Week 1: Introduction to time series analysis

- Shumway, R.H. & Stoffer, D.S. (2016). Time series analysis and its applications with R examples. springer.
  - Chapter 1: Characteristics of Time Series
- Brockwell, P.J., & Davis, R.A. (2009). Time series: theory and methods. Springer.
  - Chapter 1: Stationary Time Series

Chapter 1

What is time series analysis?

The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that these adjacent observations are **independent and identically distributed**, i.i.d in abbreviation.

Definition 1.1 *The systematic approach by which one goes about answering the mathematical and statistical questions posed by* **time correlations** *is commonly referred to as* **time series analysis**.

Time series problems may arise in different fields. The followings are a few examples:

- Many familiar time series occur in the field of economics, where we are continually exposed to daily stock market quotations or monthly unemployment figures.
- Social scientists follow population series, such as birthrates or school enrollments.
- An epidemiologist might be interested in the number of influenza cases observed over some time period.
- In medicine, blood pressure measurements traced over time could be useful for evaluating drugs used in treating hypertension.
- Functional magnetic resonance imaging of brain-wave time series patterns might be used to study how the brain reacts to certain stimuli under various experimental conditions.

Note 1.1 *There are two separate, but not necessarily mutually exclusive, approaches to time series analysis*

- *time domain approach*
- *frequency domain approach.*

The time domain approach views the investigation of lagged relationships as most important (e.g., how does what happened today affect what will happen tomorrow?), whereas the frequency domain approach views the investigation of cycles as most important (e.g., what is the economic cycle through periods of expansion and recession?).

1.1 The first step in time series analysis

The first step in any time series investigation always involves careful examination of the recorded data plotted over time. Plots enable many features of the data to be visualised, including patterns, unusual observations, changes over time, and relationships between variables. Besides, they often suggest the method of analysis as well as statistics that will be of use in summarizing the information in the data.

The R package that is used in this chapter is `astsa` and `itsmr`.

```
library(astsa)
library(itsmr)
```

Listing 1.1: Libraries for time series in R

If the dataset you have is of class `ts`, you can simply use the command `plot` to see a visualisation of the data.

```
plot(name_of_dataset, type=..., ...)
```

Listing 1.2: Time series plot

Example 1.1 (Population of USA) *The dataset* `uspop.txt` *provides the population of the U.S.A.at 10 year intervals 1790-1990. This dataset is a data-frame and just contains the value. To define it as a time series and plot the data, you can use the following code:*

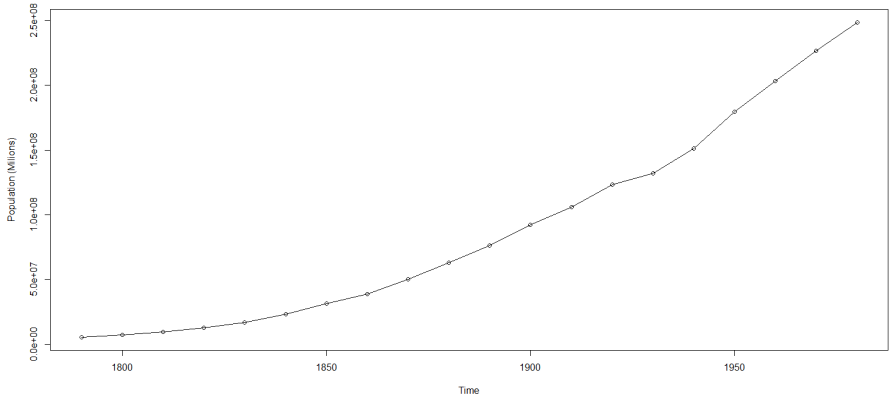

Figure 1.1: Population of the U.S.A. at ten-year intervals, 1790-1980 (U.S. Bureau of the Census).

```
data_pop=read.table("Data/uspop.txt", header=F)
class(data_pop)
data_pop=ts(data_pop, start=1790, end=1980, frequency = 0.1)
```

```
class(data_pop)
plot(data_pop, type="o", xlab="Time", ylab="Population (Milions)")
```

Listing 1.3: Time series plot of Example 1

Question: *What is the primary patterns in this time series?*
Answer: *There is a gradually increasing trend, which might be suggested to be a quadratic or exponential trend.*

Example 1.2 (Johnson & Johnson Quarterly Earnings) *The dataset* `jj` *provides quarterly earnings per share for the U.S. company Johnson & Johnson. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Use the following code to reproduce Figure 1.2:*

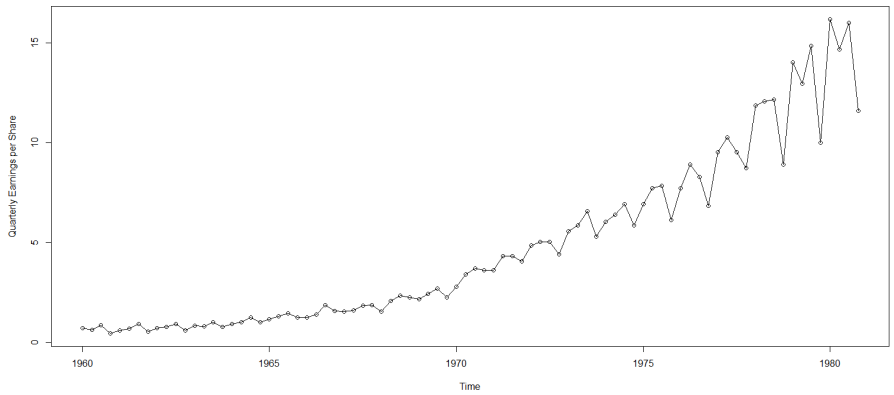

Figure 1.2: Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV

```
plot(jj, type="o", ylab="Quarterly Earnings per Share")
```

Listing 1.4: Time series plot of Example 2

Question: *What is the primary patterns in this time series?*
Answer: *There is a gradually increasing underlying trend and the rather regular variation superimposed on the trend that seems to repeat over quarters.*

Example 1.3 *[Speach data] This dataset is a small .1 second (1000 point) sample of recorded speech for the phrase aaa · · · hhh, and we note the repetitive nature of the signal and the rather regular periodicities. One current problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase aaa · · · hhh. Spectral analysis can be used in this context to produce a signature of this phrase that can be compared with signatures of various library syllables to look for a match. Use the following code to reproduce Figure 1.3:*

```
plot(speech)
```

Listing 1.5: Time series plot of Example 3

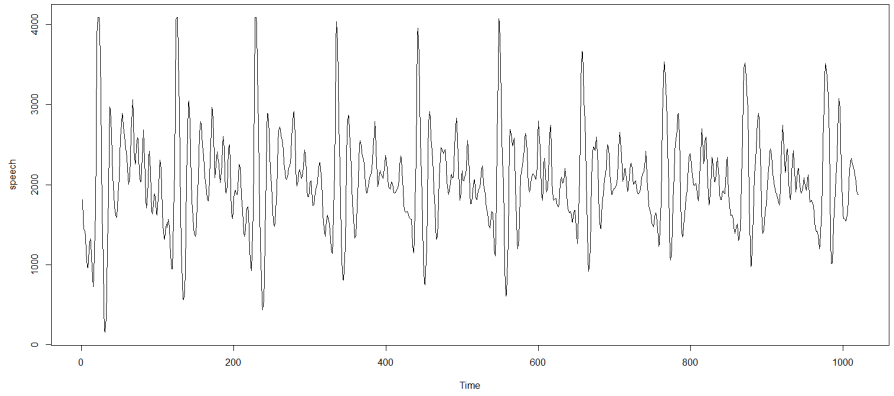

Figure 1.3: Speech recording of the syllable aaa $\cdots$ hhh sampled at 10,000 points per second with n = 1020 points

One can immediately notice the rather regular repetition of small wavelets. The separation between the packets is known as the pitch period and represents the response of the vocal tract filter to a periodic sequence of pulses stimulated by the opening and closing of the glottis.

Example 1.4 *[Sunspot data]* `sunspots.txt` *dataset provides the number of sunspots from 1770 to 1869. Use the following code to reproduce Figure 1.4:*

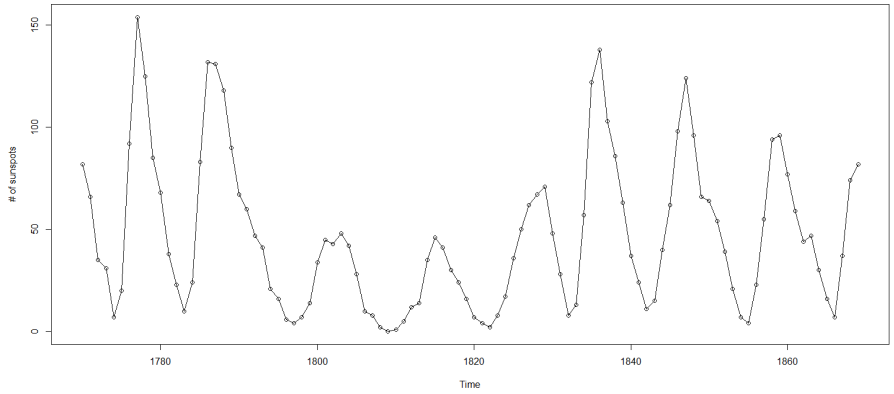

Figure 1.4: Number of sunspots from 1770 to 1869

```
data_sun=read.table("Data/sunspots.txt", header=F)
#data_sun=Sunspots (it is not a time series (numeric))
class(data_sun)
data_sun=ts(data_sun, start=1770, end=1869, frequency = 1)
class(data_sun)
plot(data_sun, type="o", xlab="Time", ylab="# of sunspots")
```

Listing 1.6: Time series plot of Example 4

Example 1.5 (fMRI Imaging) *This dataset presents data collected from various locations in the brain via functional magnetic resonance imaging (fMRI). In this example, five subjects were given periodic brushing on the hand. The stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds (n = 128). For this example, we averaged the results over subjects (these were evoked responses, and all subjects were in phase). Use the following code to reproduce this plot:*

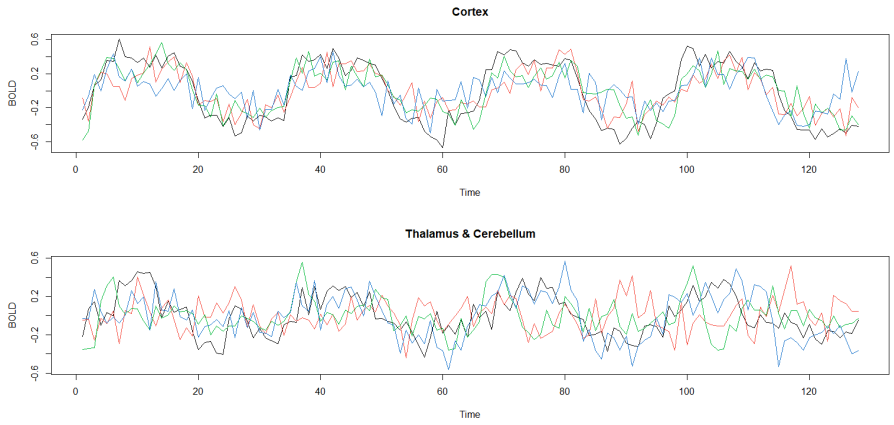

Figure 1.5: fMRI data from various locations in the cortex, thalamus, and cerebellum; n = 128 points, one observation taken every 2 s

```
par(mfrow=c(2,1))
ts.plot(fmri1[,2:5], col=1:4, ylab="BOLD", main="Cortex")
ts.plot(fmri1[,6:9], col=1:4, ylab="BOLD", main="Thalamus & Cerebellum"
    )
```

Listing 1.7: Time series plot of Example 5

1.2 Time Series Statistical Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for data. To allow for the possibly unpredictable nature of future observations, it is natural to suppose that each observation x_t is a **realized value** of a certain random variable X_t.

The time series $\{x_t,\ t \in T_0\}$ is then a realization of the family of random variables $\{X_t,\ t \in T_0\}$. These considerations suggest modelling the data as a realization (or part of a realization) of a **stochastic process** $\{X_t,\ t \in T\}$ where $T_0 \subseteq T$.

Definition 1.2 (Stochastic Process) *A stochastic process is a family of random variables* $\{X_t,\ t \in T\}$ *defined on a probability space* $(\Omega, \mathcal{F}, P)$.

Note 1.2 *In time series analysis the index (or parameter) set* T *is a set of time points, very often* $0, \pm 1, \pm 2, \ldots$, $1, 2, 3, \ldots$, $[0, \infty)$ *or* $(-\infty, \infty)$. *Stochastic processes in which* T *is not a subset of* $\mathbb{R}$ *are also of importance.*

Recalling the definition of a random variable, we note that for each fixed $t \in T$, X_t is in fact a function $X_t(\cdot)$ on the set Ω. On the other hand, for each fixed $\omega \in \Omega$, $X_\cdot(\omega)$ is a function on T.

Definition 1.3 (Realizations of a Stochastic Process) *The functions* $\{X_\cdot(\omega),\ \omega \in \Omega\}$ *on* T *are known as the realizations or sample-paths of the process* $\{X_t,\ t \in T\}$.

Note 1.3 *We shall frequently use the term time series to mean both the data and the process of which it is a realization.*

For some examples of stochastic processes, you may refer to Pages 9 and 10 of Brockwell and Davis.

The fundamental visual characteristic distinguishing the different series is their differing degrees of smoothness. One possible explanation for this smoothness is that adjacent points in time are correlated, so the value of the series at time t, say, x_t, depends in some way on the past values $x_{t-1}, x_{t-2}, \ldots$. This model expresses a fundamental way in which we might think about generating realistic-looking time series. To begin to develop an approach to using collections of random variables to model time series, consider the following example.

Example 1.6 *[White Noise] A simple kind of generated series might be a collection of uncorrelated random variables,* w_t*, with mean 0 and finite variance* σ_w^2*. The time series generated from uncorrelated variables is used as a model for noise in engineering applications, where it is called* **white noise***; denoted as* $w_t \sim wn(0, \sigma_w^2)$.
We will sometimes require the noise to be independent and identically distributed (iid) random variables with mean 0 and variance σ_w^2*. We distinguish this by writing* $w_t \sim iid(0, \sigma_w^2)$ *or by saying* **white independent noise** *or* **iid noise***. A particularly useful white noise series is* **Gaussian white noise***, wherein the* w_t *are* **independent normal** *random variables, with mean 0 and variance* σ_w^2*; or more succinctly,* $w_t \sim N(0, \sigma_w^2)$.

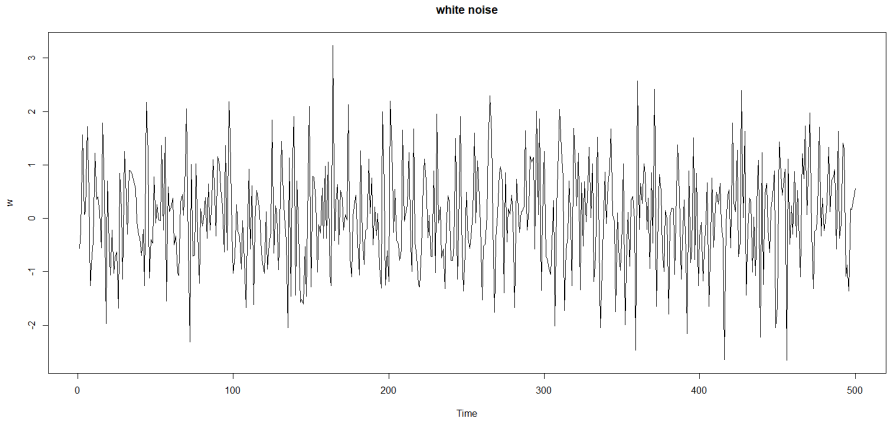

Figure 1.6: Gaussian white noise series

```r
set.seed(123)
w = rnorm(500,0,1) # 500 N(0,1) variates
plot.ts(w, main="white noise")
```

Listing 1.8: 500 observations from a Gaussian white noise

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

Example 1.7 *[Moving Averages and Filtering] We might replace the white noise series w_t by a moving average that smooths the series: For example, consider replacing w_t in Example 1.6 by an average of its current value and its immediate neighbors in the past and future. That is, let*

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1}) \tag{1.2.1}$$

Here, for each time t, we take the average of its current value and its immediate neighbors in the past and future, Figure 1.7. Inspecting the series shows a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster oscillations are taken out.

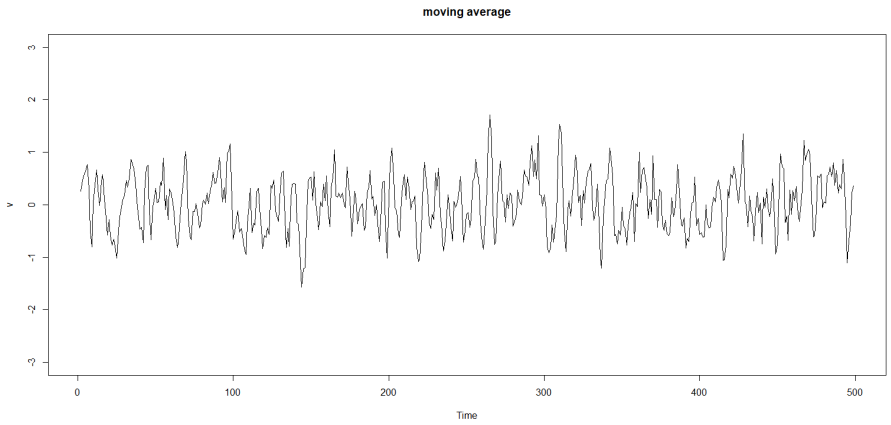

Figure 1.7: Three-point moving average of the Gaussian white noise series generated from model (1.2.1)

```r
v = filter(w, sides=2, filter=rep(1/3,3)) # moving average
plot.ts(v, ylim=c(-3,3), main="moving average")
```

Listing 1.9: moving average observations generated from Gaussian white noise series in example 6

Note 1.4 *A linear combination of values in a time series such as in Eq.* (1.2.1) *is referred to as a* **filtered series**.

The speech series in Fig. 1.3 differ from the moving average series because one particular kind of oscillatory behavior seems to predominate. A number of methods exist for generating series with this quasi-periodic behavior.

Example 1.8 *[Autoregressions] Suppose we consider the white noise series w_t of Example 1.6 as input and calculate the output using the second-order equation*

$$x_t = x_{t-1} - 0.9x_{t-2} + w_t \tag{1.2.2}$$

successively for $t = 1, 2, \ldots, 500$. Eq. (1.2.2) represents a regression or prediction of the current value x_t of a time series as a function of the past two values of the series, and, hence, the term autoregression is suggested for this model.

A problem with startup values exists here because (1.2.2) also depends on the initial conditions x_0 and x_{-1}, but assuming we have the values, we generate the succeeding values by substituting into (1.2.2). The resulting output series is shown in Fig. 1.8, and we note the periodic behavior of the series, which is similar to that displayed by the speech series in Fig. 1.3. As in the previous example, the data are obtained by a filter of white noise. The function filter uses zeros for the initial values. In this case, $x_1 = w_1$, and $x_2 = x_1 + w_2 = w_1 + w_2$, and so on, so that the values do not satisfy (1.2.2). An easy fix is to run the filter for longer than needed and remove the initial values.

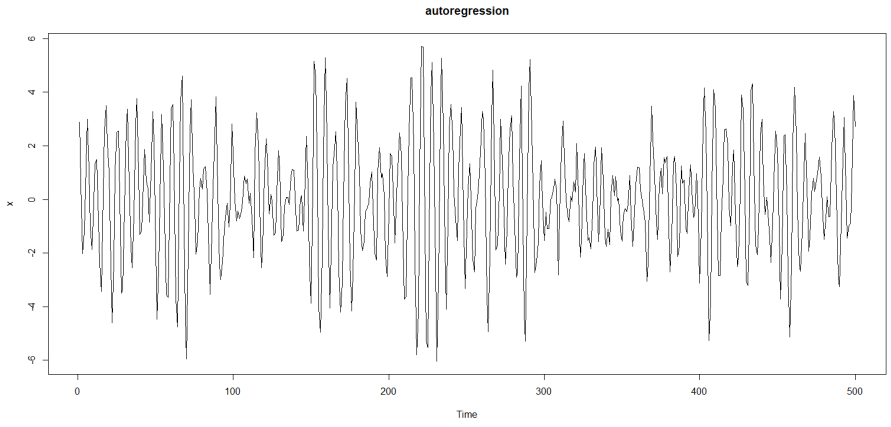

Figure 1.8: Autoregressive series generated from model (1.2.2)

```
set.seed(123)
w = rnorm(550,0,1) # 50 extra to avoid startup problems
x = filter(w, filter=c(1,-.9), method="recursive")[-(1:50)] # remove
    first 50
plot.ts(x, main="autoregression")
```

Listing 1.10: Autoregressive observations generated from Gaussian white noise series

Example 1.9 (Random walk with drift) *The random walk with drift model is given by*

$$x_t = \delta + x_{t-1} + w_t \tag{1.2.3}$$

successively for $t = 1, 2, \ldots$ *with initial condition* $x_0 = 0$, *and where* w_t *is white noise. The constant* δ *is called the drift, and when* $\delta = 0$, (1.2.3) *is called simply a random walk. The term random walk comes from the fact that, when* $\delta = 0$, *the value of the time series at time t is the value of the series at time* $t - 1$ *plus a completely random movement determined by* w_t. *Note that we may rewrite* (1.2.3) *as a cumulative sum of white noise variates:*

$$x_t = t\delta + \sum_{i=1}^{t} w_i \tag{1.2.4}$$

for $t = 1, 2, \ldots$.

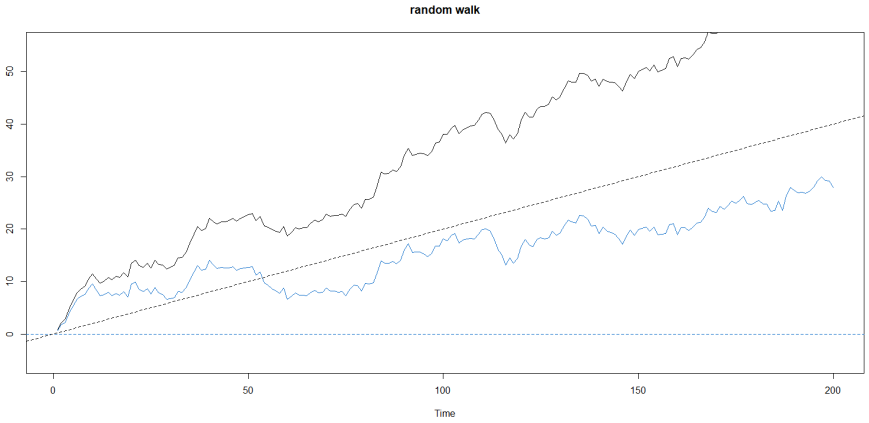

Figure 1.9: Random walk, $\sigma_w = 1$, with drift $\delta = 0.2$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and straight (dashed) lines with slope δ.

```
set.seed(145) # so you can reproduce the results
w = rnorm(200); x = cumsum(w) # two commands in one line
wd = w +.2; xd = cumsum(wd)
plot.ts(xd, ylim=c(-5,55), main="random walk", ylab='')
lines(x, col=4); abline(h=0, col=4, lty=2); abline(a=0, b=.2, lty=2)
```

Listing 1.11: Simulated data from random walk in Example 9.

Example 1.10 (Signal in Noise) *Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise. For example, it is easy to detect the regular cycle fMRI series displayed on the top of Fig. 1.5. Consider the model*

$$x_t = 2\cos\left(\frac{2\pi(t + 15)}{50}\right) + w_t \tag{1.2.5}$$

for $t = 1, 2, \ldots$, *where the first term is regarded as the signal, shown in the upper panel of Fig. 1.10. We note that a sinusoidal waveform can be written as*

$$A\cos(2\pi\omega t + \phi), \tag{1.2.6}$$

where A is the amplitude, ω *is the frequency of oscillation, and* ϕ *is a phase shift. In* (1.2.5), $A = 2$, $\omega = 1/50$ *(one cycle every 50 time points), and* $\phi = 2\pi 15/50 = 0.6\pi$.

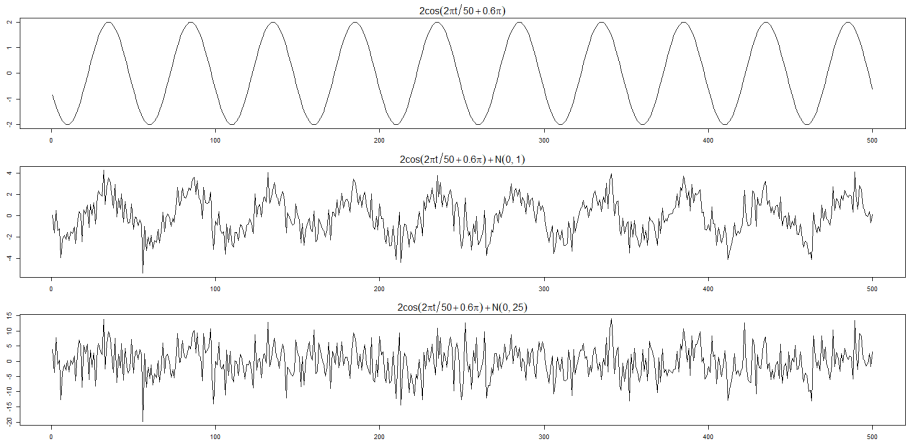

Figure 1.10: Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel)

```
set.seed(125)
cs = 2*cos(2*pi*1:500/50 + .6*pi); w = rnorm(500,0,1)
par(mfrow=c(3,1), mar=c(3,2,2,1), cex.main=1.5)
plot.ts(cs, main=expression(2*cos(2*pi*t/50+.6*pi)))
plot.ts(cs+w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,1)))
plot.ts(cs+5*w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,25)))
```

Listing 1.12: Signal in Noise data introduced in Example 10.

1.3 Stationarity and Strict Stationarity

To understand how random variables are related, we can use some tools from probability theory. One of them is the joint distribution function, which tells us the probability of observing different combinations of values. Another one is the mean, which measures the average value of each variable. And the last one is the covariance, which quantifies how much two variables vary together. These tools are also useful for time series analysis, where we study data that changes over time.

1.3.1 Measures of Dependence

A complete description of a time series, observed as a collection of n random variables at arbitrary time points $t_1, t_2, \ldots, t_n$, for any positive integer n, is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less than the n constants, $c_1, c_2, \ldots, c_n$; i.e.,

$$F_{t_1,t_2,\ldots,t_n}(c_1, c_2, \ldots, c_n) = Pr(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_n} \leq c_n). \tag{1.3.1}$$

Unfortunately, these multidimensional distribution functions cannot usually be written easily unless the random variables are jointly normal, in which case the joint density has the well-known form.

For x_t, the marginal distribution function is defined as $F_{x_t}(x) = P(x_t \leq x)$ or the corresponding marginal density function is

$$f_{x_t}(x) = \frac{\partial F_t(x)}{\partial x} \tag{1.3.2}$$

and when they exist, they are often informative for examining the marginal behavior of a series. Another informative marginal descriptive measure is the mean function.

Definition 1.4 *The mean function is defined as*

$$\mu_{x_t} = E(x_t) = \int_{-\infty}^{\infty} x f_{x_t}(x) dx \tag{1.3.3}$$

provided it exists, where E denotes the usual expected value operator. When no confusion exists about which time series we are referring to, we will drop a subscript and write μ_{x_t} *as* μ_x .

The lack of independence between two adjacent values x_s and x_t can be assessed numerically, as in classical statistics, using the notions of covariance and correlation. Assuming the variance of x_t is finite, we have the following definition.

Definition 1.5 (The Autocovariance Function) *If* $\{x_t, \ t \in T\}$ *is a process such that* $Var(x_t) < \infty$ *for each* $t \in T$*, then the autocovariance function* $\gamma_x(\cdot, \cdot)$ *of* $\{x_t, \ t \in T\}$ *is defined by*

$$\gamma_x(s,t) = Cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \quad s,t \in T. \tag{1.3.4}$$

When no possible confusion exists about which time series we are referring to, we will drop the subscript and write $\gamma_x(s,t)$ *as* $\gamma(s,t)$*. Note that* $\gamma_x(s,t) = \gamma_x(t,s)$ *for all time points s and t.*

The autocovariance measures the linear dependence between two points on the same series observed at different times.

- Very smooth series exhibit autocovariance functions that stay large even when the t and s are far apart
- choppy series tend to have autocovariance functions that are nearly zero for large separations

Note 1.5 *Recall from classical statistics that if* $\gamma(s,t) = 0$*,* x_s *and* x_t *are not linearly related, but there still may be some dependence structure between them. If, however,* x_s *and* x_t *are bivariate normal,* $\gamma(s,t) = 0$ *ensures their independence.*

Note 1.6 *For* $s = t$*, the autocovariance reduces to the (assumed finite) variance, because* $\gamma(t,t) = E[(x_t - \mu_t)^2] = Var(x_t)$.

Example 1.11 (Mean and autocovariance functions for the white noise process) *Consider the white noise series w_t. Based on the definition, $E(w_t) = 0$ and by uncorrelatedness of the observations, we have*

$$\gamma_w(s,t) = Cov(w_s, w_t) = \begin{cases} \sigma_w^2 & s = t \\ 0 & s \neq t \end{cases} \tag{1.3.5}$$

Example 1.12 (Mean autocovariance functions of a moving average series) *If w_t denotes a white noise series, then $\mu_{w_t} = E(w_t) = 0$ for all t. Smoothing the series as in Example 1.7 does not change the mean, since*

$$\mu_{v_t} = E(v_t) = \frac{1}{3}[E(w_{t-1}) + E(w_t) + E(w_{t+1})] = 0.$$

For the autocovariance function, we have

$$\begin{aligned}
\gamma_v(s,t) &= Cov(v_s, v_t) \\
&= Cov\left(\frac{1}{3}(w_{s-1} + w_s + w_{s+1}), \frac{1}{3}(w_{t-1} + w_t + w_{t+1})\right) \\
&= \frac{1}{9}\left[Cov(w_{s-1}, w_{t-1}) + Cov(w_{s-1}, w_t) + Cov(w_{s-1}, w_{t+1})\right. \\
&\quad + Cov(w_s, w_{t-1}) + Cov(w_s, w_t) + Cov(w_s, w_{t+1}) \\
&\quad \left. + Cov(w_{s+1}, w_{t-1}) + Cov(w_{s+1}, w_t) + Cov(w_{s+1}, w_{t+1})\right]
\end{aligned}$$

When $s = t$, we have

$$\begin{aligned}
\gamma_v(t,t) &= \frac{1}{9}\left[Cov(w_{t-1}, w_{t-1}) + Cov(w_t, w_t) + Cov(w_{t+1}, w_{t+1})\right] \\
&= \frac{3}{9}\sigma_w^2
\end{aligned}$$

When $s = t + 1$, we have

$$\begin{aligned}
\gamma_v(t+1,t) &= \frac{1}{9}\left[Cov(w_t, w_t) + Cov(w_{t+1}, w_{t+1})\right] \\
&= \frac{2}{9}\sigma_w^2
\end{aligned}$$

When $s = t + 2$, we have

$$\begin{aligned}
\gamma_v(t+2,t) &= \frac{1}{9}\left[Cov(w_{t+1}, w_{t+1})\right] \\
&= \frac{1}{9}\sigma_w^2
\end{aligned}$$

Similarly, $\gamma_v(t-1,t) = \frac{2}{9}\sigma_w^2$ and $\gamma_v(t-2,t) = \frac{1}{9}\sigma_w^2$. We summarize the values for all s and t as

$$\gamma_v(s,t) = \begin{cases} \frac{3}{9}\sigma_w^2 & s = t \\ \frac{2}{9}\sigma_w^2 & |s-t| = 1 \\ \frac{1}{9}\sigma_w^2 & |s-t| = 2 \\ 0 & |s-t| > 2 \end{cases} \tag{1.3.6}$$

Note 1.7 *From* (1.3.6)*, it can be seen that the smoothing operation introduces a covariance function that decreases as the separation between the two time points increases and disappears completely when the time points are separated by three or more time points. This particular autocovariance is interesting because it only depends on the time separation or* **lag** *and not on the absolute location of the points along the series.*

Example 1.13 (Mean and autocovariance functions of a random walk with drift) *Consider the random walk with drift model given in* (1.2.4)*. Since* $E(w_t) = 0$*, for all* t*, and* δ *is constant, we have*

$$\mu_{x_t} = E(x_t) = \delta t + \sum_{i=1}^{t} E(w_i) = \delta t,$$

which is a function of t*. Besides, the autocovariance function can be calculated as follows:*

$$\begin{aligned} Cov(x_s, x_t) &= Cov(\delta s + \sum_{i=1}^{s} w_i, \delta t + \sum_{j=1}^{t} w_j) \\ &= Cov(\sum_{i=1}^{s} w_i, \sum_{j=1}^{t} w_j) \\ &= \sum_{i=1}^{\min\{s,t\}} Var(w_i) \\ &= \min\{s, t\}\sigma_w^2 \end{aligned}$$

because w_t *are uncorrelated random variables. Note that, as opposed to the previous examples, the autocovariance function of a random walk depends on the particular time values* s *and* t*, and not on the time separation or lag. Also, notice that the variance of the random walk,* $Var(x_t) = t\sigma_w^2$*, increases without bound as time* t *increases. The effect of this variance increase can be seen in Fig. 1.9 where the processes start to move away from their mean functions* δt*.*

Example 1.14 (Mean Function of Signal Plus Noise) *Many practical applications depend on assuming the observed data have been generated by a fixed signal waveform superimposed on a zero-mean noise process, leading to an additive signal model of the form* (1.2.5)*. It is clear, because the signal in* (1.2.5) *is a fixed function of time, we will have*

$$\begin{aligned} \mu_{x_t} = E(x_t) &= 2\cos\left(\frac{2\pi(t + 15)}{50}\right) + E(w_t) \\ &= 2\cos\left(\frac{2\pi(t + 15)}{50}\right). \end{aligned}$$

Therefore, the mean function is just the cosine wave, which depends on t*. What is the autocavariance function in this example?*

As in classical statistics, it is more convenient to deal with a measure of association between -1 and 1, and this leads to the following definition.

Definition 1.6 (The autocorrelation function (ACF)) *Consider the series x_t. The autocorrelation function of x_t is defined as*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}, \quad s,t \in T. \tag{1.3.7}$$

The ACF measures the **linear predictability** of the series at time t. We can show easily that $-1 \leq \rho(s,t) \leq 1$ using the Cauchy-Schwarz inequality. If we can predict x_t **perfectly** from x_s through a linear relationship, $x_t = \beta_0 + \beta_1 x_s$, then the correlation will be $+1$ when $\beta_1 > 0$, and -1 when $\beta_1 < 0$.

Sometimes, we would like to measure the predictability of another series y_t from the series x_s. Assuming both series have finite variances, we have the following definition.

Definition 1.7 (The cross-covariance function) *Consider two series, x_t and y_t. Then, the cross-covariance function between these two series is defined as*

$$\gamma_{xy}(s,t) = Cov(x_s, y_t) = E[(x_s - \mu_{x_s})(y_t - \mu y_t)], \quad s,t \in T. \tag{1.3.8}$$

Definition 1.8 (The cross-correlation function (CCF)) *Consider two series, x_t and y_t. Then, the cross-correlation function is given by*

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}, \quad s,t \in T. \tag{1.3.9}$$

We may easily extend the above ideas to the case of more than two series, say, $x_{t1}, x_{t2}, \ldots, x_{tr}$; that is, multivariate time series with r components. For example, the autocovariance function in this case is

$$\gamma_{jk}(s,t) = E((x_{sj} - \mu_{sj})(x_{tk} - \mu_{tk})), \quad j,k = 1,2,\ldots,r. \tag{1.3.10}$$

In the definitions above, the autocovariance and cross-covariance functions may change as one moves along the series because the values depend on both s and t, the locations of the points in time. In some examples, the autocovariance function depend on the **separation** of x_s and x_t, say, $h = |s - t|$, and not on **where** the points are located in time. So, as long as the points are separated by h units, the location of the two points does not matter. This notion is called **weak stationarity,** when the **mean is constant**.

1.3.2 Stationary and Strict Stationarity Time Series

The preceding examples have hinted that a sort of regularity may exist over time in the behavior of a time series. We introduce the notion of regularity using a concept called stationarity.

Definition 1.9 (Strictly stationary) *The time series x_t is called strictly stationary if the probabilistic behavior of every collection of values*

$$\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\}$$

is identical to that of the time shifted set

$$\{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}$$

That is,

$$Pr(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_k} \leq c_k) = Pr(x_{t_1+h} \leq c_1, x_{t_2+h} \leq c_2, \ldots, x_{t_k+h} \leq c_k\}) \quad (1.3.11)$$

for all $k = 1, 2, \ldots$, *all time points* $t_1, t_2, \ldots, t_k$, *all numbers* $c_1, c_2, \ldots, c_k$, *and all time shifts* $h = 0, \pm 1, \pm 2, \ldots$, *(the same joint distribution)*

If a time series is strictly stationary, then all of the multivariate distribution functions for subsets of variables must agree with their counterparts in the shifted set for all values of the shift parameter h.

- when $k = 1$, (1.3.11) implies that

$$Pr(x_s \leq c) = Pr(x_t \leq c) \quad (1.3.12)$$

  for any time points s and t. Therefore, if the mean function, μ_t, of the series exists, (1.3.12) implies that $\mu_s = \mu_t$, for all s and t, and hence μ_t **must be constant**.

- When $k = 2$, we can write (1.3.11) as

$$Pr(x_s \leq c_1, x_t \leq c_2) = Pr(x_{s+h} \leq c_1, x_{t+h} \leq c_2) \quad (1.3.13)$$

  for any time points s and t and shift h. Thus, if the variance function of the process exists, then it can be implied that the autocovariance function of the series x_t satisfies

$$\gamma(s, t) = \gamma(s + h, t + h)$$

  for all s and t and h. We may interpret this result by saying the autocovariance function of the process depends only on the time difference between s and t, and not on the actual times.

Strict stationarity means intuitively that the graphs over two equal-length time intervals of a realization of the time series should exhibit similar statistical characteristics. For example, the proportion of ordinates not exceeding a given level c should be roughly the same for both intervals. This version of stationarity is too strong and is difficult to assess.

Definition 1.10 (Weakly stationary) *A weakly stationary time series,* x_t *, is a process such that*

(i) $E(|x_t|^2) < \infty$, *for all* t, *or equivalently the process has finite variance, i,e,* $Var(x_t) = \sigma_t^2 < \infty$,

(ii) the mean value function, μ_t, *is constant and does not depend on time* t, *i.e.,* $E(x_t) = \mu$, *for all* t,

(iii) $\gamma(s,t) = \gamma(s+h, t+h)$*, which means that* $\gamma(s,t)$ *depends on* s *and* t *only through their difference* $|s-t|$.

Henceforth, we will use the term **stationary** to mean **weakly stationary**; if a process is stationary in the strict sense, we will use the term strictly stationary.

Since the autocovariance function, $\gamma(s,t)$, of a stationary time series depends on s and t only through $|s-t|$, we may simplify the notation. Let $s = t+h$, where h represents the time shift or lag. Then

$$\gamma(t+h,t) = Cov(x_{t+h}, x_t) = Cov(x_h, x_0) = \gamma(h,0)$$

because the time difference between times $t+h$ and t is the same as the time difference between times h and 0. Thus, the autocovariance function of a stationary time series does not depend on the time argument t. Henceforth, for convenience, we will drop the second argument of $\gamma(h,0)$.

Definition 1.11 (Autocovariance and autocorrelation functions of a stationary time series) *The autocovariance function of a stationary time series will be written as*

$$\gamma(h) = Cov(x_{t+h}, x_t), \tag{1.3.14}$$

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{1.3.15}$$

The Cauchy-Schwarz inequality shows again that $-1 \leq \rho(h) \leq 1$ for all h, enabling one to assess the relative importance of a given autocorrelation value by comparing with the extreme values -1 and 1.

Note 1.8 *To find the relation between stationary and strictly stationary time series, note that*

- *A strictly stationary time series, with finite variance, is stationary.*
- *A stationary time series is not necessarily strictly stationary. For example, if* x_t *is a sequence of independent random variables such that it is exponentially distributed with mean one when t is odd and normally distributed with mean one and variance one when t is even, then* x_t *is stationary with* $\gamma_x(0) = 1$ *and* $\gamma_x(h) = 0$ *for* $h \neq 0$. *However since* x_1 *and* x_2 *have different distributions,* x_t *cannot be strictly stationary.*
- *One important case where stationarity implies strict stationarity is if the time series is Gaussian.*

Definition 1.12 (Gaussian process) *A process* $\{x_t\}$ *is said to be a Gaussian process if the n-dimensional vectors* $x = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})'$*, for every collection of distinct time points* $t_1, t_2, \ldots, t_n$*, and every positive integer n, have a multivariate normal distribution.*

Why a Gaussian stationary time series is strictly stationary? A multivariate normal distribution depends only on the means (known to be constant), variances (also known to be constant), and covariances of the random variables involved, and the covariances depend only on the time differences between the variables. Thus, $Cov(x_i, x_j) = Cov(x_{i+m}, x_{j+m})$ and so the multivariate normal distribution of $(x_i, x_j, x_k, \ldots)$ is the same as the multivariate normal distribution of $(x_{i+m}, x_{j+m}, x_{k+m}, \ldots)$, which is what we need in order to assert strict stationarity.

Example 1.15 (White noise process) *White noise process is stationary and the autocorrelation function is given by* $\rho_w(0) = 1$ *and* $\rho_w(h) = 0$ *for* $h \neq 0$. *If the white noise variates are also Gaussian, the series is also strictly stationary (Gaussian white noise).*

Stationary processes play a crucial role in the analysis of time series. Of course many observed time series are nonstationary in appearance. Frequently such data sets can be transformed into series which can reasonably be modelled as realizations of some stationary process. The theory of stationary processes is then used for the analysis, fitting and prediction of the resulting series.

Example 1.16 (Trend Stationarity) *If* $x_t = \alpha + \beta t + y_t$, *where* y_t *is stationary, then the mean function is* $\mu_{x_t} = \alpha + \beta t + \mu_{y_t}$, *which is not independent of time. Therefore, the process is not stationary. The autocovariance function, however, is independent of time, because* $\gamma_x(h) = Cov(x_{t+h}, x_t) = \gamma_y(h)$. *Thus, the model may be considered as having stationary behavior around a linear trend; this behavior is sometimes called* **trend stationarity**. *Can you suggest a transformation to make this time series statioanry?*

1.3.3 Autocovariance function of a stationary process and its estimation

If $\gamma(\cdot)$ is the autocovariance function of a stationary process x_t, then

$$\gamma(0) \geq 0 \tag{1.3.16}$$

$$|\gamma(h)| \leq \gamma(0), \quad \text{for all } h \in \mathbb{Z} \tag{1.3.17}$$

$$\gamma(h) = \gamma(-h), \quad \text{for all } h \in \mathbb{Z} \tag{1.3.18}$$

Besides, $\gamma(h)$ is non-negative definite, i.e., $\sum_{j=1}^{n} \sum_{k=1}^{n} a_j \gamma(j-k) a_k \geq 0$, for any $n \geq 1$ and constants $a_1, \ldots, a_n$. This ensurs that variances of linear combinations of the variates x_t will never be negative. How?

Question: What are the properties of autocorrelation function $\rho(\cdot)$?

Although the theoretical autocorrelation and cross-correlation functions are useful for describing the properties of certain hypothesized models, most of the analyses must be performed using sampled data. So, we need to estimate the mean, autocovariance and autocorrelation functions using sample. This is an important step towards constructing an appropriate mathematical model for the data.

If a time series is stationary,

- the mean function μ_t is constant so we can estimate it by the **sample mean**,

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{1.3.19}$$

  This estimator is unbiased, i.e., $E(\bar{x}) = \mu$ with $Var(\bar{x}) = \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_x(h)$. (Why?)

- the autocovariance function is estimated by the **sample autocovariance** function:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}) \tag{1.3.20}$$

  with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \ldots, n-1$.

Note 1.9 *(i) The divisor n is used rather than $(n-h)$ since this ensures that the matrix $[\hat{\gamma}(i-j)]_{i,j=1}^{n}$ is non-negative definite.*
(ii) The sample autocorrelation function (ACF) is defined in terms of the sample autocovariance function as

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad |h| < n. \tag{1.3.21}$$

The corresponding matrix $[\hat{\rho}(i-j)]_{i,j=1}^{n}$ is then also non-negative definite.

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data comes from a completely random or white series or whether correlations are statistically significant at some lags.

Property 1.1 (Large-Sample Distribution of the ACF) *Under general conditions, if x_t is white noise, then for n large, the sample ACF, $\hat{\rho}_x(h)$, for $h = 1, 2, \ldots, H$, where H is fixed but arbitrary, is approximately normally distributed with zero mean and standard deviation given by*

$$\sigma_{\hat{\rho}_x(h)} = \frac{1}{\sqrt{n}} \tag{1.3.22}$$

This property helps us to obtain a rough method of assessing whether peaks in $\hat{\rho}(h)$ are significant by determining whether the observed peak is outside the interval $\pm 2/\sqrt{n}$ (Why?). For a white noise sequence, approximately 95% of the sample ACFs should be within these limits. The applications of this property develop because many statistical modeling procedures depend on reducing a time series to a white noise series using various kinds of transformations. After such a procedure is applied, the plotted ACFs of the residuals should then lie roughly within the limits given above.

Example 1.17 (A Simulated Time Series) *To compare the sample ACF for various sample sizes to the theoretical ACF, consider a contrived set of data generated by tossing a*

fair coin, letting $x_t = 1$ *when a head is obtained and* $x_t = -1$ *when a tail is obtained. Then, construct* y_t *as*

$$y_t = 5 + x_t - 0.7x_{t-1}. \tag{1.3.23}$$

To simulate data, we consider two cases, one with a small sample size ($n = 10$) *and another with a moderate sample size* ($n = 100$).

```
set.seed(101010)
x1 = 2*rbinom(11, 1, .5) - 1 # simulated sequence of coin tosses
x2 = 2*rbinom(101, 1, .5) - 1
# filter(x1, sides=1, filter=c(1,-.7))
y1 = 5 + filter(x1, sides=1, filter=c(1,-.7))[-1]
y2 = 5 + filter(x2, sides=1, filter=c(1,-.7))[-1]
par(mfrow=c(2,2), mgp=c(1.6,.6,0), mar=c(3,3,1,1) )
plot.ts(y1, type='s'); plot.ts(y2, type='s') # plot both series
c(mean(y1), mean(y2)) # the sample means
acf(y1, ylab='ACF(X)')
acf(y2, ylab='ACF(X)')
acf(y1, lag.max=4, plot=FALSE) # 1/sqrt(10)=0.32
acf(y2, lag.max=4, plot=FALSE) # 1/sqrt(100)=0.1
# Note that the sample ACF at lag zero is always 1 (Why?).
```

Listing 1.13: Simulated data from Example 17.

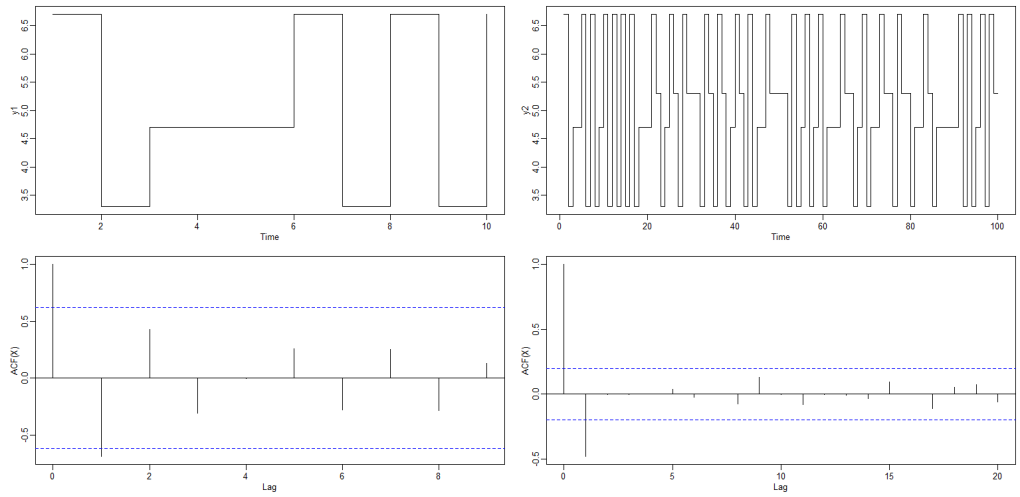

Figure 1.11: Simulated time series and associated ACF in Example 17.

The theoretical ACF of this model can be obtained using the fact that the mean of x_t *is zero and the variance of* x_t *is one. It can be shown that*

$$\rho_y(1) = \frac{-0.7}{1 + (0.7)^2} \tag{1.3.24}$$

and $\rho_y(h) = 0$ *for* $|h| > 1$. *(Show that* (1.3.24) *holds. Find* $\rho_y(-1)$. *)*

Example 1.18 (ACF of a Speech Signal) *Figure 1.12 shows the ACF of the speech series (Example 1.3).The original series appears to contain a sequence of repeating short*

signals. The ACF confirms this behavior, showing repeating peaks spaced at about 106–109 points. Autocorrelation functions of the short signals appear, spaced at the intervals mentioned above. The distance between the repeating signals is known as the pitch period and is a fundamental parameter of interest in systems that encode and decipher speech. Because the series is sampled at 10,000 points per second, the pitch period appears to be between .0106 and .0109 seconds. To compute the sample ACF in R, use acf(speech, 250).

```
library(astsa)
acf(speech, 250)
```

Listing 1.14: Speech data

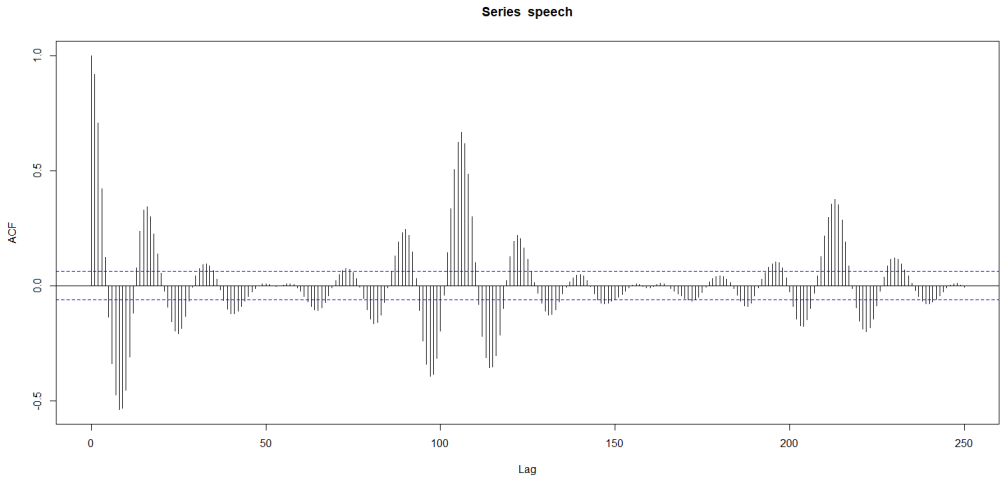

Figure 1.12: ACF of the speech series

Note 1.10 *A* **linear process**, x_t, *is defined to be a linear combination of white noise variates* w_t, *and is given by*

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty, \tag{1.3.25}$$

For the linear process, we may show that

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j \tag{1.3.26}$$

for $h \geq 0$; *recall that* $\gamma_x(-h) = \gamma_x(h)$.
Notice that the linear process (1.3.25) *is dependent on the future (*$j < 0$*), the present (*$j = 0$*), and the past (*$j > 0$*). For the purpose of forecasting, a future dependent model will be useless. Consequently, we will focus on processes that do not depend on the future. Such models are called* **causal**.

1.4 Jointly stationary processes

When several series are available, a notion of stationarity still applies with additional conditions.

Definition 1.13 *Two time series, say, x_t and y_t, are said to be jointly stationary if they are each stationary, and the cross-covariance function*

$$\gamma_{xy}(h) = Cov(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)] \tag{1.4.1}$$

is a function only of lag h.

Definition 1.14 *The cross-correlation function (CCF) of jointly stationary time series x_t and y_t is defined as*

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}} \tag{1.4.2}$$

Again, $-1 \leq \rho_{xy}(h) \leq 1$ which enables comparison with the extreme values -1 and 1 when looking at the relation between x_{t+h} and y_t.

- The cross-correlation function is not generally symmetric about zero, i.e., typically $\rho_{xy}(h) \neq \rho_{xy}(-h)$.
- $\rho_{xy}(h) = \rho_{yx}(-h)$.

Example 1.19 *[Joint Stationarity] Consider the two series, x_t and y_t, formed from the sum and difference of two successive values of a white noise process, say,*

$$x_t = w_t + w_{t-1} \quad \textit{and} \quad y_t = w_t - w_{t-1},$$

where w_t are independent random variables with zero means and variance σ_w^2. It is easy to show that $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$ and $\gamma_x(1) = \gamma_x(-1) = \sigma_w^2$, $\gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$. Also,

$$\gamma_{xy}(1) = Cov(x_{t+1}, y_t) = Cov(w_{t+1} + w_t, w_t - w_{t-1}) = \sigma_w^2.$$

Similarly, $\gamma_{xy}(0) = 0$, $\gamma_{xy}(-1) = -\sigma_w^2$. Therefore

$$\rho_{xy}(h) = \begin{cases} 0 & h = 0, \\ 1/2 & h = 1, \\ -1/2 & h = -1, \\ 0 & |h| \geq 2. \end{cases}$$

Clearly, the autocovariance and cross-covariance functions depend only on h, so the series are jointly stationary.

Definition 1.15 *The estimators for the cross-covariance function,$\gamma_{xy}(h)$, and the cross-correlation, $\rho_{xy}(h)$, are given, respectively, by the sample cross-covariance function*

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y}), \tag{1.4.3}$$

where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ determines the function for negative lags, and the sample cross-correlation function

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}} \tag{1.4.4}$$

For x_t and y_t independent linear processes of the form (1.3.25), we have the following property.

Property 1.2 (Large-Sample Distribution of Cross-Correlation) *The large sample distribution of $\hat{\rho}_{xy}(h)$ is normal with mean zero and*

$$\sigma_{\hat{\rho}_{xy}} = \frac{1}{\sqrt{n}}, \tag{1.4.5}$$

if at least one of the processes is independent white noise.

Example 1.20 *Consider the two time series introduced in Example 1.19. Figure 1.13 presents 198 observations generated from x_t and y_t and their sample autocorrelation functions. Besides, the sample cross-correlation function is displayed in Figure 1.14.*

```
library(astsa)
set.seed(123)
w = rnorm(200,0,1)
x=filter(w,sides=2,filter=c(0,1,1))[c(-1,-200)]
y=filter(w,sides=1,filter=c(1,-1))[c(-1,-200)]
par(mfrow=c(2,2), mgp=c(1.6,.6,0), mar=c(3,3,1,1) )
plot(x,type="l")
plot(y,type="l")
acf(x,30, ylab='ACF(x)')
acf(y,30, ylab='ACF(y)')
dev.off()
ccf(x,y,15, ylab='CCF(X,Y)')
```

Listing 1.15: Speech data

1.5 Vector-Valued and Multidimensional Series

We frequently encounter situations in which the relationships between a number of jointly measured time series are of interest. Hence, it will be useful to consider the notion of a vector time series $x_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, which contains as its components p univariate time series. We denote the $p \times 1$ column vector of the observed series as x_t . The row vector x_t' is its transpose. For the stationary case, the $p \times 1$ mean vector

$$\mu = E(x_t) \tag{1.5.1}$$

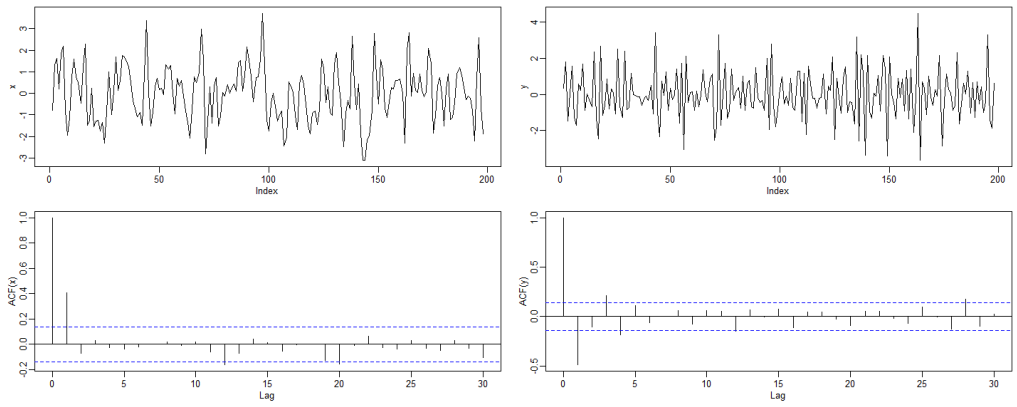

Figure 1.13: Time series x_t and y_t and their sample autocorrelation functions.

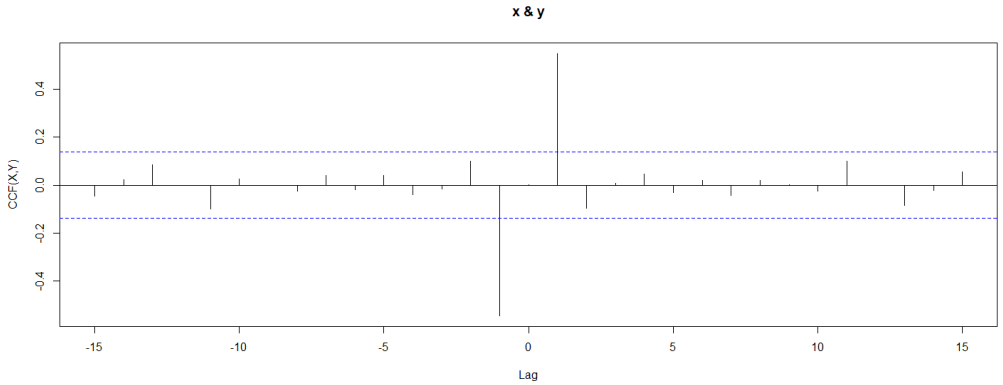

Figure 1.14: The sample cross-correlation of x_t and y_t.

of the form $\mu = (\mu_{t1}, \mu_{t2}, \ldots, \mu_{tp})$ and the $p \times p$ autocovariance matrix

$$\Gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)'] \tag{1.5.2}$$

can be defined, where the elements of the matrix $\Gamma(h)$ are the cross-covariance functions

$$\gamma_{ij}(h) = E[(x_{t+h,i} - \mu_i)(x_{tj} - \mu_j)] \tag{1.5.3}$$

for $i, j = 1, \ldots, p$. Because $\gamma_{ij}(h) = \gamma_{ji}(-h)$, it follows that

$$\Gamma(-h) = \Gamma'(h) \tag{1.5.4}$$

Now, the sample autocovariance matrix of the vector series x_t is the $p \times p$ matrix of sample cross-covariances, defined as

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})', \tag{1.5.5}$$

where

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t, \tag{1.5.6}$$

denotes the $p \times 1$ sample mean vector. The symmetry property of the theoretical autocovariance extends to the sample autocovariance, which is defined for negative values by taking

$$\hat{\Gamma}(-h) = \hat{\Gamma}'(h). \tag{1.5.7}$$

1.6 Review Notes on Multivariate Distribution*

These notes will be relied on for basic facts about joint distributions of random variables as well as properties of the multivariate normal distribution essential for MATH5845.

1.6.1 Joint Distribution and Density Functions

An n-dimensional random vector

$$X = (X_1, \ldots, X_n)'$$

has joint distribution function

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

for all real numbers $x_1, \ldots, x_n$.

The joint distribution of any sub-vector can be obtained by setting $x_i = \infty$ for the other random variables not in the sub-vector. For example, the distribution of X_1 is given by

$$F_{X_1}(x_1) = F(x_1, \infty, \ldots, \infty)$$

and the joint distribution of (X_i, X_j) by

$$F_{X_i,X_j}(x_i,x_j) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty, x_j, \infty, \ldots, \infty).$$

A random vector is said to be continuous if its distribution function can be written in terms of a non-negative density function $f(\cdot, \ldots, \cdot)$ as

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(y_1, \ldots, y_n) dy_1 \ldots dy_n$$

where

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(y_1, \ldots, y_n) dy_1 \ldots dy_n = 1.$$

Note that we can derive the density by differentiating the distribution function

$$f(x_1, \ldots, x_n) = \frac{\partial F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

1.6.2 Independence

The random variables $X_1, \ldots X_n$ are said to be independent if

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$$

for all real numbers $x_1, \ldots, x_n$. This is equivalent to

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

or

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

1.6.3 Conditional Distributions.

Let $X = (X_1, \ldots, X_n)'$ and $Y = (Y_1 \ldots, Y_m)'$ be two random vectors with joint density $f_{X,Y}$. The conditional density of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Note that if X and Y are independent

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

so that $f_{Y|X}(y|x) = f_Y(y)$ in which case knowledge of $X = x$ does not alter the probabilities assigned to outcomes for Y. Conversely, if $f_{Y|X}(y|x) = f_Y(y)$ then X and Y are independent. Similary properties hold in terms of the distribution functions.

1.6.4 Expected Values.

Let $g(X)$ be a function of the random vector X. The expected value is

$$\begin{aligned} E(g(X)) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \ldots dx_n \\ &= \int_{-\infty}^{\infty} g(x) f(x) dx. \end{aligned}$$

The mean $\mu = E(X)$ of a random variable corresponds to setting $g(X) = X$ while the variance $\sigma^2 = \text{var}(X) = E(X - \mu)^2$ corresponds to setting $g(X) = (X - \mu)^2$. The linearity property of expectation is

$$E(aX + b) = aE(X) + b.$$

Note also that

$$\text{var}(aX + b) = a^2 \text{var}(X).$$

1.6.5 Means and Covariances for Random Vectors

The mean vector is

$$\mu_X = E(X) = (E(X_1), \ldots, E(X_n))'$$

and the covariance between X_i and X_j is

$$\text{cov}(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j).$$

The correlation is

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)\ \text{var}(X_j)}}.$$

For two random vectors X and Y the covariance matrix between them is

$$\Sigma_{XY} = \text{cov}(X, Y) = E(X - EX)(Y - EY)' = E(XY') - (EX)(EY)'$$

with (i, j) element

$$(\Sigma_{XY})_{ij} = \text{cov}(X_i, Y_j).$$

When $Y = X$, $\text{cov}(X, Y)$ reduces to the covariance matrix of the random vector X. Note that if X and Y are independent then the covariance between them is the null matrix. The converse is not true in general but is true for the multivariate normal distribution - see below.

Let Y and X be linearly related as $Y = a + BX$ where a is a vector and B is a matrix (all with conforming dimensions). Then

$$\mu_Y = E(Y) = a + BE(X) = a + B\mu_X$$

and

$$\Sigma_{YY} = B\Sigma_{XX}B'.$$

Note also that any covariance matrix Σ is non-negative definite, that is $b'\Sigma b \geq 0$ for any vector b. The proof of this follows from the last identity. Let $Y = b'X$ where X has covariance matrix Σ. Then

$$0 \leq \text{var}(Y) = b'\Sigma b.$$

1.6.6 The Multivariate Normal Distribution.

The general multivariate normal density

The random vector X has the multivariate normal distribution with mean μ and non-singular covariance matrix Σ if

$$f_X(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\right\}.$$

Notation: $X \sim N(\mu, \Sigma)$.

The Bivariate Normal Density

As special case is the bivariate normal density from which most of the required insight about the multivariate normal is obtained. Let $X = (X_1, X_2)'$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

with inverse

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{bmatrix}$$

and $\det(\Sigma) = \sigma_1^2\sigma_2^2(1-\rho^2)$. Substitution in the general multivariate normal density gives

$$f_X(x) = \frac{1}{2\pi[\sigma_1^2\sigma_2^2(1-\rho^2)]^{1/2}} \exp\left\{-\frac{1}{2}Q(x_1, x_2; \sigma_1, \sigma_2, \rho)\right\}$$

with quadratic form

$$Q(x_1, x_2; \sigma_1, \sigma_2, \rho) = \frac{1}{(1-\rho^2)}\Bigg[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\Bigg].$$

Some important facts about the bivariate normal are:

1. The contours of equal density are ellipses

   $$\{(x_1, x_2) : Q(x_1, x_2; \sigma_1, \sigma_2, \rho) = k\}$$

   for any constant $k \geq 0$.

2. When the correlation $\rho = 0$ the two random variables X_1 and X_2 are independent. This can easily be concluded from the form of $Q(x_1, x_2; \sigma_1, \sigma_2, \rho = 0)$. Hence for the bivariate normal distribution, independence is equivalent to uncorrelatedness.

Standardised Bivariate Normal Density

Consider a special case of the bivariate normal density for the two standardised random variables

$$U = \frac{X_1 - \mu_1}{\sigma_1}, \quad V = \frac{X_2 - \mu_2}{\sigma_2}$$

have joint normal density with

$$\mu_U = \mu_V = 0$$
$$\sigma_U^2 = \sigma_V^2 = 1$$

so that

$$f_{U,V}(u, v) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2(1-\rho^2)}[u^2 + v^2 - 2\rho uv]\right). \tag{1.6.1}$$

Recall, for any pair of continuous random variables, the conditional density of $V|U = u$ is

$$f_{V|U}(v|u) = \frac{f_{U,V}(u,v)}{f_U(u)}$$

so that the joint density can be expressed as

$$f_{U,V}(u,v) = f_U(u) f_{V|U}(v|u). \tag{1.6.2}$$

Hence if we can find a factorization of the bivariate normal density (1.6.1) in the form (1.6.2) then we have derived the marginal density of U and the conditional density of $V|U = u$.

The key to the factorization is the completion of the square in the exponent as follows

$$\begin{aligned}\frac{u^2 + v^2 - 2\rho uv}{1-\rho^2} &= \frac{(u^2 - \rho^2 u^2) + (v^2 - 2\rho uv + \rho^2 u^2)}{1-\rho^2} \\ &= \frac{u^2(1-\rho^2) + (v-\rho u)^2}{1-\rho^2} \\ &= u^2 + \frac{(v-\rho u)^2}{1-\rho^2}.\end{aligned}$$

Substituting this into the exponent in equation (1.6.1) we get

$$\begin{aligned}f_{U,V}(u,v) &= \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2}u^2 - \frac{1}{2}\frac{(v-\rho u)^2}{1-\rho^2}\right) \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)\right]\left[\frac{1}{\sqrt{2\pi}(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{(v-\rho u)^2}{1-\rho^2}\right)\right].\end{aligned}$$

Now the first factor is the standard normal $N(0,1)$ density. The second factor is a density of a $N(\rho u, (1-\rho^2))$ random variable. We can therefore identify the first factor as the marginal density for U,

$$f_U(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

and the second factor as the conditional density for $V|U = u$,

$$f_{V|U}(v|u) = \frac{1}{\sqrt{2\pi}(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{(v-\rho u)^2}{1-\rho^2}\right).$$

This proves that the marginal density for U,the standard normal density, and that the conditional density for $V|U = u$ is normal with (conditional) mean

$$E(V|U = u) = \rho u$$

and (conditional) variance

$$\mathrm{Var}(V|U = u) = 1 - \rho^2.$$

In summary

$$U \sim N(0,1)$$

and

$$V|u \sim N(\rho u, 1-\rho^2).$$

Notes:

1. The parameter ρ is the correlation between U and V as is easily derived as follows

$$\begin{aligned}\text{corr}(U,V) &= \frac{\text{cov}(U,V)}{\sqrt{\text{var}(U)\text{var}(V)}} \\ &= \text{cov}(U,V), \quad \text{(since } U \text{ and } V \text{ have unit variance)} \\ &= \int\int uv f_{U,V}(u,v)dvdu \\ &= \int\int uv f_U(u) f_{V|U}(v|u)dvdu \\ &= \int u f_U(u)\left[\int v f_{V|U}(v|u)dv\right]du\end{aligned}$$

   But $\int v f_{V|U}(v|u)dv$ is the mean value for a random variable with the $N(\rho u, 1-\rho^2)$ density. Hence $\int v f_{V|U}(v|u)dv = \rho u$. Substituting this in the double integral we get

$$\begin{aligned}\text{corr}(U,V) &= \int u f_U(u)[\rho u]du \\ &= \rho \int u^2 f_U(u)du.\end{aligned}$$

   But $\int u^2 f_U(u)du$ is $E\left(U^2\right)$ where $U \sim N(0,1)$ and hence $\int u^2 f_U(u)du = 1$ leading to

$$\text{corr}(U,V) = \rho.$$

   This proves that for the bivariate normal density given by equation (1.6.1) the parameter ρ is the correlation between U and V.

2. When $\rho = 0$ the conditional density simplifies to

$$f_{V|U}(v|u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}v^2\right) = f_V(v)$$

   so that U, V are independent. You can also verify independence directly by considering what happens in expression (1.6.1) when $\rho = 0$. That is

$$\begin{aligned}f_{U,V}(u,v) &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}(u^2+v^2)\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}v^2\right) \\ &= f_U(u)f_V(v).\end{aligned}$$

3. For both positive and negative $\rho \neq 0$, $\text{Var}(V|U=u) = 1-\rho^2 < 1$ so that use of $U = u$ information to predict V using the conditional mean $E(V|U=u) = \rho u$ will lead to lower conditional variance for this prediction of V. That is when U and V are not independent, conditioning on one improves precision of prediction of the other.

4. As $|\rho| \to 1$ so that correlation get large in absolute value, note that the conditional variance $\text{Var}(V|U=u) \to 0$ so that once $U = u$ is known V is known with certainty to be equal to u.

Properties of the General Multivariate Normal Distribution.

Some important facts about the general multivariate normal distribution are:

1. Any subvector of a multivariate normal vector has a multivariate normal distribution.
2. If $X \sim N(\mu, \Sigma)$, B is an $m \times n$ matrix of real numbers and a is a real $m \times 1$ vector then
$$Y = a + BX \sim N(a + B\mu_X, B\Sigma B').$$
3. In particular, any linear combination $b'X$ has a univariate normal distribution.

In general, consider a multivariate normal random vector $X \sim N(\mu, \Sigma)$. Partition as

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\mu^{(j)} = E(X^{(j)})$ and $\Sigma_{ij} = E(X^{(i)} - \mu^{(j)})(X^{(j)} - \mu^{(j)})'$. Then

1. $X^{(1)}$ and $X^{(2)}$ are independent if and only if $\Sigma_{12} = 0$.
2. The conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is multivariate normal with conditional mean vector
$$E(X^{(1)}|X^{(2)} = x^{(2)}) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})$$
and covariance matrix
$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$
This is a very useful and important result.

Exercise 1.1 *Find the ACF of the Moving avearge time series introduced in Example 1.7.*

Exercise 1.2 *Is a random walk process stationary?*

Exercise 1.3 *Let* $x_t = A\cos(\theta t) + B\sin(\theta t)$ *where A and B are two uncorrelated random variables with zero means and unit variances with* $\theta \in [-\pi, \pi]$. *Is* x_t *stationary?*

Exercise 1.4 *Consider the process* $x_t = z_t + \theta z_{t-1}$, *where* $z_t \sim N(0, 1)$ *and* $|\theta| < 1$.

(a) Find the ACF of x_t.

(b) Simulate 300 observation from this process, (i) with $\theta = 0.95$ *(ii) with* $\theta = -0.95$. *Plot the sample ACF upto Lag 40 for each case.*

Exercise 1.5 *Get the ACF of the sunspot data (Example 1.4). Can you observe any special pattern?*