, q), model, it is $p + q$.[49]

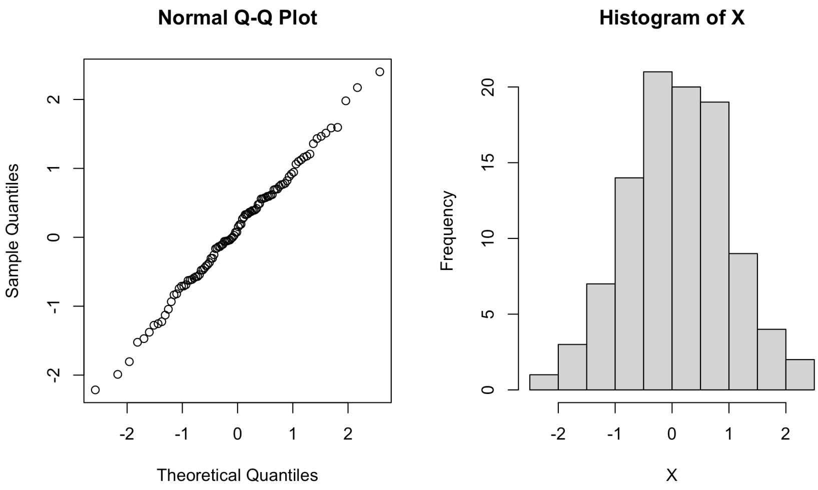
When we reject the hypothesis that the residuals are i.i.d., we are claiming that the fitted ARMA(p, q) model is incorrect.[50] If the test results are compatible with the null hypothesis, we must also verify that the residuals are normally distributed, however, either by plotting a Q-Q plot or a histogram.

In the first example, the time series $\{X_t\}$ is normally distributed.

[49] Be careful! Here, we are testing whether the sequence `MyTimeSeries`, which we know to be AR(1), could be white noise (i.i.d.), which is why we use `fitdf=0`. That is, we are assuming that it is a time series of residuals that arose naturally, not as a result of having fit an ARMA(p, q) model to the data. The `lag` parameter represents the positive value h.

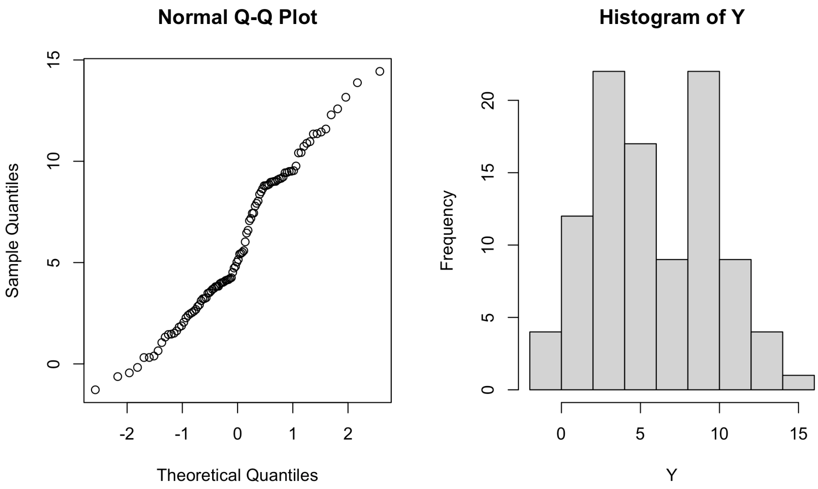
[50] We must thus remove $p+q$ degrees of freedom from h, since we had to estimate $p + q$ parameters from the data before obtaining the residual time series.

```
par(mfrow=c(1,2))
qqnorm(X)
hist(X)
```



In the second case, the time series $\{Y_t\}$ is a random walk, and it is not normally distributed.

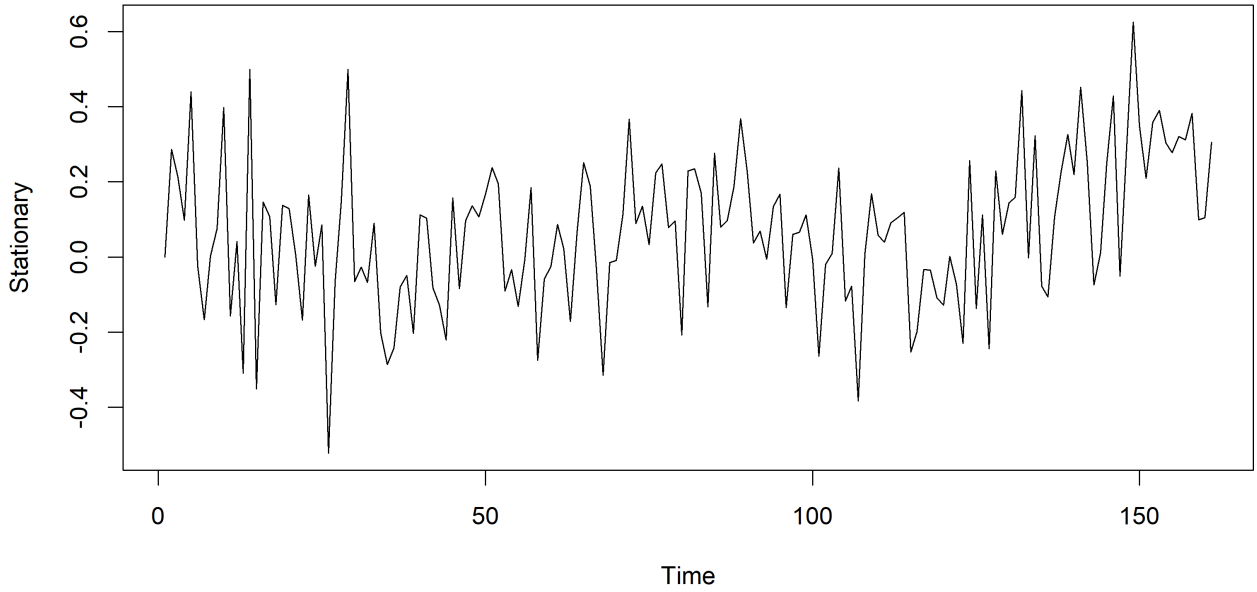
```
Y = cumsum(X)
par(mfrow=c(1,2))
qqnorm(Y)
hist(Y)
```



9.6.2 Example: Temperature

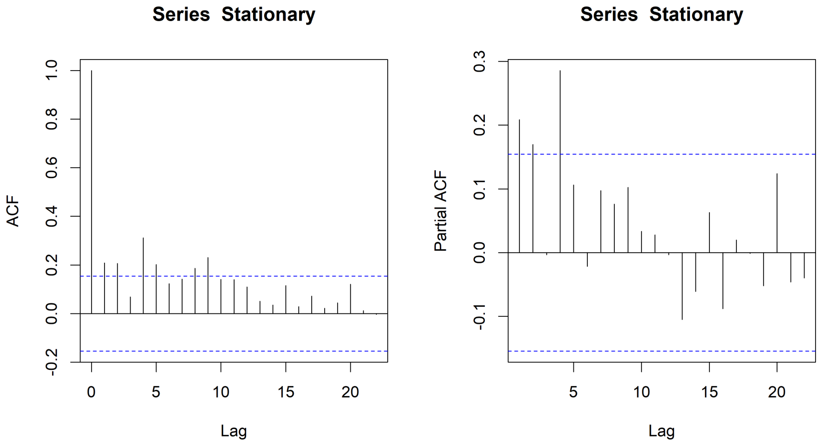
We consider the temperature data from page 499; it is clearly not stationary, so we conduct exponential smoothing on it, with smoothing parameter 0.1, yielding the time series `MySmoothedTS1`, which is then centered.

```
Stationary = Temperature - MySmoothedTS1
plot.ts(Stationary, type="l")
```



This time series certainly appears stationary.[51] Could it arise from an ARMA(p, q) model? We plot its ACF and PACF.

```
par(mfrow=c(1,2))
acf(Stationary); pacf(Stationary)
```



AR(4) seems like a reasonable model;[52] Yule-Walker agrees.

```
(fit.ar.yw <- ar(Stationary,method="yule-walker"))
```

```
Coefficients:
      1        2        3        4
 0.1745   0.1218  -0.0529   0.2855

Order selected 4  sigma^2 estimated as  0.03412
```

We compute the residuals for which $\widehat{Z}_t = \widehat{\phi}(B)X_t$ is defined.

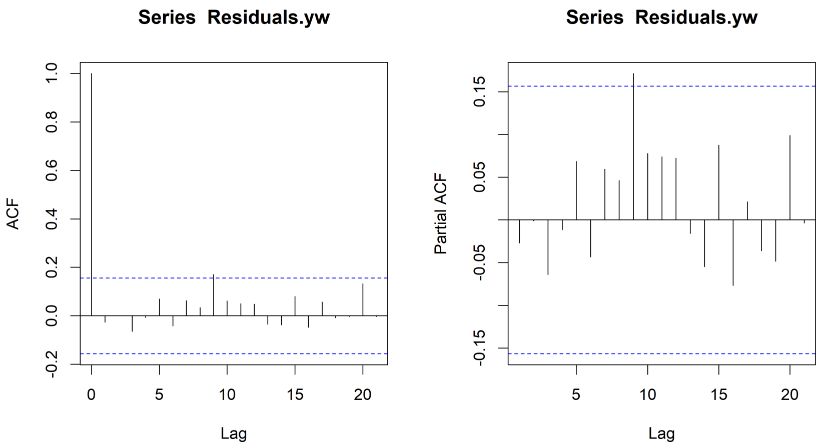
[51] Although there is a bit of growth near the end.

[52] Be sure to understand why!

```
phi.yw = fit.ar.yw$ar
n = length(Stationary)
Residuals.yw <- fit.ar.yw$resid
Residuals.yw = na.omit(Residuals.yw)
```

The ACF and PACF of the obtained residuals are as follows.

```
par(mfrow=c(1,2))
acf(Residuals.yw)
pacf(Residuals.yw)
```



There is no dependence left in the residuals (although you can argue that there is a significant lag at 9); the fit seems appropriate.

We can conduct the Box-Ljung test with $h = 5 > 4 = p + q$, say.

```
Box.test(Residuals.yw,type="Ljung",lag=5,fitdf=4)
```

```
Box-Ljung test

data:  Residuals.yw
X-squared = 1.5724, df = 1, p-value = 0.2099
```

The $p-$value is small, but not that small... does the value of h matter? What if we used $h = 4$, instead?

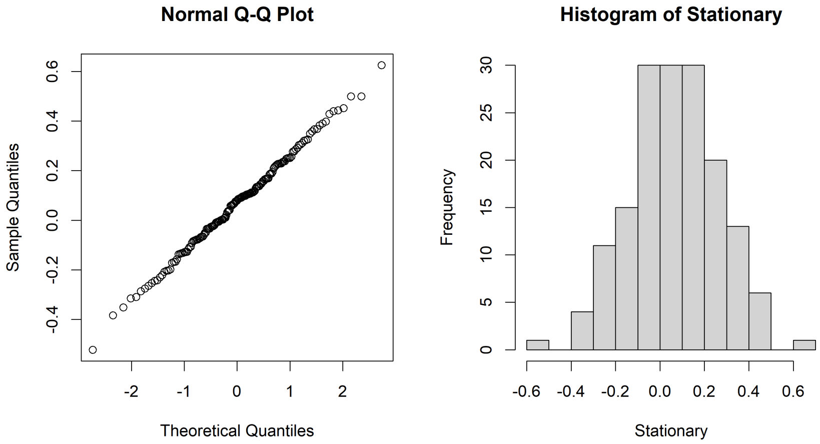
```
Box.test(Residuals.yw,type="Ljung",lag=4,fitdf=4)
```

```
Box-Ljung test

data:  Residuals.yw
X-squared = 0.79007, df = 0, p-value < 2.2e-16
```

The $p-$value is indeed much smaller than 0.05, but it's not clear how the test implementation handles the case where $h = p + q$.

Either way, we should study the normality of the residuals visually.

```
par(mfrow=c(1,2))
qqnorm(Stationary)
hist(Stationary)
```



So, what do you think? We will return this example in the next section.

One thing to note is that the Box-Ljung test is not unanimously favoured by practitioners: see the Breusch-Godfrey test for an alternative.

9.7 Maximum Likelihood Estimation

We start with a brief refresher on the topic.

9.7.1 I.I.D. Random Variables

Assume that the random variables $X_1, \ldots, X_n$ are i.i.d. with a known probability density function $f_X(x; \theta)$. The objective of **maximum likelihood estimation** (MLE) is to find the parameter θ that best fits the observed data, in the MLE sense.[53]

The **likelihood function** is

$$L(\theta) = L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} f_X(X_i; \theta).$$

The **log-likelihood function** is $\ell(\theta) = \log L(\theta) = \ln L(\theta)$. The **maximum likelihood estimator** $\widehat{\theta}_{\text{MLE}}$ is a parameter value (often unique, for commonly-used f, but it also depends on the observed data) satisfying

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta).$$

[53] This does not have to be a univariate problem; we might be interested in the parameter vector $\boldsymbol{\theta}$, depending on the context. The principle is the same, but we will be working with $\nabla_{\boldsymbol{\theta}}$ instead of the derivative $\frac{d}{d\theta}$.

Example: Exponential Distribution Assume that $X_1, \ldots, X_n$ is a random sample from an exponential distribution. Recall that $X \sim \text{Exp}(\beta)$, $\theta = \beta > 0$ if

$$f_X(x;\beta) = \begin{cases} \beta^{-1}\exp(-x/\beta), & x > 0; \\ 0, & x \leq 0 \end{cases}$$

The likelihood function is:

$$L(\beta) = \beta^{-n}\prod_{i=1}^{n}\exp(-X_i/\beta) = \beta^{-n}\exp\left(-\beta^{-1}\sum_{i=1}^{n}X_i\right),$$

and the log-likelihood is:

$$\ell(\beta) = -n\log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n}X_i.$$

To optimize ℓ, we must find its critical points with respect to β. There is only one such point, since

$$\frac{\partial\ell(\beta)}{\partial\beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n}X_i = 0 \implies \widehat{\beta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}.$$

Technically, this only tells us that $\overline{X}$ is a critical point of $\ell(\beta)$, not necessarily that it is a maximizer. But

$$\left.\frac{\partial^2\ell(\beta)}{\partial\beta^2}\right|_{\beta=\overline{X}} = -n\overline{X}^2 < 0,$$

so $\widehat{\beta}_{\text{MLE}} = \overline{X}$ is indeed a global maximizer, according to the second derivative test.

The sample mean is not only the MLE estimator for the Exponential distribution, however.

Example: Normal Distribution Assume that $Z_1, \ldots, Z_n$ is a i.i.d. sample from a normal distribution with mean μ and variance σ_Z^2. The likelihood function is

$$L(\mu,\sigma_Z) = \frac{1}{(\sqrt{2\pi})^n\sigma_Z^n}\exp\left(-\frac{1}{2\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu)^2\right),$$

and the log-likelihood is:

$$\ell(\mu,\sigma_Z) = -\frac{n}{2}\log(2\pi) - n\log\sigma_Z - \frac{1}{2\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu)^2.$$

We proceed as above, differentiating with respect to μ to find the critical points:

$$\frac{\partial\ell(\mu,\sigma_Z)}{\partial\mu} = -\frac{1}{\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu) = 0 \implies \widehat{\mu}_{\text{MLE}} = \overline{Z} = \frac{1}{n}\sum_{i=1}^{n}Z_i.$$

Substituting $\widehat{\mu}_{\text{MLE}} = \overline{Z}$ in L, differentiating with respect to σ_Z, setting to 0 and solving yields

$$\widehat{\sigma}^2_{Z;\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \widehat{\mu}_{\text{MLE}})^2,$$

demonstrating that the MLE estimators are not always **unbiased**.

9.7.2 Time Series Model

We now assume that $X_1, \ldots, X_n$ are observation from a stationary time series. Let $f_n(x_1, \ldots, x_n)$ be their joint density.[54] We further assume that the time series is **Gaussian** and **centered**.[55]

We introduce the following notation:

$$\mathbf{X}_n = (X_1, \ldots, X_n)^\top, \qquad \widehat{\mathbf{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)^\top, \qquad \mathbf{U}_n = (U_1, \ldots, U_n)^\top,$$

where $U_i = X_i - \widehat{X}_i$, $i = 1, \ldots, n$, are the **innovations**. Recall that $\Gamma_n = \text{E}[\mathbf{X}_n^\top \mathbf{X}_n] = [\gamma_X(i-j)]_{i,j=1}^n$ is the **variance-covariance matrix** of $\mathbf{X}_n$ (see Section 9.4.1).

The likelihood (the joint density of $X_1, \ldots, X_n$) is

$$L = \frac{1}{(2\pi)^{n/2}}\frac{1}{\det(\Gamma_n)^{1/2}}\exp\left(-\frac{1}{2}\mathbf{X}_n^\top \Gamma_n^{-1}\mathbf{X}_n\right),$$

where $\det(\Gamma_n)$ is the determinant. Note that the ACVF (and hence, also the covariance matrix Γ_n) depends on model parameters.

For example, if the model is AR(1), then $\gamma_X(h) = \sigma_Z^2\phi^h/(1-\phi^2)$. Thus, its variance-covariance matrix and the log-likelihood depend on the model parameters σ_Z, ϕ, so that we can write $L(\sigma_Z, \phi)$.

In this particular case, the MLE estimators are obtained by maximizing $L(\sigma_Z, \phi)$ with respect to σ_Z, ϕ. In the general case, there are no no explicit formulas to do so and everything must be conducted numerically (see Chapter 4).

It turns out that

$$\mathbf{X}_n^\top \Gamma_n^{-1}\mathbf{X}_n = \mathbf{U}_n^\top \mathbf{D}^{-1}\mathbf{U}_n,$$

where $\mathbf{D} = \text{diag}(v_0, \ldots, v_{n-1})$, for $v_i = \text{E}\left[\left(X_{i+1} - \widehat{X}_{i+1}\right)^2\right]$.[56] Thus, we have

$$\mathbf{X}_n^\top \Gamma_n^{-1}\mathbf{X}_n = \sum_{i=1}^{n}(X_i - \widehat{X}_i)^2/v_{i-1}.$$

Furthermore, $\det(\Gamma_n) = v_0 \cdots v_{n-1}$, and so the likelihood function takes the form

$$L = \frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{v_0 \cdots v_{n-1}}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2/v_{i-1}\right).$$

[54] No longer in the product form.

[55] This is an important assumption – we need to verify that it applies to the data of interest.

[56] See Section 9.9.3 for more details.

The form of L above can be used as long as we have formulas for $\widehat{X}_i$, even if those do not arise from the innovation algorithm.

Theorem: the MLE estimator $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ is **asymptotically normal**,[57] with mean $\boldsymbol{\theta}$ and variance $n^{-1}\text{V}(\boldsymbol{\theta})$, where $\text{V}(\boldsymbol{\theta})$ is a covariance matrix.

If the data arises from an ARMA(p, q) process, we would use the innovation algorithm to express $\widehat{X}_i$ in terms of the coefficients $\theta_1, \dots, \theta_q$, and then plug them into the likelihood function

$$L(\boldsymbol{\theta}) = L(\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_Z^2),$$

which can be maximized using the MLE approach, as above.

AR(1) Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where Z_t are i.i.d. normal random variables with mean 0 and variance σ_Z^2, starting with $t = 1$. Then $\widehat{X}_{i+1} = \phi X_i$ and $v_i = \text{E}[(X_{i+1} - \widehat{X}_{i+1})^2] = \sigma_Z^2$ for all $i = 1, \dots, n-1$. The likelihood function is thus:

$$L = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma_Z^n} \exp\left(-\frac{1}{2} \sum_{i=2}^{n} (X_i - \phi X_{i-1})^2 / \sigma_Z^2 \right).$$

Ignoring the constant term $\frac{1}{(2\pi)^{n/2}}$, the log-likelihood is

$$\ell = -n \log \sigma_Z - \frac{1}{2\sigma_Z^2} \sum_{i=2}^{n} (X_i - \phi X_{i-1})^2.$$

Hence,

$$\widehat{\phi}_{\text{MLE}} = \frac{\sum_{i=2}^{n} X_{i-1} X_i}{\sum_{i=2}^{n} X_{i-1}^2}$$

and

$$\widehat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=2}^{n} (X_i - \widehat{\phi}_{\text{MLE}} X_{i-1})^2.$$

If σ_Z is known, then we do not need to use the MLE estimator; we have $\theta = \phi$ and $\text{V}(\theta)$ becomes

$$\text{V}(\theta) = \text{V}(\phi) = \sigma_Z^2 (1 - \phi^2).$$

We note that the MLE estimator of ϕ (as well as its asymptotic variance) are the same as those obtained by the Yule-Walker procedure.

AR(p) In general, for AR(p) models, the Yule-Walker estimator and MLE of $(\phi_1, \dots, \phi_p)$ also agree; in both cases the asymptotic variance is

$$\text{V}(\phi_1, \dots, \phi_p) = \sigma_Z^2 \Gamma_p^{-1}.$$

For AR(2) we have seen that

$$\text{V}(\phi_1, \phi_2) = \begin{pmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{pmatrix}.$$

However, the MLE and Yule-Walker estimators of variance σ_Z^2 do not need to agree in general!

[57] See Section 9.9.2.

AR(p) Models (Revisited) For simplicity's sake, consider the AR(1) model $X_t = \phi X_{t-1} + Z_t$, where Z_t are i.i.d. normal with mean 0 and variance σ_Z^2.

We assume that $\mu = 0$; then,

$$L = \frac{1}{(\sqrt{2\pi})^n \sigma_Z^n} \exp\left(-\frac{1}{2\sigma_Z^2}\sum_{i=1}^{n} Z_i^2\right),$$

Since $Z_t = X_t - \phi X_{t-1}$, this transforms to

$$L(\phi, \sigma_Z) = \frac{1}{(\sqrt{2\pi})^n \sigma_Z^n} \exp\left(-\frac{1}{2\sigma_Z^2}\sum_{i=2}^{n}(X_i - \phi X_{i-1})^2\right).$$

The likelihood function now depends explicitly on ϕ and σ_Z, and we can continue as we did in the previous section (without having to use innovations).

This approach works for arbitrary AR(p) models, but not for MA(q) or general ARMA(p, q) models.

9.7.3 Order Selection

We have discussed a visual criterion to identify a time series follows a AR(p) or MA(q) model, as well as a formal approach (Yule-Walker). Another classical approach to ARMA (p, q) order selection is provided by the **Akaike information criteria** (AIC) method.

We consider several ARMA(p, q) models, all depending on parameter vectors $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$. The `ar()` function in `R`, for instance, has $q = 0$ and tries $p = 1, \ldots, 12$.

For each model we calculate the following expression:

$$\text{AIC} = 2\log L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_Z) - 2(p + q + 1)\frac{n}{n - p - q - 2}.$$

When $q = 0$ (i.e., when we consider AR(p) models), this reduces to:

$$\text{AIC} = 2\log L(\boldsymbol{\phi}, \sigma_Z) - 2(p + 1)\frac{n}{n - p - 2}.$$

The AIC method chooses a model with a high likelihood but penalizes models with too many parameters (i.e., if p and q are too large).[58] Another function, `arima()`, computes AIC as follows:

$$\text{AIC} = -2\log L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_Z) + 2(p + q + k + 1),$$

where k is the number of additional parameters to estimate (in our case, $k = 1$ since we estimate σ_Z and there is no seasonality); the optimal model is the one that **minimizes** that version of AIC.

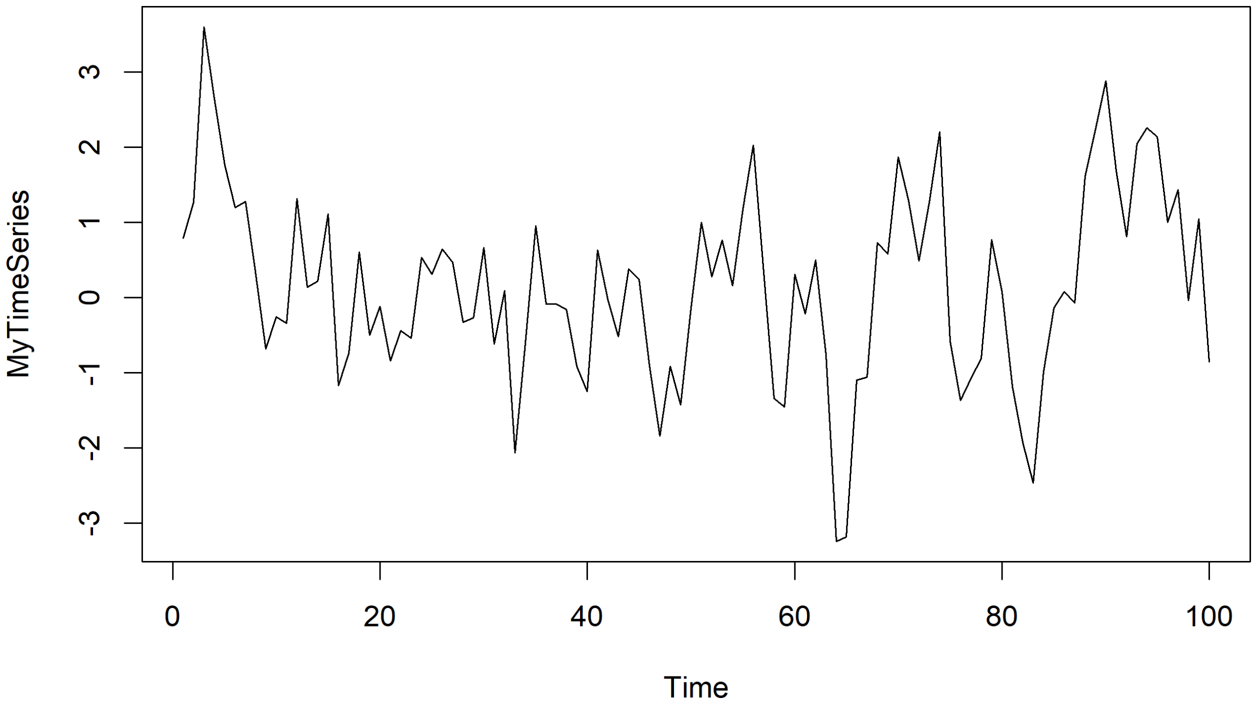
9.7.4 Examples

We consider three examples: an artificial time series, a Lake Huron time series, and a continuation of the Temperature example.

[58] Note that maximizing AIC is equivalent to minimizing −AIC.

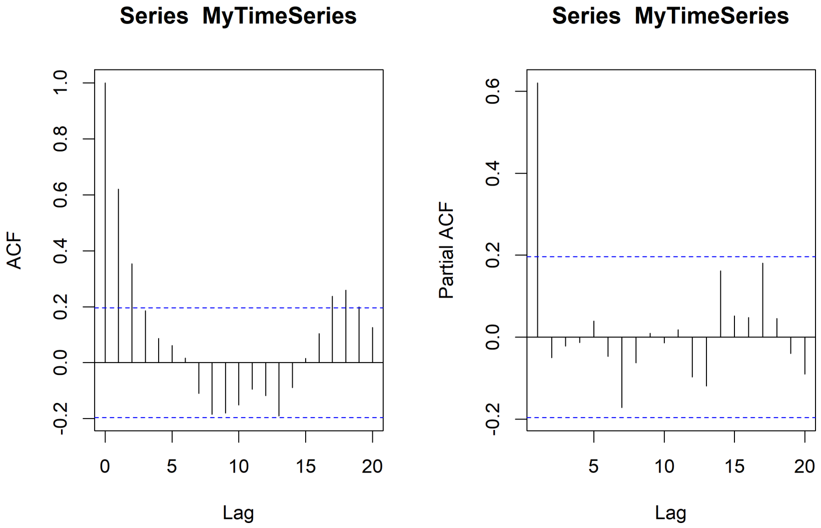
Example: Artificial Data This artificial time series appears to be stationary (more or less).

```
MyTimeSeries <- c(0.793,  1.270,  3.600,  2.649,  1.767,
                  1.198,  1.278,  0.347, -0.683, -0.255,
                 -0.338,  1.316,  0.142,  0.218,  1.118,
                 -1.170, -0.731,  0.609, -0.498, -0.118,
                 -0.839, -0.439, -0.537,  0.537,  0.314,
                  0.647,  0.470, -0.323, -0.264,  0.670,
                 -0.616,  0.092, -2.062, -0.603,  0.958,
                 -0.084, -0.083, -0.156, -0.914, -1.250,
                  0.634, -0.031, -0.519,  0.383,  0.241,
                 -0.903, -1.838, -0.912, -1.422, -0.134,
                  1.004,  0.282,  0.766,  0.164,  1.180,
                  2.030,  0.341, -1.337, -1.452,  0.313,
                 -0.212,  0.500, -0.762, -3.239, -3.179,
                 -1.094, -1.055,  0.735,  0.582,  1.869,
                  1.295,  0.492,  1.272,  2.210, -0.574,
                 -1.363, -1.076, -0.809,  0.774,  0.082,
                 -1.180, -1.925, -2.463, -0.983, -0.135,
                  0.081, -0.071,  1.612,  2.241,  2.884,
                  1.686,  0.811,  2.046,  2.260,  2.142,
                  1.003,  1.435, -0.039,  1.049, -0.855)

plot.ts(MyTimeSeries)
```



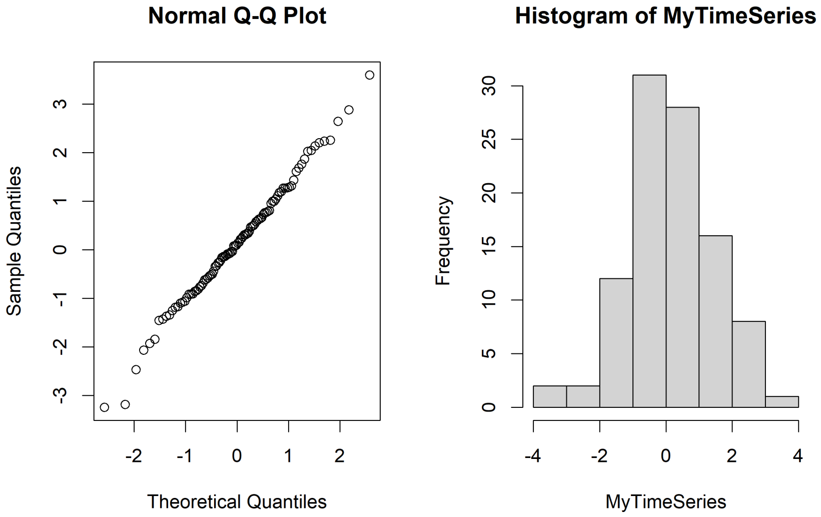
The ACF and PACF displays suggest that the data could arise from an AR(1) process.

```
par(mfrow=c(1,2))
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



We draw your attention to the structure of the ACF; a continuous stretch of positive values, followed by a continuous stretch of negative values, followed by a continuous stretch of positive values (and so on?). This could be indicative of a seasonality effect in the data (see Section 9.9.1). Are the values of the time series normally distributed?

```
par(mfrow=c(1,2))
qqnorm(MyTimeSeries)
hist(MyTimeSeries)
```



We perform model estimation using two approaches: Yule-Walker and MLE.[59]

[59] 59: We do not need normality for the former, but we do need it for the latter, which is why we took the time to verify that the time series values *could* be normally distributed.

```
(fit.ar.yw <- ar(MyTimeSeries,method="yule-walker"))
```

```
Coefficients:
     1
0.6201

Order selected 1  sigma^2 estimated as  0.9707
```

```
(fit.ar.mle <- ar(MyTimeSeries,method="mle"))
```

```
Coefficients:
     1
0.6197

Order selected 1  sigma^2 estimated as  0.9458
```

In both cases, the selected model is AR(1), but the estimated parameters are slightly different. However, the estimates of the autoregressive parameter ϕ should be be the same, regardless of the method used. What is going on?

The difference comes from the fact that the `R` implementation of the MLE approach uses a fairly complicated optimization algorithm, leading to numerical discrepancies – the differences are not significant, to be honest, which is comforting.

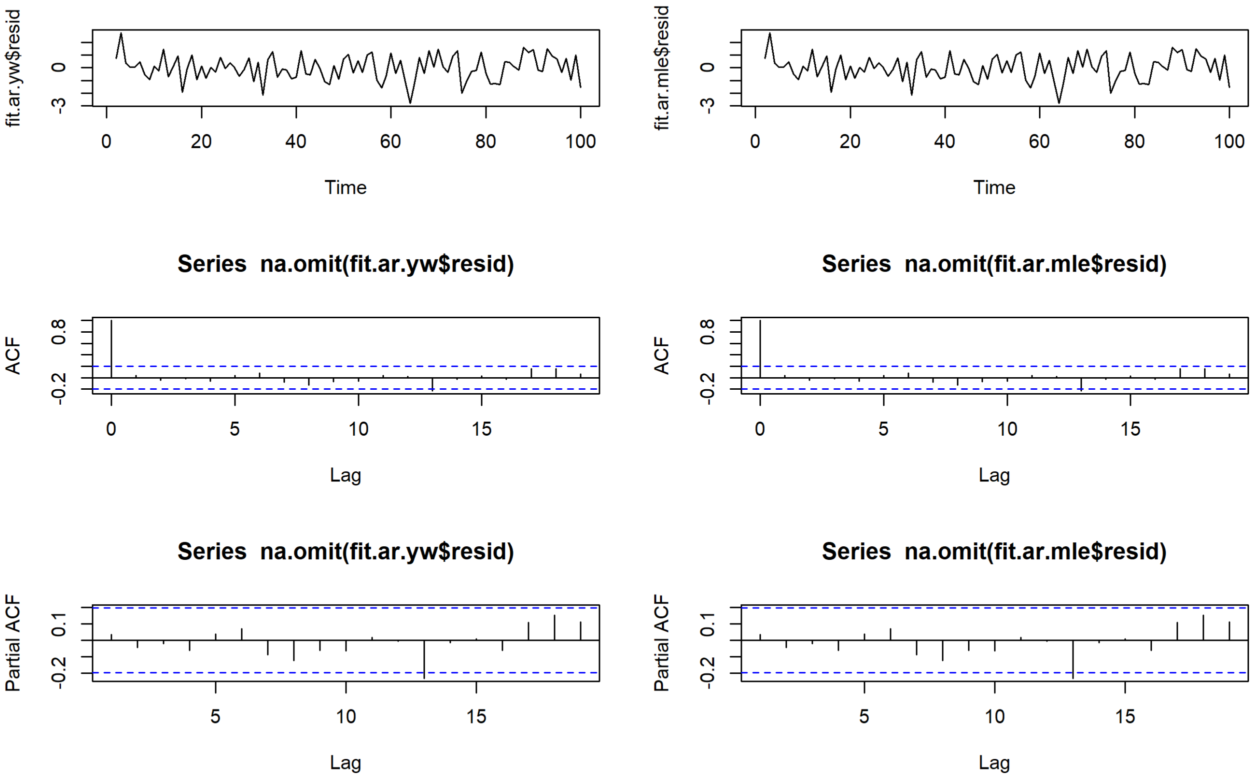
Note, however, that the estimates for σ_Z^2 are different, as they should be, since one is unbiased (Yule-Walker), whereas the other is biased (MLE).

The order and the coefficient value can be extracted using the following code – the displays are suppressed as they can be read above.

```
fit.ar.yw$order
fit.ar.mle$order
fit.ar.yw$ar
fit.ar.mle$ar
```

In order to assess the fit, we can take a look at the residuals.

```
par(mfrow=c(3,2))
plot.ts(fit.ar.yw$resid)
plot.ts(fit.ar.mle$resid)
acf(na.omit(fit.ar.yw$resid))
acf(na.omit(fit.ar.mle$resid))
pacf(na.omit(fit.ar.yw$resid))
pacf(na.omit(fit.ar.mle$resid))
```



In both cases, the residuals certainly look like they could arise from i.i.d. processes.

What would the prediction for the next value of the time series be, in both cases?

```
predict(fit.ar.yw)
```

```
$pred
Time Series:
Start = 101
End = 101
Frequency = 1
[1] -0.4737512

$se
Time Series:
Start = 101
End = 101
Frequency = 1
[1] 0.9852252
```

```
predict(fit.ar.mle)
```

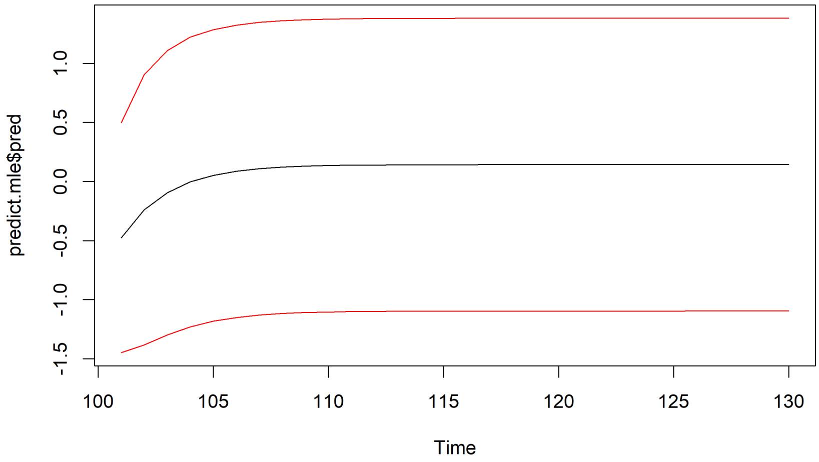
```
$pred
Time Series:
Start = 101
End = 101
Frequency = 1
[1] -0.4754649

$se
Time Series:
Start = 101
End = 101
Frequency = 1
[1] 0.9725163
```

The different predictions values stem from the fact that ϕ_{YW} is slightly different from ϕ_{MLE}.

Both models seem appropriate – which one should we choose? We select the MLE model, for no particular reason. We forecast the next 30 iterations of the model; the **confidence bands** with confidence bands obtained as

$$\text{prediction} \pm \text{standard error of prediction.}$$

```
predict.mle <- predict(fit.ar.mle,n.ahead=30)

par(mfrow=c(1,1))
y.max = max(predict.mle$pred+predict.mle$se)
y.min = min(predict.mle$pred-predict.mle$se)
plot.ts(predict.mle$pred,ylim=c(y.min,y.max))
lines(predict.mle$pred-predict.mle$se,col="red")
lines(predict.mle$pred+predict.mle$se,col="red")
```



Note that these prediction bounds are quite **wide** – the moral of this story is that **long-term forecasts are a fool's errand**, more often than not. Tread with care.

In both estimation methods, the order of the AR model is selected according to AIC (with the maximal order controlled by `order.max`).

```
fit.ar.mle$aic
```

```
        0         1         2         3         4         5
46.636914  0.000000  1.621425  3.598935  5.537506  7.360361
        6         7         8         9        10        11
 8.973913  7.460411  8.709559 10.705111 12.469661 14.417006
       12
14.506713
```

Sure enough, the lowest value (AIC minus a constant) is for AR(1).[60]

We can also use the more general `arima()` function (but we need to specify the order).

[60] How this value is computed depends on the implementation.

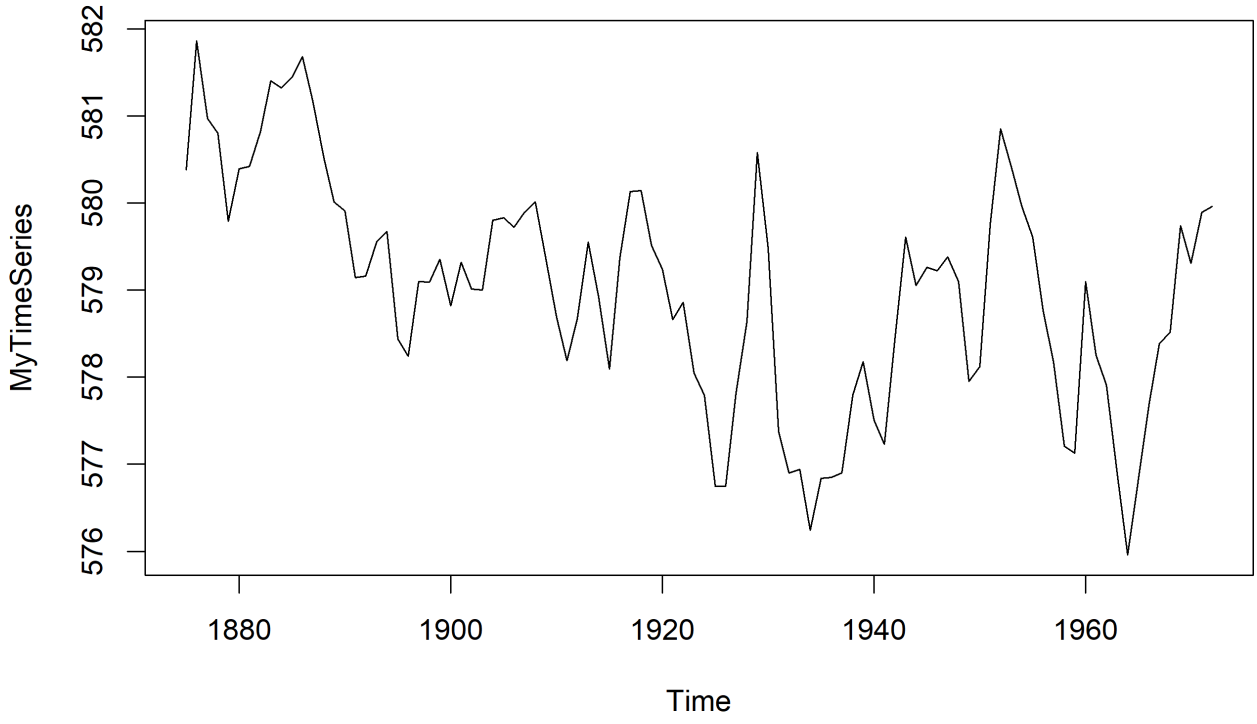
```
(fit.arma <- arima(MyTimeSeries, order=c(1,0,0)))
```

```
Coefficients:
         ar1  intercept
      0.6197     0.1430
s.e.  0.0777     0.2517

sigma^2 estimated as 0.9458:  log likelihood = -139.35,  aic = 284.7
```

The results are readily seen to be identical to those of MLE (suggesting a reason to select MLE over YW, perhaps).

Example: Lake Huron We now conduct a similar analysis with the built-in Lake Huron dataset. We start by loading and plotting the data.

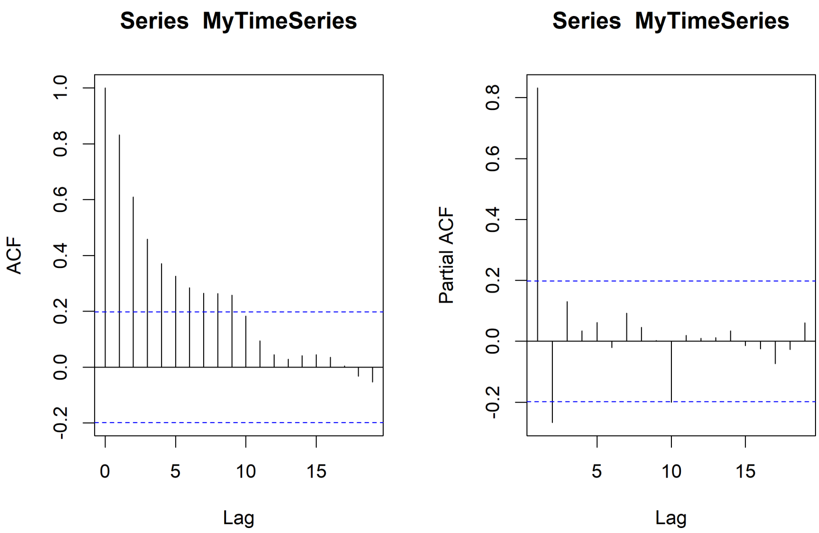
```
MyTimeSeries = LakeHuron
plot.ts(MyTimeSeries)
```



There is a downward trend in the first half of the data (from 1875 to 1925), but it seems almost accidental – if a few of these points were lower, the trend would probably appear to be horizontal. We will treat the time series as stationary, with the caveat that it might make sense to analyze the de-trended time series instead.

We can achieve a first pass at the order by looking at the ACF and PACF graphs.

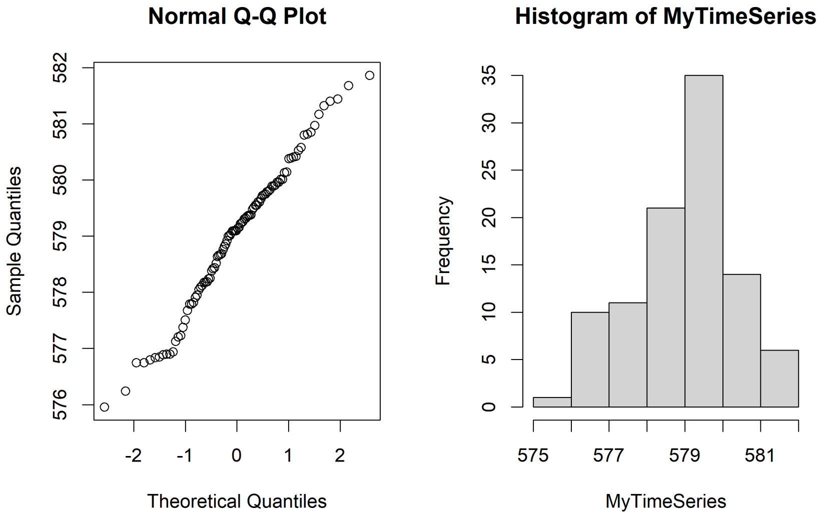
```
par(mfrow=c(1,2))
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



These plots suggest an AR(2) model, or potentially an ARMA$(1, 1)$ model.[61]

The time series appears to take on normally distributed values, as can be seen below.

```
par(mfrow=c(1,2))
qqnorm(MyTimeSeries)
hist(MyTimeSeries)
```



We start by assuming that the data is best fit by an auto-regressive model; what would its order and coefficient estimates be?

Using the Yule-Walker approach, we get the following.

[61] The ACF and PACF of an ARMA model both converge to 0, but the order (p, q) is not usually obvious... there is a lot of guess-and-check involved in the process.

```
(fit.ar.yw <- ar(MyTimeSeries,method="yule-walker"))
```

```
Coefficients:
      1        2
 1.0538  -0.2668

Order selected 2  sigma^2 estimated as  0.5075
```

The MLE approach instead yields the following.

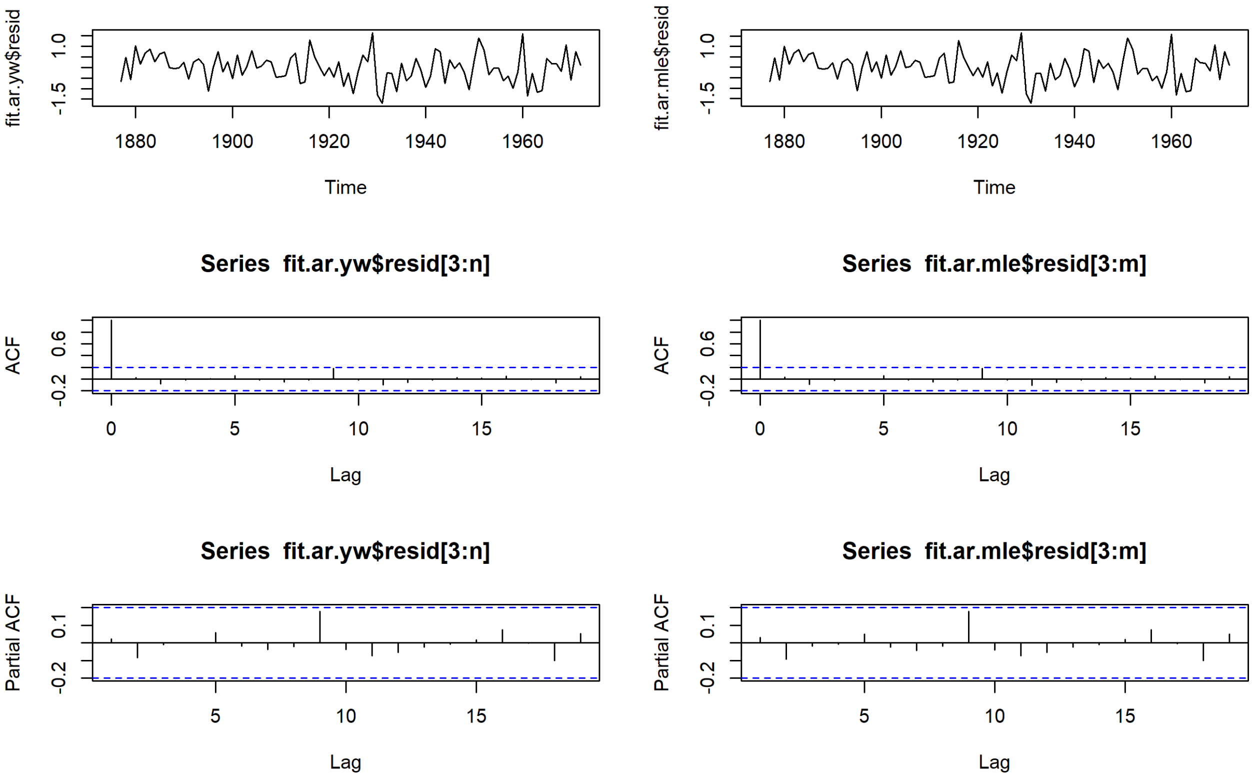
```
(fit.ar.mle <- ar(MyTimeSeries,method="mle"))
```

```
Coefficients:
      1        2
 1.0437  -0.2496

Order selected 2  sigma^2 estimated as  0.4788
```

Both of them suggest an AR(2) model, which agrees with our visual determination of the order.[62]

Are either of the fits good? We take a look at the residuals.

```
par(mfrow=c(3,2))
plot.ts(fit.ar.yw$resid); plot.ts(fit.ar.mle$resid)
n=length(fit.ar.yw$resid); m=length(fit.ar.mle$resid)
acf(fit.ar.yw$resid[3:n]); acf(fit.ar.mle$resid[3:m])
pacf(fit.ar.yw$resid[3:n]); pacf(fit.ar.mle$resid[3:m])
```



[62] The ϕ_i should be identical in both approaches, but we have already discussed that the discrepancies are due to the choice of numerical algorithms in the implementations.

The residual plots look as we would expect if the data arose from either of the two AR(2) processes (i.e., there does not appear to be dependences in the residuals). So which model should be chosen? We could pick the one with smallest AIC, or selecting the model that best "predicts" past values of the data (as done in Section 9.4.4 with the currency exchange rate data). We select the MLE model for the purpose of illustration.

In order to investigate ARMA$(1, 1)$ as a model for the data, we use the `arima()` function. We will re-fit the MLE AR(2) model in this framework, to gain access to the same set of attributes for both models.

```
(fit.arma.1 <- arima(MyTimeSeries, order=c(2,0,0)))
```

```
Coefficients:
         ar1      ar2  intercept
      1.0436  -0.2495   579.0473
s.e.  0.0983   0.1008     0.3319

sigma^2 estimated as 0.4788:  log likelihood = -103.63,  aic = 215.27
```

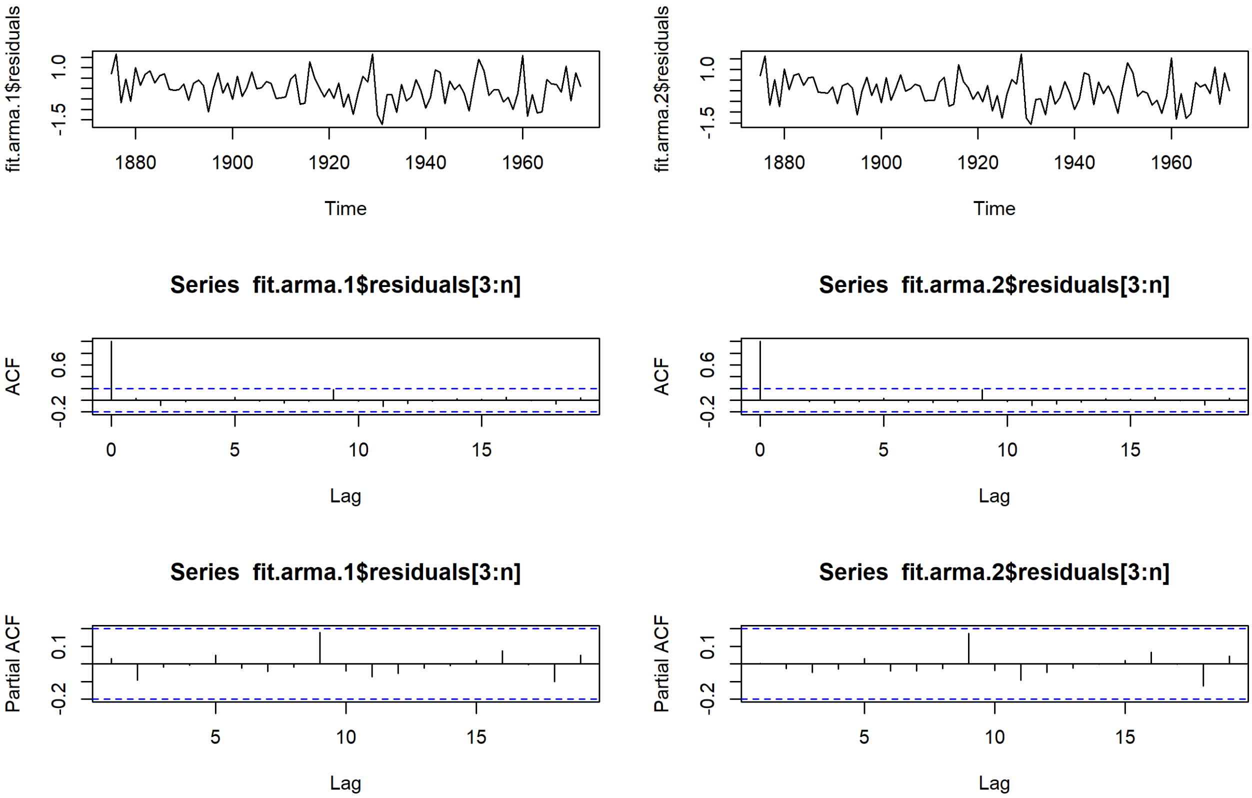
```
(fit.arma.2 <- arima(MyTimeSeries, order=c(1,0,1)))
```

```
Coefficients:
         ar1     ma1  intercept
      0.7449  0.3206   579.0555
s.e.  0.0777  0.1135     0.3501

sigma^2 estimated as 0.4749:  log likelihood = -103.25,  aic = 214.49
```

The intercept term represents the expectation $\mu = \mathrm{E}[X_t]$ of the time series. An important take-away is that there is no obvious relationship between the ϕ_1 of the AR(2) model and the ϕ_1 of the ARMA$(1, 1)$ model.

What do the residuals look like?

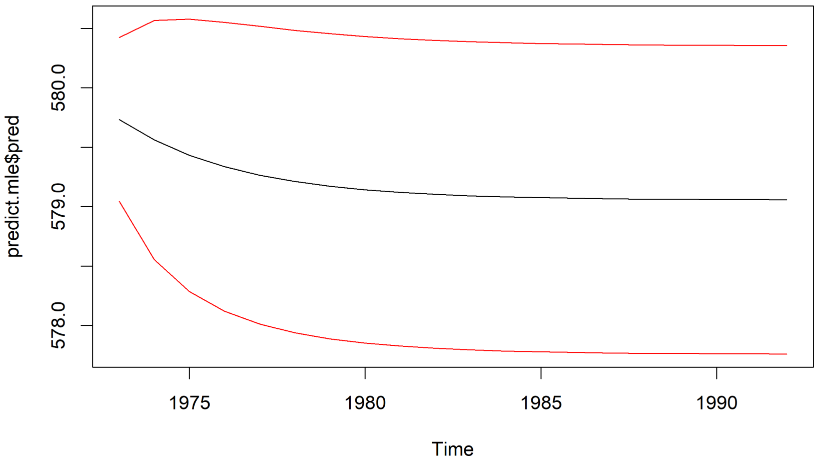
```
par(mfrow=c(3,2))
plot.ts(fit.arma.1$residuals)
plot.ts(fit.arma.2$residuals)

n = length(fit.arma.1$residuals)
m = length(fit.arma.2$residuals)

acf(fit.arma.1$residuals[3:n])
acf(fit.arma.2$residuals[3:n])
pacf(fit.arma.1$residuals[3:n])
pacf(fit.arma.2$residuals[3:n])
```



Both AR(2) and ARMA(1, 1) are acceptable; we select the latter since it has the smallest AIC.[63] We can predict the next 20 time steps.

[63] The AIC can be read off of the outputs above, but they can also be extracted directly with `fit.arma.1$aic` and `fit.arma.2$aic`.

```
par(mfrow=c(1,1))
predict.mle <- predict(fit.arma.2,n.ahead=20)
y.max = max(predict.mle$pred+predict.mle$se)
y.min = min(predict.mle$pred-predict.mle$se)
plot.ts(predict.mle$pred,ylim=c(y.min,y.max))
lines(predict.mle$pred-predict.mle$se,col="red")
lines(predict.mle$pred+predict.mle$se,col="red")
```



Note that the predictions are not as "jagged" as the original time series.[64]

[64] When seasonality is taken into account, we might expect to see some up-and-down motion in the predictions.

Example: Temperature (cont.) We consider the temperature data from pages 499 and 553; using the Yule-Walker procedure, we found that the centered stationary part of the exponentially smoothed time series (`Stationary`) was decently approximated by an AR(4) process.

We now approach the same time series via the MLE procedure. The chart on page 556 indicates that the normality assumption is reasonable. We can thus safely apply the procedure.

```
(fit.ar.mle <- ar(Stationary,method="mle"))
```

```
Coefficients:
      1        2        3        4        5  
 0.1427   0.1290  -0.0682   0.2716   0.1187  

Order selected 5  sigma^2 estimated as  0.03241
```

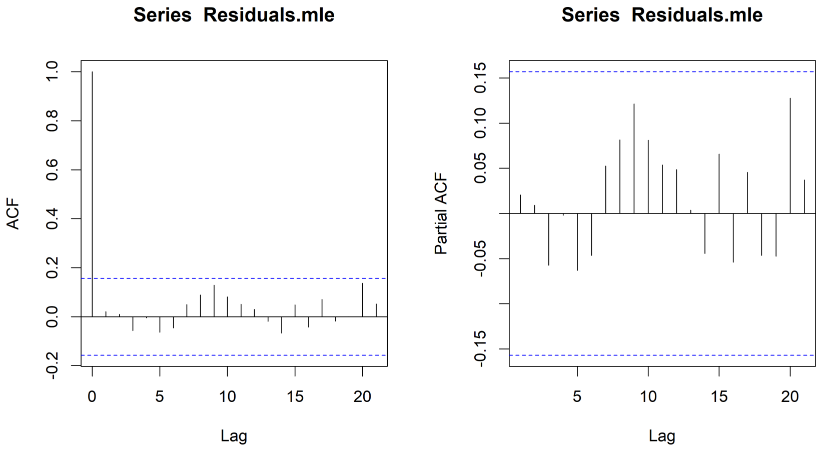
The MLE procedure selected a different order – but there is nothing wrong with that! Note that we could recover this model with the `arima` function (which also displays the standard errors for the AR coefficients).

```
arima(Stationary,order=c(5,0,0),method="ML")
```

```
Coefficients:
         ar1     ar2      ar3     ar4     ar5  intercept
      0.1427  0.1290  -0.0682  0.2716  0.1187     0.0743
s.e.  0.0786  0.0761   0.0766  0.0763  0.0798     0.0342

sigma^2 estimated as 0.03241:  log likelihood = 47.34,  aic = -80.69
```

Is the MLE fit appropriate? Do the residuals appear to be white noise?

```
Residuals.mle = fit.ar.mle$resid
Residuals.mle = na.omit(Residuals.mle)
par(mfrow=c(1,2))
acf(Residuals.mle)
pacf(Residuals.mle)
```



Yes-ish. Close enough is good enough, certainly. We accept the fit. But now we have two competing models. Which one should we choose? We can check the quality of the prediction, for instance.

```
(Squared.Error.yw = mean((Residuals.yw)^2))
```

```
[1] 0.0332902
```

```
(Squared.Error.mle = mean((Residuals.mle)^2))
```

```
[1] 0.03214238
```

The MLE approach yields a lower total error, so we might as well select the MLE model.

But why was AR(5) selected by the MLE procedure? We can compare with the AR(4) MLE model and calculate the respective AIC.

```
(fit.mle.4 <- arima(Stationary,order=c(4,0,0),method="ML"))
```

```
Coefficients:
         ar1     ar2      ar3     ar4  intercept
      0.1782  0.1196  -0.0541  0.2918     0.0715
s.e.  0.0754  0.0764   0.0765  0.0757     0.0303

sigma^2 estimated as 0.03287:  log likelihood = 46.25,  aic = -80.49
```

```
(fit.mle.5 <- arima(Stationary,order=c(5,0,0),method="ML"))
```

```
Coefficients:
         ar1     ar2      ar3     ar4     ar5  intercept
      0.1427  0.1290  -0.0682  0.2716  0.1187     0.0743
s.e.  0.0786  0.0761   0.0766  0.0763  0.0798     0.0342

sigma^2 estimated as 0.03241:  log likelihood = 47.34,  aic = -80.69
```

Note the values of log-likelihod and AIC.[65]

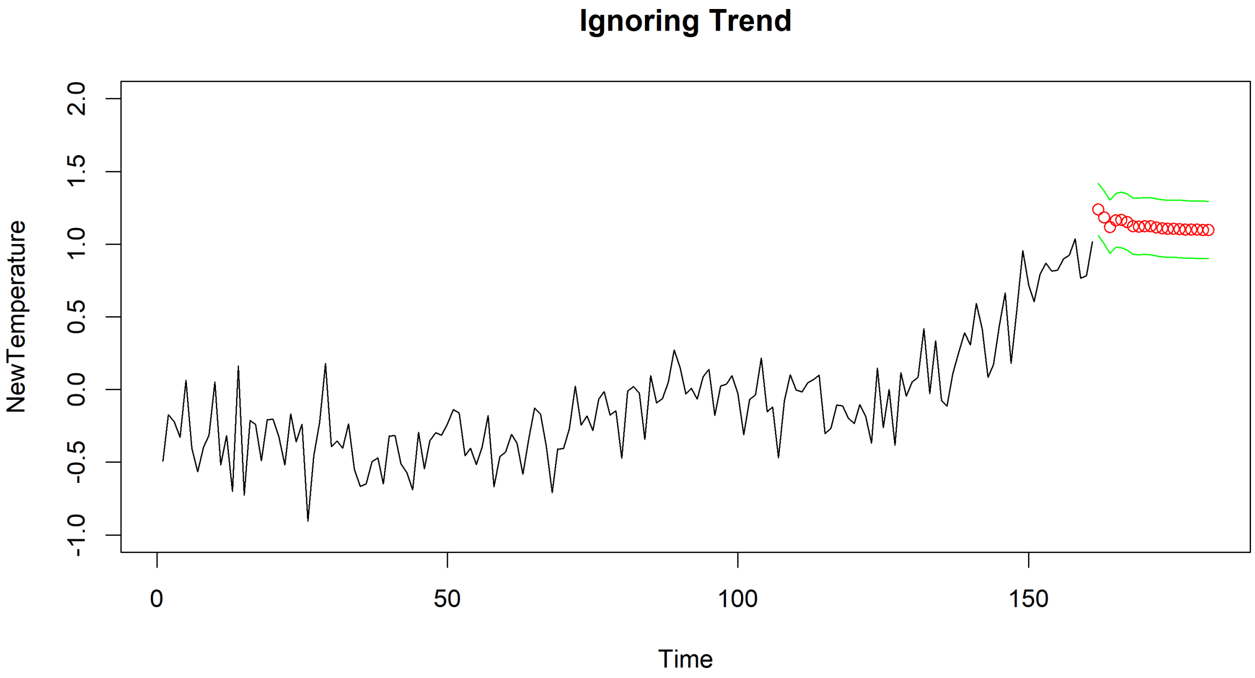
We can use the MLE model to predict the next 20 observations

```
k = 20
prediction = predict(fit.ar.mle,n.ahead=k)$pred
error = predict(fit.ar.mle,n.ahead=k)$se
```

In order to transform these `Stationary` predictions into values in the original time series, we have to add them to the Temperature data. In the next code chunk, we will ignore the trend in the original data.

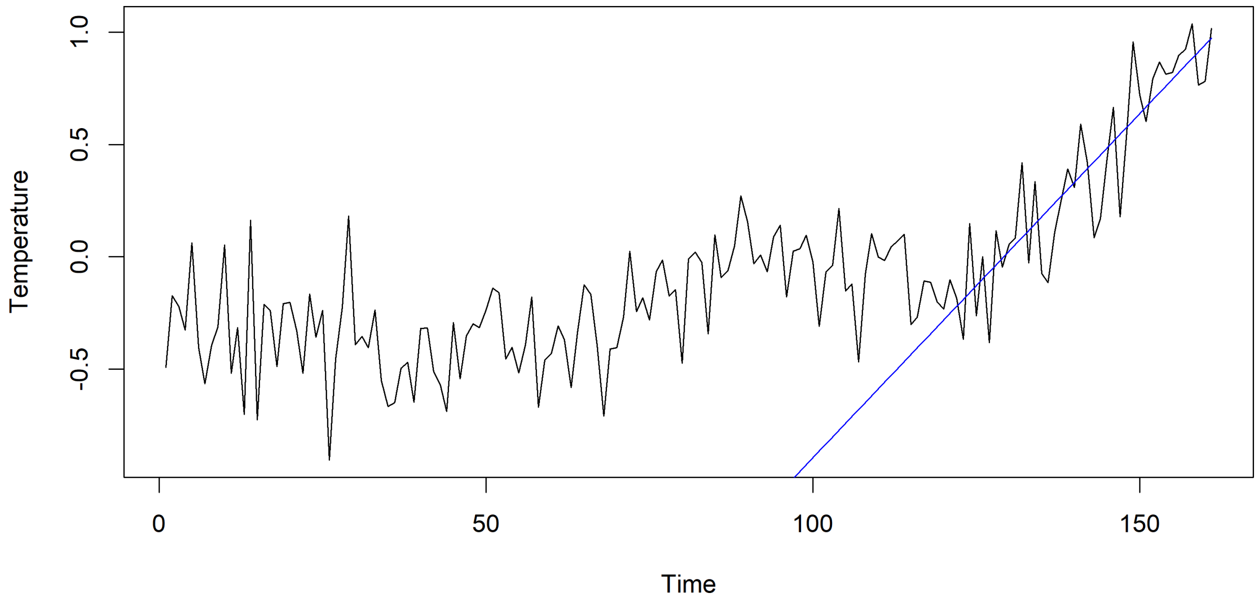
[65] It would be important to make sure that you can recover the AIC values from the log-likelihood values, with the formula.

```
n=length(Temperature)
k = 20
prediction.1 = prediction+Temperature[n]
prediction.1.upper = prediction.1 + error
prediction.1.lower = prediction.1 - error

dummy.ts = c(rep(NA,k))
NewTemperature = c(Temperature,dummy.ts)
dummy.pred=c(rep(NA,n))
PredictedStationary = c(dummy.pred,prediction.1)
PredictionUpperLimit = c(dummy.pred,prediction.1.upper)
PredictionLowerLimit = c(dummy.pred,prediction.1.lower)

par(mfrow=c(1,1))
plot.ts(NewTemperature,ylim=c(-1,2),main="Ignoring Trend")
points(PredictedStationary,col="red",type="p")
points(PredictionUpperLimit,col="green",type="l")
points(PredictionLowerLimit,col="green",type="l")
```



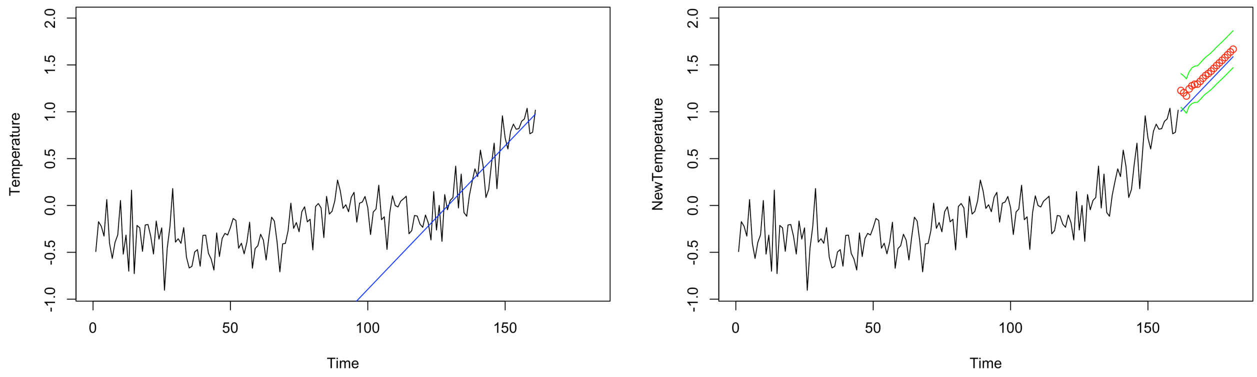
Something about this is definitely not right. The problem is that we ignored the trend in the original data, but starting in year 120 (or thereabouts), the time series follows a linear trend (more or less). We fit a linear trend to this part of the data.

```
n = length(Temperature)
Time = seq(1,n,by=1)
lin.reg = lm(Temperature[120:n]~Time[120:n])
Lin.Trend = lin.reg[[1]][1] + lin.reg[[1]][2]*Time

plot.ts(Temperature)
points(Lin.Trend,col="blue",type="l")
```



The next step is to extend the linear trend and the predictions.[66]

```
k = 20
dummy.ts = c(rep(NA,k))
NewTemperature = c(Temperature,dummy.ts)
dummy.trend = c(rep(NA,n)); Time = seq(1,n+k,by=1)
Extended.Trend = lin.reg[[1]][1] + lin.reg[[1]][2]*Time
Trend = c(dummy.trend,Extended.Trend[(n+1):(n+k)])
y.max = 2; y.min = min(Temperature)

par(mfrow=c(1,2))
plot.ts(Temperature,xlim=c(1,n+k),ylim=c(y.min,y.max))
points(Lin.Trend,col="blue",type="l")
plot.ts(NewTemperature,xlim=c(1,n+k),ylim=c(y.min,y.max))
points(Trend,col="blue",type="l")
Prediction.stationary = c(dummy.trend,prediction)
PredictedStationary = Trend+Prediction.stationary
points(PredictedStationary,col="red",type="p")
```



[66] The code for the prediction limits is not included – how could they be produced?

9.8 Nonlinear Time Series

The log-returns of financial data typically have the following properties:

- they are **uncorrelated**;
- their **squares** are correlated;
- they are **not normally distributed**.

Such features cannot be modelled by ARMA models.

9.8.1 ARCH model

A time series $\{X_t \mid t = 1, \dots, n\}$ is **autoregressive conditionally heteroscedastic of order** p, denoted ARCH(p) if

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2,$$

where Z_t are i.i.d. with mean 0 and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$ for all i. We note explicitly that the values of σ_t depend on the past values of the sequence $\{X_t\}$: $X_{t-1}, X_{t-2}, \dots$.

If $p = 1$, then

$$\begin{aligned} X_t^2 &= \sigma_t^2 Z_t^2 = \left(\alpha_0 + \alpha_1 X_{t-1}^2\right) Z_t^2 = \left(\alpha_0 + \alpha_1 \sigma_{t-1}^2 Z_{t-1}^2\right) Z_t^2 \\ &= \alpha_0 Z_t^2 + \alpha_1 Z_t^2 Z_{t-1}^2 \sigma_{t-1}^2. \end{aligned}$$

We can continue on this way by replacing σ_{t-1}^2 by its formulation, and so on. Consequently, we see that X_t^2 depends only on $Z_t, Z_{t-1}, Z_{t-2}, \dots$.[67] This is valid for all ARCH models, not only ARCH(1).

For a general p, we have

$$\mathrm{E}[X_t \mid X_{t-1}, \dots, X_{t-p}] = \mathrm{E}[\sigma_t Z_t \mid X_{t-1}, \dots, X_{t-p}] = \sigma_t \mathrm{E}[Z_t \mid X_{t-1}, \dots, X_{t-p}] = 0$$

and

$$\begin{aligned} \mathrm{Var}(X_t \mid X_{t-1}, \dots, X_{t-p}) &= \mathrm{E}[X_t^2 \mid X_{t-1}, \dots, X_{t-p}] = \mathrm{E}[\sigma_t^2 Z_t^2 \mid X_{t-1}, \dots, X_{t-p}] \\ &= \sigma_t^2 \mathrm{E}[Z_t^2 \mid X_{t-1}, \dots, X_{t-p}] = \sigma_t^2 \mathrm{E}[Z_t^2] = \sigma_t^2. \end{aligned}$$

The "conditionally heteroscedastic" in ARCH refers to this last equation.

The series $\{\sigma_t^2 \mid t \geq 1\}$ is the **volatility** of the time series; ARCH(p) is an example of a **stochastic volatility process**.

Proposition: the ARCH(1) process is stationary if and only if $\alpha_1 < 1$. A stationary solution is given by

$$X_t^2 = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i \prod_{j=0}^{i} Z_{t-j}^2.$$

In an ARCH(1) model, we have

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2.$$

[67] As a further consequence, Z_t and X_{t-1} are independent.

We can estimate the model parameters using the maximum likelihood principle. Consider the joint density

$$f_{(X_0,\ldots,X_n)}(x_0,\ldots,x_n) = f_{X_0}(x_0)\prod_{i=1}^{n} f_{X_t|X_{t-1}}(x_t \mid x_{t-1})$$

where

$$f_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{1}{\sigma_t} g(x_t/\sigma_t)\ ,$$

with $\sigma_t^2 = \alpha_0 + \alpha_1 x_{t-1}^2$ and g is the density of Z_0 (which is typically a normal or Student T distribution).

Let

$$\begin{aligned}
F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) &= P(X_t \le x_t \mid X_{t-1} = x_{t-1})\\
&= P(\sigma_t Z_t \le x_t \mid X_{t-1} = x_{t-1})\\
&= P(Z_t \le x_t/\sigma_t \mid X_{t-1} = x_{t-1})\\
&= P(Z_t \le x_t/\sigma_t) = G(x_t/\sigma_t),
\end{aligned}$$

where G is the cumulative distribution function of Z: $G(z) = P(Z \le z)$.

We can show that

$$f_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{d}{dx_t} F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{1}{\sigma_t} g(x_t/\sigma_t)\ .$$

Indeed, we start with the conditional distribution:

$$\begin{aligned}
F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) &= P(X_t \le x_t \mid X_{t-1} = x_{t-1}) = P(\sigma_t Z_t \le x_t \mid X_{t-1} = x_{t-1})\\
&= P(Z_t \le x_t/\sigma_t \mid X_{t-1} = x_{t-1})\\
&= P(Z_t \le x_t/\sqrt{\alpha_0 + \alpha_1 x_{t-1}^2 \mid X_{t-1} = x_{t-1})}\\
&= F_Z(Z_t \le x_t/\sqrt{\alpha_0 + \alpha_1 x_{t-1}^2}) = F_Z(x_t/\sigma_t)
\end{aligned}$$

and the density is

$$\frac{d}{dx_t} F_Z(x_t/\sigma_t) = \frac{1}{\sigma_t} f_Z(x/\sigma_t) = \frac{1}{\sigma_t} g(x/\sigma_t),$$

keeping in mind that $\sigma_t^2 = \alpha_0 + \alpha_1 x_{t-1}^2$.

Thus, the likelihood function has the form

$$L(\alpha_0,\alpha_1) = \prod_{t=1}^{n} \frac{1}{\sigma_t} g(X_t/\sigma_t)$$

and

$$(\alpha_0,\alpha_1) = \arg \max_{\alpha_0>0, 0<\alpha_1<1} L(\alpha_0,\alpha_1),$$

where the optimization problem is solved numerically (see Section 4).

9.8.2 GARCH Model

A time series $\{X_t \mid t = 1,\ldots,n\}$ is a **generalized autoregressive conditionally heteroscedastic** model of **order** (p,q), denoted GARCH(p,q) if the variance σ_t^2 is modeled using past squared observations X_{t-i}^2 **and** past variances σ_{t-j}^2:

$$X_t = \sigma_t Z_t, \quad \mathrm{Var}(X_t \mid X_{t-1}, \ldots, X_{t-p}) = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

where Z_t are i.i.d. with mean 0 and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for all i, j.

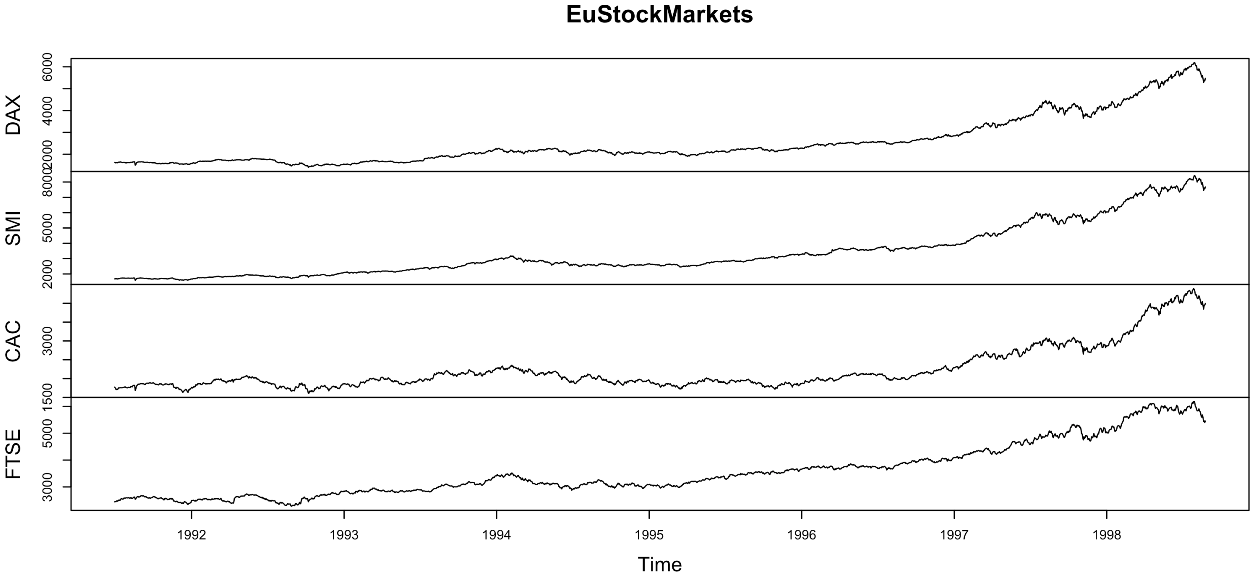
On the topic of identifying an ARCH/GARCH model in practice, [5] has this to say:

> The best identification tool may be a time series plot of the series. It's usually easy to spot periods of increased variation sprinkled through the series. It can be fruitful to look at the ACF and PACF of both X_t and X_t^2. For instance, if X_t appears to be white noise and X_t^2 appears to be AR(1), then an ARCH(1) model for the variance is suggested. If the PACF of X_t^2 suggests AR(p), then ARCH(p) may work. GARCH models may be suggested by an ARMA-type look to the ACF and PACF of X_t^2. [...] You might have to experiment with various ARCH and GARCH structures after spotting the need in the time series plot of the series.

9.8.3 Example: Stock Returns

We consider the daily closing price of Germany's DAX stock index, from 1991 to 1998;[68] the dataset is pre-built in R.

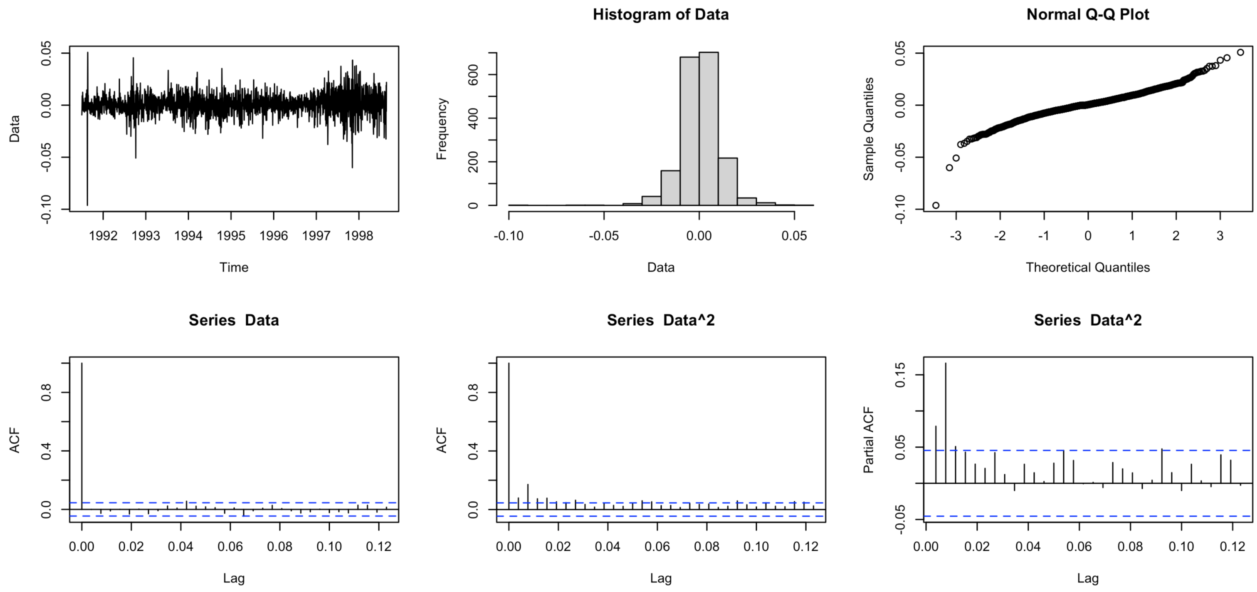
```
library(tseries)
plot(EuStockMarkets)
```



We differentiate the log-returns of the DAX index to obtain a time series which appears to be stationary, but which is not normally distributed. We display the ACF of the data, as well as the ACF and PACF of the square of the data.

[68] The observations are recorded on business days, and are also available for 3 other indices: SMI, CAC, UK FTSE.

```
Data <- diff(log(EuStockMarkets))[,"DAX"]
par(mfrow=c(2,3))
plot.ts(Data)
hist(Data); qqnorm(Data);
acf(Data); acf(Data^2); pacf(Data^2)
```



It is reasonable to fit to model the data as an ARCH(1) process.

```
fit.ARCH1 <- garch(Data,order=c(0,1))
```

```
***** ESTIMATION WITH ANALYTICAL GRADIENT *****

     I     INITIAL X(I)        D(I)
     1     1.008019e-04     1.000e+00
     2     5.000000e-02     1.000e+00

    IT   NF      F          RELDF     PRELDF     RELDX    STPPAR   D*STEP   NPRELDF
     0    1 -7.582e+03
     1    8 -7.582e+03  7.08e-06  1.27e-05  1.0e-05  9.4e+10  1.0e-06  5.95e+05
     2    9 -7.582e+03  9.60e-08  9.77e-08  1.0e-05  2.0e+00  1.0e-06  7.31e-01
     3   18 -7.584e+03  2.66e-04  4.85e-04  2.6e-01  2.0e+00  3.5e-02  7.31e-01
     4   19 -7.584e+03  1.47e-05  1.13e-05  4.4e-02  0.0e+00  7.9e-03  1.13e-05
     5   20 -7.584e+03  1.81e-06  1.67e-06  2.0e-02  0.0e+00  3.8e-03  1.67e-06
     6   21 -7.584e+03  1.51e-08  1.46e-08  1.9e-03  0.0e+00  3.6e-04  1.46e-08
     7   22 -7.584e+03  1.47e-11  1.47e-11  6.3e-05  0.0e+00  1.2e-05  1.47e-11

 ***** RELATIVE FUNCTION CONVERGENCE *****

 FUNCTION    -7.584131e+03   RELDX        6.254e-05
 FUNC. EVALS       22        GRAD. EVALS       8
 PRELDF        1.471e-11     NPRELDF      1.471e-11

     I      FINAL X(I)        D(I)          G(I)
     1    9.611161e-05     1.000e+00    -9.229e-01
     2    9.703263e-02     1.000e+00    -7.850e-05
```

The resulting GARCH object has the following attributes.

```
attributes(fit.ARCH1)
```

```
$names
 [1] "order"         "coef"          "n.likeli"      "n.used"
 [5] "residuals"     "fitted.values" "series"        "frequency"
 [9] "call"          "vcov"

$class
[1] "garch"
```

The estimated coefficient values of α_0 and α_1 are obtained as below.

```
(Coefficients <- fit.ARCH1$coef)
alpha0=Coefficients[1]; alpha1=Coefficients[2]
```

```
          a0           a1
9.611161e-05 9.703263e-02
```

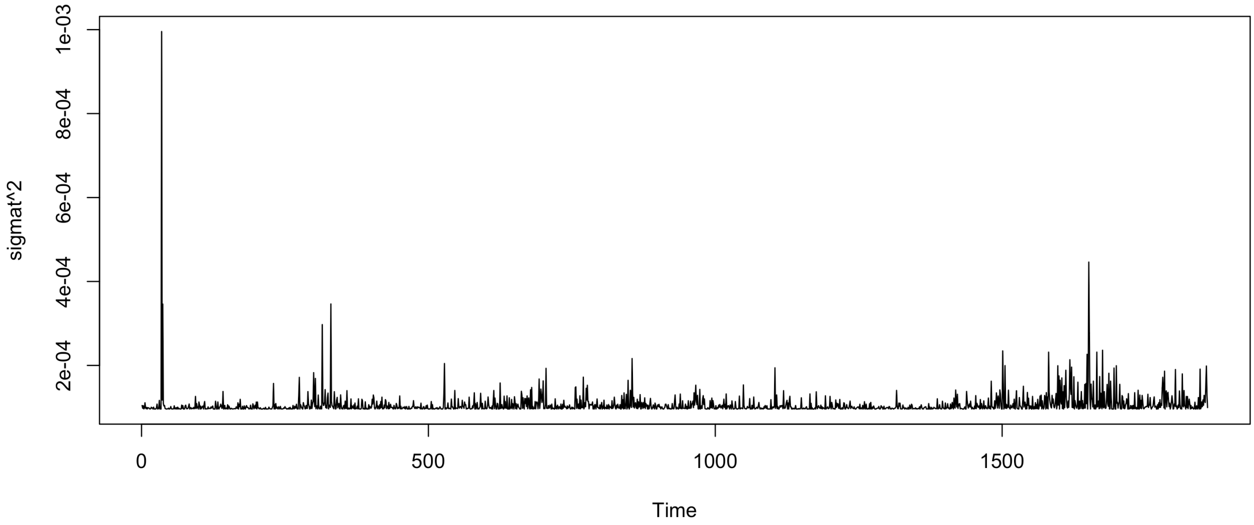
We can view the fitted values as past prediction of σ_t, the first 10 of which are as below.

```
past.prediction = fit.ARCH1$fitted.values
past.prediction[1:10]
```

```
[1]          NA 0.010225064 0.009899957 0.010196954 0.009819289 0.009911300
[7] 0.010540232 0.009966482 0.009844322 0.010001275
```

We can plot the time series $\{\sigma_t^2\}$:

```
n = length(Data)
sigmat = past.prediction[2:n]
par(mfrow=c(1,1))
plot.ts(sigmat^2)
```



The prediction of the next observation in the sequence can be obtained directly.

```
n1 = length(sigmat)
sqrt(alpha0+alpha1*sigmat[n1]^2)
```

```
        a0
0.01028445
```

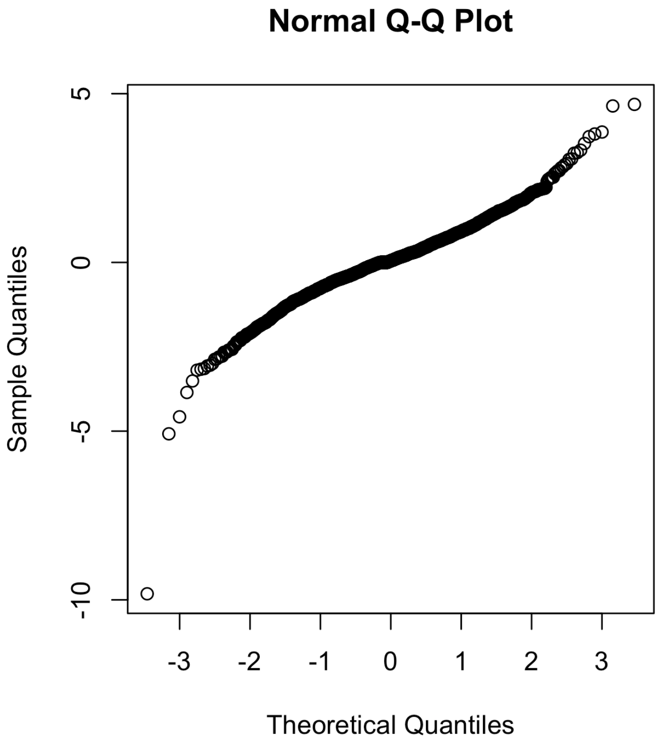
It is easy to extract the residuals, the first 10 of which are:

```
residuals <- fit.ARCH1$residuals
residuals[1:10]
```

```
[1]          NA -0.4324838  0.9094782 -0.1743871
[5] -0.4762781  1.2538257  0.5464646
[8] -0.2879321  0.6451482  0.1183917
```

We can see that the residuals are normally distributed, roughly.[69]

```
residuals = residuals[2:n]
qqnorm(residuals)
```



9.9 Miscellenous Topics

We will finish this chapter by briefly discussing three additional topics: **seasonality**, **asymptotic normality**, and **innovations**.

[69] Remember that normality of the residuals (Z_t) is not the same as normality of the data (X_t).

9.9.1 Seasonality

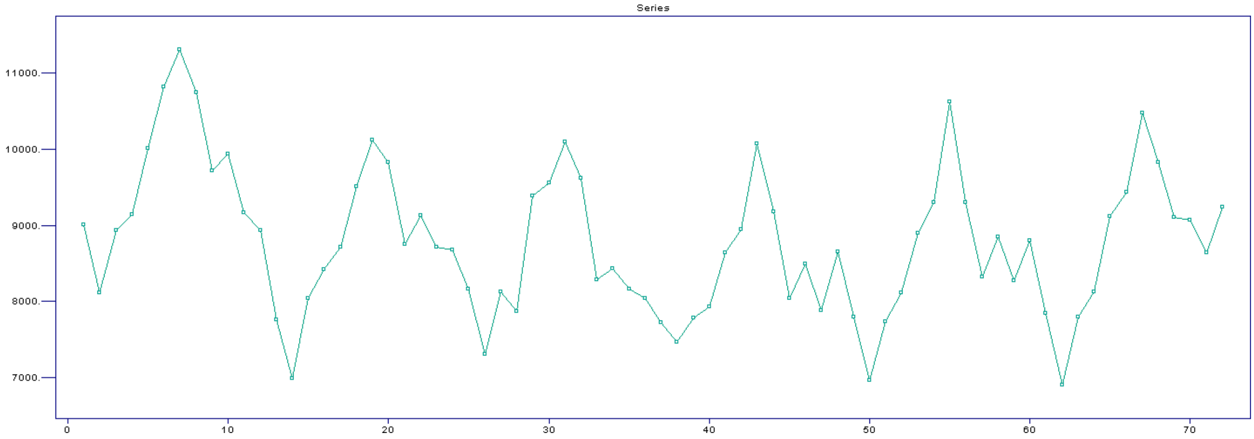
In the study of time series data, **seasonality** – a repeating pattern that occurs at regular intervals – is an important concept. For instance, we might expect a time series of the average monthly temperature in a specific location to show regularity from one year to the next. Or, assuming that an employee's salary is deposited twice monthly directly into their bank account and that expenses come out on a monthly basis form the same account, we would expect the time series of end-of-day balances in the account to follow a regular monthly pattern.

Differencing is a simple way to correct for a seasonal component: if we have identified such a component with a period of T time steps,[70] then we can remove it on X_t by subtracting from it the value X_{t-T}, yielding a time series

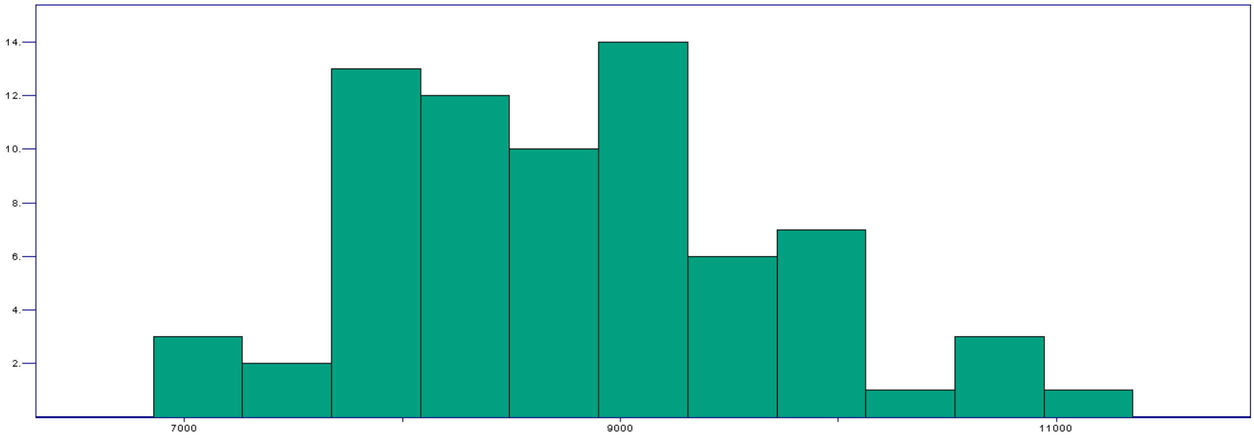
$$Y_t = \nabla_T X_t = X_t - X_{t-T}, t > T.$$

We have seen some examples of seasonal decomposition when we were using the `decompose()` function to de-trend the data and obtain the stationary (random) component for analysis (see page 504, for instance).

Example: Accidental Deaths The monthly accidental deaths figures (`USAccDeaths`) in the US from January 1973 ($t = 1$) to December 1978 ($t = 72$) are plotted below.

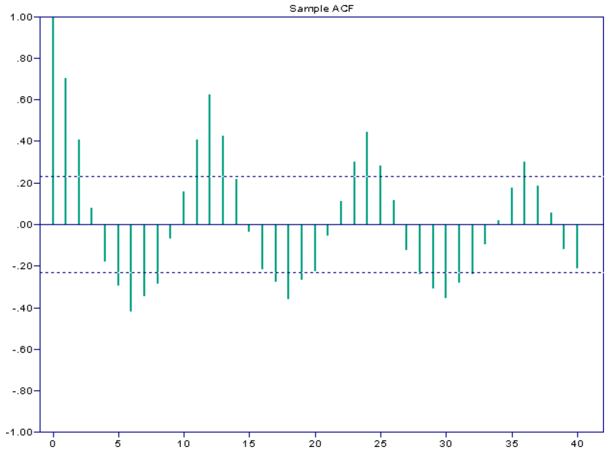


A histogram of the data is also provided.

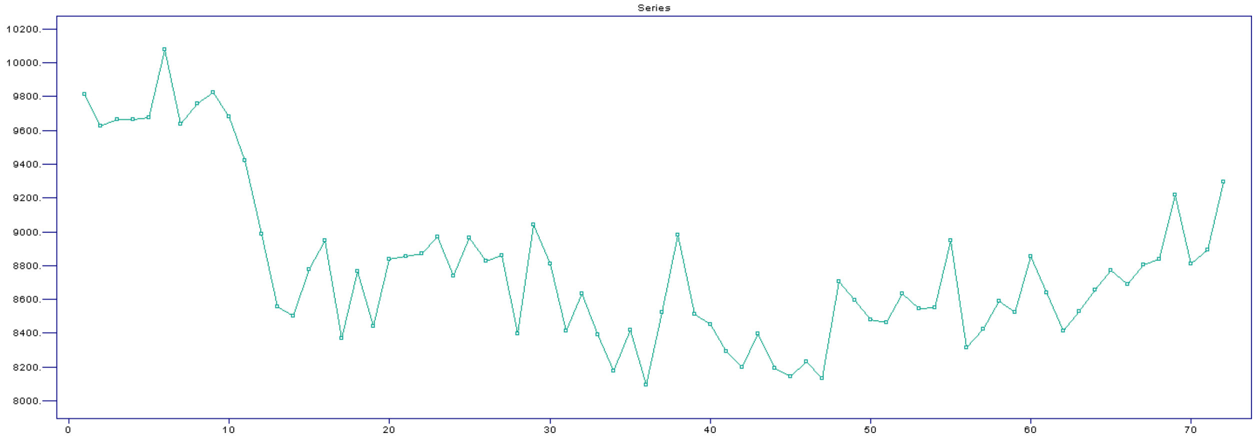


[70] By searching for regularities in the ACVF, through a Fourier analysis of the data, or using domain expertise.

The sample autocorrelation function also shows a seasonal trend with period $T = 12$.

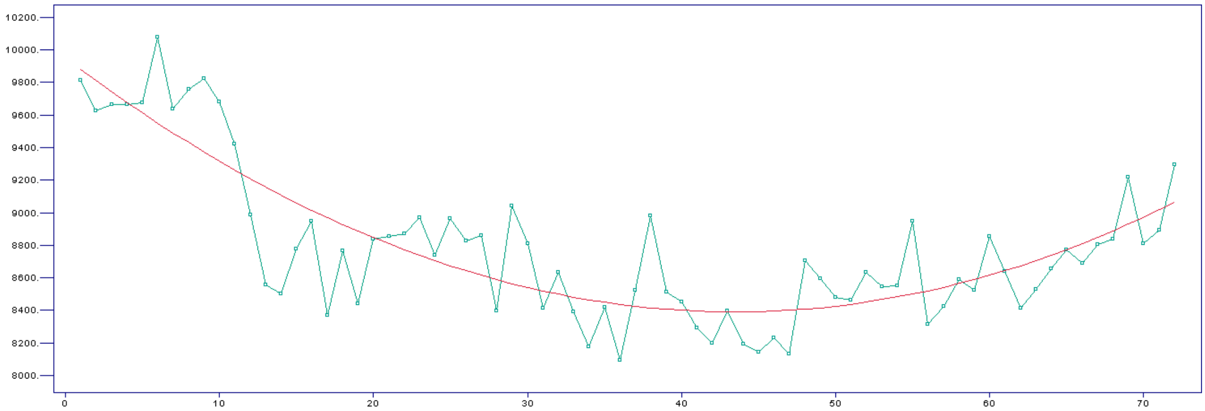


The deseasonalized deaths data is shown below.



This graphs suggests the presence of an additional quadratic component:

$$x_t = \underbrace{m_t}_{\text{local trend}} + \underbrace{s_t}_{\text{seasonal trend}} + \underbrace{Z_t}_{\text{noise}}, \quad m_t = a_0 + a_1 t + a_2 t^2.$$



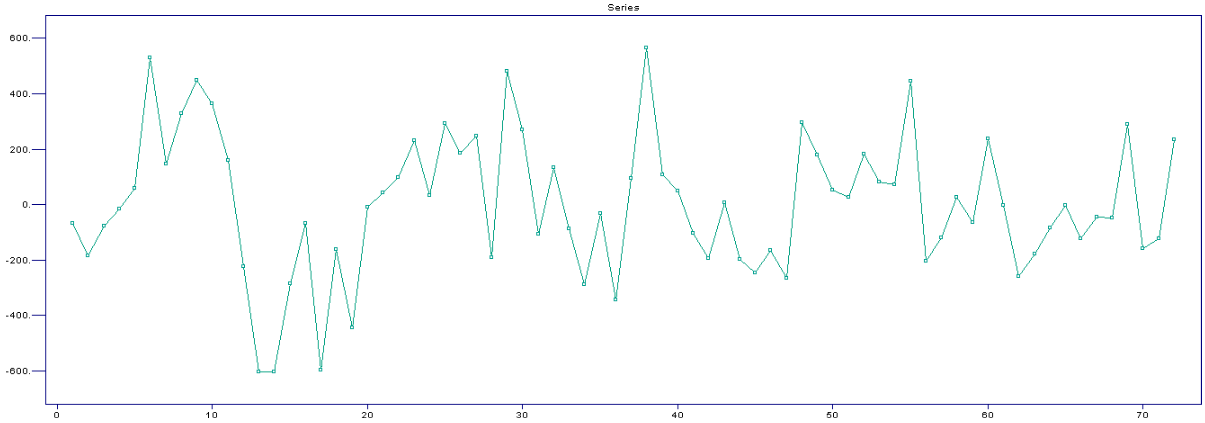
We estimate the local trend as

$$\hat{m}_t = 9951.822 - 71.817t + 0.826t^2, \quad 1 \leq t \leq 72.$$

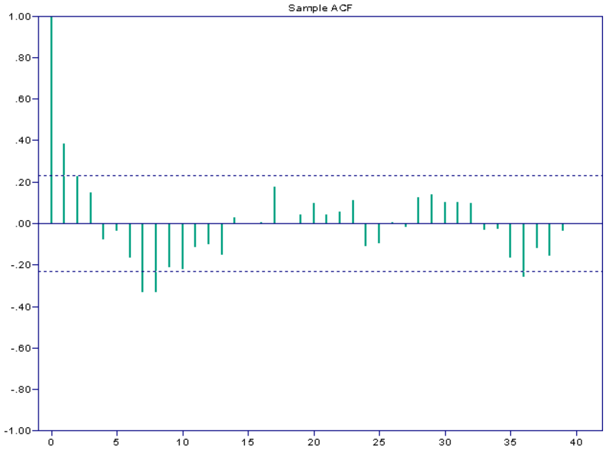
The estimated residuals (the stationary signal)

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad 1 \leq t \leq 72$$

is shown below.



The residuals do appear to be dependent, as there are long stretches of residuals with the same sign. Furthermore, 10% of the autocorrelations are outside the bounds $\pm 1.96/\sqrt{72}$, which is also an indication that we should reject the i.i.d. hypothesis.



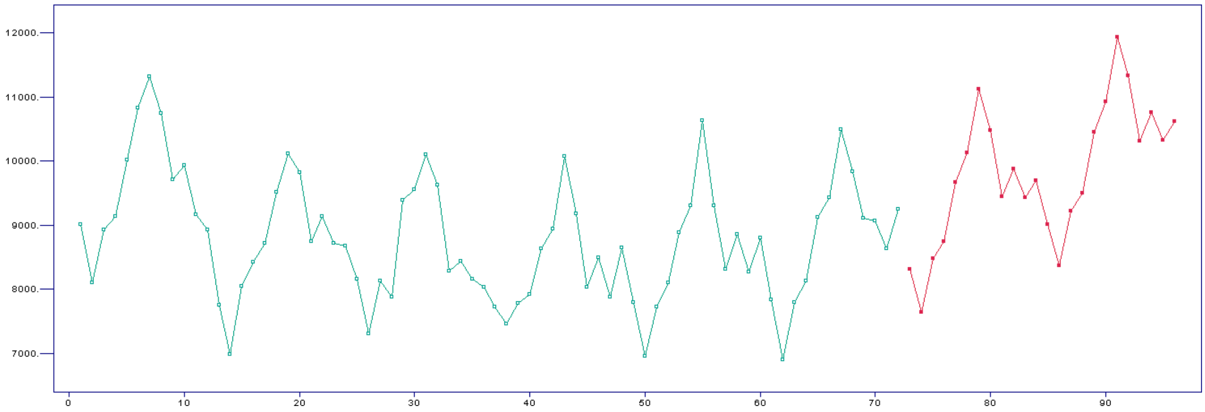
The results of the randomness tests for residuals are:

```
Ljung - Box statistic = 55.384 Chi-Square ( 20 ), p-value = .00004
Order of Min AICC YW Model for Residuals = 1
```

The sample value of the Ljung-Box statistic Q_{LB} with lag $h = 20$ is 51.84. Since the corresponding $p-$value is $0.00004 < 0.05$ we reject the i.i.d. hypothesis at a level of 0.05. The minimum-AICC Yule-Walker auto-regressive model for the data is of order 1 ($\neq 0$), which supports the evidence provided by the sample ACF and the Ljung-Box statistic against the i.i.d. hypothesis.

We forecast data for the years 1979 and 1980 (using an ARMA model) and display the prediction in red below.[71]

[71] The order is not provided.



Note the "jaggedness" of the predictions.

9.9.2 Asymptotic Normality

Asymptotic normality is an important concept in time series analysis for several reasons, some of which are outlined below.

- **Statistical Inference:** Asymptotic normality allows for the application of standard statistical tests (like $t-$tests and $z-$tests) for hypothesis testing and confidence interval construction. This simplifies the analysis by using familiar and well-understood techniques.
- **Large Sample Approximation:** Time series data often involve a large number of observations. The Central Limit Theorem suggests that the sampling distribution of many statistics will be approximately normal in large samples, making the results generalizable.
- **Parameter Estimation:** In many time series models, parameter estimates are often obtained through methods like Maximum Likelihood Estimation (MLE) or Ordinary Least Squares (OLS). Asymptotic normality of these estimators provides a basis for conducting inference about the parameters.
- **Model Validation:** When fitting models to time series data, it is important to know under what conditions the model will produce reliable forecasts. Knowing that a model's estimators are asymptotically normal helps in understanding its long-term behaviour.
- **Comparison of Models:** Asymptotic normality provides a common ground for comparing different models. This is especially useful in model selection criteria, like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), where the likelihood function plays a crucial role.
- **Robustness:** Models that possess asymptotically normal properties are often more robust to minor deviations from assumptions, like non-normality of errors in small samples.
- **Simplicity and Computation:** When the statistics of interest are asymptotically normal, it simplifies both the theoretical and computational aspects of the analysis. This allows for easier interpretation and faster computation, which is crucial in real-world applications where time and computational resources may be limited.

The **score function** of a probability density $f(x;\theta)$ is:

$$s(x;\theta) = \frac{\partial \log f(x;\theta)}{\partial\theta} = \frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}.$$

The **Fisher information** of the time series $\{X_t \mid t = 1, \dots, n\}$ is:

$$I_n(\theta) = \mathrm{Var}\left(\sum_{i=1}^{n} s(X_i;\theta)\right).$$

If the random variables are i.i.d., then the Fisher information collapses to

$$I_n(\theta) = n\mathrm{Var}(s(X_1;\theta)) = nI_1(\theta) = nI(\theta).$$

Lemma: the score function satisfies $\mathrm{E}[s(X;\theta)] = 0$.

Proof: we used the definition of the expectation to obtain:

$$\begin{aligned}\mathrm{E}[s(X;\theta)] &= \int s(X;\theta)f(x;\theta)dx = \int \frac{\partial \log f(x;\theta)}{\partial\theta}f(x;\theta)dx\\ &= \int \frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}f(x;\theta)dx = \int \frac{\partial f(x;\theta)}{\partial\theta}dx\\ &= \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \frac{\partial}{\partial\theta}[1] = 0. \qquad \blacksquare\end{aligned}$$

In the proof, we assumed that we could interchange integration and differentiation.[72] Using the above lemma, we then find:

$$I(\theta) = \mathrm{Var}(s(X;\theta)) = \mathrm{E}[s^2(X;\theta)].$$

Lemma: we have

$$I(\theta) = \mathrm{E}[s^2(X;\theta)] = -\mathrm{E}\left[\frac{\partial s(X;\theta)}{\partial\theta}\right] = -\mathrm{E}\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right].$$

Proof: first, we note that:

$$\mathrm{E}\left[s^2(X;\theta)\right] = \int \left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)dx = \int \frac{1}{f^2(x;\theta)}\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)dx = \int \frac{1}{f(x;\theta)}\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 dx.$$

Next, we see that:

$$\begin{aligned}-\mathrm{E}\left[\frac{\partial s(X;\theta)}{\partial\theta}\right] &= -\int \frac{\partial s(X;\theta)}{\partial\theta}f(x;\theta)dx = -\int \frac{1}{f^2(x;\theta)}\left(\frac{\partial^2 f(x;\theta)}{\partial\theta^2}f(x;\theta) - \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2\right)f(x;\theta)dx\\ &= -\int\left(\frac{\partial^2 f(x;\theta)}{\partial\theta^2}\right)dx + \int \frac{1}{f(x;\theta)}\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 dx = -\frac{\partial^2}{\partial\theta^2}\underbrace{\int f(x;\theta)dx}_{=1} + \mathrm{E}[s^2(X;\theta)] = \mathrm{E}[s^2(X;\theta)].\end{aligned}$$

Finally, we have:

$$\begin{aligned}\mathrm{E}\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right] &= \int \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}f(x;\theta)dx = \int \frac{\partial}{\partial\theta}\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)f(x;\theta)dx\\ &= \int \frac{\partial}{\partial\theta}\left(\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}\right)f(x;\theta)dx = \int \frac{\partial s(x;\theta)}{\partial\theta}f(x;\theta)dx = \mathrm{E}\left[\frac{\partial s(X;\theta)}{\partial\theta}\right]. \qquad \blacksquare\end{aligned}$$

[72] This holds for most reasonable density functions $f(x;\theta)$.

Example Consider $X \sim \text{Exp}(\beta)$. The density function of X is

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta},$$

with $\mathrm{E}[X] = \beta$, so that $\log f(x;\beta) = -\log(\beta) - x\beta$. The score function of X is thus

$$s(x;\beta) = -\frac{1}{\beta} + \frac{1}{\beta^2} x$$

and its derivative (w.r.t. β) is

$$-\frac{\partial s(x;\beta)}{\partial \beta} = -\frac{1}{\beta^2} + \frac{2}{\beta^3} x.$$

Hence,

$$I(\beta) = \mathrm{E}\left[-\frac{\partial s(x;\beta)}{\partial \beta}\right] = -\frac{1}{\beta^2} + \frac{2}{\beta^3}\mathrm{E}[X] = -\frac{1}{\beta^2} + \frac{2}{\beta^3}\beta = \frac{1}{\beta^2}.$$

Thus,

$$I_n(\beta) = \frac{n}{\beta^2}.$$

Note that for $\overline{X}_n = (X_1 + \ldots + X_n)/n$, we have

$$\text{Var}(\overline{X}_n) = \frac{\text{Var}(X)}{n} = \frac{\beta^2}{n},$$

so that $\text{Var}(\overline{X}_n) = I_n^{-1}(\beta)$.

This can be generalized to other distributions.

Theorem: under appropriate regularity conditions, we have

$$\frac{\widehat{\theta}_{\text{MLE}} - \theta}{\text{Var}\left(\sqrt{\widehat{\theta}_{\text{MLE}}}\right)} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where

$$\text{Var}\left(\sqrt{\widehat{\theta}_{\text{MLE}}}\right) = I_n^{-1}(\theta).$$

Proof: the MLE estimator, $\widehat{\theta}_{\text{MLE}}$, solves

$$\frac{\partial}{\partial \theta}\ell(\widehat{\theta}_{\text{MLE}}) = 0,$$

where ℓ is the log-likelihood. We apply Taylor's theorem to ℓ around $\theta = \widehat{\theta}_{\text{MLE}}$ to obtain

$$\ell(\theta) + \left(\widehat{\theta}_{\text{MLE}} - \theta\right)\frac{\partial^2}{\partial \theta^2}\ell(\theta) \approx 0.$$

Rearranging the terms, we get:

$$\sqrt{n}\left(\widehat{\theta}_{\text{MLE}} - \theta\right) = \frac{\frac{1}{\sqrt{n}}\frac{\partial}{\partial \theta}\ell(\theta)}{-\frac{1}{n}\frac{\partial^2}{\partial \theta^2}\ell(\theta)}.$$

Next, we show that the numerator converges to a normal distribution, whereas the denominator converges in probability to a constant.

Recall that

$$\ell(\theta) = \log f(X_1;\theta) + \cdots + \log f(X_n)(\theta)$$

and so

$$\frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(X_i;\theta) = \sum_{i=1}^{n} s(X_i;\theta).$$

We have already shown that $\mathrm{E}[s(X_i;\theta)] = 0$. Hence, the numerator can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,$$

where $Y_i = s(X_i;\theta)$ are i.i.d. with mean 0 and variance

$$\mathrm{Var}(s(X_i;\theta)) = \mathrm{E}[s^2(X_i,\theta)] = I(\theta).$$

Thus, we have

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathrm{E}[s^2(X_1,\theta)]\right) = \mathcal{N}\left(0, I(\theta)\right).$$

Similarly, the numerator can be written as

$$\frac{1}{n}\sum_{i=1}^{n} U_i,$$

where

$$U_i = \frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta), \qquad i = 1,\ldots,n$$

are i.i.d. random variables. From the previous Lemma, we can write

$$\mathrm{E}[U_i] = \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\right] = -I(\theta).$$

The **Law of Large Numbers**[73] then yields

$$-\frac{1}{n}\sum_{i=1}^{n} U_i \to I(\theta),$$

from which we conclude the result. ■

Example: Exponential Distribution (continued) Applying the theorem on the , we have

$$\sqrt{n}(\overline{X}_n - \beta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \beta^2\right).$$

9.9.3 Innovations

We now provide some of the details that allowed us to use innovations in Section 9.7.2. The goal is to try to determine a “good” prediction for the $n+1$th observation in the time series, which we denote by $P_n X_{n+1}$.

A by-product of the **innovation algorithm** is that we will also "predict" $X_1, \ldots, X_n$.[74]

[73] To wit: if the X_i are i.i.d. with finite mean μ, then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i = \mu.$$

There are two versions of this, the **weak law** and the **strong law**, depending on the type of convergence, but that falls outside the scope of these course notes, as does **convergence in distribution**, which basically states that the corresponding cumulative distribution functions F_n converge pointwise to a cumulative distribution function F.

[74] Of course, we do not need to predict these values since they have already been observed in practice, but we can use the **innovations**, i.e., the differences between the observed values X_i and the “predicted” values $\widehat{X}_i$ for model choice and estimation purposes.

As in Section 9.7.2, we define

$$\widehat{X}_{i+1} = P_i X_{i+1} = a_{i1}X_i + \cdots + a_{ii}X_1, \qquad i = 0, \ldots, n;$$

which is to say that $\widehat{X}_{n+1}$ is the **predicted value** for X_{n+1}, whereas $\widehat{X}_1, \ldots, \widehat{X}_n$ are the **"predicted" values** for $X_1, \ldots, X_n$.

We also define the column vectors

$$\mathbf{X}_n = (X_1, \ldots, X_n)^\top, \qquad \widehat{\mathbf{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)^\top, \qquad \mathbf{U}_n = (U_1, \ldots, U_n)^\top,$$

where $U_i = X_i - \widehat{X}_i$, $i = 1, \ldots, n$, are the *innovations* of the time series; a "good" prediction is such that these errors are small. As we have no data before $n = 1$ on which to base the prediction, we opt for $\widehat{X}_1 = \mathrm{E}[X_1] = 0$.[75]

Omitting $\widehat{X}_{n+1}$, we re-write the predictions, individually, as

$$\begin{aligned}
i = 0: &\quad \widehat{X}_1 = 0, \\
i = 1: &\quad \widehat{X}_2 = a_{1,1}X_1, \\
i = 2: &\quad \widehat{X}_3 = a_{2,1}X_2 + a_{2,2}X_1, \\
i = 3: &\quad \widehat{X}_4 = a_{3,1}X_3 + a_{3,2}X_2 + a_{3,1}X_1, \\
&\quad \vdots \\
i = n-2: &\quad \widehat{X}_n = a_{n-1,1}X_{n-1} + \cdots + a_{n-1,n-1}X_1,
\end{aligned}$$

or, simultaneously, as

$$\widehat{\mathbf{X}}_n = \mathbf{A}^* \mathbf{X}_n,$$

where

$$\mathbf{A}^* = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{1,1} & 0 & 0 & \cdots & 0 \\ a_{2,2} & a_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1,n-1} & a_{n-1,n-2} & \cdots & a_{n-1,1} & 0 \end{pmatrix}.$$

Note that the matrix is lower diagonal.

We write

$$\mathbf{U}_n = \mathbf{X}_n - \widehat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{A}^*\mathbf{X}_n = \mathbf{A}\mathbf{X}_n,$$

where $\mathbf{A} = \mathbf{I}_n - \mathbf{A}^*$. This matrix is invertible since $\det(\mathbf{A}) = 1 \neq 0$.

Let $\mathbf{C} = \mathbf{A}^{-1}$ and $\mathbf{B} = \mathbf{C} - \mathbf{I}_n$; then we can write

$$\mathbf{X}_n = \mathbf{C}\mathbf{U}_n, \quad \text{and} \quad \widehat{\mathbf{X}}_n = (\mathbf{C} - \mathbf{I}_n)\mathbf{U}_n = \mathbf{B}\mathbf{U}_n,$$

representing the "predicted" values in terms of the innovations $\mathbf{U}_n$ and the lower diagonal matrix $\mathbf{B}$ (indeed, $\mathbf{C}$ must be lower diagonal, as is $\mathbf{I}_n$, so that $\mathbf{B} = \mathbf{C} - \mathbf{I}_n$ is also lower diagonal).

[75] Remember, we are assuming that $\{X_t\}$ is a stationary time series.

We can write the second of these equations as

$$\widehat{\mathbf{X}}_n = (\mathbf{C} - \mathbf{I}_n)\,\mathbf{U}_n = \begin{pmatrix} \widehat{X}_1 \\ \widehat{X}_2 \\ \widehat{X}_3 \\ \vdots \\ \widehat{X}_n \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 0 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{pmatrix},$$

and the first as

$$\mathbf{X}_n = \mathbf{C}\mathbf{U}_n = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{pmatrix}.$$

Note that the coefficients $\theta_{k,j}$ have nothing to do with the Durbin-Levinson algorithm (see Section 9.4.2).

From the above matrix equation, we have, for instance,

$$\begin{aligned} \widehat{X}_1 &= 0, \\ \widehat{X}_2 &= \theta_{1,1}(X_1 - \widehat{X}_1), \\ \widehat{X}_3 &= \theta_{2,1}(X_2 - \widehat{X}_2) + \theta_{2,2}(X_1 - \widehat{X}_1). \end{aligned}$$

The prediction of X_3 is then based on the first and the second innovations $X_1 - \widehat{X}_1$ and $X_2 - \widehat{X}_2$.

In general, for a MA(q) model, we can write

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

For an ARMA(p, q) model, we have instead

$$\widehat{X}_{i+1} = \begin{cases} 0 & i = 0 \\ \phi_1 X_i + \cdots + \phi_p X_{i+1-p} + \displaystyle\sum_{j=1}^{i} \theta_{i,j}(X_{i+1-j} - \widehat{X}_{i+1-j}) & i \geq 1 \end{cases}.$$

The only thing left is to determine how to evaluate the coefficients $\theta_{i,j}$; this is the subject of the next theorem.

Innovation Algorithm: assume that $\{X_t\}$ is a stationary time series with mean 0. Let $v_i = \mathrm{E}[(X_{i+1} - \widehat{X}_{i+1})^2]$, $i \geq 0$, and $v_0 = \mathrm{E}[X_1^2] = \gamma_X(0)$.

Then

$$\theta_{n,n-i} = v_i^{-1}\left(\gamma_X(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j}\theta_{n,n-j}v_j\right), \qquad 0 \leq i < n,$$

$$v_n = \gamma_X(n-1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

Example Consider the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, where $\mathrm{E}[Z_t] = 0$ and $\mathrm{Var}(Z_t) = \sigma_Z^2$. Recall that $\gamma_X(0) = \sigma_Z^2(1+\theta^2)$, $\gamma_X(1) = \theta\sigma_Z^2$ and $\gamma_X(h) = 0$, $h > 1$.

We have:

- $n = 1$
  - $i = 0$: $v_0 = \gamma_X(0) = \sigma_Z^2(1+\theta^2)$, $\theta_{1,1} = v_0^{-1}\gamma_X(1) = \rho_X(1)$ and $v_1 = \gamma_X(0) - \theta_{1,1}^2 v_0$
- $n = 2$
  - $i = 0$: $\theta_{2,2} = v_0^{-1}\gamma_X(2) = 0$
  - $i = 1$: $\theta_{2,1} = v_1^{-1}\gamma_X(1)$ and $v_2 = v_n = [1 + \theta^2 - v_1^{-1}\theta^2\sigma_Z^2]\sigma_Z^2$
- general n
  - $i = 0, \dots, n-2$: $\theta_{n,j} = 0, 2 \le j \le n$,
  - $i = n-1$: $\theta_{n,1} = v_{n-1}^{-1}\gamma_X(1)$ and $v_n = [1 + \theta^2 - v_{n-1}^{-1}\theta^2\sigma_Z^2]\sigma_Z^2$

Important Property The innovations $U_1, \dots, U_n$ are **uncorrelated**: we have $\mathrm{Cov}(U_i, U_j) = 0$ for $i \neq j$.[76] Remembering that the sequence is centered, we have:

$$\Gamma_n = \mathrm{E}\left[\mathbf{X}_n\mathbf{X}_n^\top\right] = \mathrm{E}[\mathbf{C}\mathbf{U}_n\mathbf{U}_n^\top\mathbf{C}^\top] = \mathbf{C}\mathrm{E}[\mathbf{U}_n\mathbf{U}_n^\top]\mathbf{C}^\top = \mathbf{C}\mathbf{D}\mathbf{C}^\top$$

where $\mathbf{D}$ is the diagonal matrix with entries $v_0, \dots, v_{n-1}$, where the values $v_i = \mathrm{E}[U_i^2] = \mathrm{E}[(X_i - \widehat{X}_i^2)]$ are the same quantities as those in the innovation algorithm.

[76] This is not trivial to show.

9.10 Exercises

1. Show that the set $\mathcal{T}_n$ of stationary time series of length n is a vector subspace (over $\mathbb{R}$) of the set of all time series.
2. Let $\{Z_t\}$ be independent normalrandom variables with mean 0 and variance σ_Z^2. Let a, b, c be constants. Which of the following processes are stationary? Evaluate the mean and the autocovariance functions.
   a) $X_t = Z_t\cos(at) + Z_{t-1}\sin(at)$.
   b) $X_t = a + bZ_t + cZ_{t-2}$.
   c) $X_t = Z_t Z_{t-2}$.
3. Let $\{Z_t\}$ be a sequence of independent normal random variables with mean 0 and variance $\sigma_Z^2 = 1$. Consider the sequence
   $$X_t = Z_t + (Z_{t-1}^2 - 1), t = 1, 2, \dots.$$
   a) Show that $\mathrm{E}[X_t] = 0$.
   b) Show that $\mathrm{E}[X_t X_{t+h}] = 0$ for $h \neq 0$.
4. Let $\{Z_t\}$ be independent random variables with mean 0 and variance σ_Z^2. Let $\{Y_t\}$ be a stationary sequence with a covariance function $\gamma_Y(h)$. Assume that the sequences $\{Z_t\}$ and $\{Y_t\}$ are independent from each other. Define $X_t = Y_t Z_t$. Verify that $\mathrm{Cov}(X_t, X_{t+h}) = 0$ for $h \geq 1$.

5. Show that the PACF between X_1 and X_3 when removing the effect of X_2 is:

$$\rho_{1,3;2} = \frac{\mathrm{Corr}(X_1, X_3) - \mathrm{Corr}(X_1, X_2) \cdot \mathrm{Corr}(X_2, X_3)}{\sqrt{\left(1 - \mathrm{Corr}^2(X_1, X_2)\right)\left(1 - \mathrm{Corr}^2(X_2, X_3)\right)}}.$$

6. Let $\{Z_t\}$ be independent random variables with mean 0 and variance σ_Z^2. Consider the model $X_t = Z_t + Z_{t-1}$. Evaluate $\alpha(1)$ and $\alpha(2)$.
7. Let $\{Z_t\}$ be independent random variables with mean 0 and variance σ_Z^2. Determine if the following processes are stationary and causal.
   a) $X_t + 0.2X_{t-1} + 0.48X_{t-2} = Z_t$.
   b) $X_t + 1.6X_{t-1} = Z_t - 0.42Z_{t-1} + 0.04Z_{t-2}$.
8. Derive a linear representation of the general ARMA$(1, 2)$ model.
9. Derive a linear representation of the general ARMA$(1, q)$ model.
10. Derive a linear representation of the AR(2) model $X_t = \phi X_{t-2} + Z_t$.
11. Use the linear representation of ARMA$(1, 1)$ to compute its covariance function.
12. Use the recursive method to compute the covariance function of the general AR(2) model.
13. This is an exercise about simulating time series.
    a) Generate ARMA(p, q) sequence X_t. You have to choose p, q as well as the required parameters. Make sure that the chosen parameters imply existence of a stationary solution.
    b) Identify the model using ACF and PACF. Include graphs of ACF and PACF (2 graphs).
    c) Add a linear or a polynomial trend m_t. The new sequence is $Y_t = m_t + X_t$.
    d) Estimate m_t using all three methods:
       - parametric method;
       - exponential smoothing;
       - moving average smoothing with your chosen Q.
    e) For each of the three methods, plot Y_t and the estimated trend $\widehat{m}_t$ on the same graphs (3 graphs).
    f) For each of the three methods, compute $\widehat{X}_t = Y_t - \widehat{m}_t$. Plot residuals (that is $\widehat{X}_t$) (3 graphs).
    g) Analyze $\widehat{X}_t$ using ACF and PACF. Graph ACF and PACF for all three methods (6 graphs). Identify the most likely ARMA model for the data. Compare with your identification in b).
14. Download a data set from this page ↗ or use your own data set.
    a) Remove the trend using any of the methods, if needed, to obtain a stationary time series. State the chosen $\widehat{m}_t$.
    b) Plot the original sequence together with the estimated trend.
    c) Plot the stationary part, then its ACF and PACF. Comment on the results when it comes to the choice of a model.
15. Assume that Z_t are i.i.d random variables with mean 0 and variance σ_Z^2.
    a) Apply the Yule-Walker procedure to obtain $P_n X_{n+2}$ (two step prediction) for AR(1) model $X_t = \phi X_{t-1} + Z_t$, $|\phi| < 1$. Compute the corresponding $\mathrm{MSPE}_n(2)$. Can you guess a general formula for $P_n X_{n+k}$?
    b) Apply the Yule-Walker procedure to obtain $P_n X_{n+1}$ for AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Compute the corresponding $\mathrm{MSPE}_n(1)$.
16. Consider the ARMA$(1, 1)$ model $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, $|\phi| < 1$, $\theta \in \mathbb{R}$, where Z_t are i.i.d. random variables with mean 0 and variance σ_Z^2. The goal is to find the best linear predictor $P_n X_{n+1}$ of X_{n+1} based on $X_1, \ldots, X_n$.
    a) Let $n = 1$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n; 1)$ to obtain a_1 in $P_1 X_2 = a_1 X_1$.
    b) Let $n = 2$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n; 1)$ to obtain coefficients a_1, a_2 in $P_2 X_3 = a_1 X_2 + a_2 X_1$.

    Hint: We have the following formulas for the covariance function:

$$\gamma_X(0) = \sigma_Z^2 \left[1 + \frac{(\phi + \theta)^2}{1 - \phi^2}\right], \quad X(1) = \sigma_Z^2 \left[(\phi + \theta) + \frac{(\phi + \theta)^2 \phi}{1 - \phi^2}\right], \quad \gamma_X(h) = \phi^{h-1} \gamma_X(1), \qquad h \geq 2.$$

17. Consider the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, $\theta \in \mathbb{R}$, where Z_t are i.i.d. random variables with mean 0 and variance σ_Z^2. The goal is to find the best linear predictor $P_n X_{n+1}$ of X_{n+1} based on $X_1, \dots, X_n$.
    a) Let $n = 1$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n;1)$ to conclude that
    $$P_1 X_2 = \frac{\gamma_X(1)}{\gamma_X(0)} X_1 = \frac{\theta}{1+\theta^2} X_1.$$
    b) Let $n = 2$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n;1)$ to obtain coefficients a_1, a_2 in $P_2 X_3 = a_1 X_2 + a_2 X_1$.
    c) Let $n = 2$. Apply the Durbin-Levinson algorithm to get $P_2 X_3 = \phi_{2,1} X_2 + \phi_{2,2} X_1$.
18. Consider a stationary ARMA(1, 1) model
    $$(X_t - \mu) = \phi(X_{t-1} - \mu) + Z_t + \theta Z_{t-1}.$$
    Evaluate $\sum_{h=-\infty}^{\infty} \gamma_X(h)$.
19. Assume that Z_t are i.i.d random variables with mean 0 and variance σ_Z^2. Consider the AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$.
    a) Derive confidence intervals for $\widehat{\phi}_1$ and $\widehat{\phi}_2$.
    b) Assume that $n = 100$, $\widehat{\gamma}_X(0) = 3$, $\widehat{\gamma}_X(1) = 1.5$, $\widehat{\gamma}_X(2) = 0.5$. Use a) to get the confidence intervals.
20. In this question we develop Yule-Walker estimators for the AR(1) and ARMA(1, 1) models and study their numerical performance. Recall that the Yule-Walker estimator for the AR(1) model is
    $$\widehat{\phi} = \frac{\widehat{\gamma}_X(1)}{\widehat{\gamma}_X(0)} = \widehat{\rho}_X(1), \quad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\phi}\widehat{\gamma}_X(1) = \widehat{\gamma}_X(0) - \widehat{\rho}_X(1)^2 \widehat{\gamma}_X(0).$$
    a) Numerical experiment for AR(1):
        i. Load the file `Data-AR.txt` into R. This is a data set generated from a AR(1) model with $\phi = 0.8$.
        ii. Type `var(Data)` to obtain $\widehat{\gamma}_X(0)$.
        iii. Type `ACF<-acf(Data)`. Then type `ACF`. You will get $\widehat{\rho}_X(h)$, the estimators of $\rho_X(h)$. The second entry is $\widehat{\rho}_X(1) = \widehat{\phi}$.
        iv. Write the final values for $\widehat{\phi}$ and $\widehat{\sigma}_Z^2$.
        v. Compare the estimated $\widehat{\phi}$ with the true ϕ.
    b) Consider the ARMA(1, 1) model $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$, $|\phi| < 1$; the sequence X_t is causal. Apply the Yule-Walker procedure to obtain the estimators for ϕ, θ and $\sigma_Z^2 = \mathrm{Var}(Z_t)$.
    c) Numerical experiment for ARMA(1, 1):
        i. Load the file `Data-ARMA.txt` into R. This is a data set generated from a ARMA(1, 1) model with $\phi = 0.8$ and $\theta = 1$.
        ii. Identify the values of $\widehat{\phi}$, $\widehat{\theta}$, and $\widehat{\sigma}_Z^2$.
        iii. Compare the estimated $\widehat{\phi}$ with the true ϕ. Which estimate is more accurate: ARMA(1, 1) or AR(1)?
21. a) One hundred observations from AR(1) yield the following sample statistics:
    $$\overline{x} = 0, \quad \widehat{\gamma}_X(0) = 1.1, \quad \widehat{\rho}_X(1) = 0.42.$$
        i. Find the Yule-Walker estimators of ϕ and σ_Z^2.
        ii. Write the confidence interval for ϕ.
        iii. If $X_{100} = 1.5$, what is the predicted value of X_{101}? What is the squared error of this prediction?
    b) Two hundred observation from AR(2) yields the following sample statistics:
    $$\overline{x} = 3.82, \quad \widehat{\gamma}_X(0) = 1.15, \quad \widehat{\rho}_X(1) = 0.427, \widehat{\rho}_2 = 0.475.$$
        i. Find the Yule-Walker estimators of ϕ_1, ϕ_2 and σ_Z^2.
        ii. Is the estimated model causal?.
        iii. If $X_{100} = 3.84$ and $X_{99} = 3.26$, what is the predicted value of X_{101}?

22. Consider the general AR(1) model. Derive the MLE for ϕ and σ_Z^2.

23. We have already fitted an AR(4) model to US unemployment data, and estimated the parameters using the Yule-Walker procedure.

    a) Calculate the residuals, and plot their ACF and PACF. Is the chosen AR(4) model appropriate?

    b) Predict the next observation in the time series.

    c) Backcast the past observations and verify the quality of the "prediction" by plotting the original values and the "predicted" values on the same graph. Compute the squared error of that prediction.

    d) Now, pretend that the model is AR(1). Estimate the model's parameters. Repeat b)-d). State conclusions.

24. Use the Lake Huron data for this question (an in-built dataset in R).

    a) Type the following code at the prompt.

```
My.TS <- LakeHuron
help(LakeHuron)
mean = mean(My.TS)
My.Centered.TS <- My.TS - mean(My.TS)
```

    b) Fit an AR(2) model to the data using the Yule-Walker estimator. Obtain $\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\sigma}_Z^2$.

```
fit.ar <- ar(My.Centered.TS, method="yule-walker")
```

    c) Verify that the command `ar()` leads to the correct Yule-Walker estimator.

        i. At the prompt, type the following code.

```
ACF <- acf(LakeHuron)
var(LakeHuron)
```

        Read off $\widehat{\rho}_X(1)$ and $\widehat{\rho}_X(2)$ and $\widehat{\gamma}_X(0)$. Using theis information, compute $\widehat{\gamma}_X(1)$, $\widehat{\gamma}_X(2)$.

        ii. Create a vector $(\widehat{\gamma}_X(1), \widehat{\gamma}_X(2))$ and call it `gamma.vector`.

        iii. Create a matrix $\hat{\Gamma}_2$ and call it `Gamma.matrix`.

        iv. Compute $\hat{\Gamma}_2^{-1} * \gamma_{X,2}$ by typing in

```
solve(Gamma.matrix)%*%gamma.vector
```

        Compare the results with those of part b).

25. When $p \geq 2$, it can be rather difficult to identify the right p from the data. Start by loading `BadData.txt` into the R variable X.

    a) Based on the ACF and PACF of the data, argue that an AR(3) model can be reasonably chosen.

    b) Type the following code at the prompt.

```
(fit.ar <- ar(X,method="mle"))
```

    What order does `ar()` select? Denote this order by `p`.

    c) Using `p` from the step above, type the following code at the prompt.

```
(fit.arima  <- arima(X,order=c(3,0,0)))
(fit.arima1 <- arima(X,order=c(p,0,0)))
```

    Why did MLE select `p` and not 3?

26. Derive the formulas for the spectral density of MA(1) and ARMA(1, 1).

27. Assume that (X_1, X_2) is a vector of dependent normal random variables with mean 0 and variance σ^2 each. Assume that the covariance matrix is given by

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

In other words, ρ is the correlation between X_1 and X_2. Assuming that σ is known, find the maximum likelihood estimator of ρ.

28. Let $\{Z_t\}$ be an i.i.d. sequence of normal random variables with mean 0 and variance $\sigma_Z^2 = 1$. Define

$$X_t = \begin{cases} Z_t, & t \text{ even}, \\ (Z_{t-1}^2 - 1)/\sqrt{2}, & t \text{ odd}. \end{cases}$$

Find $\mathrm{E}[X_t]$, $\gamma_X(t, t+1)$ and $\gamma_X(t, t+2)$.

29. Consider the sequence

$$X_t = Z_t Z_{t-1} + 0.5 Z_{t-1},$$

where Z_t are i.i.d random variables with mean 0 and variance σ_Z^2.

a) Show that $\mathrm{E}[X_t] = 0$ for all t.
b) Compute $\gamma_X(t, t+h) = \mathrm{E}[X_t X_{t+h}]$ for $h = 0, 1, 2$.
c) Is the sequence X_t stationary? Why?

30. Assume that A and B are random variables with mean 0 and variance σ^2. Assume also that $\mathrm{Cov}(A, B) = 0$. Let $\omega \in \mathbb{R}$ and define

$$X_t = A\cos(at) + B\sin(bt), \quad a, b \neq 0.$$

Is $\{X_t\}$ stationary?

31. Consider the ARMA(2, 1) model given by

$$X_t - 0.75X_{t-1} + 0.5625X_{t-2} = Z_t + 2.25Z_{t-1}.$$

Is this process causal? Is this process stationary?

32. Consider the linear process given by

$$X_t = \sum_{j=0}^{\infty} (\phi^j + \phi^{j+1}) Z_{t-j},$$

where $|\phi| < 1$ and Z_t is an i.i.d sequence with mean 0 and variance σ_Z^2. Write the formula for $\gamma_X(h)$, $h \geq 0$.

33. Consider the ARMA(1, 2) model

$$X_t - \phi X_{t-1} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2},$$

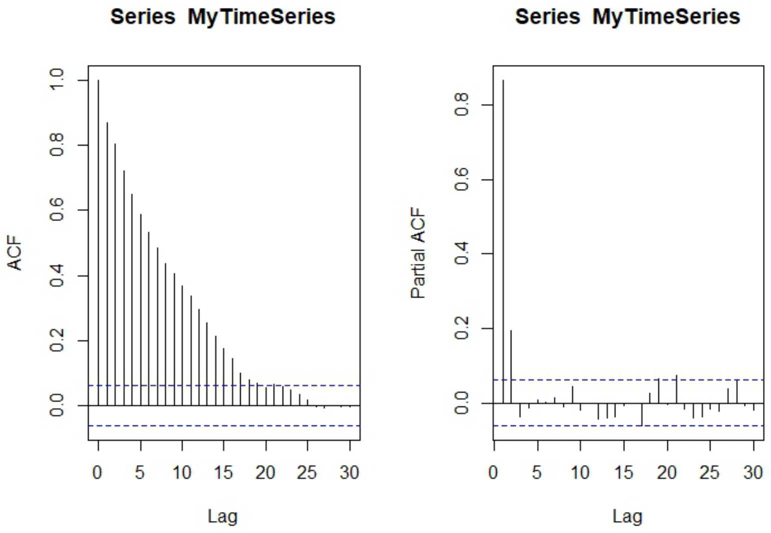
where $|\phi| < 1$, $\theta_1, \theta_2 \in \mathbb{R}$, and Z_t is an i.i.d sequence with mean 0 and variance σ_Z^2. Derive the linear representation for X_t, i.e. find the coefficients ψ_j in $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$.

34. Consider a stationary AR(3) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} = Z_t$. Use the recursive method to conclude

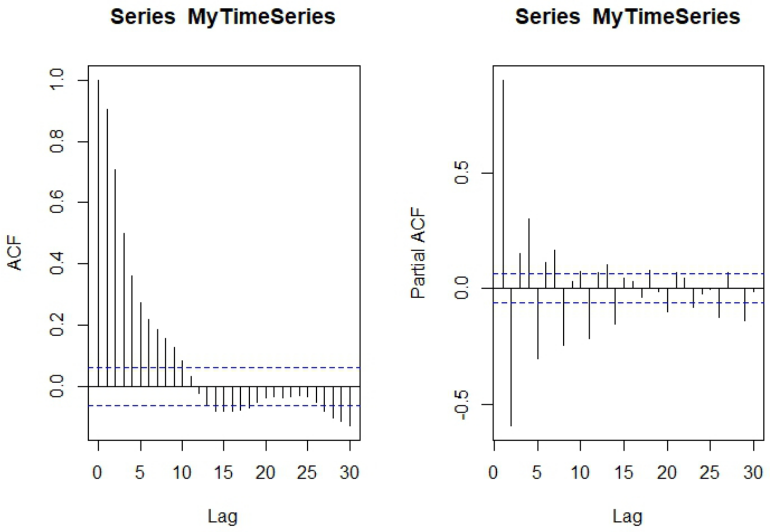
$$\gamma_X(h) = \phi_1 \gamma_X(h-1) + \phi_1 \gamma_X(h-2) + \phi_1 \gamma_X(h-3), \quad h \geq 3.$$

35. Derive the linear representation of a stationary AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$.
36. Write the non-causal linear representation of an AR(1) $X_t = \phi X_{t-1} + Z_t$ with $\phi > 1$.
37. Obtain the coefficients $\phi_{1,1}$, $\phi_{2,2}$, $\phi_{3,3}$ for the AR(1) model. Compare with the Yule-Walker procedure.
38. Obtain the coefficients $\phi_{1,1}$, $\phi_{2,2}$, $\phi_{2,1}$ for the AR(2) model.
39. If $\{X_t\}$ and $\{Y_t\}$ are two uncorrelated stationary processes, show that $\{X_t + Y_t\}$ is a stationary process. What is its ACVF?

40. Identify the ARMA model based on the ACF and PACF below.



41. Identify the ARMA model based on the ACF and PACF below.



42. Consider the AR(1) model $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and the random variables Z_t are i.i.d. with mean 0 and variance σ_Z^2. Prove that $\phi_{n,n} = 0$ for all $n \geq 2$ (recall that $\phi_{n,n}$ = partial autocovariance at lag n).

43. a) Let X and Y be random variables with $\mathrm{E}[Y^2] < \infty$. Show that $\mathrm{E}[Y \mid X]$ minimizes

$$\mathrm{MSE} = \mathrm{E}\left([Y - g(X)]^2\right)$$

over all functions g such that $\mathrm{E}\left([g(X)]^2\right) < \infty$.

b) Generalize to $X_1, \ldots, X_n$ to show that $\mathrm{E}[X_{n+1}|X_1, \ldots, X_n]$ minimizes

$$\mathrm{MSE} = \mathrm{E}\left([X_{n+1} - g(X_1, \ldots, X_n)]^2\right)$$

over all functions g such that $\mathrm{E}\left([g(X_1, \ldots, X_n)]^2\right) < \infty$.

c) If $X_1, X_2, \ldots$ are i.i.d. with $\mathrm{E}[X_i^2] < \infty$ and $\mathrm{E}[X_i] = \mu$ for all i, where μ is known, what is the minimum mean square predictor of X_{n+1} in terms of $X_1, \ldots, X_n$?

d) If $X_1, \ldots, X_n$ are i.i.d. with $\mathrm{E}[X_i^2] < \infty$ and $\mathrm{E}[X_i] = \mu$ for all i, where μ is unknown, show that the best linear unbiased estimator (BLUE) of μ is $\overline{X}$.

44. Let $\{Z_t\}$ be i.i.d. with $Z_t \sim N(0,1)$ and define

$$X_t = \begin{cases} Z_t & \text{if } t \text{ is even} \\ \frac{Z_{t-1}^2 - 1}{\sqrt{2}} & \text{if } t \text{ is odd} \end{cases}$$

   a) Show that $\{X_t\}$ is WN$(0,1)$ but not i.i.d. $(0,1)$ noise.
   b) Find E$[X_{n+1}|X_1,\dots,X_n]$ for n even and for n odd and compare the results.

45. Consider the time series

$$X_t = \underbrace{m_t}_{\text{local trend}} + \underbrace{Z_t}_{\text{noise}}$$

and the simple moving average filter with weights $a_j = (2q+1)^{-1}$ for $-q \le j \le q$.

   a) If $m_t = c_0 + c_1 t$ show that $\sum_{j=-q}^{q} a_j m_{t-j} = m_t$.
   b) If $\{Z_t\}_{t\in\mathbb{Z}}$ are i.i.d. with mean 0 and variance σ^2, show that the moving average

$$A_t = \sum_{j=-q}^{q} a_j Z_{t-j}$$

   is small in the sense that E$[A_t] = 0$ and Var$(A_t^2) = \frac{\sigma^2}{2q+1}$.

46. Compute the ACF of the model $X_t - 0.6X_{t-1} = Z_t + 1.2Z_{t-1}$, where Z_t is WN$(0,\sigma^2)$.

47. Let X_t denote a non-causal AR(1) process $X_t = \phi X_{t-1} + Z_t$ where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and $|\phi| > 1$.

   a) Denote $W_t = X_t - \frac{1}{\phi}X_{t-1}$. Show that $\{W_t\} \sim$ WN$(0,\sigma_w^2)$ and express σ_w^2 in terms of σ^2 and ϕ.
   b) Show that $Y_t = \frac{1}{\phi}Y_{t-1} + W_t$ is causal and has the same ACVF as X_t above.
   c) Find the causal form of $X_t = 1.2X_{t-1} + Z_t$ where $\{Z_t\} \sim$ WN$(0,1)$.

48. Let $\{Y_t\}$ be the AR(1) plus white noise time series defined by $Y_t = X_t + W_t$ where $\{W_t\} \sim$ WN$(0,\sigma_w^2)$, $\{X_t\}$ is the AR(1) process $X_t - \phi X_{t-1} = Z_t$, $|\phi| < 1$, $\{Z_t\} \sim$ WN$(0,\sigma_z^2)$, E$[X_s Z_t] = 0$ for all $s < t$ and E$[W_s Z_t] = 0$ for all s,t.

   a) Show that $\{Y_t\}$ is stationary and find its ACVF.
   b) Show that the time series $U_t = Y_t - \phi Y_{t-1}$ is 1−correlated and hence is an MA(1) process.
   c) Conclude from b) that $\{Y_t\}$ is an ARMA(1,1) process and express the three parameters of this model in terms of ϕ, σ_w^2 and σ_z^2.

49. Let $\{X_t\}$ be an AR(p) process defined by

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and E$[X_s Z_t] = 0$ for all $s < t$.

   a) Show that for $n > p$, the best linear predictor $P_n X_{n+1}$ is $\phi_1 X_n + \cdots + \phi_p X_{n-p}$.
   b) Compute the mean square error of this forecast.

50. Let $\{X_t\}$ be an MA(1) process defined by $X_t = Z_t - \theta Z_{t-1}, \quad t \in \mathbb{Z}$ where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and $|\theta| < 1$.

   a) Show that the best linear predictor $\tilde{P}_n X_{n+1}$ based on $\{X_j | j \le n\}$ is

$$\tilde{P}_n X_{n+1} = -\sum_{j=1}^{\infty} \theta^j X_{n+1-j}.$$

   b) Find the mean square error of $\tilde{P}_n X_{n+1}$.

51. In the innovations algorithm, show that for each $n \ge 2$, the innovation $X_n - \hat{X}_n$ is uncorrelated with $X_1,\dots,X_{n-1}$. Conclude also that the innovation $X_n - \hat{X}_n$ is uncorrelated with the innovations $X_1 - \hat{X}_1,\dots,X_{n-1} - \hat{X}_{n-1}$.

52. Let X_1, X_2, X_4, X_5 be observations from an MA(1) process defined by $X_t = Z_t - \theta Z_{t-1}$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.
    a) Find the best linear estimate of the missing value X_3 in terms of X_1, X_2.
    b) Find the best linear estimate of the missing value X_3 in terms of X_4, X_5.
    c) Find the best linear estimate of the missing value X_3 in terms of X_1, X_2, X_4, X_5.
    d) Compute the mean squared error of the previous estimates. Which one of them is the best estimate for X_3.
53. Let $\{X_t\}$ be an AR(p) process defined by $X_t = \phi X_{t-1} + Z_t$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.
    a) Show that $\sqrt{n}\frac{\hat{\rho}(1)-\rho(1)}{\sqrt{1-(\rho(1))^2}}$ has asymptotically standard normal distribution $N(0, 1)$.
    b) If $n = 100$ and $\hat{\rho}(1) = 0.64$, build an approximate 95% confidence interval for ϕ.
54. Let $\{X_t\}$ be an AR(1) process defined by $X_t = \phi X_{t-1} + Z_t$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ with the usual hypotheses. For $h = 1, 2, \ldots$, compute the $h-$step ahead forecast $P_n X_{n+h} = \hat{X}_n(h)$ in terms of $\{1, X_n, \ldots, X_1\}$ and find its mean square error.
55. Suppose that $\{X_t\}$ is a non-causal and non-invertible ARMA(1, 1) process satisfying $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, with $|\phi|, |\theta| > 1$. Define $\tilde{\phi}(B) = 1 - \frac{B}{\phi}$ and $\tilde{\theta}(B) = 1 + \frac{B}{\theta}$ and let $W_t = \tilde{\theta}^{-1}(B)\tilde{\phi}(B)X_t$.
    a) Show that $\{W_t\}$ has a constant spectral density function.
    b) Conclude that $\{W_t\} \sim \mathrm{WN}(0, \sigma_w^2)$. Give an explicit formula for σ_w^2 in terms of σ^2, θ and ϕ.
    c) Deduce that $\tilde{\phi}(B)X_t = \tilde{\theta}(B)W_t$, so that $\{X_t\}$ is a causal and invertible ARMA(1, 1) process relative to the white noise $\{W_t\}$ (see [1] for definition).
56. Let $\{X_t\}$ be the MA(1) process defined by $X_t = Z_t + \theta Z_{t-1}$ where $|\theta| < 1$ and $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. The best linear predictor of X_{n+1} based on $X_1, \ldots, X_n$ is

$$\hat{X}_{n+1} = \phi_{n,1}X_n + \cdots + \phi_{n,n}X_1,$$

where $\phi_n = (\phi_{n,1}, \ldots, \phi_{n,n})^\top$ satisfies $R_n \phi_n = \rho_n$; $\rho_n = (\rho(1), \ldots, \rho(n))^\top$. Show that

$$\phi_{n,n-j} = (1 + \theta^2 + \cdots + \theta^{2j})(-\theta)^{-j}\phi_{n,n} \quad \text{for } 1 \le j < n$$

and conclude that the PACF of the process is

$$\phi_{n,n} = -\frac{(-\theta)^n}{1 + \theta^2 + \cdots + \theta^{2n}}.$$

57. Let $\{X_t\}$ be a causal ARMA(1, 1) process of the form $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. Consider the innovation algorithm

$$\hat{X}_{n+1} = \phi X_n + \theta_{n,1}(X_n - \hat{X}_n)$$

for this process. It can be shown that the innovation algorithm coefficients $\theta_{n,1}$ can be found recursively as follows:

$$r_0 = \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}, \quad \theta_{n,1} = \frac{\theta}{r_{n-1}}, \quad r_n = 1 + \theta^2\left(1 - \frac{1}{r_{n-1}}\right).$$

   a) With the notation $y_n = \frac{r_n}{r_n - 1}$, show that

$$y_n = \theta^{-2}y_{n-1} + 1, \quad n \ge 1.$$

   b) Deduce that

$$y_n = \theta^{-2n}y_0 + \sum_{j=1}^{n} \theta^{-2(j-1)} := A(n).$$

   Determine r_n and $\theta_{n,1}$ for all $n \ge 1$.

   c) Evaluate the limits of r_n and $\theta_{n,1}$ for $|\theta| < 1$ as $n \to \infty$.

58. a) Compute and plot the spectral density of the stationary series $\{X_t\}$ satisfying

$$X_t - 0.99X_{t-3} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0,1).$$

b) Does the spectral density suggest that the sample paths of $\{X_t\}$ will exhibit approximately oscillatory behaviour? If so, then with what period?

c) Simulate and plot a realization of $X_1, \dots, X_{60}$. Does the graph of the realization support the conclusion in part b)?

d) Compute the spectral density of the filtered process

$$Y_t = \frac{1}{3}(X_{t-1} + X_t + X_{t+1})$$

and compare the numerical values of the spectral densities of $\{X_t\}$ and $\{Y_t\}$ at frequency $\lambda = \frac{2\pi}{3}$ radians per unit time. What effect would you expect the filter to have on the oscillations of $\{X_t\}$?

e) Apply the filter of part d) to the realization of part c). Comment on the result.

59. Consider the sunspot numbers $\{X_t, t = 1, \dots, 100\}$, filed as SUNSPOTS.TSM.

a) Compute the sample autocovariances $\hat{\gamma}(0)$, $\hat{\gamma}(1)$, $\hat{\gamma}(2)$ and $\hat{\gamma}(3)$.

b) Use these values to find the Yule-Walker estimates of ϕ_1, ϕ_2 and σ^2 in the AR(2) model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0,\sigma^2)$$

for the mean corrected series $Y_t = X_t - \overline{X}_t$.

c) Assuming that the data really are a realization of an AR(2) process, find 95% C.I. for $\hat{\phi}_1$ and $\hat{\phi}_2$.

d) Use the Durbin-Levinson algorithm to compute the sample PACF $\hat{\phi}_{1,1}$, $\hat{\phi}_{2,2}$ and $\hat{\phi}_{3,3}$ of the sunspot series. Is the value of $\hat{\phi}_{3,3}$ compatible with the assumption that the data are generated from an AR(2) process? Use significance level $\alpha = 0.05$.

60. Use the ARMA Process Gaussian Likelihood formula to prove that if $\{X_t\}$ is an AR(p) process with the equation $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$, $\{Z_t\} \sim \mathrm{WN}(0,\sigma^2)$, then for $n > p$, the likelihood function can be written as

$$L(\phi,\sigma^2) = (2\pi\sigma^2)^{-n/2}(\det(G_p))^{-1/2} \exp\left\{-\tfrac{1}{2\sigma^2}\left[\mathbf{X}_p^\top G_p^{-1}\mathbf{X}_p + \textstyle\sum_{t=p+1}^n Z_t^2\right]\right\},$$

where $\mathbf{X}_p = (X_1, \dots, X_p)^\top$, $\phi = (\phi_1, \dots, \phi_p)^\top$ and $G_p = \sigma^{-2}\Gamma_p = \sigma^{-2}E(\mathbf{X}_p\mathbf{X}_p^\top)$.

61. If $\{Y_t\}$ is a zero-mean causal ARMA process and X_0 is uncorrelated with Y_t for all t, show that the best linear predictor of Y_{n+1} in terms of $1, X_0, Y_1, \dots, Y_n$ is the same as the best linear predictor of Y_{n+1} in terms of $1, Y_1, \dots, Y_n$.

62. Suppose that $\{Z_t\}$ is a causal stationary AR(p) process with $\mathrm{E}[Z_t^4] < \infty$, and $Z_t = \sqrt{h_t}e_t$ where $\{e_t\} \sim$ i.i.d. $(0,1)$ and

$$h_t = \alpha_0 + \alpha_1 Z_{t-1}^2 + \cdots + \alpha_p Z_{t-p}^2, \quad \sum_{j=1}^p \alpha_j < 1.$$

a) Show that $\mathrm{E}[Z_t^2 | Z_{t-1}^2, Z_{t-2}^2, \dots] = h_t$.

b) Show that $\{Z_t^2\}$ is an AR(p) process. Identify its parameters.

Chapter References

[1] A. Aue. *Time Series Analysis* ↗. LibreTexts, 2021.

[2] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer New York, 2006.

[3] P.S.P. Cowpertwait and A.V. Metcalfe. *Introductory Time Series with R*. Use R! Springer New York, 2009.

[4] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice* ↗. OTexts, 2018.

[5] Department of Statistics. *Applied Time Series Analysis* ↗. PennState's College of Science.